

UCLA

UCLA Electronic Theses and Dissertations

Title

Symbolic Language, Embodied Worlds: Multimodal Intelligence in Humans and Machines

Permalink

<https://escholarship.org/uc/item/1z07z8ht>

ISBN

9798247947172

Author

Jiang, Yanru

Publication Date

2026-06-01

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Symbolic Language, Embodied Worlds:
Multimodal Intelligence in Humans and Machines

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Communication

by

Yanru Jiang

ABSTRACT OF THE DISSERTATION

Symbolic Language, Embodied Worlds: Multimodal Intelligence in Humans and Machines

by

Yanru Jiang

Doctor of Philosophy in Communication

University of California, Los Angeles, 2026

Professor Richard Alan Clarke Dale, Chair

Language and perception operate over fundamentally different representational regimes: language (especially in its textual form used to train models) tends to promote more discrete, symbolic processing, whereas perception and action are more continuous, gradient, and grounded in embodied experience. This distinction shapes how both humans and machines learn, generate, and evaluate information. This dissertation advances a unified account of how these representational differences influence model behavior and human–AI interaction, integrating theoretical analysis, computational modeling, and human experiments.

Chapter 1 synthesizes prior work in cognitive science, language models, and multimodal AI to argue that, while language encodes substantial information about embodied experience through statistical regularities and compositional structure, it remains constrained by its reliance on representations of prior experience. Incorporating multimodality provides access to continuous signals that support the construction of novel shared experiences, while also enhancing communicative clarity in human–AI interactions.

Chapter 2 introduces a model explainability approach that characterizes the learning trajectories of recurrent neural networks processing linguistic versus behavioral signals. Through computational modeling, this chapter demonstrates that sequential models encode linguistic

and gestural inputs in distinct ways, highlighting differences in how discrete and continuous signals support information processing and learning.

Chapter 3 examines human evaluations of AI-generated news image captions. Despite strong performance in language generation, models exhibit weaker perceptual and cross-modal capabilities. This chapter shows that cross-modal context mitigates linguistic heuristics that would otherwise bias judgments under text-only evaluation. Chapter 4 introduces *Visual Narrative Freedom*, a property emerging from loosely coupled image–text relationships in real-world multimodal communication. By systematically varying textual constraints, this chapter illustrates that higher degrees of freedom increase the diversity of generated images and can shift user preferences—sometimes favoring AI-generated images over human-selected ones in high-stakes civic domains.

Overall, these contributions establish a theoretical and empirical framework linking representational structure to model behavior and human perception. This dissertation demonstrates that multimodal intelligence is not merely an additive interface, but is fundamentally shaped by the structure of the representations through which information is encoded, influencing both system performance and human–AI interaction.

The dissertation of Yanru Jiang is approved.

Jungseock Joo

Elisa Kreiss

Hongjing Lu

Richard Alan Clarke Dale, Committee Chair

University of California, Los Angeles

2026

TABLE OF CONTENTS

List of Figures	ix
Acknowledgments	xi
Curriculum Vitae	xviii
1 Introduction	1
1.1 What Does Multimodality Add Beyond Language?	2
1.2 Classic Debates on Symbolic and Embodied Cognition	5
1.2.1 The prominence of symbolic cognition in language models	5
1.2.2 The necessity of embodied cognition via multimodality	9
1.3 Conclusion	13
2 A Model Explainability Approach for Comparing Linguistic and Behavioral Learning	16
2.1 Introduction	17
2.1.1 Background	18
2.1.2 The current study	22
2.2 Methods	25
2.2.1 Datasets	26
2.2.2 Model architecture	29
2.2.3 Deep learning simulation	31
2.2.4 Quantifying performance across epochs	34
2.3 Results	37
2.3.1 Start performance	39

2.3.2	Max performance	40
2.3.3	Time at max	41
2.3.4	End - start performance	43
2.4	Discussion	45
2.4.1	Non-monotonicity	47
2.4.2	Pairwise comparisons	48
2.4.3	Domain distinctions	48
2.4.4	Limitations	49
2.5	Conclusion	50
3	Human Heuristics in Context-Sensitive Image Captioning	52
3.1	Introduction	53
3.2	Methods	55
3.2.1	Collecting and generating image captions	56
3.2.2	Participant recruitment	58
3.2.3	Experimental design	58
3.2.4	Computational analysis	59
3.3	Results	62
3.3.1	Experiment 1	62
3.3.2	Computational analysis	63
3.3.3	Experiment 2	66
3.4	Discussion	68
3.4.1	Limitations and ethics statement	70
3.5	Appendices	70

3.5.1	Examples of top- and bottom-ranked news image captions of informational features	70
3.5.2	Regression coefficients for ground truth vs. human judgment (Models 1–3)	74
3.5.3	Regression coefficients for generation strategies (Models 4–6)	76
4	Visual Narrative Freedom under Loose Image–Text Alignment	78
4.1	Introduction	79
4.1.1	Background on AI-generated photorealistic imagery	81
4.2	Related Work	83
4.2.1	Image–text relationships in vision–language models	83
4.2.2	Image–text relationships in multimedia communication	84
4.3	Characterizing Visual Narrative Freedom	85
4.3.1	Variation in VNF across text-to-image tasks	86
4.3.2	Measuring visual narrative freedom	86
4.4	Model Experiments	88
4.4.1	Dataset	88
4.4.2	Experiment setup	89
4.4.3	Results	89
4.5	Human Evaluation	91
4.5.1	Bridging the depiction gap as a mechanism for VNF	93
4.5.2	Stylistic disadvantages of AI-generated images	96
4.6	Conclusion	97
4.6.1	Limitations and ethics statement	98
4.7	Appendices	99

4.7.1	Other typologies of image–text relationships	99
4.7.2	Descriptive statistics of the news dataset	101
4.7.3	Balancing image fidelity of human experiment stimuli	101
4.7.4	Reference deviation of Imagen 3	103
4.7.5	Visual style analysis of Imagen 3	104
4.7.6	Generation prompts	105
5	Conclusion	107
5.1	Future Directions and Open Questions	111

LIST OF FIGURES

1.1	Illustration of isomorphic mapping between semantic and perceptual representation spaces	7
2.1	A conceptual illustration of learning curves in DNNs.	19
2.2	Illustrations of the learning curve pipeline	26
2.3	Sequential data for Emotion Tweet and Gesture Classifications	29
2.4	Model architecture of RNN-LSTM	30
2.5	Examples at epoch 1 and 50 of binary classifications	35
2.6	<i>Start</i> performance for pairwise classifications	40
2.7	<i>Max</i> performance for pairwise classifications	41
2.8	Time at max (t_{max}) for pairwise classifications	43
2.9	<i>End - start</i> performance for pairwise classifications	45
3.1	Overview of experiments and analysis in Chapter 3	54
3.2	Examples of news image captions generated by different sources and different contextual strategies	56
3.3	Image captioning Experiment 1 results	62
3.4	Directionality of feature coefficients for ground truth and human responses . . .	63
3.5	Image captioning Experiment 2 results	65
3.6	Coefficients of linguistic and informational features for captions generated using <i>full-article</i> , <i>local-context</i> , or <i>shuffled-context</i> input	67
3.7	Examples of top- and bottom-ranked news image captions across informational features	73
4.1	Visual Narrative Freedom shapes human preference in T2I tasks	79

4.2	VNF gradually increases from description- to caption- to article-based generation	90
4.3	Depiction gaps in original and generated image–article pairs	93
4.4	Relationship between CLIP-based image–text similarity and human coherence judgments across news images	94
4.5	AI visual style disadvantage generalizes across T2I systems	96
4.6	Reference deviation (measured using reference-based similarity metrics) of Imagen 3-generated images across three generation conditions	103
4.7	Reference-free image quality metrics assessing the visual style of Imagen 3-generated images across three conditions	104

ACKNOWLEDGMENTS

My parents had a somewhat unorthodox way of raising me. Family dinners were often filled with very adult conversations about business, politics, and economics. Naturally, I ended up pursuing a bachelor's degree in Business Administration at the University of Southern California and working in finance as what seemed like a fairly straightforward life path. That plan, along with my “performant straight-A student” mentality, quickly fell apart during some of my most miserable and self-doubting days working as a private equity analyst, once I realized that what truly intrigued me was the depth of technicality and knowledge itself: the feeling that there existed hard-to-access ideas waiting to be discovered and understood. That instinct probably explains why I was willing to completely overhaul years of experience and preparation in finance and business to start preparing PhD applications from scratch in my early twenties. In retrospect, it became one of the best decisions I have ever made for the narrative of my own life.

I cannot stress how lucky I have been to work with and be advised by my PhD committee chair, Rick Dale. Throughout my PhD, Rick consistently prioritized my intellectual development and reminded me (or perhaps gently hinted to me) to look beyond purely pragmatic goals and genuinely enjoy the process of research itself. As much of a nice Canadian (sorry Rick) as he is, he also somehow always managed to untangle my career anxieties with remarkable clarity. When I was worried about pursuing an additional M.S. in Statistics concurrently with my PhD and whether it might negatively affect my research output, he reassured me that since I had also been a productive “sender,” the workload really should not be a concern. Later, when I was debating academia versus industry and wondering how I should allocate my time and preparation between these two very different paths, Rick wisely reassured me that by already performing well in academia, I was naturally preparing myself for the academic job market. Therefore, instead of [over-]optimizing for academia, I should use additional time to gain industry experience and explore internships.

During my third year, while drafting my comprehensive exam (the “c” paper) after shift-

ing from political communication toward cognition and communication, Rick encouraged me to lean into the more theoretical and almost philosophical literature surrounding language, embodiment, and cognition rather than avoiding it. He pointed out that since I was always competent with methodologies and technical work, I should use that foundation to seriously write through and engage with these conceptual questions in cognitive science. As difficult as that process was, it ultimately led me to discover the research questions I now want to pursue and dedicate my life to. The language-perception projects in this dissertation would not exist if Rick had not challenged me years ago to think more deeply about language and embodiment for my c-paper. That theory paper ended up shaping nearly every project that followed. Looking back, I can only say how fortunate I have been to have an advisor who truly values mentorship and intellectual growth beyond immediate pragmatic concerns, and whose advice remained meaningful far beyond the moment in which it was given.

As the faculty member who admitted me to the UCLA Department of Communication, Jungseock Joo has had a lasting impact on my development as a scholar. From very early on, he encouraged me to think carefully about the implications of choosing between academia and industry, while also pushing me to publish early and build momentum. Long before the current generative AI wave, he strongly encouraged me to take as many deep learning and technical courses as possible, which became one of the most important decisions for my development as a computational social scientist. The rigor of that training shaped my technical foundation early on and continued to benefit me throughout my PhD journey.

Coming from a computational social science background and still having never run a human experiment by my third year, I often felt insecure about my academic training. Hongjing Lu genuinely supported my interest in the intersection of human and machine cognition and guided me patiently through the fundamentals of cognitive psychology experimentation. I am especially grateful for the countless long one-on-one meetings spent going through experiment results, discussing stimuli design, and selecting appropriate statistical models step by step. Her rigor in experimental design shaped how I approach human-centered AI research and gave me the confidence to take on more complex naturalistic human-AI interaction

studies and human-vs-model experimental work across cognitive science and AI research.

I am also deeply grateful to Elisa Kreiss, who joined our department three years ago and immediately became an important mentor in my scholarly development. Elisa’s rigor, patience, hands-on mentorship style, and commitment to maintaining high standards have been incredibly influential in shaping how I think about positioning myself between cognitive science and NLP. She also somehow simultaneously brought the best board games to lab socials, where I had the chance to meet many wonderful academic friends, including Je Hoon Chee, Thomas Kwok, Yingjia Wan, Siyi Gong, and Pavel Savgira. We not only formed meaningful intellectual communities beyond disciplinary boundaries, but our “multimodal representation” reading group discussions also sought to challenge existing engineering-centered assumptions in AI and encouraged us to think more fundamentally about vision–language representation and the theoretical questions underlying foundation models.

I have also learned tremendously from my collaborators, some of whom became my closest friends throughout graduate school. Siyu Liang (also known as my “academia bestie” and [ir]responsible “drinking buddy”) has been one of my best collaborators and closest friends. Academically, we naturally balance each other: I tend to ideate at an arguably “unreasonable” speed and level of ambition, while Siyu pulls me back to reality and makes me explain every tiny implementation detail step by step to ensure rigor and clarity. Socially, we somehow always became the people bringing alcohol to parties (whether appropriate or not), aggressively organizing gatherings, “assigning” celebration events to our friends, and functioning as the logistics queens of our academic lives (Siyu is definitely the more qualified queen here!). Our remote working sessions, TA grading, and RA recruitment emails also left traces across Central America during spring breaks, capturing the typical juggling act of being productive, “work hard, play hard” PhD students. I think one of the clearest signs of deep friendship is finding someone you can repeatedly argue with, disagree with, resolve conflicts with (or sometimes not resolve them at all), and still continue choosing each other as friends afterward. It is probably a little narcissistic to borrow from *My Brilliant Friend* to describe our friendship, but Elena Ferrante’s words perfectly capture part of what I feel:

“you’re my brilliant friend, you have to be the best of all...”

Despite the extremely nonchalant vibe that Je Hoon Chee gives off, I have learned some of the most important lessons about research from him. Having him as part of my friend group, occasional running mate circle, and intellectual life challenged me to pursue more rigorous and ambitious thinking. Somewhere between pseudo-philosophical discussions at Bar Lubitsch and headache-inducing conversations after Venice Run Club sessions, we often talked about the similarities between artistic self-expression and the intellectual metaphysics of being a researcher. I always say that when it comes to research instincts, Je Hoon is one of the people I trust most. His mindset constantly challenged me not to settle for local optima and to think much bigger. As he often pointed out, your research trajectory may appear scattered to other people, while in your own head there is actually a very coherent direction tying everything together. One of the many ideas from him that stayed with me was that nobody will ever fully understand your work better than yourself. As much as you can seek help from advisors and faculty, at the end of the day you are the person responsible for pushing through the blockers and wrestling with the problem until you solve it, no matter how painful or obsessive that process becomes. Because of all this intellectual spirituality, my partner often jokes that I am basically Je Hoon’s fan girl. Honestly, he may not be entirely wrong.

The PhD program I came from highly values interdisciplinarity, and because of that, I have been extremely fortunate to spend time with cohorts and classmates from very different backgrounds. Their diversity challenged my own narrow-mindedness and pushed me to think less linearly about success, achievement, and the many different ways of living a meaningful life.

If I had to describe someone as “such a character,” that title in my program would immediately go to Grace Miao. I genuinely do not understand how one person can simultaneously enjoy piloting studies, collecting hundreds of neuro-dyadic conversations, joining five labs in one year, and successfully networking across social neuroscience, communication, and AI safety communities. We spent countless hours talking through academia, industry, visas, ca-

reer strategy, and life planning, though somehow every conversation would eventually circle back to dating stories and relationship gossip. I have also been very fortunate to collaborate with Gülşah Akçakır throughout my PhD, both on NLP methodologies and multi-agent LLM social simulations. During my typically high-intensity, fast-moving, freestyle brainstorming sessions, I was always impressed by how sharply Gülşah could identify the core issue. Her training in complex systems strongly influenced how I think about research design, structure, and measurement. I also want to thank Jun Luo and Seonhye Noh from our “abandon lab,” where we found a sense of community through celebrations, gossip, and both the sensible and nonsensical sides of academia. I am proud to see everyone finding paths that genuinely suit them.

I had the privilege of learning from many friends and mentors in industry throughout this journey. Mazi Boustani showed me the importance of bridging research and engineering; through countless hours of whiteboarding, we bounced between wild abstractions, technical rabbit holes, and moments where everything suddenly clicked into place. Danny Kim would often stop by my office to catch up on developments in AI and entertainment, while also teaching me how to navigate organizational politics and thoughtfully nudge teams to support ambitious ideas and deliverables. Eugenio Herrera-Berg showed me the empowerment and multiculturalism fostered through open-science practices, as well as the value of remaining intellectually unconventional in an increasingly crowded field that still demands genuine originality. Lastly, Yuming Hu demonstrated an exceptional ability to see the cutting edge of AI implementation in downstream applications, while leading with both strategic vision and genuine care for nurturing people and building teams.

Finally, I want to thank the many friends, collaborators, labmates, and communities I encountered throughout UCLA. On the day of my defense, I felt an overwhelming amount of love and support from people across many different disciplines and stages of life. When I joined the UCLA Department of Communication as part of only the second cohort, I initially hesitated because the program was still so new. Looking back now, pursuing my PhD here was one of the best, if not the best, decisions I have ever made.

For me, the PhD was never ultimately about credentials, but about being on this unique journey of crafting my own story. More importantly, it gave me the time and freedom to think clearly, to be brutally honest about what I genuinely care about, and to repeatedly question what kind of intellectual and personal life I wanted to build. Through this process, I realized that what matters most to me is the pursuit of hard-to-access knowledge, intellectual rigor and relevance, spirituality, and all these meaningful relationships. For that, I will always be grateful.

This acknowledgment should conclude with my lifetime partner, Julian Park. We have spent five years together, which essentially overlaps with the entirety of my PhD journey, and God knows that also means enduring much of the suffering that came with it. He handled what must have been millions of my mental breakdowns and constantly reminded me that we should properly celebrate every achievement instead of immediately treating it as simply “the next thing to move on to.” For every accomplishment, small or large, we would often shout (or more accurately, scream), “we did it” or “we strategized well.” There were endless nights spent drawing decision trees or airdropping each other rambling written notes about how to align my technical strengths, credentials, and personality with my convictions and the kind of life I wanted to build. At one point, he curiously asked me, “I’ve always wondered, would your life be more successful without me?” I always responded without hesitation: “I don’t know whether I would be more or less successful, but I would certainly be much less happy.”

When I asked him in return what his dream in life was, there was never any dramatic or overly sentimental answer. He would simply respond, very logically, “My dream is for your dream to come true,” before joking that “if your dream comes true, then my dream of becoming a ‘househusband’ also comes true.” Even today, after what felt like my millionth mental breakdown, I “forced” him to stand there and listen to yet another long spiral about being overloaded and “perpetually unfortunate” in life. Somehow, he always has this magical ability to calmly respond: “Okay, calm down. You haven’t exhausted your solution space. Let’s look at it together.” Those are the moments that make me feel deeply secure about

being on this journey with him.

Thinking back to the dinner conversations I had with my parents, many of the business instincts I was raised with unexpectedly became valuable additions to how I navigate academia. While there is immense value in pursuing knowledge through intrinsic intellectual curiosity, I also constantly remind myself to think about positioning, downstream impact, the space between proof-of-concept and productionization, and the importance of teamwork and leadership—the belief that if you gather a group of brilliant people together, you can build something much bigger than yourself.

As ambitious and career-driven as my parents are, they never once asked me to settle for external achievements or prestige alone. What always mattered was doing things in a way I could truly be proud of and remaining deeply intrinsic in my pursuit of self-actualization. In many ways, this PhD journey, and academia more broadly, fulfilled that drive more than I could have imagined.

Chapter 1 is adapted from Jiang & Dale, currently under review. Chapter 2 has been published in Jiang & Dale (2025) in *PLOS Computational Biology*. Chapter 3 has been published in Jiang, Dale, & Lu (2025) in the proceedings of Cognitive Science Society, with the full manuscript currently under review. Chapter 4 is adapted from the manuscript by Jiang, Olson, Herrera-Berg, Dale, Lu, & Kreiss, currently under review. This dissertation research was supported by UCLA’s Dissertation Year Fellowship, with my third year supported by the UCLA Graduate Council Diversity Fellowship.

CURRICULUM VITAE

2023–2025	M.S., Statistics University of California, Los Angeles
2017–2019	M.A., Communication Management University of Southern California
2014–2018	B.S., Business Cinematic Arts University of Southern California

HONORS AND AWARDS

2025–2026	Dissertation Year Award, Graduate Division, UCLA
2023–2024	Graduate Council Diversity Fellowship, Academic Senate, UCLA
2023–2024	Meta Research PhD Fellowship Finalist, Meta
2025	Cohere For AI Research Grant Program, Cohere Labs
2025	OpenAI Researcher Access Program, OpenAI
2023	C-SPAN Scholarship & Engagement, C-SPAN
2022	Top Method Paper, AEJMC
2022	First Place Student Paper, AEJMC

PUBLICATIONS

- [1] Y. Jiang*, and S. Liang*, “Capturing epistemic uncertainty in LLM-based soft labeling,” (2026), [ACL Workshop on Generation, Evaluation, and Metrics \(GEM\) Oral](#) .
- [2] Y. Jiang, and R. Dale, “Mapping the learning curves of deep learning networks,” (2025), [PLOS Computational Biology](#) **21**, e1012286 .
- [3] Y. Jiang, S. Lai, G. Lei, M. Betke, P. Ishwar, and D. Wijaya, “Exploring an alternative computational approach for news framing analysis through community detection,” (2025), [Journalism & Mass Communication Quarterly](#).
- [4] Y. Jiang, S. Liang, and J. Choi, “Synthetic survey data generation and evaluation,” (2025), [Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining](#).

- [5] Y. Jiang, R. Dale, and H. Lu, “Understanding human heuristics in context-sensitive image captioning,” (2025), [Proceedings of the Cognitive Science Society \(Vol. 47\)](#).
- [6] Y. Jiang, and G. Akçakır, “Explicit cooperation shapes human-like multi-agent LLM negotiation,” (2025), [ICWSM Workshop on Integrating NLP and Psychology to Study Social Interactions \(NLPSI\)](#).
- [7] Y. Jiang, “Emotions in Televised Debates: Proposing a multimodal deep learning approach for emotion detection,” (2025), [Partisan Rhetoric and Polarization: The Year in C-SPAN Archives Research, Vol. 10](#), pp. 213–236, Purdue University Press .
- [8] G. Miao, Y. Jiang, A. Pluta, R. Dale, F. Steen, and M. Lieberman, “A Deep Neural Network Approach for Integrating Neural and Behavioral Signals: Multimodal investigation with fNIRS hyperscanning and facial expressions,” (2024), [Proceedings of the Cognitive Science Society \(Vol. 46\)](#).
- [9] Y. Jiang, R. Dale, and H. Lu, “Transformability, Generalizability, but Limited Dif-fusibility: Comparing global vs. task-specific language representations in deep neural networks,” (2023), [Cognitive Systems Research](#).
- [10] G. Akçakır, Y. Jiang, J. Luo, and S. Noh, “Validating a Mixed-Method Approach for Multilingual News Framing Analysis: A case study of COVID-19,” (2023), [Computational Communication Research](#).
- [11] C. Shea, Y. Jiang, and W. Leung, “David vs. Goliath: Transnational grassroots outreach and empirical evidence from the #HongKongProtests Twitter network,” (2022), [Review of Communication](#).
- [12] S. Lai, Y. Jiang, G. Lei, M. Betke, P. Ishwar, and D. Wijaya, “An unsupervised approach to discover media frames,” (2022), [LREC Workshop on NLP for Political Sciences](#).
- [13] Y. Jiang, X. Jin, and Q. Deng, “Short Video Uprising: How #BlackLivesMatter content on TikTok challenges the protest paradigm,” (2022), [ICWSM Workshop on Images in Online Political Communication \(PhoMemes\)](#).
- [14] Y. Chen, Y. Shi, J. Luo, Y. Jiang, *et al.*, “How is vaping framed on online knowledge dissemination platforms?,” (2022), In: Thomson, R., Dancy, C., Pyke, A. (eds) *Social, Cultural, and Behavioral Modeling*. SBP-BRiMS 2022, vol 13558. Springer, Cham .
- [15] L. Xie, Z. Li, X. Ye, and Y. Jiang, “Environmental regulation and energy investment structure,” , (2021), [Technological Forecasting and Social Change](#)

CHAPTER 1

Introduction

What Does Multimodality Add Beyond Language?

Abstract

Language and perception operate over distinct representational regimes: language (especially in its textual form used to train models) tends to promote more discrete, symbolic processing, whereas perception and action are more continuous, gradient, and grounded in embodied experience. This distinction constrains how foundation models trained on each modality learn, generalize, and interact. Rather than focusing on downstream capabilities, this position paper examines multimodality through the lens of representation. These differences have been extensively studied in cognitive science, often framed as the debate between *symbolic* (e.g., LLMs) and *embodied cognition* (e.g., multimodal AI). Synthesizing insights from these traditions, I argue that language encodes substantial information about embodied experience through statistical regularities and compositional structure, but remains limited by its reliance on representations of prior experience. Incorporating multimodality provides access to continuous signals that support the construction of novel shared experiences, while also enhancing communicative clarity and enabling co-participation in human–AI interactions. From this perspective, multimodality is not merely an additive interface, but reflects fundamental differences in representational regimes that shape model intelligence.

1.1 Introduction

Recent advances in large language models (LLMs) demonstrate that language systems can exhibit surprisingly rich cognitive and reasoning abilities [248, 247, 41]. As foundation models increasingly integrate multiple modalities [48], a fundamental question remains: *how does multimodality contribute to intelligence beyond what can be acquired from language alone?*

At first glance, the benefits of multimodality appear straightforward. Integrating vision, audio, and other perceptual channels enables capabilities that are difficult to achieve through language alone, including spatial reasoning, perceptual grounding, and the generation of rich multimodal content such as images, speech, and music [20]. More recently, multimodality has become central to building world models—systems that capture physical dynamics and predict future states under different actions [99, 100, 133, 48]. These advantages are well-recognized and have driven rapid progress in multimodal AI systems.

In this Chapter, I take a complementary perspective. Rather than focusing on downstream capabilities or architectural design, I examine multimodality through the lens of representation, bridging AI advances with model explainability grounded in cognitive principles.

Language and perception are often characterized as operating over distinct representational regimes: language (especially in its textual form used to train models) tends to promote more discrete, symbolic processing, whereas perception and action are more continuous, gradient, and grounded in embodied experience [92, 11, 80, 266]. This distinction constrains how foundation models trained on each modality learn, generalize, and interact [173, 232, 111]. The differences between language and perception have been extensively studied in cognitive science and linguistics, often framed as the distinction between *symbolic cognition*¹, derived

¹To avoid ambiguity, it is important to distinguish symbolic cognition from classical symbolic AI. In the AI literature, symbolic AI typically refers to rule-based, explicitly programmed systems that operate over hand-specified symbols and logical rules, whereas neural or numeric AI refers to systems that learn distributed representations through gradient-based optimization [129]. In this review article, I use symbolic cognition to describe a representational property—the use of amodal, language-mediated representations—rather than a commitment to a particular computational paradigm.

from the amodal language system, and *embodied cognition*, which emphasizes the activation of multimodal perceptual systems [153, 130, 242, 38]. The distinctions drawn throughout this dissertation should not be interpreted as strict dichotomies. Language acquisition is deeply grounded in embodied and social experience, and linguistic representations encode substantial information about perception, action, and interaction accumulated across human communication and cultural use [153, 152, 79]. Likewise, perceptual modalities are not inherently “non-symbolic,” as grounded and continuous sensory signals can also become abstracted into discrete categories and structured representations, such as visual, auditory or kinetic semantics [260, 158, 154, 247, 120]. Thus, terms such as *symbolic*, *embodied*, *perceptual*, and *multimodal* are used here to describe differences in representational emphasis rather than mutually exclusive forms of cognition².

Synthesizing insights from these traditions, I argue that language encodes substantial information about embodied experience through statistical regularities and compositional structure, enabling efficient abstraction and generalization. However, it remains limited by its reliance on representations of prior experience. In contrast, perceptual modalities provide access to continuous signals that support the construction of novel shared experiences, while also enhancing communicative clarity and enabling co-participation in human–AI interactions (see Table 1.1 for a summary of the theoretical frameworks).

From this perspective, multimodality is not only about expanding the range of tasks that AI systems can perform, but about introducing qualitatively different forms of information that reshape how models encode and represent knowledge [186, 111, 16, 261, 51], generalize to new domains, and interact with other agent and humans [9]. Thus, my perspective redirects attention from multimodal capabilities as application-level features to multimodality as a representational and information-theoretic property.

This perspective has practical implications. Many current challenges—such as inconsis-

²The partially interchangeable use of terms such as *multimodal*, *embodied*, and *perceptual* in this chapter reflects how prior scholarship has often employed these concepts within broader debates surrounding symbolic and embodied cognition [46, 275]. More precise distinctions between these terms are summerized in the conclusion section of this chapter (Section 1.3).

tencies in cross-modal safety alignment driven by superficial modality-specific cues rather than shared representations [77], spurious correlations between visual and textual modalities in Visual Question Answering tasks [246], and divergent failure patterns in textual versus visual analogical reasoning [13]—can be understood as mismatches or imbalances in how information is represented and integrated across language and perception.

Addressing these issues requires not only better architectures and benchmarks, but a more precise understanding of modality-specific representations and their interactions. In the following sections, I first review relevant theoretical frameworks on *symbolic* and *embodied cognition*, examining how these representational differences manifest in contemporary language and multimodal models. The concluding section of this chapter synthesizes these theoretical perspectives and connects them to the empirical and methodological contributions presented in the rest of the dissertation.

Table 1.1: A theoretical comparison of symbolic and embodied approaches to AI, with key arguments presented in each paradigm.

Symbolic Cognition	Embodied Cognition
Language encodes sufficient embodied information to represent, communicate, and problem-solve.	Perception intertwines with language comprehension.
Key Arguments	
Language can isomorphically represent concepts through structured mappings to perceptual experience.	Embodied systems facilitate continuous learning and flexible reasoning at greater capacities.
Contemporary language models can extract and compose high-level constructs from linguistic input.	Multimodality enhances interpersonal communication for embodied agents.

1.2 Classic Debates on Symbolic and Embodied Cognition

The question of whether intelligence can arise from language alone or requires perceptual and embodied grounding has long been debated in cognitive science [130, 242, 38]. This debate centers on two contrasting perspectives: *symbolic cognition*, which views language as an amodal system where the meaning of a word can be purely derived from the statistical computation of its associations with other words, and *embodied cognition*, which contends that the comprehension of an item involves the activation of perceptual and embodied experiences [153].

Considering the classic example of a rose, symbolic cognition argues that the meaning of “rose” is derived from associations with words like “flower”, “red”, “thorny”, “romance”, and “love”. Embodied cognition posits that the meaning of “rose” is activated through its texture, color, scent, and the occasions on which people tend to perceive it [46, 275]. To distinguish these perspectives, scholars have used terms like “amodal”, “symbolic”, and “semantic” to describe the former system, and “(multi)modal”, “embodied”, and “perceptual” to describe the latter.

1.2.1 The prominence of symbolic cognition in language models

Before turning to the limitations of language-only systems, I first examine the strongest case for symbolic cognition in contemporary models. A long-standing line of work in cognitive science and computational linguistics suggests that language itself encodes rich information about the physical and social world [23, 76, 131, 117, 228].

This perspective rests on two core claims. First, linguistic representations are shaped by statistical regularities that reflect perceptual and experiential structure in the world, leading to an approximate isomorphism between language and physical concepts [153]. Second, recent advances in LLMs demonstrate that modern transformer architectures can extract and recombine compositional structure from text to form high-level constructs for advance reasoning abilities [88, 219].

1.2.1.1 Isomorphic mapping between language and the world

The symbolic interdependency hypothesis proposes that language inherently encodes embodied structure through the correspondence between linguistic and perceptual structures [153, 152]. This view holds that language evolves as a communicative shortcut for navigating the physical and social world, such that statistical regularities in text reflect regularities in perception and experience. As a result, language representations can serve as an effective proxy for perceptual understanding, and these correspondences can be recovered by statistical models [228].

Empirical evidence from language models increasingly supports this claim. Across languages, unsupervised machine translation can achieve precision levels exceeding 85%, suggesting that semantic spaces share approximately isomorphic structure [228]. More recently, Gurnee and Tegmark [79] demonstrate that Llama 2 embeddings can linearly map to representations of space (world, United States, and New York City) and time (historical figures, artwork, and news) across multiple scales. Cross-modal alignment has been observed in speech and written language [36], color perceptual encoding [1], grammatical gender biases [28, 272], and between language models (e.g., BERT, GPT-2, and OPT) and image models (e.g., ResNet-152) through weak linear realignment over limited image-text pairs [228].

These cross-linguistic and cross-modal findings support a reasonable belief that language representations are (near-)isomorphically to the physical world, such that mappings between linguistic and perceptual systems can emerge even in the absence of direct supervision (see Figure 1.1 for an illustrative example) [228, 160].

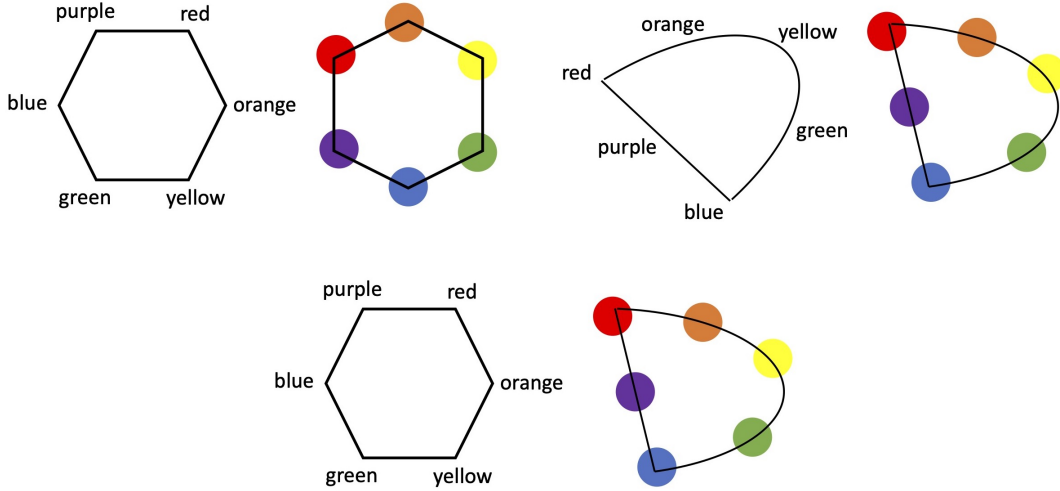


Figure 1.1: Three pairs of potential semantic and perceptual vector spaces indicate the possibilities of isomorphic mapping between the two systems. Upper left: The mapping is nearly impossible if the color terms are positioned in perfectly equidistant positions on a sphere (i.e., the linguistic and physical representations of colors have perfectly parallel distributions) [160], making any mapping unidentifiable. Upper right: The ideal distribution of semantic and embodied representations can facilitate one-to-one isomorphic mapping between the two. Bottom: The more realistic representational distribution in our current AI models and the real world—linguistic and embodied systems are not parallel but also not one-to-one. Each system situates with a unique positioning, with some amount of correlation between the two systems enabling realignment [228].

1.2.1.2 Compositional extraction in transformer models

Beyond representing individual concepts, language exhibits compositional structure—the ability to construct complex meanings from simpler constituents [129]. This property is central to human cognitive productivity, enabling an open-ended space of combinations from finite primitives [47, 17, 95, 129, 159]. Modern language models, particularly transformers, have demonstrated remarkable capacity for extracting and leveraging this compositional structure [88].

Evidence from cognitive modeling suggests that compositional representations emerge through hierarchical organization rather than mere word co-occurrence. Human language processing has been argued to rely on multilevel “chunk-and-pass” mechanisms, in which local dependencies are abstracted into progressively higher-level representations, from lexical units to phrases and discourse [34, 85]. This mechanism allows language to be quickly overwritten and interfered with by incoming new information during communication, supporting flexible comprehension and memory [34, 85, 63].

Artificial neural networks are particularly well suited to extracting this form of structure. A widely cited explanation for the success of deep learning is its ability to learn hierarchical representations through layered nonlinear transformations [134]. In vision, CNNs have been shown to learn representations that align with neural processing stages from early sensory to higher-level cortical areas [2, 165, 132]. In language, even simple recurrent networks capture aspects of compositionality by modeling long-range dependencies beyond local co-occurrence statistics [43, 33, 35, 59].

Among deep learning architectures, transformer models offer a particularly powerful substrate for compositional structure extraction. Previous work shows that different transformer layers correspond to distinct levels of linguistic processing in the human brain, ranging from phonemic and lexical representations to higher-level semantic and discourse features [88, 102, 74]. Layer-wise analyses further reveal temporal alignment between transformer representations and evolving neural activity during comprehension [74].

Importantly, language models trained on next-word prediction objectives show substantial alignment with neural and behavioral measures of human language processing [219, 74]. In particular, autoregressive models appear to capture aspects of incremental prediction and contextual sensitivity observed during human comprehension, consistent with theories that predictive processing plays a key role in language understanding [3, 230, 74]. Rather than implying that language models mechanistically implement human cognition, findings from brain–model correspondence studies suggest that predictive language modeling may provide a useful computational framework for understanding aspects of human language processing.

1.2.2 The necessity of embodied cognition via multimodality

Embodied cognition posits that language and other cognitive processes are inseparable from perception and action systems, as human cognition integrates symbolic representations with grounded experience [11, 46, 249, 187, 227].

Despite the near-isomorphic mapping between language and the external world [228], linguistic representations remain compressed abstractions of experience, omitting much of the perceptual detail present in real-world signals [92]. In contrast, perceptual modalities provide access to richer, more continuous information from which many abstractions are derived [11, 80, 266].

Consider a scenario in which we need to identify who is seated to the left of a person wearing a yellow shirt. Solving this problem purely in symbolic terms—by specifying object attributes, spatial coordinates, and relational constraints—requires discretizing what is inherently a continuous spatial configuration. In doing so, important relational structure may be lost or only approximately captured (e.g., small angular differences or partial occlusions). By contrast, human perceptual representations preserve the continuous organization of the scene, allowing spatial relations to be directly accessed rather than reconstructed.

This asymmetry suggests that some aspects of embodied experience may be difficult to fully specify through language alone, particularly when fine-grained perceptual, spatial, motor, or affective information must be compressed into discrete symbolic descriptions [71]. One way to formalize this view is to treat discreteness as a special case of continuous representation [190], suggesting that novel or fine-grained knowledge may exist outside the space of discrete linguistic encoding [232].

Multimodality therefore does not merely expand the range of tasks that intelligent systems can perform, but broadens the representational resources available for learning and interaction. In particular, embodied signals support (1) the acquisition of non-verbalizable, experience-based knowledge and (2) richer channels of information transmission that enhance interpersonal coordination.

Notably, the language–perception asymmetry discussed here is bidirectional. As the previous section highlighted, language enables forms of abstraction, categorization, and compositional generalization that may be difficult to derive from perceptual experience alone [156]. Linguistic symbols efficiently compress and communicate relational structure across contexts, allowing humans to reason about hypothetical, temporally distant, or non-present entities in ways that raw perceptual signals do not directly support [60, 277, 115]. In this view, language and embodied perception provide complementary rather than competing representational resources.

1.2.2.1 Embodied knowledge in multimodal systems

Despite language being approximately isomorphic to the external world, not every concept is explicitly verbalized in human communication. One shortcoming of symbolic system is that most knowledge initially develops and is acquired through embodiment [182, 183]. Creating an intellectual agent reliant solely on language constrains its ability to “learn-to-learn” concepts rooted in embodiment. Throughout our developmental stages, I progressively grasp the idea of more abstract concepts, such as gravity and inertia, out of more primitive perceptions, like motion and stability [57]. More specifically, understanding the concept of a “chair” relies on an understanding of seating [161]. The difficulty of achieving one-to-one cross-language mapping arises because language is deeply rooted in sensorimotor or bodily processes that shape how I perceive and construct the world [206].

There is ample evidence from neuroscience and behavioral studies that the perceptual and motor systems regularly interact with the brain for cognitive development [11, 206, 44]. Studies show that processing action words (e.g., “kick,” “pick,” “lick”) activates somatotopic regions of the motor cortex corresponding to the relevant body parts [82]. Physical properties like shapes or gestures can only be fully comprehended through motor experience. [245] demonstrate that I distinguish the shape between a “circle” and an “ellipse” based on how I draw them (i.e, a circle involves a constant movement, whereas an ellipse involves slowing down at the extremities). This body of evidence suggests that conceptual representations

are not purely symbolic but inherently multimodal, with meaning emerging from integration across perceptual and motor systems [128].

Furthermore, embodied knowledge often remains implicit—not easily articulated or formalized. Humans routinely rely on physical intuition to navigate the world, forming judgments without explicit symbolic reasoning by drawing on accumulated perceptual and motor experience [56, 55]. Such intuitions (e.g., recognizing that certain car shapes suggest speed or that specific animal features signal danger) are readily deployed but difficult to articulate, and attempts to formalize them often reveal inconsistencies [71]. Even when physical laws can be stated explicitly, humans frequently solve intuitive physics problems by leveraging prior perceptual–motor encounters (i.e., exemplars), without explicitly understanding or verbalizing the underlying physical rules (i.e., the symbolic system) [97, 65]. This suggests that much of our physical understanding is encoded in perceptually grounded representations derived from exemplars rather than crystallized in language [96, 98]. While language models can learn statistical regularities in textual descriptions of physical concepts [247, 79], human cognition may also rely on perceptual, motor, and other multimodal experiences that are only partially articulated through explicit language [94, 205].

1.2.2.2 Enhancing interactive capacities through multimodal communication

As AI systems become more embedded in interactive environments, their role extends beyond problem-solving to “interpersonal” and collaborative settings [9, 14, 224], transcending distinctions between different types of actors, whether human–AI, AI–AI, or hybrid interactions. This introduces a qualitatively distinct dimension of intelligence: interpersonal coordination, which relies heavily on multimodal cues beyond language.

Human communication involves complex co-creation processes for effective message encoding and decoding. While speaking, our gestures and gazes constantly inform our conversation partner about where we are directing our attention as we share linguistic content [43]. We modulate volume, pitch, and speech rate during verbalization to draw attention. Multimodal channels enhance the decoding accuracy of the fleeting communicative events,

serving as a backup plan in case any of the channels fail to effectively deliver the message, thus preventing repetition and misunderstandings [123]. Additionally, they optimize the amount of information that can be shared concurrently to maximize time efficiency [122].

Similarly, embodied agents can leverage multimodal channels, such as gaze and micro-gestures, to signal attentional focus and communicate intent more effectively, drawing on shared nonverbal social cues and reference resolution to support cooperative interaction. Although language is highly structured and explicit, with innate syntactic compositionality that facilitates abstraction and short-term memory, it also imposes a relatively high cognitive load compared to other motor signals. Combining this linearly structured, efficient channel (i.e., language) with less stable but complementary nonverbal signals enables multimodal AI systems to coordinate coherent courses of action and dynamically assemble composite agents [62, 43].

Current language-only systems face persistent challenges in inferring user intent without access to nonverbal and contextual cues that humans routinely use during interaction [21, 9, 42, 18]. This limitation does not imply that text-based communication itself is ineffective: humans routinely communicate successfully through text-only channels. However, humans bring substantial embodied, social, and situational priors accumulated through multimodal experience, as well as access to contextual information that language models often lack, allowing them to compensate for missing cues and infer implicit intent from sparse linguistic signals [40, 9]. Without shared common ground grounded in co-presence, language models often struggle to select appropriate levels of detail, pace responses to user needs, or align with users’ implicit expectations [9].

More broadly, situated context constrains and clarifies communicative intent in ways that are difficult to fully encode in symbolic form [22, 18, 9, 42, 110]. Nonverbalizable social heuristics—such as turn-taking, reciprocity norms, and longer-term relational expectations—are rarely explicitly scripted in interpersonal interaction and instead are shaped through embodied, multimodal interaction, where timing, contingency, and responsiveness across encounters play an essential role [4, 21]. Shared environments further condition interaction: conversa-

tional norms differ across settings such as business meetings versus informal gatherings, and this contextual information is directly available to agents co-present in physical space [22, 14]. Without access to such situational grounding, LLMs face inherent limitations in interpreting intent and appropriately constraining interaction.

1.3 Conclusion

This dissertation addresses a central question: *why do models and humans differ in their competencies in language and perceptual processing, given the distinct properties of these signals, and how does this asymmetry affect downstream human–AI interaction?*

This introductory chapter (Chapter 1) advances a representational perspective on multimodality, arguing that language and perception contribute fundamentally different forms of information to intelligent systems: language is discrete and symbolic, whereas perception is continuous and grounded in embodied experience. Rather than treating multimodality as a set of capabilities, this perspective emphasizes that it provides access to continuous signals that support non-verbalizable, experience-based knowledge and enhance communication through richer channels.

Following the conceptual framework introduced in Chapter 1, this dissertation uses terms such as *symbolic* and *embodied* to describe differences in representational emphasis rather than strict cognitive dichotomies. In this context, *symbolic* primarily refers to representations mediated through language and discrete compositional structure, particularly in the textual training data used in modern language models. By contrast, *embodied* refers to representations grounded in perceptual and action-oriented signals that are typically more continuous, temporally correlated, and situated in sensory experience. Although perceptual inputs are ultimately discretized during model training (e.g., pixels, patches, or tokens), they originate from continuous physical and sensory processes that retain richer spatial and temporal structure than textual representations alone. Similarly, *multimodality* refers to systems that integrate information across representational formats, including language, vision, audio,

and behavioral signals. Therefore, multimodal systems are not simply adding modalities, but combining complementary forms of information that may compensate for limitations in any single representational channel, thereby providing richer and relatively more continuous representations than language-only systems.

The theoretical arguments presented in this chapter provide a foundation for investigating multimodal intelligence in both computational systems and human–AI interactions. The following chapters combine methodological and empirical approaches to examine how differences between linguistic and perceptual modalities shape model behavior and human interaction with multimodal AI.

Chapter 2 introduces a model explainability framework to characterize differences in the signal properties of linguistic and perceptual inputs. Using sequential models (e.g., recurrent neural networks), it traces learning trajectories across tasks such as sentence and gesture classification. This approach reveals qualitative differences in how models process and learn from discrete linguistic signals versus continuous behavioral signals.

Building on these differences, Chapters 3 and 4 examine multimodal tasks in naturalistic human-AI interaction contexts. While embodied cognition broadly encompasses perceptual, motor, social, and kinesthetic processes, the empirical chapters of this dissertation focus primarily on language-perception interactions—particularly photographic imagery in news contexts—as tractable and ecologically grounded case studies for examining how different representational regimes shape learning, generation, and evaluation in humans and AI systems. Chapter 3 investigates image captioning, showing that while models exhibit strong linguistic fluency, they remain limited in perceptual and cross-modal reasoning. As a result, cross-modal context mitigates linguistic heuristics that would otherwise bias human judgments under text-only evaluation.

Chapter 4 extends this analysis to image generation, identifying a phenomenon of loosely coupled image–text alignment that is common in real-world multimodal communication but often overlooked in training paradigms. This looseness gives rise to *Visual Narrative Freedom*, whereby text-to-image systems can generate diverse and plausible visual interpretations.

Under such conditions, generated images can become highly persuasive to human observers, even in high-stakes domains, despite known stylistic limitations relative to real images.

Focusing on language–perception systems as a starting point provides important methodological advantages for empirical multimodal AI research. Vision-language tasks enable relatively controlled experimental designs with scalable multimodal datasets, standardized evaluation settings, and high-fidelity stimuli enabled by recent advances in foundational vision-language and image generation systems. These properties make language–vision interaction a scalable and ecologically grounded domain for studying broader questions about multimodal and embodied intelligence. At the same time, the implications of the findings presented in Chapters 3 and 4 primarily concern perceptual modalities, particularly vision, rather than the broader range of embodied processes, including motor, social, auditory, and kinesthetic dimensions. The broader implications of this work are intended to motivate future research on multimodal intelligence beyond language–vision systems alone.

Together, this dissertation demonstrates that multimodal intelligence is not merely an additive interface, but is fundamentally shaped by the structure of the representations through which information is encoded. From a model perspective, this work provides a cognitive science–grounded and information-theoretic account of why models acquire competencies differently from language and perceptual data, and why gaps persist between perceptual and language tasks. From a human–AI interaction perspective, this work highlights both vulnerabilities—such as reliance on linguistic cues—and strengths, including robustness to cross-modal information, as well as the conditions under which generative models can exploit loosely coupled image–text relationships in real-world multimodal communication. Overall, this work bridges cognitive science and artificial intelligence by providing a unified account of how differences in linguistic and perceptual representations shape model behavior and human–AI interaction.

CHAPTER 2

A Model Explainability Approach for Comparing Linguistic and Behavioral Learning

Abstract

There is an important challenge in systematically interpreting the internal representations of deep neural networks (DNNs). Existing techniques are often less effective for non-tabular tasks, or they primarily focus on qualitative, ad-hoc interpretations of models. In response, this chapter introduces a cognitive science-inspired, multi-dimensional quantification and visualization approach that captures two temporal dimensions of model learning: the “information-processing trajectory” and the “developmental trajectory.” The former represents the influence of incoming signals on an agent’s decision-making, while the latter conceptualizes the gradual improvement in an agent’s performance throughout its lifespan. Tracking the learning curves of DNNs enables researchers to explicitly identify the model appropriateness of a given task, examine the properties of the underlying input signals, and assess the model’s alignment (or lack thereof) with human learning experiences. To illustrate this method, we conducted 750 runs of simulations on two temporal tasks: gesture detection and sentence classification, showcasing its applicability across different types of deep learning tasks. Using four descriptive metrics to quantify the mapped learning curves—*start*, *end* - *start*, *max*, *t_{max}*—, we identified significant differences in learning patterns based on data sources and class distinctions (all p ’s $< .0001$), the prominent role of spatial semantics in gesture learning, and larger information gains in language learning. We highlight three key insights gained from mapping learning curves: *non-monotonic progress*, *pairwise compar-*

isons, and *domain distinctions*. We reflect on the theoretical implications of this method for cognitive processing, language models and representations from multiple modalities ¹.

2.1 Introduction

Over the past decade, deep learning and neural networks have achieved remarkable performance in prediction and classification tasks in various domains, from machine translation and object recognition, to autonomous driving and reinforcement learning [134, 218]. As powerful representational learning tools, deep neural networks (DNNs) can capture complex patterns in data [134], yet understanding the nature of the information embedded in their multi-dimensional representations remains a challenge. Researchers have raised concerns about how DNN embeddings represent knowledge and how to holistically interpret these high-dimensional features [251, 141, 176, 215]. Over-reliance on such models for decision-making could be detrimental in both research and applied settings due to their complexity and lack of explainability. This is particularly concerning when models rely on biased training data that does not generalize well to target tasks [167, 243]. Without a deep understanding of these models' underlying properties, they may fail to align with the goals of their human designers.

Recently, there has been a combined effort from cognitive science and deep learning to utilize representational learning to address model explainability concerns. These practices have become more prominent due to the success of large-scale models [144], especially large-language models [25]. For example, in the case of language models, examining the internal processes of the BERT Transformer-based architecture has shown that it may recapitulate common natural language processing (NLP) pipelines [207, 237, 238]. Chang and Bergen [26] found that the frequency and n-gram structure of word tokens significantly alters the training for language Transformer models learning these words.

¹All supplementary materials can be accessed at <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012286>.

Inspired by this prior work, this paper outlines a technique for examining the learning trajectories of deep learning models, in particular recurrent neural networks (RNNs). There is historical precedent for our approach, too. McClelland, Rogers and others have studied the underlying knowledge of neural networks by tracking them as they learn [162, 164]. Despite this classic work in cognitive science, it is uncommon to see deep learning models that track progress as a way to unpack what is learned (e.g., going beyond simple RMSE curves; but see also [25], for a counterexample). We term this tracking a “learning curve,” as it resembles research on how human learners process incoming information and improve decision-making through iterations under different situations. A unique benefit with simulation is the possibility to examine many dimensions and measurements of the neural network over time. In the next section, we review recent work on interpreting DNNs and related models. We then introduce our approach based on learning curves.

2.1.1 Background

To date, various techniques have been proposed to interpret DNNs. For example, many model interpretability techniques provide task-specific and local explanations, such as saliency maps, attention maps, or Layer-wise Relevance Propagation (LRP), which interpret or visualize the localized influence of a region on the output. These ad-hoc approaches can be unstable, as even a minor change in a single pixel or hyperparameter can substantially affect the local relationships between input signals and output data [145]. Additionally, these local explanations fail to offer a global understanding of whether the selected architecture is well-suited to the task or how the DNN models experience learning. On the other hand, feature attribution methods like SHAP [203] and LIME [155], while conceptually easier to interpret and model-agnostic, tend to work more intuitively with interpretable dimensions typically found in structured tabular data. These methods are less effective for unstructured tasks, such as image and audio data, where feature dimensions are harder to define or interpret. While most techniques emphasize the qualitative interpretation of a model, such as what has been learned or captured by the DNN, the absence of quantifiable measurements makes

cross-model comparisons in unstructured tasks particularly challenging [262, 141, 176]. We further provided a systematic description of different model explainability methods, along with their pros, cons, and use cases, as well as a comparison to our proposed approach in Table 2.1.

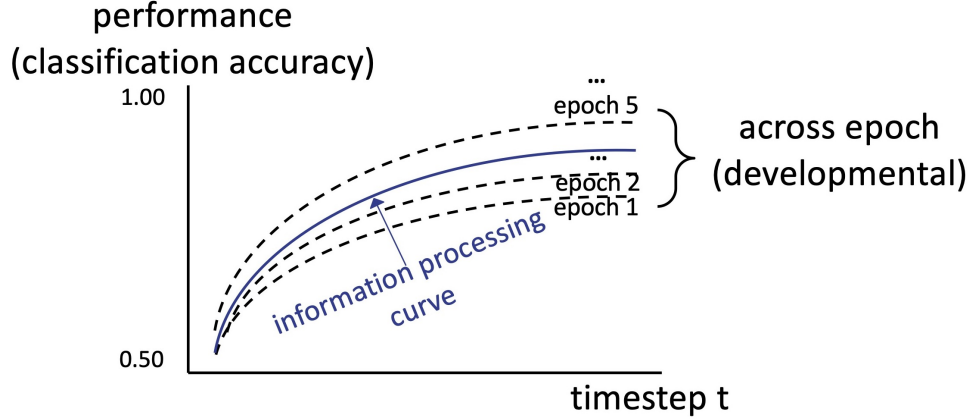


Figure 2.1: **A conceptual illustration of learning curves in DNNs.** The solid curve represents an individual information processing trajectory (across timesteps). For example, as an agent receives more signals over time in one session, its prediction of the gesture increases. The bundle of dotted lines represents a developmental trajectory (across epochs). For instance, this agent improves its ability to predict incoming gestures after repeatedly encountering similar patterns. In a typical binary classification task, the starting point at $t=0$ is expected to be near 0.50 (and gradually increase to 1.00 across timesteps and epochs) for all epochs because no signals have been provided at the first timestep for solving the underlying task. Hence, all information processing curves in this figure have the same starting point to reflect this pattern. In practice, certain classification tasks might contain structural information even at $t=0$ (e.g., gesture classification contains spatial information, such as the location of keypoints, at the initial timestep, which could help solve the gesture task from the very beginning).

Table 2.1: **DNN Explainability Methods Comparison.** A systematic overview of different DNN explainability methods, along with their comparison to the proposed learning curve approach.

Method	Description	Advantages	Limitations	Use Cases
Learning Curve (Proposed Method)	Track the model’s learning experience over time by utilizing classification performance of embeddings.	Captures changes in embeddings during training. Insight into performance evolution across two temporal dimensions.	Limited to temporal tasks and DNN architectures. Computationally expensive.	Tasks involving temporal aspects and learning experiences.
Saliency Maps	Visualizes regions in image input that most influence the output.	Simple and intuitive visual explanation. Easy to implement.	Prone to noise. Lacks global interpretation.	Image recognition tasks and image-based models.
Feature Attribution	Infers relevance of input features to output.	Conceptually easy to grasp. Model-agnostic approach.	More suited for structured data. Less intuitive for unstructured data.	Tabular data and structured tasks.
Layer-wise Relevance Propagation (LRP)	Breaks down classification decision into contributions of input elements.	Applicable to a wide variety of DNN models.	Lacks global interpretability. Requires complex configuration.	Image recognition, NLP, structured data.

In a foundational article on deep learning, LeCun and the colleagues [134] characterized DNNs as representation-learning methods utilizing multilayered large neural network-style models. Representations can be viewed as mental objects capturing semantic properties either observable or unobservable [192]. Different from traditional machine learning models, DNNs display remarkable flexibility and efficiency in encoding lower-level input signals, such as pixels, audio frequencies, or word tokens, into multidimensional vectors at a sophisticated level through multilayered nonlinear transformations [214]. These transformations generate multiple levels of representations that learn hierarchies of features at each layer [134]. The continuous numerical vectors (or hidden vectors) learned at each level are commonly referred to as “embeddings” and serve as dense representations of the original input data. Following this construction, numerous studies have demonstrated the correspondence between DNN-generated and real-world distributed representations among words and sentences [74, 102], speeches [75], images [255], objects [103] and scenes [78, 226]. Representational learning in DNNs offers fundamental contributions to cognitive science, as it can inform how cognitive systems process and organize knowledge, facilitating the comparison of learning processes between humans and machines [87, 86, 112].

Recently, several studies in cognitive science and neuroscience have highlighted the importance of integrative modeling between computation, human brains and behaviors [219]. Beyond comparisons of static end-point knowledge, scholars have begun exploring the potential correspondence in learning and information processing between DNNs and human cognitive systems, given that the current performance of DNNs can already approximate human performance across various domains [81, 233, 276] (see [25] for a review). For instance, the representations (i.e., embeddings) extracted from multilayered DNNs have shown significant accuracy in predicting neural and behavioral responses in humans throughout the hierarchy of learning and processing. This evidence spans multiple neural-behavioral measurements (e.g., fMRI, EEG, ECoG), modalities (e.g., visual, auditory, and language processing), and model architectures (ranging from simple embedding models like GloVe to more complex neural networks such as RNNs, convolutional neural networks (CNNs), and transformer

models (for further details, see [164, 255, 74, 219, 30, 229]).

Given the extensive alignment observed between DNN embeddings and neural-behavioral activities, and their presumed meaningful representational mapping with human cognitive systems, an analysis of how they emerge in learning would seem important to understand these relationships. Goldstein and colleagues [74] identify the temporal correspondence between layer-by-layer embeddings in GPT-2 and evolving neural activities in language areas. However, this temporality is restricted to layerwise representations (from low-level to high-level representations) rather than how streams of signals have been received and processed by models or brains [276]. Our aim in this paper is to use temporal analysis in a systematic way by separating and tracking the time course of a network’s learning across classification tasks, thereby enhancing the understanding of the emergence of representations in DNNs.

2.1.2 The current study

The goal of this chapter is to unpack the “learning curve” of DNNs through a sequence of hidden representations when the model encounters any temporal processing tasks across its training. To do so, we sample the network’s performance by using its embedding vectors to classify groups of items in its training input. This allows us to map out the progression of the network’s discriminations across these groups – how the network’s internal knowledge, in the form of embedding vectors, evolves during training.

The model architecture we focus on is RNNs due to their capacity to model sequential data and time-dependent tasks [236], such as text generation, speech recognition and stock market prediction. Although other deep learning architectures, such as CNNs and Transformers, also have the capacity to process time-series data, RNNs are explicitly designed for processing sequential data, as they effectively capture temporal dependencies through their recurrent connections. CNNs excel in tasks such as image recognition by applying the kernel trick, where convolutional filters are used to extract spatial features from local regions of the input. While CNNs can be adapted to handle sequential data using techniques like 1D convolutional layers, they lack the inherent ability to capture temporal ordering in the data,

as they process each segment of input independently. Transformers, on the other hand, rely on self-attention mechanisms to weigh the importance of different input tokens in parallel, allowing them to capture dependencies between distant elements (i.e., tokens). This feature makes Transformers better suited for tasks where relationships between tokens are not strictly ordered in time, and their parallel processing nature is less ideal for learning tasks that require explicit temporal progression. In sum, the strict sequential processing nature of RNNs makes them an intuitive architecture for studying cognitive processing that involves dynamic, changing contexts [114].

In this chapter, we propose a generalizable interpretability approach that maps the global learning curves of RNNs based on changes in classification performance between embeddings and output across timesteps of the input data. This performance reflects the predictive capacity or the amount of signal captured by the embeddings in predicting the final output. In RNN modeling, each timestep generates an embedding (e.g., in language processing, each timestep represents a word). These embeddings progressively incorporate information from the beginning up to timestep t . By examining changes in classification performance between these embeddings and the output, we can capture how the information evolves over time. To provide clarity, we use the term “learning curve” in this chapter to denote the holistic approach and intention behind mapping the underlying processing and developmental journey of DNNs. The method we propose separates two parts of the learning curve, one based on overall training, and another based on processing within input items during training. First, the “developmental trajectory” signifies the long-term learning process of DNNs, which is simulated by the increasing number of epochs (i.e., complete passes through an entire training dataset). Second, the “(information) processing trajectory” refers to the momentary accumulation of information across all timesteps (within an epoch), extractable from the performance of the RNN layer (see Fig 2.1 for a conceptual illustration). We detail each of these further below.

As Fig 2.1 illustrates, the resulting visualization consists of learning curves with timesteps on the x-axis and performance on the y-axis. Each epoch contains an information process-

ing trajectory, and the set of curves together forms a developmental trajectory. Due to the multidimensional nature of this plot, while it provides rich qualitative information about the entire model learning experience, it can be challenging for researchers to comprehend all the constructed curves at once. Therefore, beyond the visual presentation and qualitative inspection of the multi-dimensional learning curves, this chapter further defines four measures: start performance (*start*, the initial capacity of each information processing), max performance (*max*, the maximum performance of each information processing), time at max (t_{max} , when the current information processing reaches the maximum performance), and end – start performance ($end - start$, the overall performance gain in this information processing) – to facilitate quantitative comparisons across tasks and datasets. Our study particularly focuses on these relatively easy-to-comprehend descriptive statistics (as opposed to more complex metrics like regression coefficients) to simplify the understanding of the already multidimensional and granular nature of the constructed learning curves.

To illustrate the generalizability of our approach, we chose two distinct classification tasks: (i) sentence classification and (ii) gesture detection. These tasks differ widely in terms of their modalities. We predicted this would lead to variation in the underlying data generation processes and associated cognitive processing for each modality. In particular, gesture and body movements primarily result from the coordinated contraction and relaxation of muscles, with signals produced at later timesteps derived from the previous timesteps with relatively high autocorrelation [16]. On the other hand, verbal language, being a predominantly semantic modality, exhibits degrees of surprisal and arbitrariness that enhance the cognitive capacity of language processing [261, 51, 191]. Therefore, the expected developmental and information processing trajectories will likely exhibit distinguishable patterns across the two different tasks when the DNN system processes them respectively.

As we discuss in detail in later sections, this method has a few benefits. Through tracking the learning trajectory of a neural network, researchers can explicitly identify the appropriateness of a model for a given task as well as examine the properties of underlying input signals. This approach could also serve as a standalone visualization to map the accumu-

lation of the underlying signals processed, which can facilitate research on deep learning modeling and signal processing across various modalities. Finally, mapping the learning curve of DNNs has the potential to assist future computational cognitive and neuroscience research and address whether the learning experiences of models also correspond to (or fail to correspond to) the temporal processing in human cognition in addition to the emphasis on static representations in the current literature.

In the following sections, we will provide a step-by-step method for visualizing the learning curves of neural networks, illustrate how to holistically interpret signal processing in them and quantitatively compare these curves across two different datasets.

2.2 Methods

This research proposes a model-interpretability method that can extract the learning curve of sequence-based deep learning networks (e.g., RNNs). Inspired by cognitive science, the method measures the learning trajectory and underlying knowledge extracted by such networks. To illustrate the method, we use two temporal tasks: gesture detection and sentence classification as examples. this chapter therefore demonstrates that the method could inform a range of deep-learning tasks.

Our multi-stage pipeline includes three main steps. First, we trained the RNN-LSTM model to generate a sequence of embeddings for a temporal task. Next, we used KNN classifiers to calculate the performance across all extracted embeddings, which we refer to as learning curves (see (2) in Fig 2.2). Finally, we defined four metrics to quantify these multi-dimensional embeddings for more systematic comparison and significant tests between tasks, classes, the null hypothesis, and development trajectories. The visualization of this multi-stage procedure can be found in Fig 2.2.

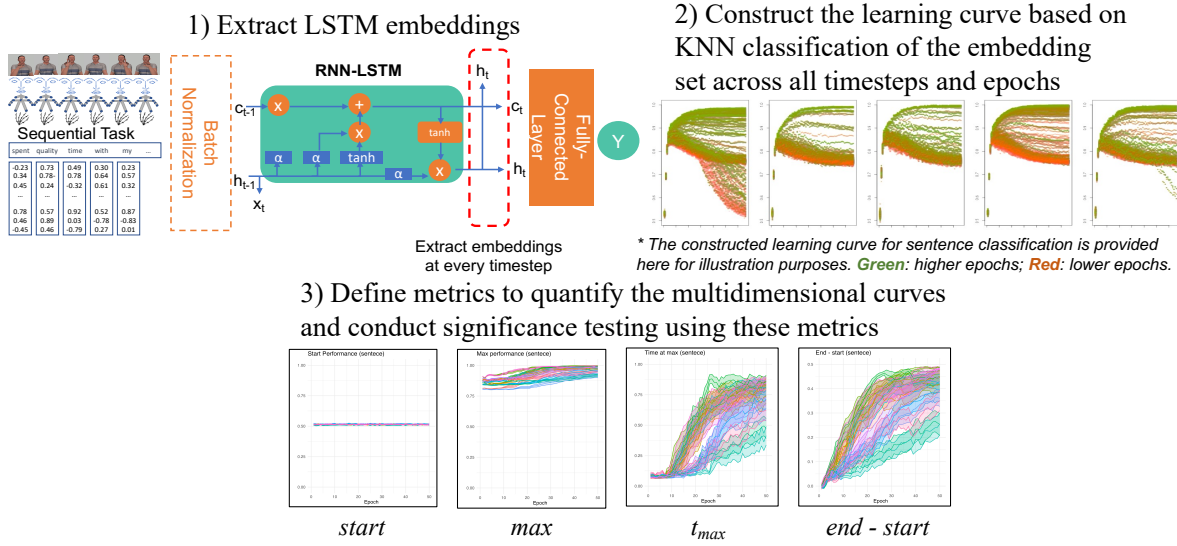


Figure 2.2: **A flowchart of the multi-stage pipeline.** Illustrations of the multi-stage pipeline, from extracting embeddings to constructing learning curves and quantifying the curves using four defined metrics. NB: Gesture figures are adapted from Wikimedia Commons [223].

2.2.1 Datasets

this chapter utilized two datasets: the Identity-free Video Dataset for Micro-Gesture Understanding and Emotion Analysis (iMiGUE) by [150] and the Emotion dataset from Hugging Face by [216] to examine the possibility of mapping learning curves for tasks involving temporality. The two temporal tasks (i.e., gesture detection and sentence classification) are distinct in terms of their modalities, lengths, and the steps required to extract and preprocess the features, thereby enhancing the diversity of data to illustrate the learning curve analysis. We detail the preparation of each dataset separately below.

2.2.1.1 Emotion tweet classification (sentence classification)

The Emotion dataset [216] consists of 20,000 English Twitter messages with six basic emotions (e.g., anger, fear, joy, love, sadness, and surprise) by adopting Plutchik’s [193] wheel of emotions, Ekman’s [58] six basic emotions, and hashtags in tweets. Tweets were annotated through noisy labels and distant supervision introduced by [73].

To prepare the emotion sentences for recognition by the RNN layer, we first applied the “basic English” tokenizer from torchtext to tokenize each tweet. Then, we used GloVe (Global Vectors for Word Representation), a word vectorization technique that does not rely on local word context statistics (local-context information), to vectorize each token in a sentence. GloVe was preferred over other token vectorization techniques like word2vec [171] due to its design to capture the universal meaning of each token/word, rather than the word’s meaning within a specific sentence or context. We opted for GloVe embeddings with 300-dimensional semantic features because it strikes a balance between capturing sufficient information and maintaining computational efficiency [112].

In theory, the input sequences of RNN are not required to have the same length. In practice, these sequences are padded with zeros or trimmed to the same length to optimize the computation in PyTorch. Accordingly, all tokenized tweets were padded to a consistent length of 66 tokens, which corresponds to the length of the longest tweet sample. The shape of each Emotion tweet follows a 300 dimensions $\times$ 66 timesteps (see Fig 2.3).

2.2.1.2 Gesture classification

The iMiGUE is a high-quality dataset that contains 18,499 identity-free, ethnically-diverse, gender-balanced samples of 32 psychologically-meaningful micro-gestures (such as scratching an arm, adjusting the hair, or touching an ear [58]). These gestures were all curated from interview clips with athletes at post-match press conferences. Unlike other gesture and emotion datasets, which are typically drawn from staged performances or movie clips, the iMiGUE provides samples of actual gestures from real-life situations. This poses a more

realistic, though more challenging, recognition task for deep learning networks [264, 52].

Due to copyright restrictions, the dataset includes only skeleton keypoints (rather than original interview clips) extracted from OpenPose, a multi-person computer-vision system that can simultaneously extract keypoints of the body, hands, face, and feet [24]². In total, 25 body, 70 facial, and 21 left hand and 21 right hand keypoints were extracted for each frame using OpenPose. Keypoint data were stored in the following format: $[x_0, y_0, c_0, x_1, y_1, c_1, \dots]$, in which (x, y) represents the coordinate of each keypoint and c indicates the confidence score of each keypoint-coordinate prediction. The confidence score was excluded in the following processing steps.

Although the gesture clips have just 39 frames on average, they have a higher standard deviation at 84 frames with a 75th percentile of 72 frames. We therefore set the padded length of the input sequence to 150 frames to ensure that our input data contained sufficient information for training a classifier. Thus, the input size for each gesture clip was 274 units $(137 \text{ keypoints} \times 2) \times 150 \text{ timesteps}$ (see Fig 2.3).

²The advantages of using key point data are twofold. First, keypoint data conserves significant computing power; since each second of image sequences (i.e., matrices of pixels) may contain as many as 30 or 40 frames, running deep learning models on image sequences can be prohibitively expensive. Second, it provides better interpretability and understandability. Instead of being possible only through abstract information at the image-frame level, gesture detection can be operationalized as a sequential movement in keypoints across frames [113].

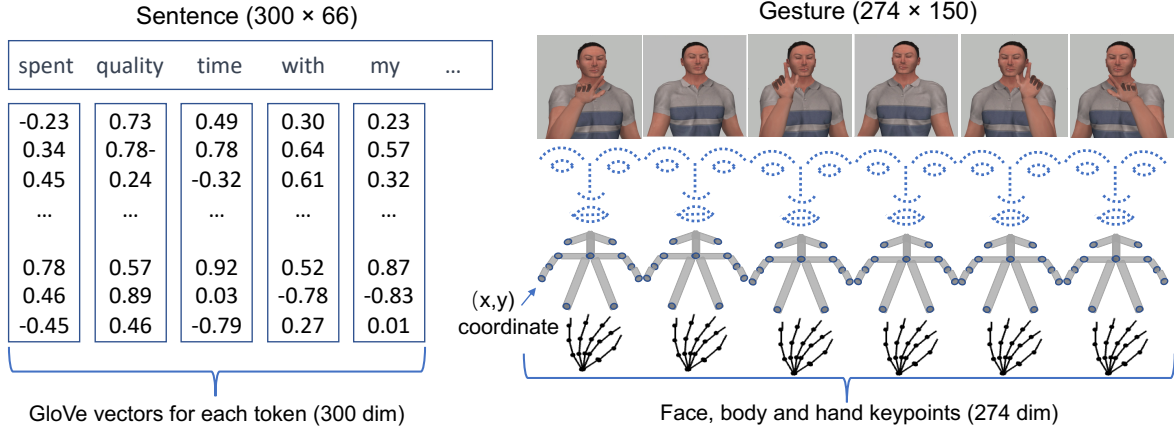


Figure 2.3: Sequential data for Emotion Tweet (left) and Gesture Classifications (right). NB: Gesture figures are adapted from Wikimedia Commons [223].

It is important to note that our aim here is not to approach benchmark performance, but rather to examine successful learning. The learning curve analysis will show how that successful learning emerges, and which stimulus discriminations seem to underlie that emergence. Since even the state-of-the-art neural network can achieve only 55% accuracy on this multi-classification task [150], we further grouped the 32 micro-gestures into six categories (body, head, hand, body-head, head-hand movements and an absence of gestures) to ensure that our RNN-LSTM was indeed “learning” when we attempted to map its learning curve.

2.2.2 Model architecture

We selected a model architecture (see Fig 2.4) that is standardized for sequential stimulus processing in the following way. First, we had a batch normalization layer, a primary RNN-LSTM layer, and then a fully connected layer to connect the final dense embeddings with the output classes (for classification). We offer details below.

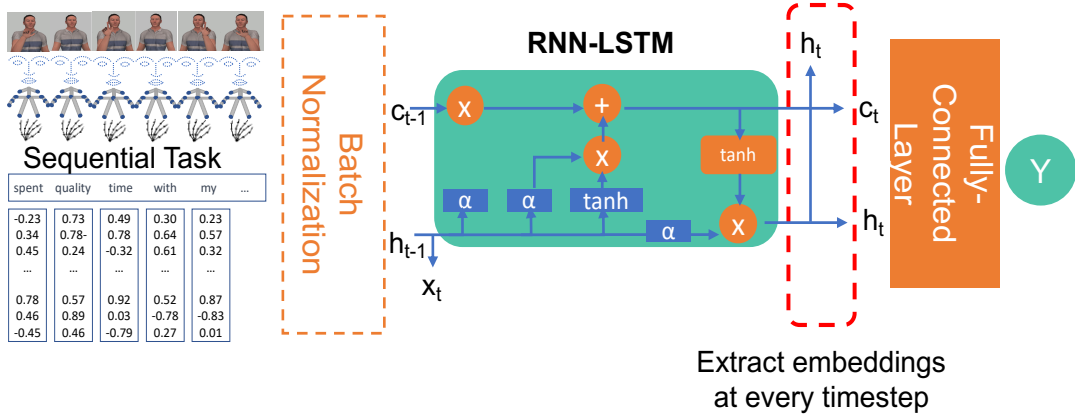


Figure 2.4: Model architecture of RNN-LSTM. NB: Gesture figures are adapted from Wikimedia Commons [223].

2.2.2.1 RNN-LSTM

RNN is capable of modeling sequential data and time-dependent tasks [236]. Its architecture represents an iterative function that takes an input sequence (x) and an internal state (h) from the previous timestep ($t - 1$) to predict the current timestep (t), then updates the state as follows:

$$h_t = f(x_{t-1}, h_{t-1}) \quad t = \{0, 1, 2, \dots, T - 1\} \quad (2.1)$$

As the formula illustrates, each timestep t should theoretically reflect the information from 0 to $t - 1$. We selected the RNN model for gesture detection because it can process temporal information under the assumption that the body movement in each timestep depends on signals in the previous timesteps.

While an RNN could leverage the context between elements by maintaining its internal state while processing the entire sequence, the “Vanilla RNN” layer experienced the vanishing-gradient problem during model training [222]. Therefore, long short-term memory

(LSTM), which is represented by the function f in Equation 2.1, was introduced here as an additional state variable, called the cell state, for controlling specific information that needed to be kept or updated while processing the entire sequence [113]. LSTM effectively reduced the vanishing-gradient problem encountered by RNN [222].

2.2.2.2 Feedforward network

The construction of our RNN-LSTM neural network follows common practice in deep learning. First, to enhance training stability and speed up convergence, we applied batch normalization, specifically the BatchNorm1d layer, to the input data at each timestep, with dimensions $\text{batch_size} \times \text{timestep} \times \text{input_dimension}$ ($1 \times 150 \times 274$ for iMiGUE and $1 \times 66 \times 300$ for Emotion). This means that each timestep is treated independently, and normalization is applied across all data points in the batch for each timestep. Specifically, this technique normalizes the inputs to have a mean of 0 and a variance of 1 within each timestep, standardizing the distribution of inputs across the data points in each batch during training. Then, an RNN-LSTM was applied to the normalized input to convert temporal information to a dense embedding at each timestep. An RNN-LSTM with an equal hidden dimension was selected to simplify the tracking of embeddings in the later stage. Finally, a fully connected layer, without any activation, was applied to the embeddings in the last timestep t to predict output labels.

2.2.3 Deep learning simulation

To ensure the generalizability of our learning curve mapping approach, we performed 15 rounds of simulation for both gesture and emotion detection tasks, each of which included 15 (6C2) pairwise binary classifications. We conducted 25 repetitions (reps) for each set of simulations (sims) to mitigate the idiosyncrasies of specific processing and developmental patterns we are extracting under each pairwise condition. In total, we collected $25 \times 15 \times 2$ (reps $\times$ sims $\times$ tasks) = 750 runs of simulation data. Having an adequate number of simulation runs also enables us to observe clustering tendencies in the convergence and

divergence of trajectory patterns across various tasks and different classes.

In each simulation, 20% of the shuffled samples were used as the test data, while the remaining data were further split into 80% training data and 20% validation data. While validation data were used for reporting each epoch’s model performance, test data were used for reporting the final model performance on unseen data.

The number of epochs indicates the number of times an entire dataset has been passed forward and backward through the neural network. We set the number of epochs to 50 for both tasks, given that most of the variation in learning tends to unfold during the early stages of development. In S3 Text and S4 Text we also illustrate how using 20 and 100 epochs leads to similar observed patterns.

Because deep learning algorithms are very sensitive to unbalanced datasets, we applied data augmentation to the training and validation datasets. Specifically, the minority class was up-sampled to match the number of data points in the majority class, which ensured a balanced dataset [108].

Default initiation from PyTorch was used to standardize the model specification across simulations. All simulation runs were trained using the Adam optimizer with a learning rate of 0.001, and the loss function used was the cross-entropy loss for all pairwise classifications. All the other learning rate hyperparameters were kept at their default values.

Since deep learning is, computationally speaking, very expensive to train, the model was run on Nvidia RTX 4090 to expedite the processing. We set the batch size to 64 for gesture recognition and 256 for emotion classification to achieve an optimal balance between training speed and performance.³

To map a learning curve for each simulation, we first extracted embeddings of all LSTM timesteps from the hidden layer. We then applied multiple interpretable machine-learning models between these embeddings and the corresponding output labels to understand how the model’s confidence is updated during the LSTM timesteps.

³The large batch size for emotion classification is due to each of its input data points using less space compared to the gesture input.

2.2.3.1 Embedding extraction

We extracted the LSTM array of all timesteps for all batches across all epochs on the test data to ensure that we were extracting embeddings from all developmental stages and the final model was adequately trained on the targeted task. Since we specified the same dimension of the input data for the LSTM hidden layer, for the gesture detection task, we obtained 150 LSTM arrays from 150 timesteps, each array having the size of $1 \times \text{hidden_dimension}$ (1×274). Similarly, we obtained 66 LSTM arrays in sentence classification, each with a size of 1×300 . We then stored the corresponding output labels of those arrays as an output array. Those LSTM arrays share one output array since they are embeddings at different timesteps of the same data points.

For binary classification, the labels were encoded as 0 and 1. Once the model had been trained and the embeddings extracted, we stored and processed those embeddings in RAM, which has a greater storing capacity than GPU.

2.2.3.2 Mapping the learning curves

To approximate the processing curve of a model’s confidence across the LSTM timesteps, this chapter used a popular machine learning algorithm, k-nearest neighbors (KNN) for identifying embedding “separables.” This algorithm classifies an object by using a majority vote of its neighbor data points [189, 184]. Because embeddings are seen as a high-dimensional physical (i.e., location-wise) projection of input data (as opposed to a multivariate representation [134]), distance-based models, such as KNN and support vector machine (SVM), are popular choices for examining DNN embeddings in previous studies. Specifically, we applied KNN to the data point (x_t, y) , where x_t is the LSTM embedding at t timestep and y is the corresponding output label and calculated the KNN accuracy across all timesteps (information processing trajectory) and all epochs (developmental trajectory), and thus captured the learning curve of the model.

2.2.4 Quantifying performance across epochs

These learning curves examine how the LSTM’s embeddings classify the stimulus as it is incrementally presented to the network. Visually, each curve is plotted as the proportion of correct classifications across the item. These constructed learning curves can be analyzed in various ways, either through qualitative interpretation of the visualizations or by applying statistical analysis for quantification and hypothesis testing. In this chapter, we utilized the four descriptive statistics introduced earlier to assess the multidimensional KKN classification performance of the RNN-LSTM (i.e, how the discriminability of the model is evolving) across its training on two datasets, reconstructing and analyzing the unfolding learning dynamics of these distinct learning tasks. Specifically, for each such curve, we extracted four simple descriptive metrics: *start*, *end* - *start*, *max*, *t_{max}*. In Fig 2.5, we illustrate these descriptive statistics from an example trajectory of a single classification run for each dataset.

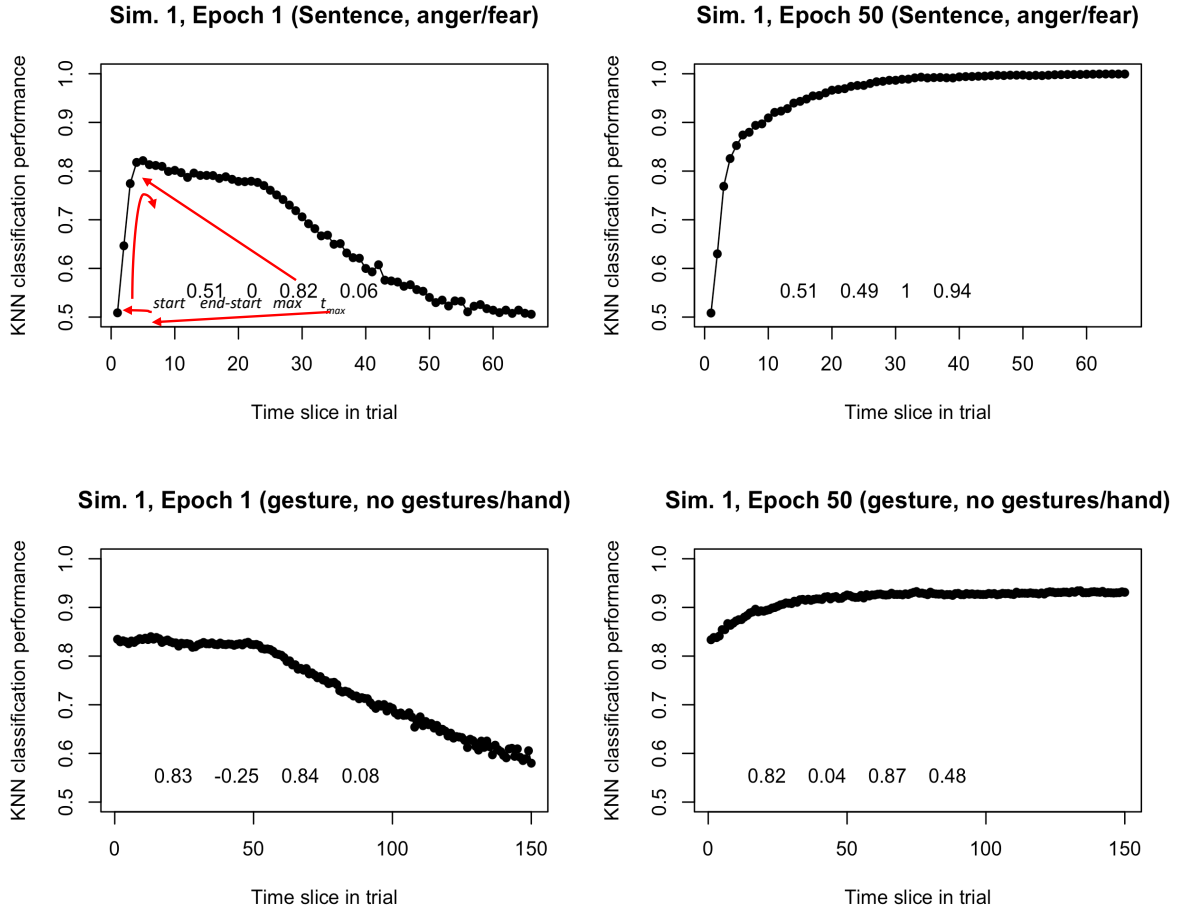


Figure 2.5: **Examples at epoch 1 and 50 of binary classification across a test input for one run of the network.** From epoch 1 to epoch 50, the LSTM’s embeddings are able to classify successfully over the test item, and we can characterize this success as a change to its performance using four metrics described in the main body of the text ($start$, $end - start$, max , t_{max}). Top: Example item from the sentence task. Bottom: Example item from the gesture task.

First, we defined the maximum performance (max) and the percentage of time at which that maximum was achieved in the presented item (t_{max}) over the whole timesteps. We included t_{max} alongside max because we discovered that max alone is not sufficient to determine performance, especially earlier in training. This is because the network may exhibit unstable performance, dropping across subsequent time slices, which may be indicative of a

non-monotonic learning trend seen in related domains [208]. This also suggests the network has learned something about the *initial* segments of an item, but the later segments wash out its performance as it has not yet encoded these later features. To capture this trend, we use t_{max} to assess when that maximum was achieved within each information processing trajectory – specifically, the percentage of the time slice at which the observed maximum occurred. In Fig 2.5 above, we show an illustration of this in the top left. The network achieved a performance of 0.82 on this particular item, but failed to sustain this performance as it dropped to near 0.50 as the sequence unfolds (at epoch 1). In the top right, the performance improves approximately monotonically across time slices (by epoch 50), and t_{max} is achieved near the final bin of the training item.

Additionally, we assess the performance at the start and end of the presented stimulus (*start*, *end*). Performance at the *start* may indicate the relative gains that can be expected from a stimulus item. As shown in the bottom panels of Fig 2.5, the gesture input already has performance above chance (~ 0.80) after the very first segment of the stimulus item. This suggests the model rapidly exploits spatial information in gesture (i.e., the location of keypoint coordinates can already vaguely differentiate gesture classes before sequential movement information is fed into the model). A model that achieves near-perfect performance at *start* and sustains it does not need to be exposed to the subsequent stimulus. The last measure we use is the subtraction of *end* - *start* performance. A high value on this measure suggests the network gets substantial information gains across a test item. For example, even at epoch 50 in the bottom right, the *end* - *start* of the gesture item is substantially lower than the simulation trained on the sentence task.

These measures can be plotted across epochs. Each learning curve now indicates how an item is being processed across an LSTM’s overall training. The measures are relevant to two timescales in the network’s behavior we conceptualized earlier: developmental (or learning) and information-processing timescales. For example, across epochs, movement along the *start* measure represents the initial performance at the item’s first time slice at the beginning of each information processing session. The *end* - *start* measure within an epoch

can describe the relative information gains during each processing session from the item’s full presentation. The *max* represents the best performance achievable in a session, indicating the network’s current information processing capacity. Finally, if t_{max} is low, it suggests that the network’s maximum performance of the current session is hindered by subsequent timesteps, implying that additional training may be necessary.

2.3 Results

We took the output from the KNN classification in Python, described above, and designed a sequence of R scripts to measure, visualize and quantify the trends in these four measures (*start*, *end* - *start*, *max*, t_{max}). R’s suite of visualization tools provided a convenient arena within which to view trends across epochs, and in these analysis scripts we also built linear models to statistically test the significance of these trends. All of the scripts in our methods are available at GitHub here: <https://github.com/JoyceJiang73/Learning-Curves/>. The R scripts only require input of simulation CSV data that contain as fields: binary classification labels, time slice, epoch, and performance measure (e.g., KNN classification performance).

We have reported the trained RNN model performance for the average of all pairwise simulations of sentence and gesture classification (epochs = 50) in S1 Text, including metrics such as validation accuracy, validation loss, test accuracy, test loss, recall, precision, and F1. These results confirm that almost all pairwise classifications have been sufficiently trained, indicating that the learning curves are capturing the learned experiences.

For each of the four metrics, we conducted a series of significance tests using regression models to assess distinctness across data sources, reshuffled null data, classes, and epochs. Specifically: (1) A linear model was used to assess whether the learning curves of the two datasets differed significantly, (2) A linear model evaluated whether the learning curve for each training classification task significantly deviated from a superimposed (randomly reshuffled) null hypothesis across the four metrics, (3) Linear models were used to assess whether the learning curves for each dataset were distinct by class and across epochs, and (4) An

ANOVA test was performed to determine whether class differences contributed more to developmental progression than epoch alone. All the significance tests (Adjusted R-squared) are reported in Table 2.2, and the corresponding visualizations for each metric are provided in their respective sections.

Table 2.2: A series of significance tests using regression models to assess distinctness across data sources, reshuffled null data, classes, and epochs.

Model	<i>start</i>	<i>max</i>	<i>t_{max}</i>	<i>end-start</i>
Data Source Difference				
Data Source	0.9814****	0.0211****	0.0034****	0.3923****
Baseline Null				
Sentence Reshuffle	0.0001	0.0042****	0.0120****	0.0027****
Gesture Reshuffle	0.0824****	0.0722****	0.0058****	0.0340****
Epoch and Class Differences				
Sentence Class	0.1022****	0.2736****	0.0951****	0.1301****
Gesture Class	0.9231****	0.8250****	0.2495****	0.3808****
Sentence Epoch*Class	0.1028****	0.6215****	0.4910****	0.5232****
Gesture Epoch*Class	0.9551****	0.9732****	0.6315****	0.6530****
Contribution of Class Differences to Epoch (Developmental) Progression				
Sentence Epoch*Class - Epoch	0.1045****	0.3010****	0.1160****	0.1437****
Gesture Epoch*Class - Epoch	0.9336****	0.8539****	0.2868****	0.4282****

*Significance tests using regression models to assess distinctness across data sources, reshuffled null data, classes, and epochs for the four metrics (*start*, *end - start*, *max*, *t_{max}*). Adjusted R^2 values are reported along with the significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$).

2.3.1 Start performance

In Fig 2.6, we show the LSTM performance at the *start* of an item across epochs of training. In general, the sentence dataset shows low, near-chance performance, while gesture classifications are already well above chance performance. This chance-level performance for sentence items stays consistent across the entire training period, though the gestural dataset shows some improvement. For the gesture dataset, this suggests that the network has some information about a classification before much of a training item is even shown to the network. It would indicate that gestural data has spatial information in the point coordinates of the body and is exploited by the network at the very first time bin. With language, it takes time for word embedding vectors to be integrated in the network.

To confirm these trends, we tested a linear model that predicted *start* performance by training data, showing that dataset accounted for about 98.14% ($p < .0001$) of the variance seen in Fig 2.6. The classification of the gesture data accounts for 95.51% ($p < .0001$) of variance internally to that dataset, whereas for sentence classifications accounts only for 10.28% (lower but also significant, $p < .0001$). The learning curves reveal that the first word of a language task has low diagnostic accuracy for a classification, but the spatial variance over gestural classifications is much more informative.

The *start* performance for sentence null is the only metric that shows no significant difference between learning from ordered sentence classes and reshuffled data (among all significance tests; $R^2 = 0.0001$, $p < .0879$). This is expected, as all sentence binary classifications begin with a consistent initial performance of 0.50 (random guessing for binary classification), which remains unchanged even with randomly reshuffled sentence data. In contrast, the *start* performance significantly differs for ordered gesture classes versus reshuffled data because the various gesture pairs contain different levels of spatial information ($R^2 = 0.0824$, $p < .0001$), affecting their initial performance.

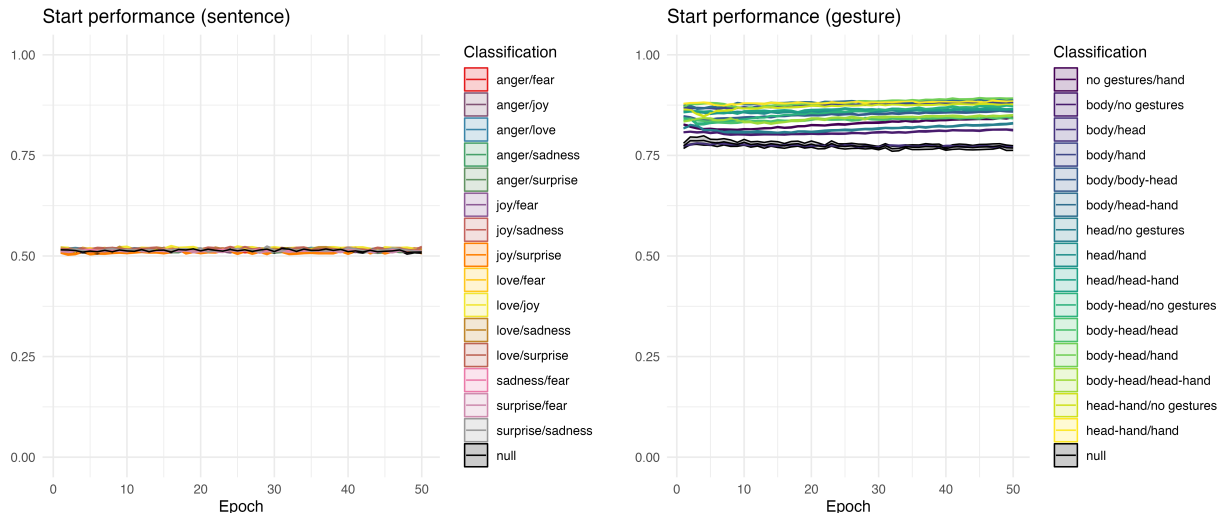


Figure 2.6: ***Start* performance for pairwise classifications.** *Start* metric over epochs, indicating that in the sentence task, the classification at the first time slice remains stable near chance, whereas the LSTM’s embeddings for the gesture task are distinct across stimulus types, and also show slightly more improvement over training (while already being well above chance relative to the sentence task).

2.3.2 Max performance

Curiously, if one only investigated maximum performance across training, these classification tasks could be regarded as relatively similar in their behavior. As shown in Fig 2.7, both sentence and gesture datasets yield a classification performance that is high, between 0.75 and 1.0 depending on the classification. In the gesture dataset, there are more “difficult” classifications, shown by outliers in *max* performance across training. This can be helpful in diagnosing representational challenges in the network’s training, marking what pairs of training items may be more difficult to distinguish than others.

Again, as with the *start* measure, the *max* performance shows greater variance associated with classifications in the gesture case ($R^2 = 0.8250$, $p < .0001$) than the sentence case ($R^2 = 0.2736$, $p < .0001$). The difference between these two datasets is not as pronounced as in the *start* measure, as a linear model shows that only 2.11% ($p < .0001$) of the variance is

associated with the dataset in a linear model predicting *max* performance observed within a trial. Despite the small value, it is nevertheless significant and driven by the relatively higher performance in the sentence task.

The *max* performance also differs for both sentence ($R^2 = 0.0042$, $p < .0001$) and gesture tasks ($R^2 = 0.0722$, $p < .0001$) when the network learns from ordered datasets compared to randomly shuffled ones. The sentence conditions appear to pick up more spurious signals, ultimately approaching 1.00 as the epochs increase, while the gesture condition remains consistently unlearned across epochs. The learning pattern for t_{max} across epochs (see next section) provides additional insights into these trends.

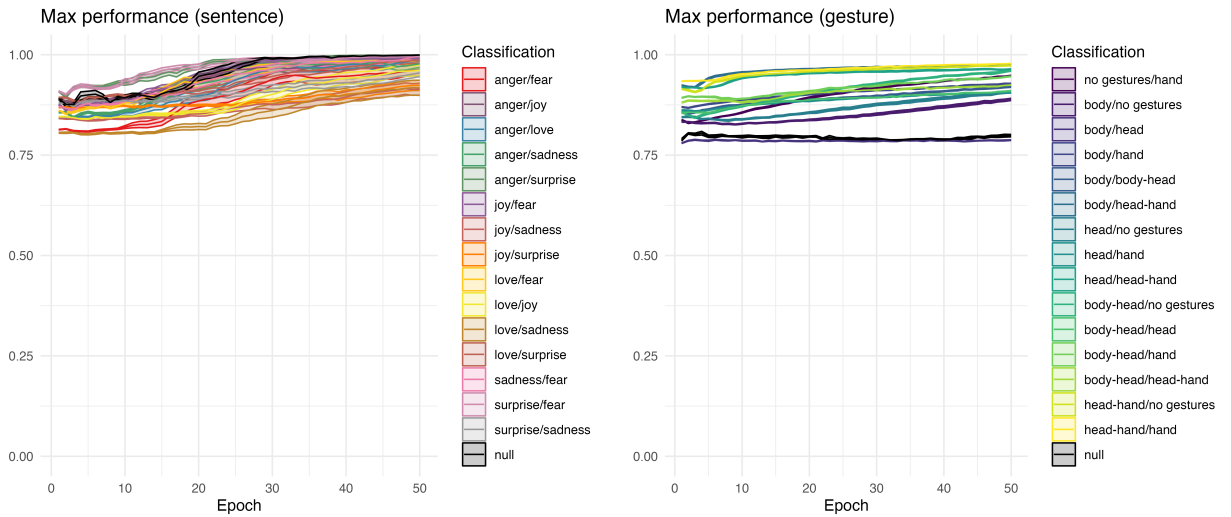


Figure 2.7: **Max performance for pairwise classifications.** Both sentence and gesture tasks show an increase in *max* performance over training epochs. On average this asymptotes near perfect performance but can vary depending on binary classification.

2.3.3 Time at max

We would expect that completed training in the LSTM should show that maximum performance should appear near the *end* of an item, as this would suggest that the network has extracted useful information in the whole presentation. Indeed, this is indicated by lower time-at-max values earlier in the training. During the first few epochs, the maximum

value occurs proportionally earlier in a training item, similar to the example shown above in Fig 2.5. However in both sentence and gesture datasets, networks slowly extract features across the stimulus items for classification, as shown in Fig 2.8. The time gradually rises for all classifications, though it can also vary widely and tends to be more irregular in gesture.

In a linear model predicting t_{max} from data source, sentence and gesture are only slightly different, with 0.34% ($p < .0001$) of the variance accounted for. Classifications in sentence and gesture both relate significantly in a linear model predicting time at max, with sentence classifications accounting for 9.51% ($p < .0001$) of the variance and gesture classifications 24.95% ($p < .0001$).

Interpreting t_{max} alongside max , we confirm that the learning experience of a language task is more reliant on spurious signals for reshuffled data compared to ordered data, as max continuously increases to reach 1.00 while t_{max} remains sporadic. The gesture condition remains consistently unlearned across epochs for reshuffled data, as the max performance can occur at any point along the processing trajectory (any timestep). Both t_{max} values indicate that no meaningful delayed signals were developed when the network was trained on superimposed reshuffled data. The time at max differs significantly for reshuffled versus ordered data in both sentence ($R^2 = 0.0120$, $p < .0001$) and gesture ($R^2 = 0.0058$, $p < .0001$) conditions.

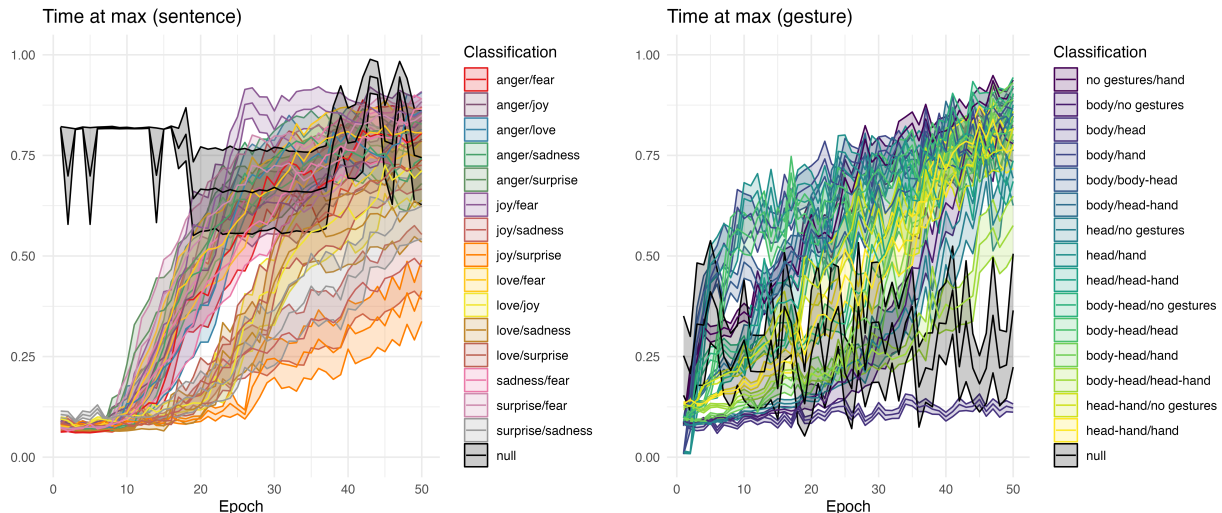


Figure 2.8: **Time at max (t_{max}) for pairwise classifications.** Both sentence and gesture tasks show an increase in the proportion of time at which maximum performance is observed. Indicative of information gain, this rises in both – suggesting the LSTM comes to better integrate the full test items for its performance. However this is much more orderly in the sentence task than the gesture task.

2.3.4 End - start performance

Finally, we wish to get a sense of how the LSTM models are extracting information within an information processing session. One way to do this is to discern how much higher its performance is at the end of a session compared to its performance initially. This is where we see the greatest difference between the datasets. As shown as Fig 2.9, sentence shows substantial information gain across stimulus items. By the end of training, the start and end of trials involves considerable difference, and the positive value of this difference shows that the network improves its performance significantly across presentations. With gesture, the situation is quite distinct. In most classifications, there is indeed a rise. But because gesture is informative even at the *first* time bin of an item, there isn't much information gain remaining. Indeed even the highest instances of this value hover near 0.10. Therefore, while performance on gesture is higher at first, the network may find the task more difficult

in integrating that spatial information over time to improve performance.

As in the *start* measure, the difference between datasets is quite large. When data source is used to predict the *end* - *start* measure, 39.23% ($p < .0001$) of the observed variance in this measure can be associated with data source, with sentence associated with much higher information gain than gesture results. When looking at each data source separately, only 13.01% ($p < .0001$) of the measure is associated with the classification pairs in sentence. In the gesture dataset, this association is 38.08% ($p < .0001$), showing much higher contribution of the class differences in gesture.

Both sentence's ($R^2 = 0.0027$, $p < .0001$) and gesture's ($R^2 = 0.0340$, $p < .0001$) *end-start* values differ significantly between ordered and reshuffled data. Consistent with *max* performance gradually reaching 1.00, the information gain of the sentence network increases incrementally from 0.00 to 0.50 when trained on reshuffled data. Combined with insights from t_{max} , we suggest that although learning may appear to occur even with reshuffled data, it is likely driven by spurious signals picked up during the processing session. Additionally, the *end-start* performance for gesture gradually improved from highly negative to slightly negative but was not able to reach 0.00 during the early epochs (e.g. epochs smaller than 50), which explains the flat *max* performance for the reshuffled gesture condition.

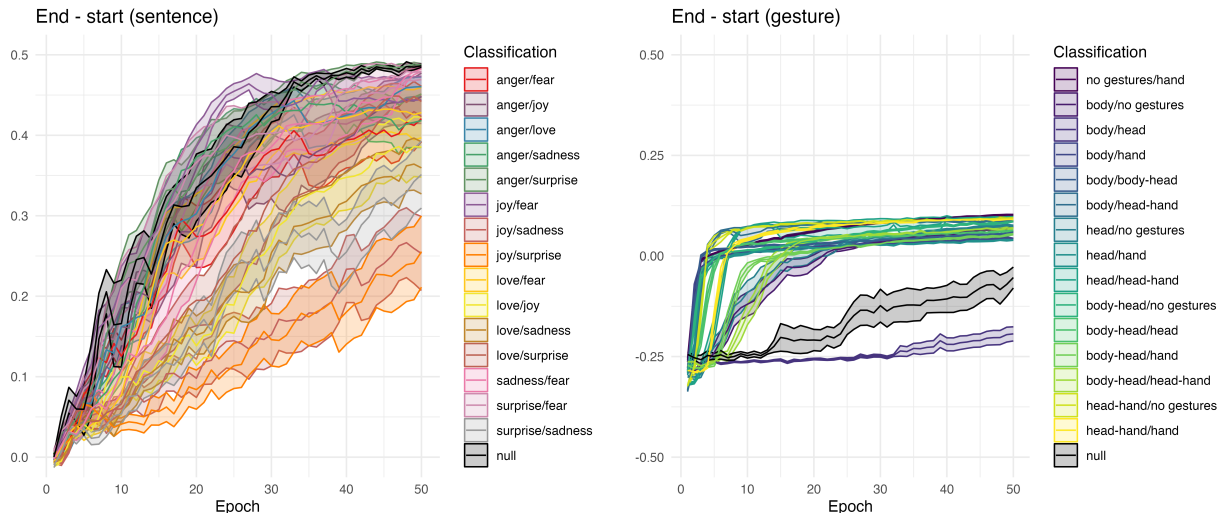


Figure 2.9: ***End - start*** performance for pairwise classifications. The information gain from start to end of stimulus item shows a stark difference between datasets. The sentence task shows that the LSTM’s performance rises as it integrates data from start to end. With gestures, the story is more complicated, showing less improvement overall, and in one classification task, it is almost entirely unlearned.

2.4 Discussion

Recognizing the importance and challenges involved in systematically unpacking the internal representations of DNNs, this chapter introduced a multi-dimensional quantification and visualization approach, “learning curves,” which can capture two temporal dimensions of a model learning experience. First, it captures the “information processing trajectory,” how the network is doing as it processes test items. Second, it captures the “developmental trajectory,” describing how this processing is changing over training epochs. The former represents the influence of incoming signals on an agent’s decision-making, which is operationalized by the timestep within a single epoch of an RNN. The latter conceptualizes the gradual improvement in an agent’s decision-making abilities throughout its lifespan, operationalized by the iteration of epochs. The learning curve approach we illustrate in our two datasets shows that we can quantify and qualitatively investigate both of these dynamics

within the same analysis, utilizing four descriptive metrics: *start*, *end - start*, *max*, *t_{max}*.

Based on a series of significant tests (see Table 2.2), we first demonstrated that there is a data source difference between sentence and gesture classification in the overall learning patterns across all four metrics (all p 's < .0001). Additionally, the learning experiences of both datasets' ordered classifications are distinct from their respective reshuffled null baselines, except for the *start* of the sentence classification. This is because both ordered (aggregated across all classes) and unordered binary sentence classifications will have an initial classification performance of 0.50 due to random guessing. This is corroborated by the significant test of the *start* metric for sentence class and gesture class: sentence pairwise classifications show very minor differences, with R-squared 0.10 (p < .0001), whereas gesture classes show much greater divergence, with R-squared above 0.90 (p < .0001), due to the initial spatial information contained in gesture coordinates, even at timestep 0, as mentioned in the previous results section. The significant difference between the reshuffled null baseline for ordered vs. unordered gesture classification is also due to different gesture class pairs having different prior spatial information, deviating from the 0.50 random guess scenario observed in sentence classification.

The modality distinction is further evident in the *end - start* metric (where sentence classification shows a much larger information gain than gesture classification, with gesture's information gain is more gradual). Finally, the ANOVA test capturing the R-squared difference between epoch $\times$ class versus epoch-only indicates significant distinctions between pairwise classes and not being distinguished by class (all p 's < .0001). This reveals that within the same multiclassification task, certain classes can be more difficult or easier to separate. Taken together, based on the analysis of these four measures derived from the learning curves across two distinct datasets, we highlight three insights gained from mapping these curves: *non-monotonicity*, *pairwise comparisons*, and *domain distinctions*.

2.4.1 Non-monotonicity

First, we observed a non-monotonic trend in the learning experiences of DNNs. In other words, learning does not always show an increasing monotonic improvement because networks sometimes reveal temporary decrements in their performance across training. This is a characteristic recognized in previous literature as a key advantage for their advancement in addressing the most challenging tasks. For example, in applying neural network models to language, the presence of non-monotonic learning was taken as evidence that networks “reorganize” their knowledge as they learn. This may mean networks are acquiring more efficient representations of a problem space and the drop in performance is indicative of a transition into that more efficient representation. Classically, in the case of models learning language, they show a decrement in performance when they “discover a rule” in grammars [211]. Our results are a testament to the observation that both representational consolidation and “catastrophic forgetting” remain as important issues in DNN learning [118].

Specifically in our results, we found that RNNs exhibit different preferences for early versus late cues when addressing various sequential tasks. For instance, in the sentence task, we noted a more pronounced performance improvement occurring in the later stages of information processing (i.e., model development). In contrast, gesture learning tends to show quicker progress, with more variability across epochs and repetitions of simulations, suggesting that DNNs tend to rely on shortcuts, such as naive cues related to keypoint coordinates, for gesture classification, rather than focusing on high-level movement sequences. This shortcut-based “learning” also is evident in the higher initial performance (*start*) for gesture classification (> 0.75). On the other hand, sentence classification begins at around 0.50 (the at-chance probability) and exhibits greater performance gains in later epochs, indicating that the models classify based on high-level semantic sequences.

2.4.2 Pairwise comparisons

Additionally, although multiclassification tends to exhibit collective model performance, our between-class pairwise comparisons reveal the presence of outliers within multiclassification. For example, “joy/fear”, “surprise/fear”, and “sadness/fear” demonstrate higher information gains across epochs compared to other emotion pairs, suggesting that these classes are further apart from each other. The “body/head” classification appears to experience learning challenges, possibly because these two movements have difficulty being completely separated, as the head’s movement in naturalistic data may inevitably coincide with that of the body due to joint coordination.

The combination of learning curve mapping and instance measures thus serves as an effective approach for “auditing” representations in multiclassification problems. The proposed pipeline improves model explainability beyond holistic evaluation of classification performance and ad-hoc attention visualization by unpacking pairwise class learning patterns to reveal any pairs that are unsuccessful in being discriminated or involve delayed knowledge gain. This granular examination allows modelers to better investigate a model’s appropriateness for the underlying task, as well as the properties of the processed input signals, reaffirming the value of our learning curve conceptualization.

2.4.3 Domain distinctions

With these findings, it is tempting to infer that there are domain distinctions among different modality classification tasks. Gesture learning, for instance, may rely on autocorrelated signals (as body movements result from the coordinated contraction and relaxation of muscles), potentially emphasizing spatial semantics as early cues ⁴, while language learning relies on higher degrees of surprisal, irregularities and arbitrariness, as suggested by previous literature (see [16, 261, 51, 191]). These domain distinctions are valuable to cognitive science in un-

⁴The spatial information arising from the gesture, specifically the keypoint coordinates in the gesture dataset, can vaguely help the model distinguish between gesture classes (e.g., hand vs. no gesture). However, this information remains a low-level cue, as the model lacks an understanding of what a complete hand gesture sequence or a no-gesture sequence looks like.

derstanding how humans process and distinguish between various modes of communication, shedding light on the neural mechanisms underlying the flexibility and adaptability of the human mind when processing different forms of information and communication modalities.

Though it is intuitive and tempting to explain these domain distinctions here, we cannot yet assert that these trends would hold for *all* sentence or gesture (keypoint) classification tasks, only the ones we investigate here. However the learning curve results would seem to align with an intuition of how linguistic symbols would be sequentially integrated into a neural network in contrast to the highly auto-correlated spatial information contained in gesture performance. Still, we cannot infer a broad generalization about “language vs. gesture” and only leave it as a potential path for future investigation. Indeed, this may be an additional benefit of a method like the one we present here. Learning curves could finely ascertain these domain distinctions, and expanding the set of data to test may permit generalization in future work. This may have theoretical implications itself. The learning curve analysis may provide information about the distinct sources of information from varied modalities. When neural models (or human brains, presumably) integrate distinctive sources of information, they may strengthen understanding of complex multimodal data by leveraging their unique information-processing and developmental benefits.

2.4.4 Limitations

This chapter provides conceptual and operational illustrations of applying learning curve methods to RNNs, though both the illustration and the proposed method come with certain considerations. First, we selected two distinct datasets to demonstrate how learning curves offer insights into different data and modality learning experiences. While illustrating significant domain distinctions, our study is limited in making broad generalizations between “language vs. behavioral.” Future cognitive and behavioral research interested in such conceptual generalization can apply our proposed method to various distinct language tasks (e.g., language translation) and behavioral datasets (e.g., facial expression detection) to achieve broader generalizations across different modalities. Additionally, as highlighted

at the beginning, our method could have implications for multimodal DNN modeling, given that our simulations utilize datasets with distinct modalities. For the current scope, we focus on how our proposed method facilitates the understanding of distinguishable modalities as an initial step, while future work can explore adopting this method for multimodal DNNs and datasets (e.g., audiovisual emotion classification) to assess how multimodal learning differs from unimodal learning.

From a methodological standpoint, as detailed in the model explainability technique comparison (Table 2.1), our proposed method is currently limited to temporal tasks and DNN architectures, specifically RNNs. However, it has the potential to be relatively easily adapted to various tasks, modalities, multiclassification, and different performance measures, beyond the binary classification used in this chapter. We have provided illustrations of the multiclassification learning curve based on sentence classification in S2 Text. Additionally, because our method generates measures at every timestep and epoch, it may be more computationally expensive compared to other explainability methods. However, this arises from the tradeoff between obtaining a more granular measure across all training steps versus only obtaining snapshot visualizations of a single neural network layer or attribution scores for all input features. As documented in the DNN explainability comparison (Table 2.1), each technique has its strengths and considerations, and researchers can exercise discretion in choosing the appropriate method. The learning curve method excels in providing a more systematic understanding of the model’s learning process across models and modalities.

2.5 Conclusion

There is a long tradition in cognitive science and computational neuroscience of examining the internal representations of models [255, 162, 74, 219]. One reason for this is to determine if a model’s features or processes reflect processes of the human mind. Such models can be informative for inferring properties of human mental processing, and so have direct theoretical implications. For example, Elman [59] and others [32] showed that RNNs can learn patterns sufficiently complex to resemble human grammar. By examining the inter-

nal activations of these recurrent networks, they showed that these systems are driven by graded, statistical features. Words are not discrete “symbols” but scalar vectors conditioned by linguistic context in time [61]. This was taken to challenge theories that see language as a purely abstract and symbolic recursive process [83].

Inspired by prior work, the learning curve method we proposed uses temporal mapping (information processing and developmental trajectories) to help modelers more comprehensively understand the underlying learning and decision-making processes of a complex model architecture without delving into the intricacies of interpreting its internal representations directly [164, 276]. This kind of systematic and quantitative approach has recently gained popularity in both computational cognitive science and deep learning communities [276, 176, 195, 168, 109], as it facilitates multidimensional comparisons across models and modalities, which were previously seen as challenging for DNN-like models. The current study illustrates multiple techniques for analyzing model learning experiences and highlights three insights across different communication modalities based on these analyses. Future studies can utilize this learning curve mapping approach to enhance model interpretability studies by evaluating a model’s appropriateness for the task at hand, examining the properties of the underlying input signals, and assessing the model’s alignment (or lack thereof) with human learning experiences, which is also a critical consideration for computational cognitive science and neuroscience research.

CHAPTER 3

Human Heuristics in Context-Sensitive Image Captioning

Abstract

Multimodal generative AI is transforming how people create and interpret digital content. Building on prior findings that humans rely on intuitive but often misleading linguistic heuristics when evaluating AI-generated text, we examine how these biases manifest in context-rich, cross-modal news image captioning. Through a series of experiments and computational analyses, we first show that language-based heuristics persist in AI caption evaluation, with high-quality context further biasing judgments. Feature analyses identify a set of informational cues—specifically, image-caption alignment and caption descriptiveness—that may mitigate such biases. A follow-up context manipulation experiment confirms that these informational cues reliably guide judgments, but more strongly for caption preference than for AI detection, where participants continue to rely on superficial linguistic cues. Together, these findings suggest that while flawed linguistic heuristics persist, informational cues can improve evaluation in multimodal contexts, though task differences highlight the need to carefully frame evaluation questions when assessing generative content¹

¹All data and code are available at <https://github.com/JoyceJiang73/context-sensitive-imgcap>.

3.1 Introduction

Multimodal generative AI systems are transforming how people create and consume digital content [29, 146, 147]. Modern vision-language models (VLMs) like GPT-4o [180], Gemini [70], and Qwen-VL [6] have democratized sophisticated image description and storytelling capabilities once exclusive to human experts [147, 257, 53].

The availability of these tools introduces concerns about how people detect and interpret the AI-generated content. Prior research has documented systematic biases in how people judge AI outputs. In text, people often rely on flawed linguistic heuristics—such as impersonal or monotonous tone, or the presence of rare and lengthy words—which makes AI-generated content difficult to detect, raising concerns about deception and manipulation [39, 121, 126, 54, 250, 105, 270]. Similar biases appear in visual domains, where synthetic faces are often rated as more trustworthy than authentic ones [172, 178, 221, 241]. Building on these previous findings, we examine whether these linguistic heuristics persist, and if they can be mitigated, when both visual and textual information are presented.

Across a series of experiments and computational analyses, we investigated this question through naturalistic, context-rich scenarios in news image captioning. Image captioning is a challenging perceptual-linguistic task that requires machines to recognize visual elements (such as objects, actions, and scenes), understand their compositions and semantic relationships, and generate coherent descriptions that align with human cognition [253, 15]. News image captioning places particularly high demands on world knowledge, requiring inference about people, locations, events and intentions beyond visually grounded entities [217]. Thus, access to contextual information from accompanying articles plays a key role in shaping caption quality and informativeness [177, 124, 125].

Our analyses focus on two dimensions in news caption evaluation: linguistic style and informativeness. Linguistic style reveals whether AI systems can mimic human writing patterns in ways that exploit unreliable linguistic heuristics, extending prior text-based findings to multimodal generative tasks [105, 270]. Informativeness captures cross-modal rea-

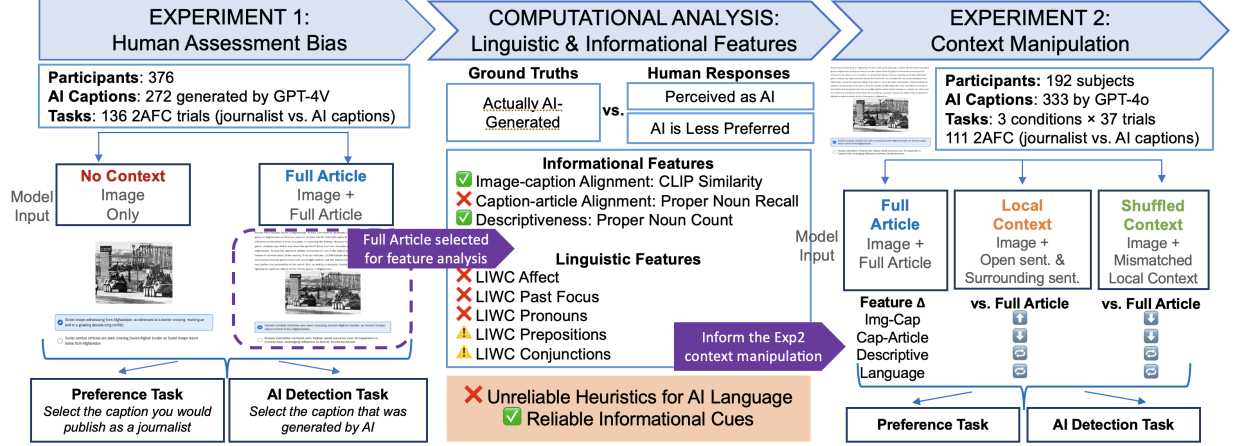


Figure 3.1: **Overview of Experiments and Analysis.** (1) **Experiment 1** extends prior findings on human assessment bias from AI-generated text to cross-modal image captioning, using a 2×2 design crossing context conditions (*No Context*, *Full Article*) with assessment tasks (Preference, AI Detection). (2) Captions generated with *Full Article* context and corresponding human responses were analyzed computationally to identify informational and linguistic features associated with ground-truth caption source versus human judgments in both tasks. Reliably informative features (✓) and flawed (✗) or non-indicative features (Δ) motivated the design of Experiment 2. (3) **Experiment 2**, informed by feature analysis, compares three context-based conditions—*Full Article* (as in Experiment 1), *Local Article*, and *Shuffled Article*—and validates changes in feature associations (increase, decrease, or no change) for preference and AI detection judgments.

soning—whether captions enable audiences to extract information beyond the image itself through accurate cross-modal alignment, retrieval of contextual details from accompanying articles, and identification of proper nouns such as people, places, or events [217, 256]. This design uniquely reveals that flawed language-based heuristics persist even when visual and contextual information are available, yet richer informational cues have the potential to mitigate the biases that undermine text-only evaluations of AI systems.

3.2 Methods

To examine how people evaluate AI-generated content in multimodal settings, we constructed AI-generated image captions under varying levels of news context: *no context*, *full article*, *local context*, and *shuffled context* (see examples of the four conditions in Figure 3.1, overview of the experimental pipeline; Figure 3.2, illustration of stimulus construction).

Experiment 1 considered only the first two conditions—*no context* versus *full article*—to assess how the presence of article context influences caption quality judgments and audience reactions, and whether the language-based heuristics documented in text-only evaluation settings persist in AI-generated image caption evaluation. We further conducted computational analyses to identify which linguistic and informational features distinguished caption sources and were associated with human judgment patterns.

Experiment 2 focused on the three conditions with context present—*full article*, *local context*, and *shuffled context*—to test whether manipulating the structure of contextual information could predictably shape audience preferences and AI-detection accuracy by selectively altering linguistic and informational cues. In the *full article* condition (used in both Experiments 1 and 2), captions were generated from the complete accompanying news article. The *local context* condition restricted input to the excerpt referencing the image, enhancing local image-text alignment. The *shuffled context* condition paired these excerpts with mismatched article segments, disrupting cross-modal alignment while preserving descriptiveness and linguistic style in the generated captions.

We generated captions using GPT-4V and GPT-4o across these contextual conditions, comparing them against journalist-written captions from the Voice of America (VOA) dataset [138]. Across experiments, 568 participants evaluated a total of 605 AI image-caption pairs, judging either which caption they preferred (preference task) or which was AI-generated (detection task).

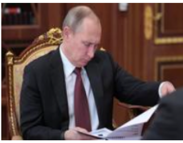
News Image	VOA Caption	No Context	Full Article	Local Context	Shuffled Context
	A Syrian national flag waves as vehicles move slowly on a bridge during rush hour.	A Syrian flag billows over Damascus, capturing a moment of calm in a city seeking recovery after years of conflict.	Amidst government offensives, al-Qaida's ranks may swell in Syria as moderate rebel forces crumble and jihadists consolidate power in contested areas.	Syrian flag flies over Damascus as Russian-backed offensives reshape northern Syria, raising concerns of increased jihadist recruitment amid collapsing moderate rebel forces.	U.S. strikes Syrian airfield after suspected chemical attack, Trump shifts policy amid international outcry, Damascus in focus as tensions rise in the region.
	Boys observe as VOA cameraman Rahim Gul Sarwan prepares to shoot video in Achin district in eastern Afghanistan.	Documenting Daily Life: Journalist captures the realities of village life amid ongoing security challenges in rural Afghanistan.	VOA team reports from the frontlines of Achin, Afghanistan amidst IS presence and local tensions, highlighting the resilience and risks of journalism.	VOA journalist Rahim Gul Sarwan captures footage in Eastern Afghanistan's Achin district, navigating complex challenges and ensuring coverage reaches beyond the IS stronghold.	U.S. Secretary of State John Kerry heads to Kyiv amid escalating tensions in Crimea, as NATO-Ukraine Commission meeting is set to address the crisis.
	Russian President Vladimir Putin reads papers during a meeting in the Kremlin.	Russian President Vladimir Putin attentively reviews documents during a high-stakes meeting, highlighting the gravity of policy decisions affecting international relations.	Hopes for U.S.-Russian cooperation rise after Trump-Putin call, yet experts advise caution on the path ahead.	Russian President Vladimir Putin reviews documents in the Kremlin, Moscow, January 16, 2017, hopes rise for improved U.S.-Russia relations under Trump's presidency, despite cautious optimism in Moscow.	Russia monitors Syria situation as Turkish troops face challenges in strategic al-Bab offensive.

Figure 3.2: Examples of news image captions generated by different sources and different contextual strategies. The first column includes images; the second column includes journalist-generated captions drawn from the VOA dataset [138]. The last four columns include captions generated by AI models with different context conditions. Specifically, the *full-article*, *local-context*, and *shuffled-context* caption examples in Experiment 2 were generated by GPT-4o, whereas the *no-context* caption examples in Experiment 1 were generated by GPT-4V.

All procedures were approved by the UCLA Institutional Review Board (IRB Protocol #24-000019; February 2024). Experiments were preregistered on AsPredicted (Experiment 1: <https://aspredicted.org/kww6-qqs-d.pdf>; Experiment 2: <https://aspredicted.org/ytqs-6yhq.pdf>).

3.2.1 Collecting and generating image captions

Stimuli (see examples in Figure 3.2) were derived from the VOA dataset compiled by Li et al [138], which includes 1,014 images and 199 articles. Each article contains one to eight images, each with an original caption authored by a VOA journalist. From this corpus, one image-caption pair was sampled per article, yielding 191 image-caption-article triplets with balanced topic coverage and exclusion of graphic content (e.g., depictions of violence

or death) per IRB guidelines.

In Experiment 1, AI-generated captions were produced using GPT-4V (“gpt-4-vision-preview”) between November 2023 and February 2024 via OpenAI’s API [180]. In the *no-context* condition, the model generated captions based solely on the image; in the *full-article* condition, it received both the image and the full corresponding article. The model was prompted to generate concise captions (around 25 words), matching the average length of human-authored captions, and styled in the tone of VOA reporting. All captions were generated using default temperature and system settings. Caption quality and style were consistent across batches. We did not observe substantial changes in caption quality or style across generation batches.

To control for potential bias in the *no-context* condition—where the model lacked textual context to identify named entities (e.g., individuals or locations)—we annotated the concreteness of each caption. Only triplets in which both human- and AI-generated captions either included or omitted such entities were retained, yielding a final set of 136 matched triplets for Experiment 1.

For Experiment 2, we selected 111 triplets from the same VOA dataset that provided explicit information on each image’s placement within the article, which was necessary for constructing the *local-context* condition. AI-generated captions were produced in April 2025 using GPT-4o [180] via the OpenAI API. Three contextual conditions were implemented as follows: (1) *Full-article* context, identical to the with-article condition in Experiment 1, where the model received the full accompanying article text along with the image; (2) *local context*, where the model was provided only the article’s lead sentence, plus the sentence preceding and the sentence following the location where the image appeared in the original article (cf. Young et al. [256]; Kreiss et al. [125]); and (3) *shuffled context*, where the local context segments were randomly reassigned across stimuli, yielding mismatched textual context for each image (cf. Kreiss et al. [125]). To ensure all captions could be generated despite this misalignment, any failed generations were reassigned until captions were successfully produced for all stimuli.

3.2.2 Participant recruitment

All participants were recruited through departmental subject pools at UCLA, with no overlap across experimental tasks or conditions. All participants provided informed consent electronically prior to participation and were informed of their right to withdraw at any time without penalty and of the confidentiality of their responses. For Experiment 1, the preference task was conducted in April 2024, and the detection task was collected in two waves (May 2024 and February 2025) to balance sample sizes with the preference task for subsequent computational and regression analyses. No systematic differences in participant behavior were observed between the two waves of the detection task. For Experiment 2, all participants were recruited between April and May 2025. Our sample represented a relevant population for the study, given participants' familiarity with digital media and AI-generated content.

A pilot study ($n = 39$ in the *full-article* condition; $n = 41$ in the *no-context* condition) yielded an effect size of $d = 0.62$ for the difference in AI-caption preference across conditions, suggesting that a sample of similar or larger size would provide $\geq .80$ power at $\alpha = .05$.

Based on this estimate, in Experiment 1, 201 participants were recruited for the preference task ($n = 100$ *no-context*, $n = 101$ *full-article*) and 212 for the detection task ($n = 101$ *no-context*, $n = 111$ *full-article*), ensuring adequate power after exclusions. Following exclusion criteria based on attention-check performance, 183 undergraduates completed the preference task ($n = 96$ *no-context*, $n = 87$ *full-article*) and 193 completed the detection task ($n = 93$ *no-context*, $n = 100$ *full-article*). For Experiment 2, 96 undergraduates were initially recruited for the preference task and 98 for the detection task. After excluding participants who failed more than one attention check, 96 participants remained in each task.

3.2.3 Experimental design

An overview of the experimental design and task flow is provided in Figure 3.1. In Experiment 1, participants were randomly assigned to complete either the preference or detection

task, and to one of two conditions: *full-article* or *no-context*, resulting in a fully between-subjects design. Each participant completed only one task under one condition, completing all 136 stimuli in randomized order. On each trial, participants were shown an image (and article, if applicable) along with two captions—one written by a journalist and one generated by an AI model. In the preference task, participants selected the caption they would publish as a journalist. In the detection task, they identified which caption they believed was generated by AI. To account for the longer reading time in the with-article condition, participants in that condition were encouraged to spend approximately 25–30 seconds skimming the article before responding, though no time limit was enforced. The estimated completion time is 30 minutes for the *no-context* condition and 60 minutes for the with-article condition. Three attention-check trials were randomly embedded throughout the experiment; participants who failed more than one were excluded from analysis. Response data and basic demographic information were recorded.

In Experiment 2, participants were again randomly assigned to complete either the preference or detection task, maintaining a fully between-subjects design at the task level. Within each task, all three contextual conditions (*full-article*, *local*, and *shuffled*) were presented in a mixed design: each participant viewed captions generated under all three conditions but saw each image only once. Each participant completed 111 randomized trials (i.e., 37 per condition), with stimuli randomly sampled such that each participant saw a unique subset of items while ensuring balanced presentation frequencies across the sample. On each trial, participants viewed an image alongside its accompanying full article and two captions—one written by a journalist and one generated by the AI model under one of the three contextual conditions. Instructions, practice trials, attention checks, and timing guidelines were identical to those used in Experiment 1.

3.2.4 Computational analysis

We extracted a broad set of linguistic and cross-modal informational features from journalist- and AI-generated captions, building on the human heuristic framework of Jakesch et al. [105]

for AI verbal self-representation, adapted to cross-modal news captioning. Linguistic features were derived from LIWC [188], a validated psycholinguistic lexicon that captures grammatical, stylistic, and psychological markers commonly used to characterize AI-generated text, social media discourse, and news writing [105, 270, 107, 110, 212].

Informational features quantified cross-modal alignment and caption descriptiveness. To measure descriptiveness, we computed the number of *Proper Nouns* in each caption as a general indicator of how many specific entities (e.g., people, places, organizations) were named. We also derived more specific counts of *who* (PERSON, NORP, ORG) and *where* (FAC, GPE, LOC) entities using spaCy’s named-entity recognition, following journalistic conventions of “who, what, where, when” [256, 217]. The selected *who* and *where* entity types are more visually grounded than *when* or *what*, making them strong indicators of cross-modal integration, which has been shown to benefit from article context [146, 271]. To assess semantic alignment, we computed *CLIPScore* (openai/clip-vit-large-patch14), which measures how well a caption and image embed in the same multimodal semantic space [89]. For caption-article alignment, we calculated *BERTScore Recall* (bert-base-uncased), which quantifies token-level semantic overlap weighted toward recall [269], and *Proper Noun Recall*, which captures the proportion of proper nouns in the article that are mentioned in the caption. We provide examples of top- and bottom-ranked news image captions for each informational feature in Appendix 3.5.1 and Figure 3.7.

All features except *CLIPScore* and *BERTScore Recall* were normalized by caption length to control for verbosity. All features were then z-standardized prior to analysis. Given the high dimensionality of the feature set, we applied bidirectional stepwise selection (R’s `step()` function) to identify the subset of features that best predicted outcomes, minimizing model AIC. Under informational features, the *Proper Noun Count* was observed to be more robust during feature selection than the *who* and *where* counts as a measure of descriptiveness. *Proper Noun Recall* was also more robust than the *BERTScore Recall*. The final selected features included LIWC-derived linguistic categories (*Affect*, *Past Tense*, *Pro-nouns*, *Conjunctions*, *Causation*, *Differentiation*, *Quantifiers*, *Adverb*), *Proper Noun Count*

(descriptiveness), *Proper Noun Recall* and *CLIPScore* (alignment), and *Word Count*.

We fit three logistic regression models to examine how linguistic and informational features were associated with caption source (ground truth), perceived AI generation, and caption preference (see Figure 3.4). Model 1 (Actually AI-Generated) was a logistic regression predicting whether a caption was AI-generated (AI-generated =1) versus journalist-written; coefficients from this model characterize the informational features and AI-related linguistic markers that distinguish AI-generated from journalist-written captions and provide a reference for interpreting human responses. Model 2 (Perceived AI) was a mixed-effects logistic regression predicting whether participants perceived a caption as AI-generated (perceived AI =1), with random intercepts for participants to account for individual differences in attribution. Model 3 (Less Preferred) was a mixed-effects logistic regression predicting whether participants preferred the journalist-written caption over the AI-generated caption (preferred journalist =1). For Model 3, preference responses from Experiment 1 were reverse-coded so that higher values indicate lower preference for AI-generated captions, aligning the directionality with Model 2, such that higher values in both models reflect greater attribution to AI.

Because human judgments rely on subtler and noisier cues than the objective source of a caption, absolute coefficient magnitudes were smaller in Models 2 and 3 than in Model 1. Even so, the large number of observations in Models 2 and 3 (derived from 136 stimuli per participant) provided sufficient power to detect reliable associations. Accordingly, our interpretation focuses on similarities and differences in coefficient direction between the ground-truth model and human judgments rather than on absolute effect sizes. Results are shown in Figure 3.4, with full coefficient tables provided in Appendix Table 3.1.

To validate that the context manipulation on generated captions in Experiment 2 produced the intended effect, we conducted the same feature extraction and coefficient analyses used in Model 1 across three captions from three generative strategies (Models 4-6), which corresponded to the result shown in Figure 3.6. and Appendix Table 3.2.

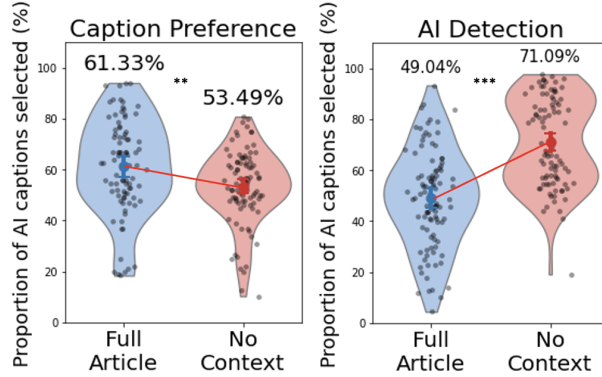


Figure 3.3: **Experiment 1 results.** Proportion of AI-generated captions selected by participants in the preference task (left) and the AI-detection task (right), comparing the *no-context* and *full-article* conditions. Each dot represents one participant’s proportion of AI-caption selections across images.

3.3 Results

3.3.1 Experiment 1: Audiences show preference for AI captions and struggle to detect them

We find that well-documented human biases in evaluating AI-generated content [126, 39, 121, 54] extend to multimodal contexts (Figure 3.3). Even in the *no context* condition—where AI models are at an informational disadvantage—participants preferred AI-generated captions 53% of the time and misclassified them as human-written in 29% (i.e., 71% AI detection accuracy) of cases.

Under the *full article* condition, when the AI model received the accompanying article and image to generate captions, participants’ preference for AI-generated captions increased significantly from 53% to 61% ($p = .002$), and their misclassification rate as human-written rose from 29% to 51% ($p < .001$). Context thus makes AI captions both more appealing and less detectable, amplifying audiences’ biases that likely stem from linguistic heuristics when evaluating multimodal generative content. Our findings are consistent with evidence from the previous Wikipedia image captioning study [125]. These results suggest that the heuris-

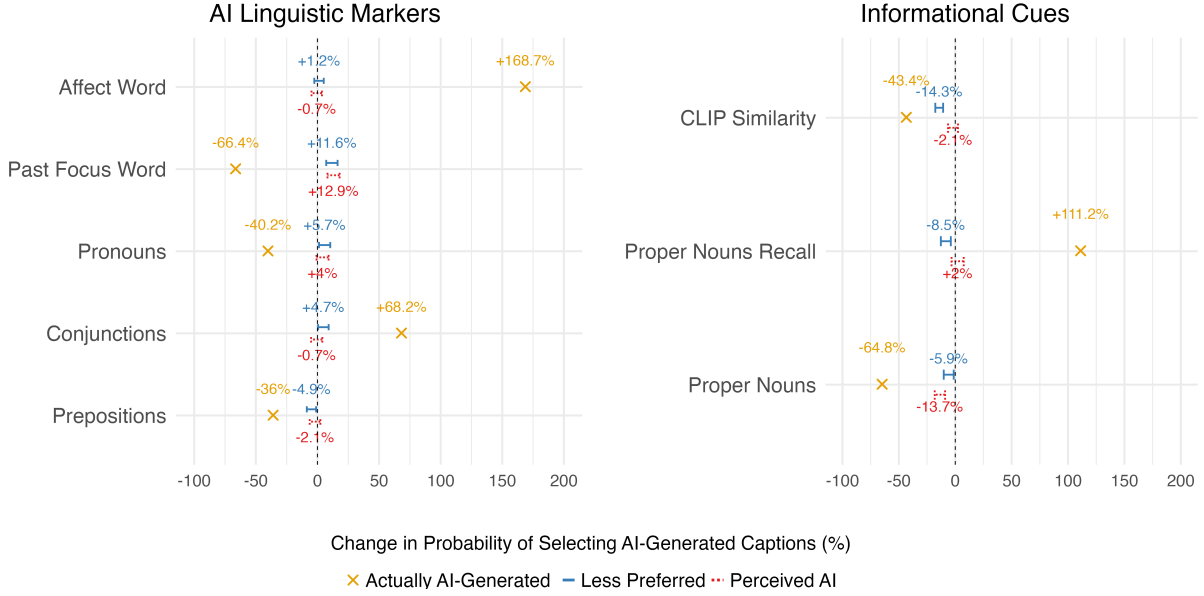


Figure 3.4: Directionality of feature coefficients for ground truth (actually AI-generated, orange $\times$) and human responses (blue and red bars), shown as deviations from an odds ratio (OR) of 1 in logistic regression models. Values represent $(OR - 1) \times 100\%$, where positive percentages indicate higher odds of being AI-generated (or choosing AI-generated) and negative percentages indicate lower odds relative to chance. For example, $+50\%$ means the feature is associated with 50% higher odds, whereas -50% means 50% lower odds. Error bars for human responses indicate 95% confidence intervals (CIs) around the estimates. Complete regression coefficients are reported in Appendix Table 3.1.

tics shaping human judgments of AI-generated content across domains such as professional communication, hospitality, dating, and journalism, [39, 121, 126, 54, 250, 105, 270] likely persist even when both visual and contextual cues are present.

3.3.2 Computational Analysis: Flawed linguistic heuristics persist but informational cues offer reliability

To understand why audiences struggled to distinguish and often preferred AI captions, we computationally analyzed caption features (Figure 3.4). Building on frameworks for analyzing AI-generated text [105, 270, 202], we performed feature selection and identified a consis-

tent set of linguistic features, drawing on categories from the Linguistic Inquiry and Word Count (LIWC) lexicon [188], including *Affect*, *Past Tense*, *Pronouns*, *Conjunctions*, *Causation*, *Differentiation*, *Quantifiers*, and *Adverb* categories, and *Word Count*. Informativeness was measured along two dimensions: cross-modal alignment (image-caption alignment via *CLIPScore* and caption-article alignment via *Proper Noun Recall*) and caption descriptiveness (measured through *Proper Noun Counts*). This feature analysis focuses exclusively on scenarios where AI captions were generated with article context, namely the *full article* condition, given their higher quality and better resemblance to real-world news captioning.

We fit three logistic regression models to examine how linguistic and informational features were associated with captions that were (1) actually AI-generated (ground-truth source), (2) perceived as AI-generated, and (3) less preferred by participants.

Odds ratios from the ground-truth model (orange $\times$ in Figure 3.4) revealed distinct linguistic patterns, which we term AI linguistic markers: AI-generated captions contained more affective language and conjunctions but showed less past focus, fewer pronouns, and fewer prepositions than journalist-written captions. On the other hand, AI captions showed weaker image-caption alignment (i.e., lower *CLIPScore*) and reduced descriptiveness (i.e., lower *Proper Noun Counts*), yet substantially higher article recall (i.e., greater *Proper Noun Recall*), reflecting the model’s direct access to article text during generation.

Results from the two human-response models revealed systematic divergences from the ground-truth linguistic markers. In the perceived AI-generated model (red **dashed bars** in Figure 3.4), participants were more likely to classify captions as AI-generated when they contained greater past-tense usage and higher pronoun frequency—features that, in reality, characterized journalist-written captions rather than AI-generated ones. Other linguistic features that reliably distinguished AI from human captions in the ground-truth model were largely non-diagnostic for participants. A similar pattern emerged in the preference model (blue **solid bars** in Figure 3.4): captions with greater past orientation and pronoun use were less preferred, while most remaining linguistic features did not significantly influence preference judgments.

In contrast, informational features showed strong and consistent effects across both human-response models. Participants reliably associated stronger image–caption alignment and greater descriptiveness with both human authorship and higher preference, aligning with the true caption source. These cues, however, were weighted differently across tasks: preference judgments placed greater emphasis on image–caption alignment, whereas AI-detection judgments relied more heavily on descriptiveness. Notably, higher article–caption alignment—where AI captions objectively outperformed journalist captions due to direct access to article text—was also associated with increased preference, revealing how seemingly positive features can paradoxically favor AI-generated content.

These feature analyses suggest that while participants continue to rely on flawed linguistic cues when assessing AI-generated text, informational cues—particularly image–caption alignment and caption descriptiveness—are more reliably aligned with true caption sources and may help mitigate such bias. Critically, these informational cues are not fixed properties of captions but are directly shaped by the textual context available to the AI system at generation time.

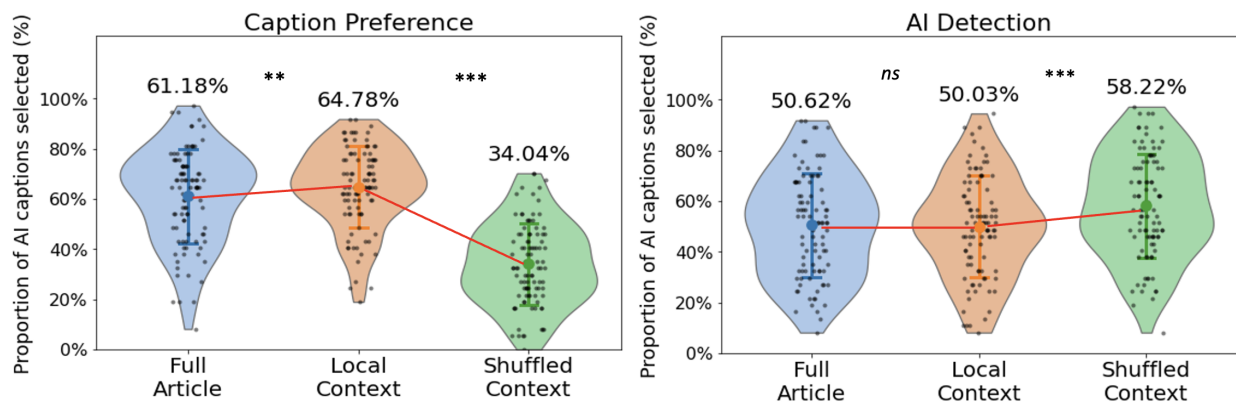


Figure 3.5: **Experiment 2 results.** Proportion of AI-generated captions selected by participants in the preference task (left) and the AI-detection task (right) across three generation conditions: *full article*, *local context*, and *shuffled context*. Each dot represents one participant’s proportion of AI-caption selections across images.

3.3.3 Experiment 2: Contextual manipulation exploits human evaluation biases

Accordingly, altering the context provided to AI systems should predictably shape caption features and, in turn, influence both human preferences and AI detectability. We tested this in Experiment 2 using captions generated under three conditions: *full-article context* (as in the previous experiment), *local context* (sentences surrounding the image plus the article’s lead sentence), and *shuffled context* (mismatched text drawn from unrelated articles).

In Experiment 2, *local-context* captions were preferred 64% of the time over journalist-written captions—significantly higher than *full article* (60%, $p = .001$) and far exceeding *shuffled context* (33%; $p < .001$; Figure 3.5). Enhanced image–caption alignment likely drove this preference hierarchy, with *local context* mitigating VLMs’ perceptual weaknesses by focusing on surrounding sentences that contained the most visually relevant information [256]. Strikingly, captions generated from entirely incorrect, *shuffled context* were still preferred 33% of the time, revealing how surface-level descriptiveness can mislead judgement.

In contrast, detection accuracy in Figure 3.5 showed minimal sensitivity to context quality: *local context* (50%), *full article* (51%), and *shuffled context* (58%) yielded only modest differences, with a significant effect observed only for the *shuffled context* condition compared with the other two (both p ’s $< .001$). Recall that in the previous experiment, detection accuracy reached 71% (i.e., a 29% misclassification rate) when *no context* was provided. Captions generated from *shuffled context* were detected well below this level, consistent with prior Wikipedia-based comparisons of *random* and *no-context* conditions [125]. This pattern suggests that the surface descriptiveness introduced by misleading proper nouns makes inaccurate captions much more difficult to identify than those lacking such cues.

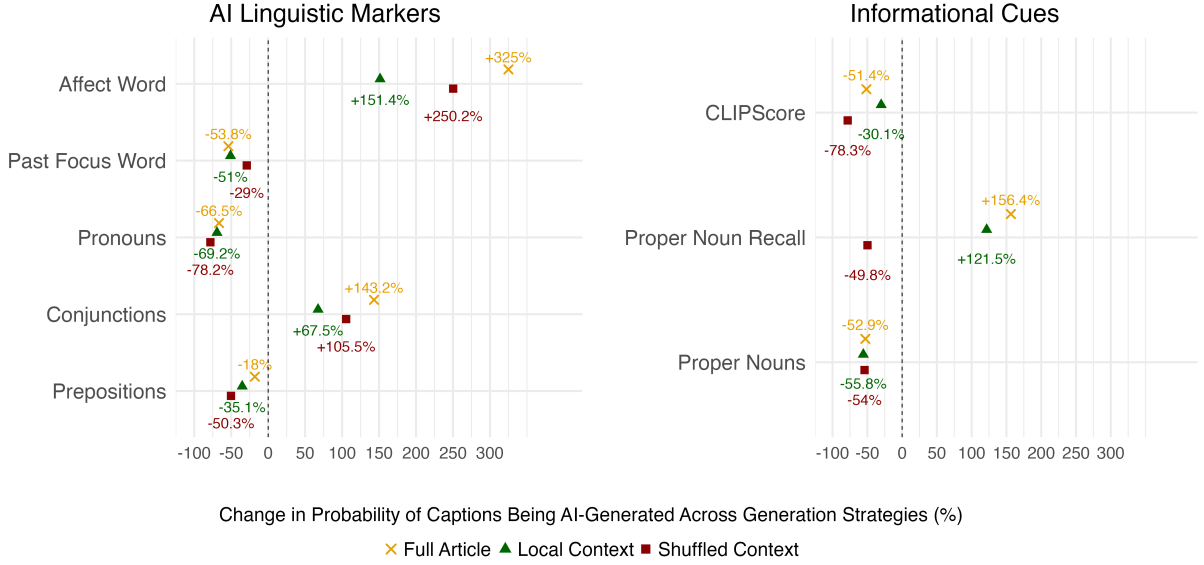


Figure 3.6: Coefficients of linguistic and informational features for captions generated using *full-article*, *local-context*, or *shuffled-context* input, based on 111 generated stimuli in each condition. Complete regression coefficients are reported in Appendix Table 3.2.

A mixed-effects analysis revealed a significant main effect of task ($p < .001$) and a significant task $\times$ context interaction ($p < .001$), confirming that contextual strategies substantially affected caption preferences while influencing detection accuracy in a qualitatively different manner. Specifically, the *local context* and *full article* conditions proved indistinguishable in the detection task ($p > .05$) despite clear preference differences. These results indicate that audience judgments of AI detectability are more susceptible to contextual manipulations than their caption preferences, highlighting a persistent vulnerability to well-formed but contextually misleading AI-generated content when task framing shifts.

To assess whether AI-generated captions produced under *local context*, *full-article*, and *shuffled context* conditions reflected the intended manipulation of captions’ informational features, we additionally examined odds ratios from logistic regression models independently predicting caption source across these three conditions (Figure 3.6). These models confirmed that the context manipulation operated as intended: AI linguistic markers remained relatively consistent across conditions, while cross-modal alignment features were effectively

altered. *Local context* narrowed the image-alignment gap between AI and journalist captions from -51% to -30%. *Shuffled context* successfully disrupted alignment, producing the lowest scores for both image-caption and article-caption measures while maintaining descriptiveness at the same level as other strategies. This selective disruption isolates a mechanism by which AI captioning can preserve surface-level quality even as contextual accuracy deteriorates, creating conditions under which misleading captions may be especially difficult for audiences to detect.

3.4 Discussion

Extending previous research on people’s limitations in evaluating AI-generated content in text-only domains, such as news, autobiographical writing, recipes, and poetry [39, 121, 126, 54, 250, 105, 270], our findings show that these limitations persist when language is embedded within richer multimodal contexts. In news image captioning, where text is expected to cohere with visual information, participants continued to rely on unreliable linguistic heuristics when judging AI-generated captions. Consistent with recent work using Wikipedia data [125], we also found that providing additional context enhances the overall quality of captions. This improvement amplifies people’s existing bias toward AI-generated text: participants preferred AI captions even more and could identify them as AI-generated only at chance levels.

Combining experimental evidence with computational analyses, we identify a set of reliable informational cues—particularly image-caption alignment and caption descriptiveness—that can mitigate these judgment limitations. Modern language models are increasingly adept at imitating human writing styles, a trend reflected in participants’ inability to distinguish AI- from journalist-written text, both in our study and in prior work [39, 121, 126, 54, 250, 105, 270]. However, aligning images with textual meaning remains a far greater challenge for the current AI systems. Previous research has documented persistent weaknesses in the perceptual understanding and cross-modal reasoning of these models, in contrast to their strong performance in purely textual domains [174, 240, 235, 148]. Achieving

high image–caption alignment in news contexts further requires models to draw on extensive world knowledge—recognizing public figures, locations, and events, or at least selecting the correct entities from a large set of plausible names mentioned in an article [217, 256].

Crucially, these weaknesses also point to a practical opportunity: while stylistic language cues are easily mimicked, misalignments between images and text are substantially harder for AI systems to conceal. Image–caption alignment therefore serves as a potentially actionable diagnostic anchor for human evaluators, media literacy interventions, and automated auditing tools. Likewise, descriptiveness in professional journalism reflects an established “who–where–when–what” convention [256] that current VLMs may not have fully internalized, in part because they are not trained under norms emphasizing factual completeness and contextual grounding [202].

Together, these gaps explain why informational cues can mitigate linguistic biases in AI language assessment, providing a more reliable basis for evaluating multimodal AI content. In practice, guiding media consumers to assess whether a caption or image description accurately reflects the named entities referenced in the article, whether visual details plausibly correspond to the depicted people, locations, and events, and whether the caption adheres to journalistic conventions may help counter the persuasive power of fluent yet contextually misleading AI-generated content.

Importantly, our results show that the effectiveness of these informational cues depends strongly on task framing. When judging which caption they preferred, participants drew on an intuitive mental model integrating alignment and descriptiveness. When asked instead to detect whether a caption was AI-generated, they defaulted to linguistic heuristics and were easily misled by surface-level descriptiveness, even when alignment was disrupted. This pattern suggests that audiences are more accustomed to evaluating linguistic style—an intuitive but unreliable signal driven by prominent exposure to AI written text—than to assessing cross-modal coherence in AI-generated content. While people have picked up the association of certain language patterns with AI authorship, they often lack awareness of how image-text alignment contributes to content authenticity. As a result, judgments of AI

detectability are more susceptible to contextual manipulations than caption preferences, as participants selectively underutilize informational cues even when they are accessible.

3.4.1 Limitations and ethics statement

We used the VOA news dataset because it enables us to locate where each image appears within the original article, which was essential for implementing the local-context condition. While we expect the findings to generalize to larger news corpora such as GoodNews [19] and VisualNews [146], future work could test robustness across more diverse news outlets and content domains, including encyclopedic and informal text formats (e.g., Wikipedia or blogs) [125, 27]. Additionally, our analyses focused on two closely related VLMs, GPT-4V and GPT-4o. Although these were considered state-of-the-art at the time of the experiments, future research should extend comparisons to a broader range of models, including open-source variants and prompt designs. Finally, our results are limited to the current generation of language models and to people’s existing heuristics for evaluating AI-generated captions. Nonetheless, prior work has consistently shown that modern VLMs remain weaker in perceptual and cross-modal reasoning compared to their strong textual reasoning abilities and to humans, who are robust perceptual and multimodal reasoners [174, 240, 235, 148]. We therefore expect these performance gaps to persist in the near term, suggesting that informational and cross-modal features will continue to serve as reliable cues for mitigating human judgment biases toward generative content.

3.5 Appendices

3.5.1 Examples of top- and bottom-ranked news image captions of informational features

All examples are drawn from journalist-generated VOA news image captions collected by Li et al. [138]. Figure 3.7 shows examples of top- and bottom-ranked news image captions across the three informational features introduced here: image-caption alignment, caption-

article alignment, and caption descriptiveness.

Image-caption alignment A high image–caption alignment (i.e., high CLIPScore) caption typically describes the event literally depicted in the image (e.g., *a U.S. representative at a news conference, both shown in the picture*). In contrast, a low image–caption alignment (i.e., low CLIPScore) caption may mention key elements (e.g., *a Facebook page*) not visible in the image, or actors (e.g., *protesters*) who are not clearly discernible.

CLIPScore: Given an image I and caption C , image and text embeddings, $f_I(I)$ and $f_T(C)$, are extracted from a pretrained CLIP model (ViT-L/14) and ℓ_2 -normalized. Image–caption alignment is computed as the cosine similarity between the two embeddings:

$$\text{CLIPSim}(I, C) = \frac{f_I(I) \cdot f_T(C)}{\|f_I(I)\| \|f_T(C)\|}.$$

Following common practice, cosine similarity scores are linearly rescaled for interpretability:

$$\text{CLIPScore}(I, C) = 100 \times \text{CLIPSim}(I, C),$$

with higher values indicating stronger semantic correspondence between the image and caption.

Caption-article alignment For high caption–article alignment, as measured by high Proper Noun Recall, the caption contains several named entities (e.g., *Tower Bridge, London, Donald Trump, U.S.*) that also appear in the accompanying news article. In contrast, a low caption–article alignment (i.e., low Proper Noun Recall) caption may include only a single proper noun (e.g., *U.S.*) despite the article mentioning many others (e.g., *U.S. President Barack Obama, Hillary Clinton, Cameroon*).

Proper Noun Recall: Let $\mathcal{P}_A$ denote the set of proper nouns extracted from the article text and $\mathcal{P}_C$ denote the set of proper nouns extracted from the caption. Proper Noun Recall is computed as

$$\text{PNRecall} = \frac{|\mathcal{P}_C \cap \mathcal{P}_A|}{|\mathcal{P}_A|},$$

with higher values indicating greater coverage of article-specific named entities in the caption. A value of 1 indicates that all article proper nouns are mentioned in the caption, whereas a value of 0 indicates no overlap.

Caption descriptiveness Caption descriptiveness, measured by the proportion of proper nouns (i.e., proper nouns normalized by caption word count), reflects how detailed a caption is without considering whether those entities accurately correspond to the image content. A highly descriptive news image caption contains multiple named entities (e.g., *Republican, Donald Trump, Palm Beach County, GOP, Lincoln Day Dinner, Mar-a-Lago Club*) that provide additional context to help the audience understand the broader news event beyond the visual elements. In contrast, a low-descriptiveness caption contains few or no proper nouns and tends to be more generic.

Proper Nouns: Computed as the proportion of proper nouns in the caption. Let N_{PN} denote the number of proper nouns and N_{words} denote the total number of words in the caption. The measure is defined as

$$\text{Proper Nouns} = \frac{N_{\text{PN}}}{N_{\text{words}}},$$

which normalizes entity counts by caption length to enable comparisons across captions of different lengths, with higher values indicating more descriptive captions due to the inclusion of specific actors, locations, and institutions (i.e., named entities) rather than generic scene descriptions.

Image-Caption Alignment

High CLIPScore = 0.95



Richard Holbrooke, the U.S. special representative to Pakistan and Afghanistan, listens to a question during a news conference.

Low CLIPScore = 0.51



Protesters say even though officials were arrested, the Facebook page where exam questions and answers are leaked is still running.

Caption-Article Alignment

High Proper Noun Recall = 0.31

...Julie Chasin, a 42-year-old teacher originally from New York who has lived in **London** for a decade, joined the **Tower Bridge** protest. I hope he's kept in check. Supporters celebrate **Trump's** supporters, who admire his experience in business and see him as an outsider who will take a fresh approach to politics, were also marking his inauguration in **London**. Some 200 of them gathered under the gaze of a life-sized cardboard figure of **Trump** to celebrate at an exclusive club near the Ritz hotel. As he completed the oath, the room broke into a standing ovation complete with high-five celebrations. In Moscow, Russians hoping **Trump** will usher in a new era of detente welcomed his inauguration. Russian nationalists held an all-night party at what used to be the main Soviet-era post office in Moscow. Painted matryoshka dolls, or Russian nesting dolls, bearing the faces of then-U.S. Republican presidential nominee **Donald Trump** and Russian President Vladimir Putin are displayed at a souvenir shop in central Moscow, Russia, Nov. 7, 2016...



Campaigners hold a banner reading "Build Bridges Not Walls" from **Tower Bridge** in **London** to protest **Donald Trump's** inauguration as U.S. president.

Low Proper Noun Recall = 0.03

In Kenya, where President Barack Obama has familial roots, many were hoping Hillary Clinton could carry on his legacy, but were prepared for disappointment. Sara Asafu-Adjaye, a Ghanaian social activist, says she would have loved to have seen the first female U.S. president. Charles Eboune, an international relations expert in Cameroon, was dismayed Clinton's decades of experience in Washington was not enough to carry her to victory. Change in style Now, Africans have to adjust to the fact that a president with a very different style and priorities from both Obama and Clinton is going to take office. More than a few Africans are worried about the policy changes Trump may implement. African leaders and politicians in Kenya, South Africa, Uganda, Tanzania, Burundi, the Democratic Republic of Congo and Somalia, to name a few, have already sent or tweeted messages to Trump, congratulating him on his victory.



Guests attend an election results watch breakfast at the home of U.S. Ambassador Robert Godec.

Caption Descriptiveness

High Proper Noun Ratio = 0.65



Republican presidential candidate **Donald Trump** speaks during the **Palm Beach County GOP Lincoln Day Dinner** at the Mar-A-Lago Club.

Low Proper Noun Ratio = 0



Relatives and friends mourn at the coffin containing the remains of a fire victim at a gym where bodies were brought for identification.

Figure 3.7: Examples of top- and bottom-ranked news image captions across informational features. Upper: image-caption alignment; middle: caption-article alignment; bottom: caption descriptiveness. Examples are adapted from journalist-generated VOA news image captions collected by Li et al. [138].

3.5.2 Regression coefficients for ground truth vs. human judgment (Models 1–3)

Corresponding to Figure 3.4 in the main text, logistic regression models (Table 3.1) predict (1) whether a caption was actually generated by AI, (2) whether a caption was perceived as AI-generated, and (3) whether an AI-generated caption was less preferred in paired comparisons. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Odds ratios (ORs) and 95% confidence intervals (CIs) were computed by exponentiating the logistic regression coefficients and their bounds: $OR = \exp(\beta)$ and $95\% \text{ CI} = [\exp(\beta - 1.96 \times SE), \exp(\beta + 1.96 \times SE)]$. An OR greater than 1 indicates a positive association with the outcome (e.g., $OR = 1.05$ corresponds to 5% higher odds of a caption being AI-generated in Model 1, being perceived as AI-generated in Model 2, or an AI-generated caption being less preferred in Model 3), whereas an OR less than 1 indicates a negative association.

Table 3.1: Regression coefficients for models predicting (1) actually AI-generated content, (2) captions perceived as AI-generated, and (3) preference for journalist-generated captions (i.e., AI-generated captions are less preferred).

Feature	(1) Actually AI	(2) Perceived as AI	(3) Less Preferred
Informational Features			
<i>Alignment</i>			
CLIPScore	0.57 (0.38, 0.82)**	0.98 (0.93, 1.02)	0.85 (0.82, 0.89)***
Proper Noun Recall	2.11 (1.40, 3.34)***	1.02 (0.97, 1.07)	0.91 (0.87, 0.96)***
<i>Descriptiveness</i>			
Proper Nouns	0.35 (0.22, 0.54)***	0.86 (0.82, 0.91)***	0.94 (0.90, 0.99)*
AI Linguistic Markers			
Affect Words	2.69 (1.82, 4.12)***	0.99 (0.95, 1.04)	1.01 (0.97, 1.05)
Past Focus Words	0.34 (0.19, 0.54)***	1.13 (1.08, 1.18)***	1.12 (1.07, 1.16)***
Pronouns	0.60 (0.40, 0.86)**	1.04 (0.99, 1.09)	1.06 (1.01, 1.10)*
Conjunctions	1.68 (1.19, 2.44)**	0.99 (0.95, 1.04)	1.05 (1.00, 1.09)*
Prepositions	0.64 (0.44, 0.91)*	0.98 (0.94, 1.02)	0.95 (0.92, 0.99)*
Human Heuristics			
Word Count	0.79 (0.53, 1.15)	0.90 (0.86, 0.95)***	0.67 (0.64, 0.70)***
Causation Words	1.20 (0.86, 1.72)	0.95 (0.91, 0.99)*	0.91 (0.87, 0.94)***
Nominal Subjects	0.79 (0.54, 1.15)	1.01 (0.97, 1.06)	1.10 (1.06, 1.15)***
Quantifiers	0.98 (0.70, 1.40)	0.96 (0.92, 1.01)	0.96 (0.92, 1.00)
Adverbs	0.94 (0.69, 1.30)	0.98 (0.93, 1.03)	0.99 (0.95, 1.03)
Model Fit			
Constant	0.96 (0.67, 1.35)	3.05 (2.47, 3.77)***	0.86 (0.75, 0.98)*
Observations	272	12,648	13,056
Log-likelihood	-110.87	-6,752.67	-8,206.56
AIC	249.74	13,535.33	16,443.12

3.5.3 Regression coefficients for generation strategies (Models 4–6)

Corresponding to Figure 3.6 in the main text, logistic regression models (Table 3.2) predict whether a caption was generated by AI (i.e., GPT-4o) under three contextual conditions: (4) full-article context, (5) local context, and (6) shuffled context. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Odds ratios (ORs) and 95% confidence intervals (CIs) were computed by exponentiating the logistic regression coefficients and their bounds: $OR = \exp(\beta)$ and $95\% CI = [\exp(\beta - 1.96 \times SE), \exp(\beta + 1.96 \times SE)]$. An OR greater than 1 indicates a positive association with the outcome (e.g., $OR = 1.05$ corresponds to 5% higher odds of a caption being AI-generated in Models 4–6), whereas an OR less than 1 indicates a negative association.

Table 3.2: Regression coefficients for models predicting (4) actually AI-generated content from the full article, (5) actually AI-generated content from a local article context, and (6) actually AI-generated content from a shuffled context.

Feature	(4) Full Article	(5) Local Context	(6) Shuffled Context
Informational Features			
<i>Alignment</i>			
CLIPScore	0.49 (0.28, 0.79)**	0.70 (0.48, 0.99)*	0.22 (0.10, 0.43)***
Proper Noun Recall	2.56 (1.63, 4.42)***	2.22 (1.44, 3.63)***	0.50 (0.20, 1.17)
<i>Descriptiveness</i>			
Proper Nouns	0.47 (0.26, 0.80)**	0.44 (0.27, 0.69)***	0.46 (0.23, 0.85)*
AI Linguistic Markers			
Affect Words	4.25 (2.52, 7.76)***	2.51 (1.65, 3.95)***	3.50 (1.91, 7.03)***
Past Focus Words	0.46 (0.23, 0.79)*	0.49 (0.29, 0.78)**	0.71 (0.43, 1.15)
Pronouns	0.34 (0.18, 0.58)***	0.31 (0.17, 0.51)***	0.22 (0.09, 0.46)***
Conjunctions	2.43 (1.51, 4.07)***	1.67 (1.15, 2.50)**	2.05 (1.26, 3.52)**
Prepositions	0.82 (0.51, 1.30)	0.65 (0.43, 0.96)*	0.50 (0.26, 0.89)*
Other Covariates			
Word Count	1.27 (0.85, 1.91)	1.57 (1.09, 2.35)*	2.11 (1.25, 3.83)**
Causation Words	1.43 (0.93, 2.27)	1.08 (0.75, 1.58)	1.26 (0.76, 2.20)
Nominal Subjects	1.08 (0.66, 1.79)	1.17 (0.80, 1.74)	1.36 (0.77, 2.48)
Quantifiers	0.94 (0.63, 1.39)	0.84 (0.56, 1.22)	0.81 (0.46, 1.34)
Adverbs	0.68 (0.38, 0.97)*	0.70 (0.43, 0.98)*	0.64 (0.39, 0.93)*
Model Fit			
Constant	0.95 (0.61, 1.46)	0.89 (0.61, 1.27)	0.92 (0.53, 1.61)
Observations	222	222	222
Log-likelihood	-75.43	-99.05	-56.65
AIC	178.86	226.11	141.29

CHAPTER 4

Visual Narrative Freedom under Loose Image–Text Alignment

Abstract

Mapping linguistic to visual representations is a central objective in vision-language model (VLM) development. Consequently, current VLM benchmarks and training objectives predominantly optimize for literal, descriptive correspondence between modalities. However, in real-world multimedia contexts, text and images are rarely strictly redundant; they are often loosely coupled, interacting at higher narrative levels. To capture this dynamic, we introduce *Visual Narrative Freedom* (VNF): the degree to which a generative system can produce multiple plausible visual realizations from an underdetermined textual input. In this chapter, we evaluate text-to-image (T2I) generation across a continuum of linguistic constraints, conditioning models on visual descriptions, captions, and full news articles. Using perceptual, structural, and semantic metrics, we first demonstrate that as textual constraints loosen (and VNF increases), T2I models produce images that are significantly more diverse from ground-truth reference photographs. Human evaluation further reveals that providing generation models with greater VNF significantly increases the likelihood that their outputs are preferred over authentic news photographs, despite the known stylistic limitations of AI imagery. Ultimately, these findings suggest that the creation of persuasive multimodal misinformation is more imminent than evaluations of T2I systems based solely on text descriptions may indicate. This highlights the need for VLM evaluation frameworks that better



Figure 4.1: **Visual Narrative Freedom shapes human preference in T2I tasks.** Left: Example inputs for a news instance, with the reference image and generated images under each condition. Right: As VNF increases, preference shifts toward AI-generated images (Section 4.5); gray dots represent individual participants, with means and bootstrapped 95% CIs.

capture the underdetermined nature of real-world image-text alignment¹

4.1 Introduction

Mapping linguistic representations to visual representations and integrating the two modalities has been at the center of vision–language model (VLM) development [273, 137, 116]. In today’s models, many existing approaches align images and text through shared embeddings or cross-modal supervision [139, 64], often optimizing for literal correspondence between textual descriptions and visual content [194]. However, such formulations overlook a fundamental aspect of multimodal communication: images and text do not merely translate information but interact at higher narrative and symbolic levels [275, 12].

Across domains such as journalism, advertising, and social media, images often complement or symbolically represent textual narratives rather than redundantly mirroring them [275, 151, 125, 29]. As a result, the relationship between text and image may be loosely coupled: a single textual narrative can plausibly correspond to multiple visual realizations.

¹All data and code are available at <https://app.box.com/s/6cy5nh47o2l57skhj0i4hkw4uq2ym1s>.

This flexibility is especially consequential for generative systems, which are not constrained by physical access, timing, or the availability of real-world scenes [37].

While recent work has started to investigate these loose image-text correspondences, research has been largely constrained to image-to-text (I2T) models and applications [93, 125, 90]. In this work, we instead frame this loose coupling as a fundamental characteristic of multimodal generation and demonstrate the direct consequences it has for text-to-image (T2I) settings. We refer to this property as *visual narrative freedom* (VNF): the degree to which a system can produce multiple plausible visual realizations for the same textual context. VNF emerges when the mapping from text to image is underdetermined, allowing generative models to instantiate different visualizations of the same narrative.

Visual journalism provides a high-stakes context in which these dynamics become visible [231, 185]. Images accompanying news articles often serve narrative or framing functions rather than strictly documenting the events, locations, or figures described in the text. At the same time, news photographs carry evidentiary authority and are often perceived as eyewitness records of reality [8, 204, 210]. These characteristics make news media an informative setting for studying the downstream impact of generative systems when image-text relationships are loosely constrained. Even when generated images achieve high perceptual fidelity, they may retain stylistic cues that distinguish them from traditional photojournalism [231, 239], potentially creating a baseline stylistic disadvantage.

As Figure 4.1 illustrates, news articles serve as an especially interesting resource to investigate the effects of VNF, since they inherently provide naturalistic examples of text that vary across three VNF levels for each individual image: *descriptions* (which can be generated from the image directly when absent), *captions*, and full news *articles*. This allows us to examine how varying levels of textual constraint shape image diversity and how these differences interact with audience preferences across levels of VNF when images are presented alongside articles.

In summary, our main contributions are: (1) We show that loosely coupled image-text relationships is not solely an image-to-text (I2T) problem but generalizes to evaluating text-

to-image (T2I) systems, making it a bidirectional Vision-Language alignment issue. (2) We introduce VNF, a measurable generative property that characterizes the range of plausible visual realizations from a given textual context. (3) Through pre-registered experiments in news media, we demonstrate that VNF increases audience preference for images presented alongside articles, even when AI-generated imagery carries stylistic disadvantages.

Overall, our results highlight that varying visual narrative freedom can significantly moderate the persuasiveness of multimodal outputs. Concretely, the finding that people start to reliably prefer AI-generated images over real news images when models are given sufficient VNF to exploit opens important questions about the risk factors of AI-generated multimodal misinformation.

4.1.1 Background on AI-generated photorealistic imagery

Advances in T2I generative models have enabled the production of highly detailed and photorealistic images across domains such as art, education, marketing, and journalism [244, 157, 149]. Early systems such as DALL-E [198, 197] and Stable Diffusion [209] established the feasibility of generating visually coherent scenes from natural language prompts. More recent models such as Google’s Imagen [213, 104] demonstrate substantial improvements in photorealism, compositional fidelity, and semantic alignment while supporting longer and more complex textual inputs [169, 45, 200]. Despite not capturing physical events through optical processes like traditional photography, AI-produced photorealistic imagery increasingly approaches the perceptual quality of professional photography by modeling statistical regularities in large-scale vision–language corpora [185, 135, 244].

T2I generation systems synthesize images from natural language prompts by learning joint visual–text representations [197]. Recent advances integrate diffusion-based generative modeling [91, 209] with large-scale contrastive language–vision pretraining (e.g., CLIP) [194]. Diffusion architectures enable high-resolution synthesis through iterative denoising [91] while contrastively trained encoders align textual inputs with visual embeddings learned from web-scale image–text corpora [194]. This integration substantially improves prompt adherence,

compositional grounding, and visual fidelity compared to earlier GAN-based approaches [201, 267, 254, 10].

Among T2I applications, photorealistic synthesis constitutes a particularly demanding setting [244]. Unlike photography, which records physical events through optical processes, generative systems approximate realism by modeling statistical regularities in large-scale visual corpora rather than simulating physical light transport [244]. Improvements in model scale and data have narrowed perceptual gaps between synthetic and photographic imagery [185, 45, 178]. Widely circulated fabricated images such as a fake earthquake in the Pacific Northwest, the Pope in a puffy Balenciaga coat, and staged depictions of Donald Trump being arrested have heightened public concern about deception and manipulation [185, 172, 84].

Despite high perceptual fidelity, AI-generated images retain identifiable stylistic priors. Previous studies document hyper-smooth textures, compositional homogenization, factual artifacts, uncanniness, and demographic exaggerations [45, 200, 181]. In journalistic contexts, where images serve evidentiary and communicative roles as perceived eyewitnesses that anchor shared reality and signal authenticity [8, 204, 210], such stylistic regularities may substantially shape audience evaluations. On the other hand, imperfections associated with real-world image capture—such as lighting irregularities, resolution constraints, or environmental noise—can serve as cues of authenticity, in contrast to the synthetic smoothness of generative images [231].

Developments in AI-generated photorealistic imagery establish two core conditions that make the present study meaningful and timely: contemporary T2I systems can generate images with near-photographic perceptual fidelity; however, they retain stylistic signatures that remain perceptible to viewers. Together, these developments raise important questions about whether such AI visual priors are systematic, whether audiences can detect them, and how these subtle cues may be interpreted and evaluated in high-stakes civic domains such as journalism, where authenticity carries strong normative expectations [84].

4.2 Related Work

Our work connects two strands of literature on image–text relationships. The first examines how VLMs operationalize relationships between images and text, typically assuming tightly aligned descriptive correspondence. The second comes from multimedia communication and journalism, where images and text often interact more loosely in real-world settings. Building on these perspectives, we introduce VNF as a property that emerges when generative models operate under loosely constrained textual inputs.

4.2.1 Image–text relationships in vision–language models

Across vision–language tasks, the dominant paradigm assumes relatively tight semantic alignment between images and text [64, 116]. Image captioning datasets typically describe visible entities, attributes, and actions within an image [124], and many T2I benchmarks similarly rely on prompts that explicitly specify the visual scene to be generated [10]. This assumption is reflected in both training objectives and dataset construction. Large-scale contrastive frameworks such as CLIP [194], ALIGN [106], CLOOB [68], and DeCLIP [140] learn multimodal representations by maximizing correspondence between paired images and captions. Widely used benchmarks, including MSCOCO [143], Flickr30K [259], Visual Genome [127], and Conceptual Captions [220], similarly pair images with text that directly describes visible content.

Recent work has begun to explore richer forms of cross-modal interaction. For example, Hessel et al. [90] introduce the New Yorker Captioning Challenge, which emphasizes generating humorous captions that capture indirect image–text relationships through cultural and experiential references. The Visual Commonsense Reasoning dataset [263] evaluates models’ ability to infer motivations and social context from images, while Concadia [125] distinguishes between *descriptions*, which substitute for images by focusing on visible features, and *captions*, which accompany images to provide contextual information. At the architectural level, BriVL [101] introduces a weakly correlated vision–language pretraining

framework, relaxing the assumption that paired image–text data must exhibit strict semantic correspondence, and Zur et al. [274] propose a CLIP variant specifically finetuned to distinguish visual descriptions from captions.

However, most of these studies remain situated in I2T contexts, primarily due to their task formulation [e.g., 90, 274], and T2I model development appears largely unaffected by these insights [69, 119, 31]. In this work, we specifically focus on the T2I context where a deep exploration of the consequences of loosened image-text alignment is still lacking. Our results highlight the importance of treating varying image-text relationships in VLMs as a fundamental point of variation that needs to be considered when developing evaluation frameworks.

4.2.2 Image–text relationships in multimedia communication

While VLM benchmarking generally assumes a tight correspondence between image and text modalities, real-world multimedia contexts frequently (perhaps, mostly) exhibit relationships that extend well beyond literal description [136]. For example, social media posts frequently pair images with captions that frame interpretation rather than strictly describe visual content [166]. In news media, images accompanying an article may illustrate broader themes, frame particular actors, or evoke emotional responses without depicting the specific events described in the text [151, 179].

Prior work in media theory and cognitive science has proposed frameworks describing how images and text interact in multimodal communication. For example, Barthes [12] distinguishes between *anchorage* (text as anchor) and *relay* (text as complement) relationships, while cognitive accounts such as Zwaan [275] emphasize the extent to which language use is embedded in the environment. Although these frameworks provide valuable theoretical insights, they are difficult to map directly onto computational measurements.

An alignment-based perspective can operationalize real-world image–text relationships as a spectrum of semantic alignment [29], ranging from negative or near-zero correlation (e.g., satirical or contrasting contexts), to weak associations (e.g., symbolic or stock imagery), to

strong correspondence (e.g., images depicting entities or events described in the text). This formulation enables the use of existing cross-modal similarity metrics, such as CLIPScore [89], to study loosely coupled image–text relationships. Additional typologies are reviewed in Appendix 4.7.1.

This view connects directly to loosely coupled image–article pairs in news media, often described as the *depiction gap*: the difference between what an image depicts and why it is connected to the accompanying article [151]. Depiction gaps arise because news images may be selected for thematic relevance, narrative framing, or practical constraints such as image availability and newsroom sourcing limitations [29, 179].

Extending this perspective to text-to-image generation, image–text looseness translates into variation in the semantic constraints imposed by the textual input on the range of plausible visual realizations, which we term Visual Narrative Freedom. Section 4.5.1 further shows that image–text similarity measures (e.g., CLIPScore) can be leveraged to capture depiction gaps. Although CLIPScore is based on a model optimized for strong semantic correspondence, we use it here to capture variation of alignment in real-world news image–article pairings (not as a measure of VNF, which is a generator property), enabling analysis of how T2I systems exploit this looseness to produce persuasive multimodal content.

4.3 Characterizing Visual Narrative Freedom

To characterize Visual Narrative Freedom (VNF) in T2I systems arising from loose image–text alignment, we analyze generated images under three textual input conditions: *description*, *caption*, and *article*. Prior work suggests that these inputs vary in their semantic constraints [19, 125, 151], allowing us to examine how reduced constraint shapes the range of plausible visual realizations produced by generative models. To facilitate direct cross-conditional comparison, all conditions are extracted from the same original news article and corresponding image.

4.3.1 Variation in VNF across text-to-image tasks

These three text conditions with varying semantic constraints provide an interesting testbed to test whether we can characterize persistent effects of VNF on T2I system output.

Descriptions represent the most restrictive condition, explicitly specifying visual attributes such as subjects, environment, and scene composition, leaving relatively little ambiguity about the intended image [93, 19, 125].

Captions provide moderate constraint, summarizing the depicted event while omitting many visual details and allowing for multiple plausible realizations [19, 125, 199].

Articles introduce the least restrictive condition and represent the novel condition introduced in our study. This condition is enabled by recent advances in T2I generators capable of processing longer textual inputs [265]. Full articles provide rich narrative context but rarely specify the precise visual scene accompanying the story, affording greater interpretive latitude for image generation [151, 179].

4.3.2 Measuring visual narrative freedom

Because VNF concerns the range of visual interpretations arising from the same textual context, we measure variation directly in generated images. Using computational similarity metrics, we quantify this variation by comparing plausible generated image candidates from the same textual input to one another (*Intra-Prompt Diversity*) or to a reference photograph representing the ground-truth image pairing instance (*Reference Deviation*).

Across these measurements, if VNF increases with looser textual constraints, we expect the diversity of generated images across candidates and their visual deviation from the original image to increase progressively from *description*- to *caption*- to *article*-based generation. At the same time, images produced by the same T2I system should maintain a consistent photographic style. For the human experiments, we validate this stylistic consistency using reference-free metrics (Appendix 4.7.5).

Intra-Prompt Diversity measures the diversity of visual realizations produced from the

same prompt. For each textual input, multiple candidate images are generated and compared using pairwise similarity. Higher scores indicates greater diversity in visual interpretations and thus greater VNF. Formally, diversity is computed as the average pairwise distance (i.e., $1 - \text{sim}$) across generated candidates:

$$\frac{1}{N(N-1)} \sum_{i \neq j} (1 - \text{sim}(I_i, I_j))$$

where N is the number of generated images for a given prompt, and I_i, I_j denote generated images from the same prompt.

Reference Deviation measures how far generated images deviate from the original photograph paired with each news instance, which serves as one possible visual realization of the narrative. Formally, we compute deviation as

$$1 - \text{sim}(I_{\text{generated}}, I_{\text{reference}})$$

where higher scores indicate greater divergence from the original depiction and thus greater VNF.

Similarity Metrics To capture visual differences (i.e., $1 - \text{sim}$) across multiple levels of representation—from low-level perceptual features to high-level semantic concepts—we consider four complementary similarity metrics that span across perceptual, structural, semantic and conceptual features [72]:

- **LPIPS (Learned Perceptual Image Patch Similarity)** computes distances between intermediate feature activations from a pretrained AlexNet network, calibrated to correlate with human perceptual similarity judgments [268].
- **DISTS (Deep Image Structure and Texture Similarity)** compares structural correlations and texture statistics derived from feature maps of a pretrained VGG network, capturing both spatial layout and texture similarity [50].

- **VGG Similarity** computes cosine similarity between image-level embeddings obtained by globally pooling pretrained VGG feature maps via adaptive average pooling, capturing high-level visual structure [225].
- **CLIP Similarity** measures semantic similarity using cosine distances between image embeddings produced by the CLIP ViT-L/14 model in a joint vision–language representation space aligned with textual concepts [194].

4.4 Model Experiments

To characterize the latitude of VNF in real-world applications, we construct a balanced set of 108 news instances drawn from five major outlets. In order to complement our automatic metrics with in-depth human judgments (Section 4.5), we opted for prioritizing a highly curated, naturalistic dataset and minimize the noise typical of larger ones. Since each instance is generated under three constraint conditions, this yields a targeted set of 324 unique text-image pairs.

4.4.1 Dataset

To assess the VNF potential of T2I systems, we curated a dataset from two established multimodal news corpora: VisualNews [146], including four major outlets (*BBC*, *The Guardian*, *USA Today*, and *The Washington Post*), and *VOA News* from M2E2 [138]. Each selected item contained a news image, caption, and full article. To ensure experimental consistency, account for LongCLIP’s [265] context window (248 tokens), and control participant engagement time, we constrained articles to 50–200 words, required captions to contain at least 15 words to avoid underspecified prompts, and sampled instances to roughly reflect the distribution of outlets and major news topics. Topic distributions and token counts are reported in Appendix 4.7.2.

For each instance, we derived an additional tightly constrained condition by generating a

highly detailed image description using GPT-5². Unlike standard short alt-text descriptions commonly found online, these descriptions were deliberately constructed to be maximally specific, explicitly specifying subjects, environment, lighting, mood, and camera perspective. The exact prompt is provided in the Appendix 4.7.6.

4.4.2 Experiment setup

For each news instance, we generated images under three conditions (description-, caption-, and article-based) using prompts derived from the same underlying content. Each prompt produced 5 candidates per model, yielding an expected total of 4,860 images (108 instances $\times$ 3 conditions $\times$ 5 candidates $\times$ 3 models)³. Exact prompts are provided in Appendix 4.7.6.

Model Selection For this analysis, we prioritized generation speed and high-level compositional and semantic variation over perceptual fidelity. We considered three models: **Nano-Banana-2**, **Realistic-Vision-V3.0**, and **Qwen-Image**. **Nano-Banana-2** is a proprietary T2I model from the Imagen family (used for high-fidelity generation in human evaluations). **Realistic-Vision-V3.0** is a Stable Diffusion-based model fine-tuned for photorealistic generation; we used 50 inference steps and replace its CLIP backbone with LongCLIP (ViT-L/14) to support longer textual inputs (up to 248 tokens). **Qwen-Image** is an open-weight T2I model; we used the DFloater11/Qwen-Image-DF11 checkpoint, a quantized variant optimized for memory-efficient deployment via CPU offloading, with 30 inference steps and a `true_cfg_scale` of 3.0. All other configurations used model defaults.

4.4.3 Results

Similarity metrics (LPIPS, DISTS, VGG, and CLIP) introduced in Section 4.3.2 are used for both intra-prompt diversity and reference deviation measurements, capturing variation across perceptual, structural, semantic, and conceptual dimensions.

²GPT-4o was used for the human experiment (Section 4.5), as GPT-5 had not yet been released.

³We obtained 4,771 images due to **Nano-Banana-2** refusals and the open-source model occasionally generating news article layouts rather than images under the article-based condition.

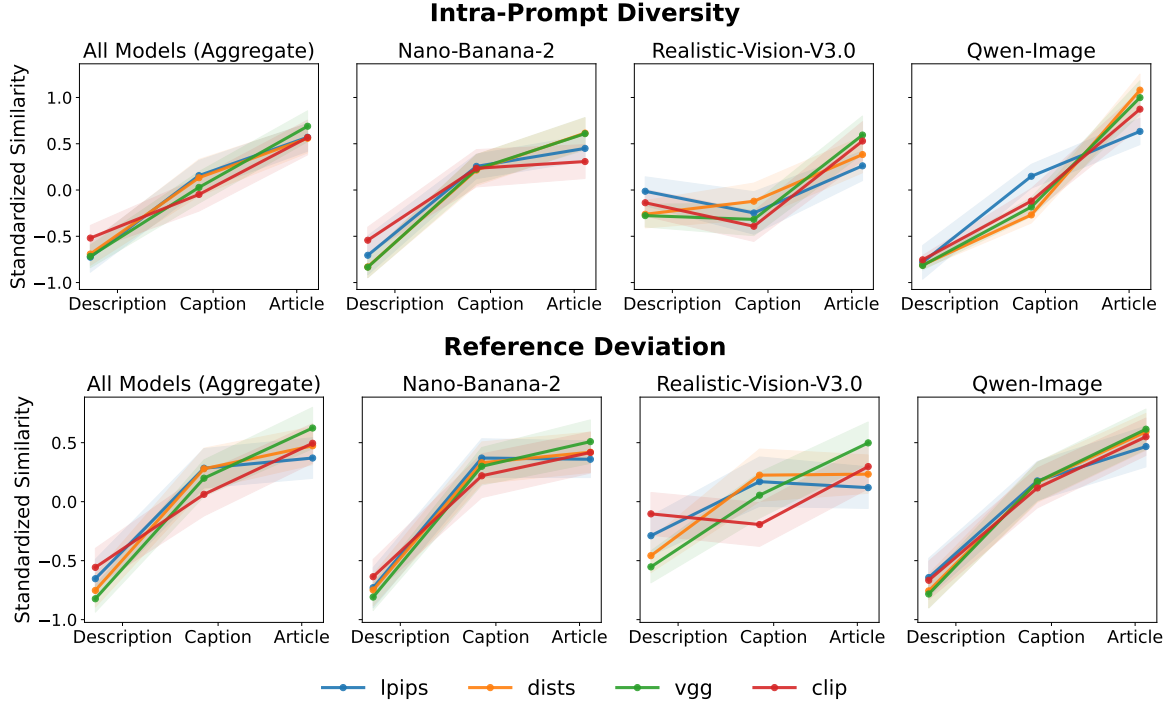


Figure 4.2: VNF gradually increases from description- to caption- to article-based generation. Metrics are z -scored across 108 instances per condition; ribbons denote 95% CIs; results are shown per model and in aggregate (across candidates and models per prompt).

Intra-Prompt Diversity As shown in Figure 4.2, at the aggregated level (i.e., diversity computed across image candidates for each prompt across all three models), diversity increases from description to caption to article, indicating greater variation under less constrained prompts. At the model level, this pattern is most consistent for **Qwen-Image** and **Nano-Banana-2**, which exhibit a monotonic increase in diversity across all metrics. Interestingly, **Realistic-Vision-V3.0** shows a partial deviation: caption-based generations exhibit slightly higher similarity than description-based generations, although article-based generations still yield the highest diversity.

Reference Deviation According to Figure 4.2, at the aggregated level, the three generation conditions exhibit a clear trend: the average deviation between generated images and reference images increases from description- to caption- to article-based generation. Examining individual models, this pattern is most consistent in **Qwen-Image** and **Nano-Banana-2**,

both of which show a strong monotonic increase across all metrics. **Realistic-Vision-V3.0**, however, exhibits a slightly inconsistent pattern across metrics, with only DISTS and VGG showing a monotonic increase. Nevertheless, similar to the intra-prompt diversity results, its article-based generations still show the largest reference deviation.

Across all representation levels, we observe a consistent increase in both intra-prompt diversity and reference deviation as prompts expand from description to caption to article, suggesting that looser image-text coupling affords greater VNF within the same T2I model. This VNF property emerges across the full spectrum of visual representations, spanning perceptual, structural, semantic, and conceptual dimensions, thereby enabling the generator to produce progressively more diverse visual interpretations of the underlying event.

4.5 Human Evaluation

As we show computationally, VNF increases from description- to caption- to article-based generation. When combined with the high visual fidelity of modern T2I systems, this property may have significant implications in high-stakes civic domains such as news media. To investigate this, we recruit human participants to assess preferences for images generated under the three conditions, each compared against the journalist-selected reference. These evaluations reveal that models leveraging VNF can shape how audiences perceive and prefer generated imagery.

For this evaluation, we use a state-of-the-art proprietary T2I system to generate high-fidelity, photorealistic news images, enabling us to measure preferences across conditions in an ecologically valid setting while minimizing low-level stylistic confounds introduced by weaker generators. All procedures were approved by the UCLA Institutional Review Board (#24-0019-AM-001; May 28, 2025). Experiments were preregistered on AsPredicted <https://aspredicted.org/8kd9ss.pdf>.

Stimuli and Model Selection All image descriptions were generated via **gpt-4o**, following the same process documented in Section 4.4.1. We use **Imagen 3** as the image gen-

erator, a state-of-the-art publicly accessible photorealistic T2I system capable of producing coherent human figures and naturalistic scenes [104, 213]. We further conducted extensive controls and validation to ensure balanced image fidelity (Appendix 4.7.3) and to confirm that reference deviation increases from description- to article-based generation under Imagen 3 as textual constraints are relaxed (Appendix 4.7.4). Additional stylistic analyses of the generated images are provided in Appendix 4.7.5.

Procedure We employed a within-subject design comparing description-, caption-, and article-based generation conditions using a two-alternative forced choice paradigm. On each trial, participants viewed a news article alongside two images: one original journalist-selected image and one generated by a T2I model, and selected which they would publish for the article. This design enables direct comparison under a shared article context, yielding a relative preference measure interpretable against the 50% chance level.

Each participant viewed images generated under all three conditions but saw each article only once. Participants completed 108 randomized trials (i.e., 36 per condition), with stimuli fully balanced across the three conditions. To encourage adequate engagement with the article content, each trial first presented the article alone with a suggested minimum viewing time of approximately 30 seconds. Participants then proceeded to a second page displaying both the article and its corresponding image for coherence assessment. Two practice trials preceded the main task, and three attention checks were embedded throughout.

Participants 111 participants were recruited from the UCLA student subject pool (Jan 2026), with no overlap with other studies in this chapter. After excluding participants who failed more than one attention check, provided incomplete responses, or were affected by duplicated stimulus sampling in Qualtrics, 98 participants remained for analysis.

Results According to Figure 4.1, under the forced-choice setup ($0.50 =$ equal preference between real and AI images), description- ($M = 0.440$) and caption-based ($M = 0.436$) conditions were below chance (both $p < .001$), indicating a preference for real images, whereas the article-based condition was above chance ($M = 0.531$, $p = .010$), indicating a preference for T2I-generated images. A logistic mixed-effects model confirmed that only the article-

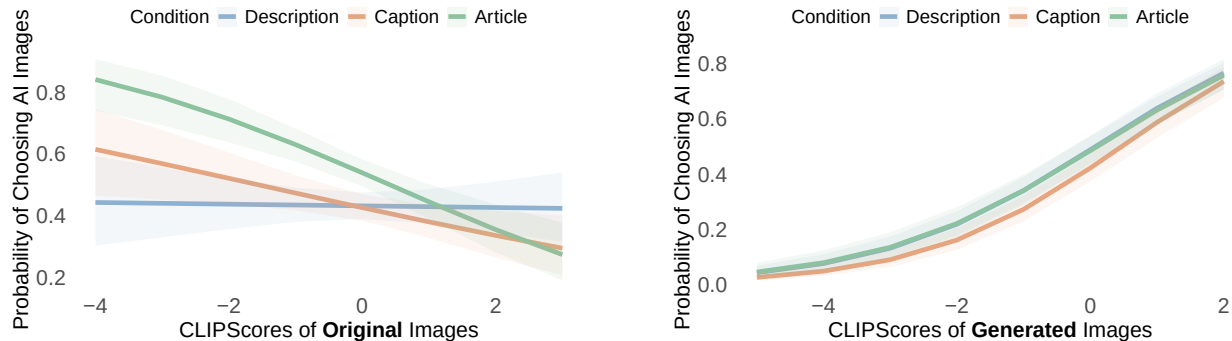


Figure 4.3: Left: depiction gap in original image–article pairs. Right: T2I systems narrowing the gap through increased alignment.

based condition was significantly more likely to be preferred than the description-based condition ($p < .001$), with no difference for captions ($p = .659$).

Overall, as VNF increases from description to article, AI-generated images become more preferred and ultimately surpass journalist-selected images, whereas under more constrained conditions (description and caption), journalist-selected images remain preferred. This suggests that high-VNF settings enable generated images to shift audience preference in their favor. To better understand the drivers of these preferences, we examine two properties: (1) depiction gaps in real-world image–article pairs that allow T2I systems to produce images that better align with audience expectations, and (2) an inherent stylistic disadvantage of AI-generated images that suppresses preference overall, making gains under higher-VNF settings particularly notable.

4.5.1 Bridging the depiction gap as a mechanism for VNF

One explanation for why T2I systems produce more preferred images under high-VNF conditions is the presence of a *depiction gap* in real-world news: journalist-selected images are often only loosely aligned with article content due to sourcing constraints and availability [151]. Under article-based generation, T2I systems can generate a broader range of plausible images, some of which may better match the article context, potentially bridging this depiction gap.

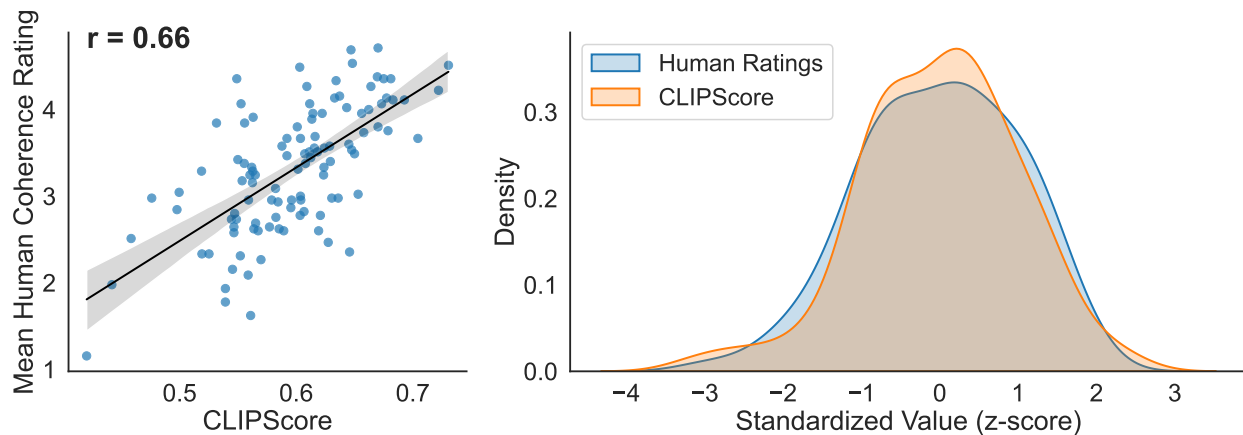


Figure 4.4: Relationship between CLIP-based image-text similarity and human coherence judgments across news images. Left: correlation between CLIPScore and mean human coherence ratings ($r = 0.66$). Right: standardized distribution of CLIPScore and human ratings across images.

To test this mechanism, we first establish that variation in image-article alignment exists in *original* news pairs using human coherence ratings. We then validate that CLIPScore (LongCLIP ViT-L/14) provides a reliable proxy for alignment. Finally, we use this measure to examine how alignment relates to preference for both *original* and *generated* images using logistic mixed-effects models with CLIPScore interacted with condition.

4.5.1.1 Human coherence rating vs. CLIPScore

Task Participants evaluated the coherence of original news image-article pairs on a 5-point Likert scale (1 = very incoherent, 5 = very coherent). Each participant rated all 108 stimuli. To encourage adequate engagement with the article content, each trial first presented the article alone with a suggested minimum viewing time of approximately 30 seconds. Participants then proceeded to a second page displaying both the article and its corresponding image for coherence assessment. Two practice trials preceded the main task.

Participants and Exclusions We recruited 50 participants from the UCLA student subject pool during October to November 2025. Three attention checks were embedded

throughout the study. Participants failing more than one attention check were excluded from analysis. After exclusions, each image–article pair received 45 valid ratings.

Experiment Results The mean image–article coherence rating for the original news items was 3.31 ($SD = 1.25$). Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC). The aggregated ratings showed excellent reliability ($ICC_{1k} = 0.96$, 95% CI [0.94, 0.97]), indicating strong agreement across annotators.

Validating CLIPScore as an Alignment Measure CLIPScore were strongly correlated with human coherence judgments (Pearson $r = 0.66$, $p < .001$; Spearman $\rho = 0.64$, $p < .001$), indicating that CLIP captures meaningful variation in perceived news image–article alignment (see Figure 4.4).

4.5.1.2 Results

Original Images (Depiction Gap Baseline) For original images (Figure 4.3, left), alignment does not predict preference under the description-based condition ($\beta = -0.011$, $SE = 0.074$, $p = .881$). Under looser conditions of caption-based and article-based conditions, the effect becomes significantly negative (caption: $\beta = -0.181$, article: $\beta = -0.366$, both $p < .001$), indicating that lower alignment of original images increases the likelihood that AI-generated images are selected for the preference judgment.

Generated Images (Bridging the Gap) For generated images (Figure 4.3, right), alignment is a strong positive predictor of preference ($\beta = 0.615$, $SE = 0.051$, $p < .001$), with no significant interaction effects. Generated images with higher article alignment are consistently more likely to be selected over original images, regardless of generation condition.

These results support the depiction gap hypothesis: when variation in original image–article alignment exists, alignment between generated images and the article becomes a strong driver of preference, suggesting that T2I systems leverage VNF to produce images that better align with article content, potentially narrowing the depiction gap and increasing audience preference.

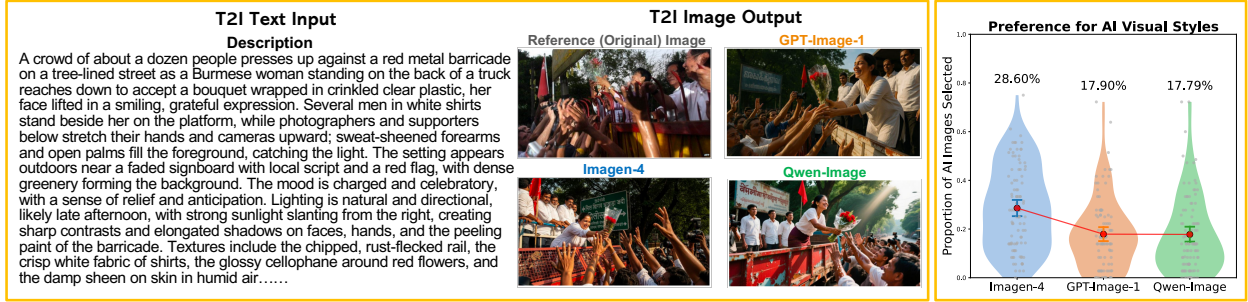


Figure 4.5: AI visual style disadvantage generalizes across T2I systems. Left: the same GPT-5-generated description used across all generators, with the corresponding real image and model-generated images. Right: violin plots of participant responses; gray dots represent individual participants, with means and bootstrapped 95% CIs.

4.5.2 Stylistic disadvantages of AI-generated images

To assess whether the observed AI stylistic disadvantage generalizes beyond **Imagen 3**, we replicated the description-based generation condition using GPT-5-generated descriptions and additional TTI models.

Stimuli and Model Selection We synthesize images using three state-of-the-art T2I models: **Imagen 4**, **GPT-Image-1**, and **Qwen-Image**. This design tests whether the observed stylistic effects generalize across model families with different training paradigms, architectures, and visual priors. **Imagen 4** is the latest iteration of the Imagen family and serves as a within-family replication of **Imagen 3** [213, 104]. **GPT-Image-1** is a proprietary image generation model from OpenAI, representing a distinct system with its own learned visual characteristics. **Qwen-Image** follows the setup in Section 4.4.2, with two modifications: inference steps are increased to 50 and `true_cfg_scale` is set to 4.0 to improve fidelity.

Procedure We employed a within-subject design comparing description-based generation across three TTI models, **Imagen 4**, **GPT-Image-1**, and **Qwen-Image**, using the same forced-choice paradigm. In each trial, participants were shown a news article alongside two images: the original image and one generated by one of the three models. Each participant viewed images generated by all three models but saw each article only once. Participants

completed 108 randomized trials (i.e., 36 per model), following the same procedure as in the previous experiment. All remaining procedures mirrored the prior study.

Participants and Exclusions A total of 114 participants were recruited from the UCLA student subject pool (Feb-March 2026), with no overlap with other studies reported in this chapter. After exclusions based on the same criteria, 97 participants remained for analysis.

Results One-sample t -tests showed that all models were significantly below chance (Imagen 4: $M = 0.286$, $p < .001$; GPT-Image-1: $M = 0.179$, $p < .001$; Qwen-Image: $M = 0.180$, $p < .001$), indicating a consistent preference for journalist-selected images over generated images. This highlights a systematic stylistic disadvantage of TTI-generated images under tightly constrained description-based prompts. Further analyzing model differences using a logistic mixed-effects model with random intercepts for subjects and images, both GPT-Image-1 ($\beta = -0.794$, $SE = 0.066$, $z = -12.06$, $p < .001$) and Qwen-Image ($\beta = -0.812$, $SE = 0.066$, $z = -12.29$, $p < .001$) were significantly less likely to be preferred than Imagen 4. This indicates that Imagen 4 yields comparatively higher-fidelity visual outputs, supporting our use of the Imagen family for high-fidelity image generation.

4.6 Conclusion

The dominant paradigm in modern VLMs typically assumes tight semantic alignment between images and text, whereas real-world multimodal contexts often exhibit image-text relationships at higher narrative and symbolic levels. This semantic looseness motivates us to examine how different levels of textual constraint shape Visual Narrative Freedom (VNF) of Text-to-Image systems. We conduct model experiments across three constraint settings (description-, caption-, and article-based generation) that vary along their hypothesized visual narrative freedom and introduce two metrics, *Intra-Prompt Diversity* and *Reference Deviation*, to characterize how these conditions reflect VNF. Our results show a gradual increase in the diversity of visual realizations as input constraints are relaxed from description- to article-based generation.

Our human evaluation reveals important societal implications. T2I systems can leverage VNF not only to produce images preferred over those generated under more tightly constrained conditions, but also to produce imagery that is more preferred—and potentially more persuasive—than journalist-selected images in high-stakes civic contexts. These audience-facing findings underscore the need to reconsider how current VLM paradigms conceptualize image–text alignment, as the looseness of real-world multimodal relationships, combined with current VLM capabilities, can meaningfully shape content production and perceived persuasiveness.

4.6.1 Limitations and ethics statement

Our analysis is subject to several limitations. Due to computational constraints, VNF measurements are restricted to T2I models with relatively fast inference; future work could extend this analysis to a broader range of proprietary and open-source systems. Additionally, our dataset focuses on news media [146, 138] and could be extended to other domains with similar multimodal structure, such as blog posts [220] and Wikipedia-based datasets [125], though this would require additional data collection (e.g., crawling paired articles and images).

Importantly, our study does not aim to introduce new model architectures, training methods, or prompting strategies that enhance current T2I systems’ ability to generate photorealistic images. All experiments rely on existing proprietary or open-source models with minimal prompting and configuration. Instead, our goal is to raise awareness of how current evaluation paradigms may underestimate the risks associated with increasingly photorealistic and persuasive generated imagery, particularly in journalism and other high-stakes civic domains. By identifying and characterizing these risks, this work contributes to the development of more robust and responsible evaluation frameworks that better capture the underdetermined nature of real-world image–text relationships.

4.7 Appendices

4.7.1 Other typologies of image–text relationships

Media theory and cognitive science have proposed a range of frameworks to characterize nuanced image–text relationships. These typologies are included here for conceptual reference and to situate our computational operationalization within broader theories of multimodal communication.

Early work by Barthes [12] distinguishes several forms of interaction between text and images:

- *Anchorage*. Text guides the viewer’s interpretation of an image by narrowing its possible meanings. Because images are often inherently polysemous, accompanying text can anchor interpretation by specifying what the viewer should attend to or how the image should be understood. For example, a news headline may frame a photograph of a crowd as either a “protest” or a “celebration”.
- *Illustration*. In this relationship, the image primarily serves to depict or visualize information already conveyed in the text. The image functions as a visual accompaniment that reinforces the textual message rather than adding new information. For example, an article describing a natural disaster may include a photograph that visually illustrates the damage described in the text.
- *Relay*. Image and text exhibit a bidirectional, interdependent relationship, contributing complementary information and jointly constructing meaning. Neither modality alone fully conveys the message; instead, understanding emerges from their interaction. This relationship is common in comics, advertisements, and visual storytelling, where text and imagery each provide distinct narrative elements.

Cognitive science further elaborates how language interacts with perceptual systems. Zwaan [275] proposes a typology of language–perception interactions that varies in the degree to which language is grounded in the perceptual environment:

- *Demonstration.* Language directly refers to perceptually present objects or events in the environment. The meaning of the utterance relies heavily on shared perceptual context. For example, a speaker pointing to an object and saying “this one” requires the visual environment to resolve the reference.
- *Instruction.* Language guides actions within a perceptual environment, often specifying how to manipulate objects or perform tasks. Examples include procedural instructions such as assembling furniture or following a recipe, where language coordinates perception and action.
- *Projection.* Language evokes mental simulations of perceptual experiences that are not directly present but remain grounded in embodied perception. For example, reading “the glass shattered on the floor” can trigger a mental simulation of the visual and auditory event.
- *Displacement.* Language refers to events, objects, or situations that are removed from the immediate perceptual environment, such as past events or distant locations. While still interpretable through prior perceptual experience, these references require more abstract mental representation.
- *Abstraction.* Language operates at a conceptual level with minimal direct grounding in perceptual experience. Abstract concepts such as “democracy,” “justice,” or “economic growth” rely more on symbolic reasoning than perceptual simulation.

For empirical analysis in this chapter, we adopt an alignment-based perspective in which image–text relationships are conceptualized along a continuous spectrum of semantic alignment [29], operationalized using cross-modal similarity metrics (e.g., CLIPScore) [89].

- *Negative correlation.* Image and text convey conflicting or contradictory meanings, often seen in satirical or misleading contexts (e.g., a celebratory crowd image paired with an article about labor violations or fatalities at the same event).

- *Zero correlation.* Image and text are semantically unrelated, with no meaningful conceptual or contextual connection (e.g., a generic city skyline image used for an article on international trade policy or economic regulation).
- *Low-positive correlation.* Image and text share loose thematic or symbolic relevance without depicting specific entities or events, commonly observed in finance and public health reporting (e.g., a stock image of a handshake for diplomatic or business negotiations, or a generic hospital scene for healthcare coverage).
- *High-positive correlation.* Image and text are strongly aligned, with images directly depicting the entities or events described, frequently in politics and breaking news (e.g., a photograph of a named political figure at a press conference, or an image of a wildfire corresponding to a reported event).

4.7.2 Descriptive statistics of the news dataset

We report descriptive statistics of our news dataset in Table 4.1, showing topic distributions that broadly reflect the underlying outlets and major news categories, along with average token counts by topic. We observe no substantial differences in length across topics, and all textual inputs fall within the LongCLIP context window (i.e., 248 tokens).

4.7.3 Balancing image fidelity of human experiment stimuli

Because the same generated images are reused in human evaluations, we control for low-level perceptual confounds between real and generated images. In particular, we balance resolution, aspect ratio, and compression to ensure that participant judgments are not driven by superficial visual cues. Full details are as follows:

First, generated image aspect ratios were matched to the original photographs where possible, using the closest available options provided by each T2I system. All images were then center-cropped when necessary to match aspect ratios when T2I outputs do not precisely align with the originals.

News Topics	N	Description	Caption	Article
Politics & International Affairs	56	217.2	28.1	229.0
Science, Technology & Environment	20	200.4	30.6	222.8
Culture & Lifestyle	15	199.7	28.7	215.0
Sports	11	220.0	25.6	210.0
Economy & Business	6	214.7	25.8	226.2
Total/Average	108	210.4	27.8	220.6

Table 4.1: Distribution of news topics and average token lengths for article text, captions, and generated image descriptions in the dataset. Token counts follow the LongCLIP tokenizer (CLIP BPE based on the GPT-2 vocabulary; max context length 248 tokens).

Second, because archival news images often appear at relatively low resolution—a salient cue of authenticity—we manually retrieve the highest-resolution publicly available versions via Google Image Search. We avoid applying any post hoc resolution enhancement, preserving realistic visual differences between circulated news images (often lower resolution) and generated images (typically higher resolution).

Third, to reduce residual resolution disparities, we apply controlled JPEG compression to both original and generated images. Compression is iteratively applied (quality = 0.9 per iteration) until file sizes fall below 200KB, ensuring consistent loading performance in Qualtrics while maintaining perceptual comparability.

All stimuli are manually inspected by the research team to ensure no salient resolution differences remain beyond inherent stylistic variation. Implausible or nonfactual generations (e.g., a floating truck in the ocean) are excluded or regenerated prior to inclusion. These refinements help minimize superficial perceptual cues that participants might otherwise rely on when assessing preferences between real and generated images.

The same procedures are applied across both human experiments in Section 4.5.

4.7.4 Reference deviation of Imagen 3

For the high-fidelity image model **Imagen 3** used in the preference experiment across VNF conditions, we find that reference deviation increases (i.e., similarity decreases) progressively and significantly from description- to caption- to article-based generation across perceptual, structural, semantic, and conceptual measures (Figure 4.6). This pattern indicates that, as textual constraints are relaxed, **Imagen 3** generates images that increasingly diverge from journalist-selected images across visual representations, thereby expanding its capacity to leverage VNF and potentially influencing audience preferences.

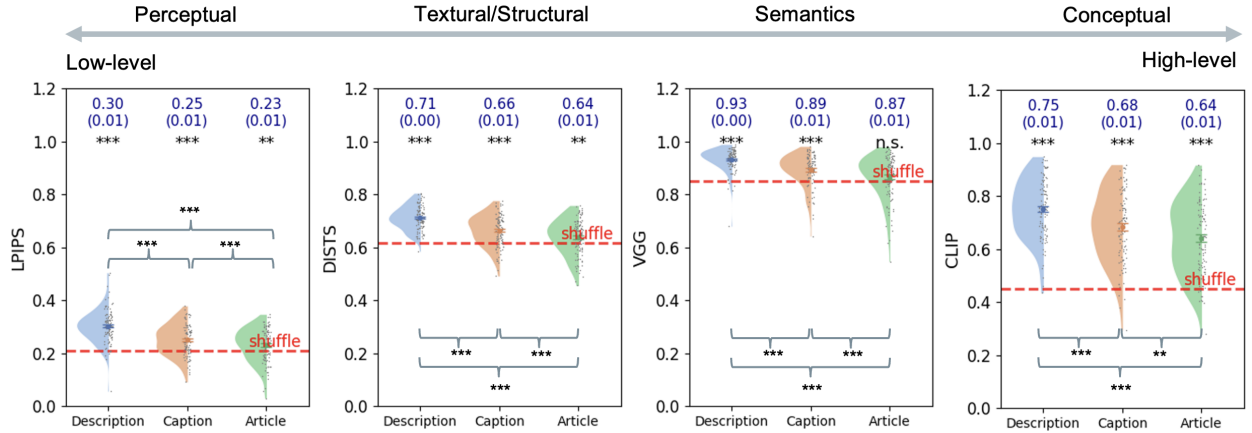


Figure 4.6: Reference deviation (measured using reference-based similarity metrics) of **Imagen 3**-generated images across three generation conditions, spanning lower-level perceptual to higher-level conceptual similarity [72]. Half-violins show distributions over 108 stimuli (gray dots), with means and bootstrapped 95% CIs. Lower similarity indicates greater deviation from the original image. The shuffled baseline (red dashed line) reflects similarity between original images and randomly paired images, serving as a lower-bound reference. Significance markers ($***p < .001$, $**p < .01$, $*p < .05$) denote comparisons to the original (top) and between conditions (bottom).

4.7.5 Visual style analysis of Imagen 3

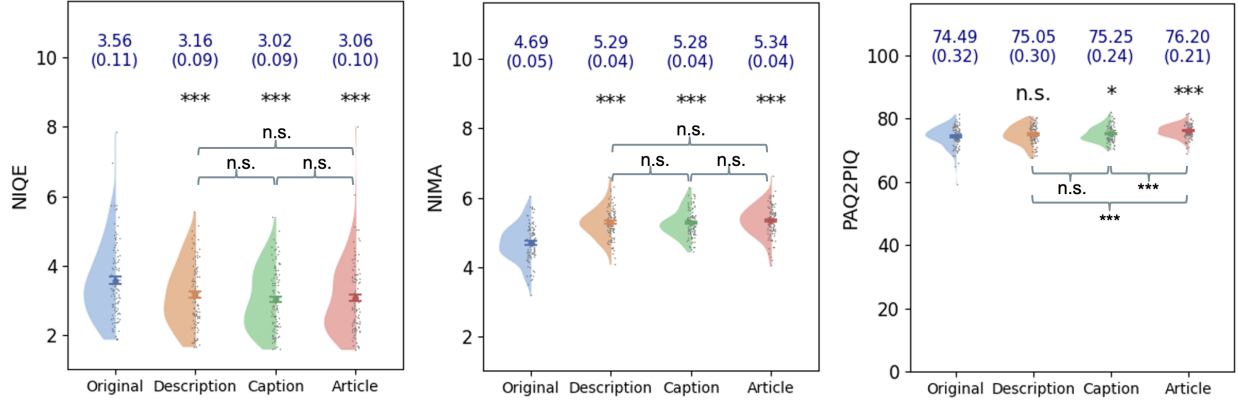


Figure 4.7: Reference-free image quality metrics assessing the visual style of Imagen 3-generated images across three conditions. Three complementary metrics are included [72]. Half-violins show distributions over 108 stimuli (gray dots), with means and bootstrapped 95% CIs. Significance markers ($***p < .001$, $**p < .01$, $*p < .05$) denote comparisons to the original (top) and between conditions (bottom).

We assess the intrinsic visual characteristics of generated images using reference-free visual quality assessment (VQA) metrics. These metrics measure perceptual realism and aesthetic quality without requiring a ground-truth image. Because all images are generated by the same T2I model (Imagen 3), we expect a consistent photographic style across generation conditions. We employ three complementary metrics capturing image distortion, perceived aesthetic and perceptual realism (Figure 4.7) [72].

Measurements: Reference-free VQA

- **Natural Image Quality Evaluator (NIQE)**: Measures deviations from natural image statistics by modeling NSS with a multivariate Gaussian and computing distance to the natural-image manifold; higher scores indicate lower image quality (i.e., greater distortion) [175].
- **Neural Image Assessment (NIMA)**: Predicts human aesthetic ratings using a CNN

trained on the AVA dataset, outputting a distribution over aesthetic scores; higher scores indicate higher aesthetic quality [234].

- **Patch Quality to Picture Quality (PAQ2PIQ):** Estimates perceptual image quality using a CNN trained on large-scale human judgments of image realism; higher scores indicate higher perceptual quality [258].

Results Visual style assessments using reference-free VQA metrics remain largely stable across the three generation conditions for **Imagen 3** (with most pairwise comparisons not significant), although they differ significantly from real news photographs (Figure 4.6). This indicates that, while generated images vary substantially in their perceptual, structural, and semantic content relative to the reference image (as shown in Figure 4.6, Section 4.7.4), they retain a consistent photographic prior imposed by the generator.

Interestingly, across three complementary VQA metrics, a consistent pattern emerges: **Imagen 3** outputs are computationally measured as more perceptually realistic and less distorted than real images. This points to a potential hyperrealism effect, consistent with prior work documenting hyper-smooth textures and compositional homogenization in AI-generated photorealistic imagery [200, 181].

4.7.6 Generation prompts

Image description generation (GPT-4o/GPT-5)

“Describe the image in a way that could serve as a photorealistic news image prompt for an AI model. Focus on the following elements: the subjects (their appearance, approximate number, and actions), the environment (location and background), the mood (emotional tone), the lighting (source, direction, and intensity), shadows, and realistic details (textures, clothing, facial expressions). Provide a realistic estimate of the number of visible people or subjects, avoiding exaggeration. Also describe the camera angle or perspective (e.g., high angle,

low angle, over-the-shoulder). Write in a descriptive, journalistic tone as a single uninterrupted paragraph. Do not include labels, bullet points, or line breaks.”

T2I prompting conditions For each news instance, we generated images under three conditions using aligned prompts derived from the same underlying content:

- **Description-based:** “Generate a photorealistic news image based on the following description.”
- **Caption-based:** “Generate a photorealistic news image based on the following caption.”
- **Article-based:** “Generate a photorealistic news image representing the main subject of the following article. No text, headlines, or labels in the image.”

The explicit prohibition of textual overlays in the article-based condition was necessary because T2I systems frequently embed article text directly within images when under-instructed.

CHAPTER 5

Conclusion

This dissertation addresses a central question at the intersection of cognitive science and artificial intelligence: *why do models and humans differ in their competencies in language and perceptual processing, and how does this asymmetry shape downstream human-AI interaction?* Across theoretical analysis, computational modeling, and human experiments, it argues that language and perception operate over fundamentally different representational regimes. Language is discrete and symbolic, whereas perception is continuous and grounded in high-dimensional embodied experience. This distinction helps explain both the strengths of modern AI systems in language tasks and their persistent limitations in perceptual and cross-modal reasoning.

Chapter 1 situates this perspective within the broader literature on *symbolic* and *embodied cognition*, connecting cognitive theory with contemporary developments in language and multimodal AI. It argues that novel or fine-grained knowledge may exist outside the space of discrete linguistic encoding, and that incorporating multimodality provides access to continuous signals that support the construction of shared experiences and enhance communicative clarity.

Building on this foundation, Chapter 2 examines how models learn from linguistic and behavioral signals, showing that differences in discreteness, continuity, and statistical structure lead to distinct learning dynamics. The discreteness, surprisal, irregularity, and arbitrariness of language contribute to its efficiency in encoding information and to the impressive cognitive abilities observed in large language models (LLMs). In contrast, the smoothness, continuity, and autocorrelated structure of perceptual signals make perceptual tasks

more demanding for current AI systems. While these properties align more closely with human strengths, allowing humans to effortlessly identify object boundaries and spatial groupings, computer vision models often require substantial data and computational resources to achieve comparable performance. This gap between discrete and continuous signal processing motivates further investigation into how human and model capabilities may deviate or complement each other.

Importantly, Chapter 2 should not be interpreted as making broad claims about “language versus behavioral” modalities in general. The original study was designed primarily as a model interpretability and learning-dynamics analysis across datasets with different modalities, rather than as a tightly controlled modality comparison. In particular, differences between the emotion tweet classification and gesture detection datasets likely reflect not only modality distinctions, but also dataset-specific characteristics, sample sizes, task structures, and representational choices. For example, gesture signals themselves may be represented in multiple forms, such as raw video frames or extracted body keypoints (the representation used in our study), each of which may produce distinct learning dynamics. More controlled comparisons could apply the proposed framework to multimodal datasets containing aligned text, audio, and visual signals (e.g., multimodal emotion recognition datasets [252, 49]), allowing modality effects to be isolated more systematically from dataset idiosyncrasies. Future cognitive and communication research can further extend this framework across a broader range of linguistic, perceptual, and behavioral tasks to better understand how learning dynamics vary across representational formats and modalities.

In the era of multimodal foundation models—such as multimodal LLMs, vision-language models (VLMs), and image generation systems—this asymmetry raises new questions about the sufficiency and limitations of language and vision, as well as their interaction in both model outputs and human evaluation. To examine these questions in naturalistic and high-stakes settings, Chapter 3 and 4 focus on the domain of news media, where datasets naturally integrate images, captions, and articles. This setting enables controlled behavioral experiments alongside model evaluations.

Chapter 3 extends previously observed flawed linguistic heuristics in human evaluation of AI-generated text to cross-modal settings, specifically news image captioning. While these unreliable heuristics persist in multimodal tasks, cross-modal context provides reliable cues that mitigate such biases, particularly given the gap between models’ language proficiency and their cross-modal reasoning.

Chapter 4 further examines this gap by highlighting a mismatch between current vision–language training paradigms—typically optimized for tight semantic alignment—and real-world multimodal communication, where image–text relationships are often loosely coupled and operate at a narrative level. Crucially, such relationships are often underdetermined, admitting multiple plausible mappings between text and images. Under these conditions, text-to-image systems can exploit underconstrained textual inputs to generate photorealistic imagery that may be more persuasive than original journalistic images.

Chapter 3 and 4 focus on language–perception systems, particularly language–vision models, as a starting point for investigating asymmetries between linguistic and perceptual representations in humans and machines. This focus is motivated in part by recent advances in foundational multimodal LLMs, image generation systems, and large-scale image–text datasets, which together provide comparatively scalable, tractable, and ecologically valid settings for both computational modeling and human–AI interaction experiments. Because the empirical work centers primarily on visual modalities, the implications of these findings should be interpreted primarily in relation to language–vision systems rather than embodiment more broadly. Nevertheless, the broader goal of this dissertation is to motivate future research on multimodal intelligence beyond language–vision systems alone.

In sum, these chapters reveal several critical insights. First, as introduced in Chapter 1, understanding the representational properties of language and perception enables a more principled and interpretable analysis of model strengths and weaknesses. On the one hand, these properties help explain why language models can encode spatial and temporal structure across multiple scales [79], and why cross-modal alignment can often be achieved through relatively weak linear transformations [228]. On the other hand, they also provide

a unifying account of a range of inconsistencies and vulnerabilities in vision–language models. These include cross-modal safety failures driven by superficial modality-specific cues rather than shared representations [77]; the tendency for VLMs to perform well on high-level vision benchmarks while struggling with low-level perceptual tasks that are trivial for humans [196]; spurious correlations between visual and textual modalities in visual question answering [246]; and the “mirage reasoning” phenomenon, where models generate detailed visual descriptions and reasoning traces despite the absence of actual visual input [5]. Collectively, these behaviors can be understood as mismatches or imbalances in how information is represented and integrated across language and perception.

Second, despite impressive progress in language-based tasks, a substantial gap remains between model and human performance in perceptual reasoning and cross-modal integration, including physical understanding and visual interpretation [67, 66]. In human–AI interaction, perceptual and cross-modal cues often provide more reliable signals for identifying AI-generated content or preferring expert-produced outputs. At the same time, this asymmetry reflects not only model limitations but also a form of complementarity: models excel at symbolic abstraction, whereas humans retain advantages in perceptual grounding and contextual interpretation. This suggests that their strengths may be most effectively leveraged in combination, such as in human–AI co-piloting, rather than treated as interchangeable.

Lastly, prevailing multimodal paradigms rely heavily on tightly aligned image–text pairs, whereas real-world communication—such as in visual journalism, social media, and educational materials—often involves looser, more flexible relationships between modalities [151, 29]. In these contexts, images (or text) may illustrate, contextualize, or loosely evoke aspects of a narrative rather than directly correspond to it. Advancing multimodal AI therefore requires moving beyond strict alignment assumptions, both in model architectures and in benchmarks, to better capture the diversity and complexity of real-world image–text relationships. More broadly, the field would benefit from shifting away from singular notions of alignment toward modeling distributions of valid interpretations, better reflecting the ambiguity and flexibility inherent in real-world multimodal communication.

Overall, this dissertation bridges cognitive science and artificial intelligence by providing a unified representational account of multimodality. It highlights that multimodal intelligence is not merely an additive interface, but is fundamentally shaped by the structure of representations through which information is encoded, influencing how both humans and machines learn, generate, and evaluate information.

5.1 Future Directions and Open Questions

This dissertation lays the foundation for treating multimodality not as an additive interface, but as a consequence of underlying representational differences across modalities. Building on this unified account of language–perception asymmetry, this work opens several promising directions for future research in model architecture, evaluation, and human–AI interaction.

1. Extending representational analysis to intermediate and multimodal signals. The model explainability framework introduced in Chapter 2 can be extended to modalities that lie between discrete and continuous representations, such as speech. Speech signals exhibit both structured discreteness (e.g., phonemes) and continuous acoustic variation [163, 142], making them a natural testbed for studying intermediate representational regimes. Applying this framework to speech and conversational data (see our previous work for an example [170]) may help characterize how models integrate heterogeneous signals and how representational properties influence downstream multimodal performance. More broadly, such signal-based analyses could provide a principled way to anticipate model behavior from data properties, and to study additive versus synergistic effects across modalities, including combinations of linguistic, perceptual, and physiological signals.

2. Modeling multimodal relationships beyond strict alignment. Chapter 4 demonstrates that real-world multimodal communication—particularly in domains like news media—is often characterized by loose, underdetermined relationships between images and text [29, 151]. Future work can build on this by systematically modeling different types of image–text relationships (e.g., descriptive, complementary, narrative, or symbolic), draw-

ing on frameworks from media theory and cognitive science [12, 275]. Developing datasets and experimental paradigms that explicitly vary these relationships would enable more precise evaluation of both human and model performance under varying degrees of semantic constraint.

Longer-form textual contexts, such as news articles, science communication, and creative writing, likely contain additional latent structures that were not explicitly modeled in the present work, particularly when paired with photorealistic or illustrative imagery. These structures may include narrative framing, discourse organization, topical emphases, and stylistic regularities that shape how readers interpret multimodal content. While Chapter 4 operationalized variation primarily through differences in textual constraint (description, caption, article), future work could more systematically disentangle which linguistic and discourse-level properties contribute to multimodal flexibility, persuasion, or narrative freedom. Extending this paradigm to other domains with rich multimodal structure, such as Wikipedia [125], blogs [220], or artwork [7], would further test the generality of these findings.

3. Learning and evaluating distributions of multimodal interpretations. Current VLMs are largely trained under objectives that assume a single correct alignment between image and text, often via contrastive or self-supervised learning. However, real-world multimodal relationships frequently admit multiple valid interpretations [64, 116]. A key direction is therefore to move from modeling pointwise alignment to modeling distributions over plausible image–text mappings. This includes developing training objectives, representations, and evaluation metrics that capture ambiguity, narrative flexibility, and contextual dependence. Such work could also examine how model representations shift when trained on tightly aligned versus loosely coupled data, and whether models can learn to “reinterpret” the same visual input under different textual contexts.

4. Human–AI complementarity and co-piloting. The asymmetry between language and perception suggests that humans and models possess complementary strengths rather than interchangeable capabilities. Rather than treating AI systems as substitutes for human judgment, future research should explore interaction paradigms that explicitly

leverage this complementarity, such as human–AI co-piloting systems. These approaches combine human perceptual grounding and contextual reasoning with models’ strengths in symbolic abstraction and generative capacity, while emphasizing interface and workflow design that preserves human cognitive agency. Beyond improving task performance, such systems may also support richer forms of interpretation by exposing users to alternative, potentially AI-native perspectives, diverse communicative styles, and viewpoints shaped by different cultural and experiential backgrounds.

Together, these directions outline a broader research agenda centered on a representational account of language–perception asymmetry in humans and models. This perspective has important implications for human–AI interaction, including human–AI co-piloting and the preservation of human cognitive agency, while also advancing our understanding of the narrative structure and expressive power of multimodal communication in real-world settings.

BIBLIOGRAPHY

- [1] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Conference on Computational Natural Language Learning*, 2021.
- [2] Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L. Gallant. Pixels to voxels: Modeling visual representation in the human brain. *ArXiv*, abs/1407.5104, 2014.
- [3] Gerry T. M. Altmann and Yuki Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264, 1999.
- [4] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: A survey. *IEEE Transactions on Autonomous Mental Development*, 1(1):12–34, 2009.
- [5] Mohammad Asadi, Jack W. O’Sullivan, Fang Cao, Tahoura Nedaei, Kamyar Rajabifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *ArXiv 2603.21687*, 2026.
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *ArXiv*, 2023.
- [7] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *International Conference in Computer Vision*, 2021.
- [8] Marcus Banks. True to life: Authenticity and the photographic image. In Thomas Fillitz and Jamie A. Saris, editors, *Debating Authenticity: Concepts of Modernity in Anthropological Perspective*, pages 160–171. Berghahn Books, New York, 2013.

- [9] Gagan Bansal, Jennifer Wortman Vaughan, Saleema Amershi, Eric Horvitz, Adam Fourney, Hussein Mozannar, Victor Dibia, and Daniel S. Weld. Challenges in human-agent communication. *ArXiv 2412.10380*, 2024.
- [10] Samah Saeed Baraheem, Trung-Nghia Le, and Tam V. Nguyen. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review*, 56(10):10813–10865, 2023.
- [11] Lawrence W. Barsalou. Perceptual symbol systems. *Behavior and Brain Sciences*, 22:577–660, 1999.
- [12] Roland Barthes. *Image, Music, Text*. Fontana Press, London, 1977. Translated by Stephen Heath.
- [13] Claas Beger, Ryan Yi, Shuhao Fu, Kaleda Denton, Arseny Moskvichev, Sarah W. Tsai, Sivasankaran Rajamanickam, and Melanie Mitchell. Do ai models perform human-like abstract reasoning across modalities? *ArXiv 2510.02125*, 2026.
- [14] Kathleen Belhassein, Víctor Fernández-Castro, Amandine Mayima, Aurélie Clodic, Elisabeth Pacherie, Michèle Guidetti, Rachid Alami, and Hélène Cochet. Addressing joint action challenges in hri: Insights from psychology and philosophy. *Acta Psychologica*, 222:103476, 2022.
- [15] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures (extended abstract). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4970–4974, 2017.
- [16] George Herbert Betts. The brain and nervous system. In *The mind and its education*, pages 31–50. D. Appleton & Company, 3rd, rev. and enlarged edition, 1923.
- [17] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, mdl priors, and

- object recognition. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.
- [18] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics.
- [19] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467, 2019.
- [20] Rishi Bommasani et al. On the opportunities and risks of foundation models. *ArXiv 2108.07258*, 2022.
- [21] Cynthia Breazeal, Cory D. Kidd, Andrea L. Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 708–713. IEEE, 2005.
- [22] Harry Bunt. Dialogue pragmatics and context specification. In *Abduction, Belief and Context in Dialogue*, 2000.
- [23] Curt Burgess and Kevin Lund. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12(2&3):177–210, 1997.
- [24] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [25] Tyler A Chang and Benjamin K Bergen. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350, 2024.
- [26] Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *ArXiv 2308.15419*, 2023.
- [27] Soravit Changpinyo, Pankaj Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021.
- [28] Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240, 2021.
- [29] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A. Bateman, and Ralph Ewerth. Understanding image-text relations and news values for multimodal news analysis. *Frontiers in Artificial Intelligence*, Volume 6 - 2023, 2023.
- [30] Mostafa Haghiri Chehreghani. The embeddings world and artificial general intelligence. *Cognitive Systems Research*, 84, 2022.
- [31] Jaemin Cho, Yushi Hu, Jason M. Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.
- [32] Morten H Christiansen and Nick Chater. Connectionism, learning and meaning. *Connection Science*, 4:227–227, 1992.
- [33] Morten H Christiansen and Nick Chater. Connectionist natural language processing: The state of the art. *Cognitive Science*, 23(4):417–437, 1999.

- [34] Morten H Christiansen and Nick Chater. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62, 2016.
- [35] Morten H Christiansen and Maryellen C MacDonald. A usage-based approach to recursion in sentence processing. *Language Learning*, 59(Suppl 1):126–161, 2009.
- [36] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 7354–7365, 2018.
- [37] Massimiliano Ciammaichella. This person does not exist. representation theories and practices of a desired face. In Daniele Villa and Franca Zuccoli, editors, *Proceedings of the 3rd International and Interdisciplinary Conference on Image and Imagination*, pages 544–551, Cham, 2023. Springer International Publishing.
- [38] Andy Clark. *Being There: Putting Mind, World, and Body Back Together*. MIT Press, 1997.
- [39] Elizabeth Clark, Tal August, Sofia Serrano, Nanyun Haduong, Suchin Gururangan, and Noah A. Smith. All that’s “human” is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7282–7296. Association for Computational Linguistics, 2021.
- [40] Herbert H. Clark and Susan Brennan. Grounding in communication. In *Perspectives on socially shared cognition*, 1991.
- [41] Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H Christiansen. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256, 2023.

- [42] Christine Cuskley, Rebecca Woods, and Molly Flaherty. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083, 08 2024.
- [43] Rick Dale and Christopher T. Kello. “how do humans make sense?” multiscale dynamics and emergent meaning. *New Ideas in Psychology*, 50:61–72, 2018.
- [44] Rick Dale, Christopher T. Kello, and P. Thomas Schoenemann. Seeking synthesis: The integrative problem in understanding language and its evolution. *Topics in Cognitive Science*, 8(2):371–381, 2016.
- [45] Maria-Teresa De Rosa Palmini and Eva Cetinic. Synthetic history: Evaluating visual representations of the past in diffusion models. In *Proceedings of the International Conference on Learning Representations*, 2026. arXiv preprint arXiv:2505.17064.
- [46] Manuel de Vega, Arthur Glenberg, and Arthur Graesser, editors. *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford, 2008. Online edition, Oxford Academic, 22 Mar. 2012.
- [47] Abhishek M Dedhe, Hayley Clatterbuck, Steven T Piantadosi, and Jessica F Cantlon. Origins of hierarchical logical reasoning. *Cognitive science*, 47(2):13250, 2023.
- [48] Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with jepa world models. *ArXiv 2601.00844*, 2025.
- [49] M. J. Dileep Kumar, M. Suresh Rao, and K. C. Narendra. Multimodal emotion recognition: A comprehensive survey of datasets, methods, and applications. *IEEE Access*, 13:201067–201097, 2025.
- [50] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

- [51] Mark Dingemanse, Damián E Blasi, Gary Lupyan, Morten H Christiansen, and Padraic Monaghan. Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, 19(10):603–615, 2015.
- [52] Sidney K. D’mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3), feb 2015.
- [53] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.
- [54] Liam Dugan, Daphne Ippolito, Ananth Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 12763–12771. AAAI Press, 2023.
- [55] Kevin Dunbar. The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In Dedre Gentner, Keith J Holyoak, and Boicho N Kokinov, editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 313–334. The MIT Press, 2001.
- [56] Kevin Dunbar and Isabelle Blanchette. The in vivo/in vitro approach to cognition: The case of analogy. *Trends in Cognitive Sciences*, 5(8):334–339, 2001.
- [57] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.
- [58] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.
- [59] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.
- [60] Jeffrey L. Elman. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582, 2009.

- [61] Jeffrey L. Elman. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33 4:547–582, 2009.
- [62] N. J. Enfield. *Relationship thinking: Agency, enchrony, and human sociality*. Oxford University Press, 2013.
- [63] Thomas A. Farmer, Morten H. Christiansen, and Padraic Monaghan. Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103(32):12203–12208, 2006.
- [64] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, Hao Sun, and Ji-Rong Wen. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.
- [65] Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, 41(5):1152–1201, 2017.
- [66] Shuhao Fu, Philip J Kellman, and Hongjing Lu. Hierarchical abstraction drives human-like 3-d shape processing in deep learning models. *PLOS Computational Biology*, 22(3):e1014047, 2026.
- [67] Shuhao Fu, Andrew Jun Lee, Yixin Anna Wang, Ida Momennejad, Trevor Bihl, Hongjing Lu, and Taylor Whittington Webb. Evaluating compositional scene understanding in multimodal generative models. *Transactions on Machine Learning Research*, 2025.
- [68] Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, and Sepp Hochreiter. Cloob: Modern hopfield networks with infoloob outperform clip. In *Proceedings of the 36th International Conference on Neural Information Processing*

- Systems (NeurIPS)*, NeurIPS '22, pages 1487–1505, Red Hook, NY, USA, 2022. Curran Associates Inc.
- [69] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, page 89–106, Berlin, Heidelberg, 2022. Springer-Verlag.
 - [70] Gemini-Team-Google et al. Gemini: A family of highly capable multimodal models. *ArXiv 2312.11805*, 2024.
 - [71] Dedre Gentner and Arthur B Markman. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56, 1997.
 - [72] Abhijay Ghildyal, Yuanhan Chen, Saman Zadtootaghaj, Nabajeet Barman, and Alan C. Bovik. Quality prediction of ai generated images and videos: Emerging trends and opportunities. *ArXiv 2410.08534*, 2024.
 - [73] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.
 - [74] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.
 - [75] Xue L. Gong, Alexander G. Huth, Fatma Deniz, Keith Johnson, Jack L. Gallant, and

- Frédéric E. Theunissen. Phonemic segmentation of narrative speech in human cerebral cortex. *Nature Communications*, 14(1):4309, 2023.
- [76] Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*, 36(2):193–202, 2004.
- [77] Yangyang Guo, Fangkai Jiao, Liqiang Nie, and Mohan Kankanhalli. The vllm safety paradox: Dual ease in jailbreak attack and defense. In *Annual Conference on Neural Information Processing Systems*, 2025.
- [78] Shikha Gupta, Krishan Sharma, Dileep Aroor Dinesh, and Veena Thenkanidiyoor. Visual semantic-based representation learning using deep cnns for scene recognition. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(2), may 2021.
- [79] Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.
- [80] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.
- [81] Uri Hasson, Samuel A. Nastase, and Ariel Goldstein. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.
- [82] Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307, 2004.
- [83] Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298 5598:1569–79, 2002.
- [84] Liv Hausken. Photorealism versus photography. ai-generated depiction in the age of visual disinformation. *Journal of Aesthetics & Culture*, 16(1):2340787, 2024.

- [85] John A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford University Press, 11 2004.
- [86] Martin N Hebart and Chris I Baker. Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, 180(Pt A):4–18, 2018.
- [87] Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS One*, 14(10):e0223792, 2019.
- [88] Micha Heilbron, Kristina Armeni, Jan-Mathijs Schoffelen, and Floris P. de Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.
- [89] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIP-Score: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [90] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [91] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [92] Charles F. Hockett. The origin of speech. *Scientific American*, 203(3):88–96, 1960.

- [93] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
- [94] Donald D Hoffman. *Visual Intelligence: How We Create What We See*. W. W. Norton, 2000.
- [95] Donald D Hoffman and Whitman A Richards. Parts of recognition. *Cognition*, 18(1-3):65–96, 1984.
- [96] Douglas R Hofstadter. Epilogue: Analogy as the core of cognition. In Dedre Gentner, Keith J Holyoak, and Boicho N Kokinov, editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 499–538. MIT Press, 2001.
- [97] Keith Holyoak, Dedre Gentner, and Boicho Kokinov. The place of analogy in cognition. In *The Analogical Mind: Perspectives from Cognitive Science*, page 119. 2001.
- [98] Keith J Holyoak. Analogy and relational reasoning. In Keith J Holyoak and Robert G Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*, pages 234–259. Oxford University Press, 2012.
- [99] Yining Hong, Huang Huang, Manling Li, Li Fei-Fei, Jiajun Wu, and Yejin Choi. Learning from trials and errors: Reflective test-time planning for embodied llms. *ArXiv 2602.21198*, 2026.
- [100] Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *ArXiv 2601.03782*, 2026.
- [101] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jing Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Danyang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chu Jin,

- Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji rong Wen. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *ArXiv*, abs/2103.06561, 2021.
- [102] Alexander Huth, Wendy de Heer, Thomas Griffiths, and et al. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458, 2016.
- [103] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, Dec 2012.
- [104] Imagen-Team-Google et al. Imagen 3. *ArXiv 2408.07009*, 2024.
- [105] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023.
- [106] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR, 2021.
- [107] Shan Jiang and Christo Wilson. Linguistic signals under misinformation and fact-checking. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018.
- [108] Yanru Jiang. Automated nonverbal cue detection in political-debate videos: An optimized rnn-lstm approach. In *International Conference on Human-Computer Interaction*, pages 32–40. Springer, 2023.
- [109] Yanru Jiang. Unlocking the complexity of deep learning in neuroimaging: Insights into within- and cross-modality representational similarities of fmri-autoencoders. Technical report, 2024. Working paper.

- [110] Yanru Jiang and Gülşah Akcikir. Explicit cooperation shapes human-like multi-agent llm negotiation. *ICWSM Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI)*, 2025.
- [111] Yanru Jiang and Rick Dale. Mapping the learning curves of deep learning networks. *PLOS Computational Biology*, 21(2):e1012286, 2025.
- [112] Yanru Jiang, Rick Dale, and Hongjing Lu. Transformability, generalizability, but limited diffusibility: Comparing global vs. task-specific language representations in deep neural networks. *Cogn. Syst. Res.*, 83(C), mar 2024.
- [113] Jungseock Joo, Erik Bucy, and Christian Seidel. Computational communication science — automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal Of Communication*, 13:23, 2019.
- [114] Michael I Jordan. Serial order: A parallel distributed processing approach. In J. W. Donahoe and V. P. Dorsel, editors, *Neural-network models of cognition: Biobehavioral foundations*, pages 471–495. North-Holland/Elsevier Science Publishers, 1997. (Reprinted in modified form from "Institute for Cognitive Science Technical Report 8904," 1986).
- [115] Annette Karmiloff-Smith. Constraints on representational change: Evidence from children’s drawing. *Cognition*, 34(1):57–83, 1990.
- [116] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [117] Walter Kintsch, Danielle S. McNamara, Sheila Dennis, and Thomas K. Landauer. Lsa and meaning: In theory and application. In Thomas K. Landauer, Danielle S. McNamara, Sheila Dennis, and Walter Kintsch, editors, *Handbook of Latent Semantic Analysis*, pages 467–479. Lawrence Erlbaum Associates Publishers, 2007.

- [118] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwińska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
- [119] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: an open dataset of user preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [120] Sotaro Kita and Karen Emmorey. Gesture links language and cognition for spoken and signed languages. *Nature Reviews Psychology*, 2(7):407–420, 2023.
- [121] Nils Köbis and Lieke D. Mossink. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in Human Behavior*, 114:106553, 2021.
- [122] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier, editors, *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, 2006. Springer.
- [123] Stefan Kopp and Ipke Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
- [124] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical*

- Methods in Natural Language Processing*, pages 4685–4697, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [125] Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. Concadia: Towards image-based text generation with a purpose. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
 - [126] Sarah Kreps, R. Miles McCain, and Miles Brundage. All the news that’s fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117, 2022.
 - [127] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
 - [128] Jan-Rouke Kuipers, Martijn van Koningsbruggen, and Guillaume Thierry. Semantic priming in the motor cortex: evidence from combined repetitive transcranial magnetic stimulation and event-related potential. *Neuroreport*, 24(12):646–651, 2013.
 - [129] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.
 - [130] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.
 - [131] Thomas K. Landauer and Susan T. Dumais. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

- [132] Mufti N Laskar, Luis G Sanchez Giraldo, and Odelia Schwartz. Deep neural networks capture texture sensitivity in v2. *J Vis*, 20(7):21–1, Jul 2020.
- [133] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022.
- [134] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.
- [135] Asko Lehmuskallio, Jukka Hkkinen, and Janne Seppnen. Photorealistic computer-generated images are difficult to distinguish from digital photographs: a case study with professional photographers and photo-editors. *Visual Communication*, 18(4):427–451, 2019.
- [136] Jay Lemke. Multiplying meaning. *Reading science: Critical and functional perspectives on discourses of science*, 87, 1998.
- [137] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online, June 2021. Association for Computational Linguistics.
- [138] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online, July 2020. Association for Computational Linguistics.
- [139] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*, 134(3):103, 2026.

- [140] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022.
- [141] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar, editors, *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pages 196–212, Montreal, Canada, 11 Dec 2015. PMLR.
- [142] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368, 1957.
- [143] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.
- [144] Tal Linzen and Marco Baroni. Syntactic structure from deep learning. *ArXiv*, abs/2004.10827, 2020.
- [145] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, jun 2018.
- [146] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online

and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

- [147] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22002–22016, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [148] Ruoyu Liu, Jiahui Geng, Andrew J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. arXiv preprint arXiv:2410.21333.
- [149] Vivian Liu, Han Qiao, and Lydia Chilton. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery.
- [150] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10626–10637, 2021.
- [151] Andreas Lommatzsch, Benjamin Kille, Özlem Özgöbek, Yuxiao Zhou, Jelena Tešić, Cláudio Bartolomeu, David Semedo, Lidia Pivovarova, Mingliang Liang, and Martha Larson. Newsimages: addressing the depiction gap with an online news dataset for text-image rematching. In *Proceedings of the 13th ACM Multimedia Systems Conference*, MMSys '22, page 227–233, New York, NY, USA, 2022. Association for Computing Machinery.

- [152] Max M Louwerse. Symbolic or embodied representations: A case for symbol interdependency. In *Handbook of Latent Semantic Analysis*, pages 107–120. Routledge, 2007.
- [153] Max M Louwerse. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302, 2011.
- [154] Hongjing Lu, Qing Liu, Nicholas Ichien, Alan Loddon Yuille, and Keith J. Holyoak. Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. In *Annual Meeting of the Cognitive Science Society*, 2019.
- [155] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*, 2017.
- [156] Gary Lupyan and Benjamin Bergen. How language programs the mind. *Topics in Cognitive Science*, 8(2):408–424, 2016.
- [157] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pre-trained text-to-image transformers for story continuation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 70–87, Berlin, Heidelberg, 2022. Springer-Verlag.
- [158] Manasi Malik and Leyla Isik. Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14(1):7317, 2023.
- [159] Eric Mandelbaum, Yarrow Dunham, Roman Feiman, Chaz Firestone, EJ Green, Daniel Harris, Melissa M Kibbe, Benedek Kurdi, Myrto Mylopoulos, Joshua Shepherd, et al. Problems and mysteries of the many languages of thought. *Cognitive Science*, 46(12):e13225, 2022.
- [160] Diego Marconi. *Lexical Competence*. MIT Press, Cambridge, MA, 1997.
- [161] Ludovic Marin and Ghiles Mostafaoui. Social-motor experience and perception-action learning bring efficiency to machines. *The Behavioral and Brain Sciences*, 40, 2017.

- [162] David C McClelland. Managing motivation to expand human freedom. *American psychologist*, 33(3):201, 1978.
- [163] James L McClelland and Jeffrey L Elman. The trace model of speech perception. *Cognitive Psychology*, 18(1):1–86, 1986.
- [164] James L McClelland and Timothy T Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, 2003.
- [165] Emalie McMahon et al. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047.e8, 2023.
- [166] Mehmet I. Mehmet and Rodney J. Clarke. B2b social media semantics: Analysing multimodal online meanings in marketing conversations. *Industrial Marketing Management*, 54:92–106, 2016.
- [167] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.
- [168] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.
- [169] Roland Meyer. Images from images: Generative ai and the reconfiguration of the ‘photographic’. *Photography and Culture*, 18(3-4):281–290, 2025.
- [170] Grace Qiyuan Miao, Yanru Jiang, Ashley Binnquist, Agnieszka Pluta, Francis Steen, Rick Dale, and Matthew Lieberman. A deep neural network approach for integrating neural and behavioral signals: Multimodal investigation with fnirs hyperscanning and facial expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

- [171] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.
- [172] Emily J. Miller, Benjamin A. Steward, Zaire Witkower, Clare A. Sutherland, Eva G. Krumhuber, and Alex Dawel. Ai hyperrealism: Why ai faces are perceived as more real than human ones. *Psychological Science*, 34(12):1390–1403, 2023.
- [173] Melanie Mitchell. *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux, New York, NY, 2019.
- [174] Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *CoRR*, 2023.
- [175] Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013.
- [176] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [177] Nandita Shankar Naik, Christopher Potts, and Elisa Kreiss. CommVQA: Situating visual question answering in communicative contexts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13362–13377, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [178] Sophie J. Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.
- [179] Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. The connection between the text and images of news articles: New insights for multimedia

- analysis. In Nicoletta Calzolari, Frédéric B  chet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H  l  ne Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4343–4351, Marseille, France, May 2020. European Language Resources Association.
- [180] OpenAI et al. Gpt-4 technical report. *ArXiv 2303.08774*, 2024.
- [181] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, Academic Mindtrek ’22, page 192–202, New York, NY, USA, 2022. Association for Computing Machinery.
- [182] J. Kevin O’Regan. *Why Red Doesn’t Sound Like a Bell: Understanding the Feel of Consciousness*. Oxford University Press, Oxford, 2011.
- [183] J Kevin O’Regan and Alva No  . A sensorimotor account of vision and visual consciousness. *The Behavioral and Brain Sciences*, 24(5):939–973; discussion 973–1031, 2001.
- [184] Ryan G Pacheco, Dale A. Hamilton, and Barry Myers. Knn vs svm: A comparison of algorithms. 2017.
- [185] Sejin Paik, Sarah Bonna, Ekaterina Novozhilova, Ge Gao, Jongin Kim, Derry Wijaya, and Margrit Betke. The affective nature of ai-generated news images: Impact on visual journalism. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2023.
- [186] Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041, 06 2023.
- [187] Diane Pecher and Rolf A. Zwaan, editors. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge University Press, New York, 2005.

- [188] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.
- [189] Leif E. Peterson. K-nearest neighbor. *Scholarpedia*, 4:1883, 2009.
- [190] Steven T Piantadosi. Modern language models refute chomsky’s approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 15:353–414, 2023.
- [191] Steven Pinker and Michael T Ullman. The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463, 2002.
- [192] David Pitt. Mental representation. In *Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, 2020.
- [193] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.
- [194] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.
- [195] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6078–6087, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [196] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 18–34, December 2024.

- [197] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Chuan Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint, abs/2204.06125*, 2022.
- [198] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.
- [199] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. The BreakingNews dataset. In Anya Belz, Erkut Erdem, Katerina Pastra, and Krystian Mikolajczyk, editors, *Proceedings of the Sixth Workshop on Vision and Language*, pages 38–39, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [200] Amon Rapp, Chiara Di Lodovico, Federico Torrielli, and Luigi Di Caro. How do people experience the images created by generative artificial intelligence? an exploration of people’s perceptions, appraisals, and emotions related to a gen-ai text-to-image model and its creations. *International Journal of Human-Computer Studies*, 193:103375, 2025.
- [201] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1060–1069. PMLR, 2016.
- [202] Alex Reinhardt, Ben Markey, Michael Laudénbach, Kachata Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122, 2025.
- [203] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*

- International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.
- [204] Sandra Ristovska. The need for visual information policy. *Surveillance & Society*, 18(3):418–421, 2020.
 - [205] Irvin Rock. *The Logic of Perception*. MIT Press, 1983.
 - [206] Frank Röder, Ozan Özdemir, Phuong D H Nguyen, Stefan Wermter, and Manfred Eppe. The embodied crossmodal self forms language and interaction: A computational cognitive review. *Frontiers in Psychology*, 12:716671, 2021.
 - [207] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.
 - [208] Timothy T Rogers, David H Rakison, and James L McClelland. U-shaped curves in development: A pdp approach. *Journal of Cognition and Development*, 5(1):137–145, 2004.
 - [209] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
 - [210] Daniel Rubinstein and Katrina Sluis. A life more photographic: Mapping the networked image. *Photographies*, 1(1):9–28, March 2008.
 - [211] David E. Rumelhart and James L. McClelland. On learning the past tenses of english verbs. In *Psycholinguistics: Critical Concepts in Psychology*, volume 4, pages 216–271. Taylor & Francis, 1986.
 - [212] Koustuv Saha and Amit Sharma. Causal factors of effective psychosocial outcomes

- in online mental health communities. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 2020.
- [213] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc.
- [214] Ruslan Salakhutdinov and Geoffrey E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics*, 2007.
- [215] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.
- [216] Elvis Saravia, Hsien-Chi Toby Liu, Yi-Hsiang Huang, Jun Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3687–3697, 2018.
- [217] Habiba Sarhan and Simon Hegelich. Understanding and evaluating harms of ai-generated image captions in political images. *Frontiers in Political Science*, 2023.
- [218] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828, 2014.
- [219] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

- [220] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018.
- [221] Bin Shen, Brandon Richard Webster, Alice J. O’Toole, Kevin W. Bowyer, and Walter J. Scheirer. A study of the human perception of synthetic faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.
- [222] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *ArXiv*, abs/1808.03314, 2018.
- [223] Joe Shishido. Joe shishido gesture 2020-01-21.gif. January 21 2020. Licensed under Creative Commons. Accessed: 2024-11-02.
- [224] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.
- [225] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Association for Computing Machinery, 2015.
- [226] Chiranjibi Sitaula, Sunil Aryal, Yong Xiang, Anish Basnet, and Xuequan Lu. Content and context features for scene image representation. *Know.-Based Syst.*, 232(C), nov 2021.
- [227] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.
- [228] Anders Søgaard. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, 33(1):33–54, 2023.

- [229] Milo Stanojević, Jonathan Brennan, Donald Dunagan, Mark Steedman, and John Tracy Hale. Modeling structure-building in the brain with ccg parsing and large language models. *Cognitive science*, 47 7:e13312, 2022.
- [230] Adrian Staub and Jr. Clifton, Charles. Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2):425–436, 2006.
- [231] Edina Strikovic and Hannes Cools. Reality re-imag(in)ed. mapping publics’ perceptions and evaluations of ai-generated images in news contexts. *Digital Journalism*, 0(0):1–22, 2025.
- [232] Christopher Summerfield. *These Strange New Minds: How AI Learned to Talk and What It Means*. Viking, New York, 2025.
- [233] David Sussillo, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18:1025–1033, 2015.
- [234] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27:3998–4011, 2017.
- [235] Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, Ramya Namuduri, Bodun Hu, Juan Diego Rodriguez, Puyuan Peng, and Greg Durrett. Chartmuseum: Testing visual reasoning capabilities of large vision-language models, 2026.
- [236] Ahmed Tealab. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, 3(2):334–340, 2018.
- [237] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

- [238] Ian Tenney, Patrick Xia, Berlin Chen, Adam Wang, Adam Poliak, R. Thomas McCoy, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.
- [239] T. J. Thomson, Ryan J. Thomas, and Phoebe Matich. Generative visual ai in news organizations: Challenges, opportunities, perceptions, and policies. *Digital Journalism*, 13(10):1693–1714, 2025.
- [240] Shengbang Tong, Ziqi Liu, Yifan Zhai, Yadong Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578. IEEE, 2024.
- [241] Rachel Tucciarelli, Nicholas Vehar, Sapphira Chandaria, and Manos Tsakiris. On the realness of people who do not exist: The social processing of artificial faces. *iScience*, 25(12):105441, 2022.
- [242] Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind, revised edition: Cognitive science and human experience*. MIT press, 2017.
- [243] Daniel Varona and Juan Luis Suárez. Discrimination, bias, fairness, and trustworthy ai. *Applied Sciences*, 12(12), 2022.
- [244] Henriikka Vartiainen and Matti Tedre. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity*, 34(1):1–21, 2023.
- [245] Paolo Viviani and Natale Stucchi. Biological movements look uniform: Evidence of motor-perceptual interactions. *Journal of Experimental Psychology: Human Perception and Performance*, 18(3):603–623, 1992.
- [246] Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia*, 26:8609–8624, 2024.

- [247] Taylor Webb, Shuhao Fu, Trevor Bihl, Keith J. Holyoak, and Hongjing Lu. Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Communications*, 14(1):5144, 2023.
- [248] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.
- [249] Margaret Wilson. Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9:625–636, 2002.
- [250] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338, March 2025.
- [251] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *CoRR*, abs/2004.14545, 2020.
- [252] Yiqian Xie. Meed: A multimodal conversational dataset for learning-related emotion recognition in collaborative learning. *IEEE Dataport*, 2026.
- [253] Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, 546:126287, 2023.
- [254] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324, 2018.
- [255] Daniel L K Yamins et al. Performance-optimized hierarchical models predict neural

- responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–8624, 2014.
- [256] Xuwen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. Journalistic guidelines aware news image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5162–5175, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [257] Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2503.07906.
- [258] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3572–3582, Los Alamitos, CA, USA, June 2020. IEEE Computer Society.
- [259] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78, 2014.
- [260] Yiling Yun and Hongjing Lu. Human semantic representations of social interactions from moving shapes. *ArXiv 2509.20673*, 2025.
- [261] Alessandra Zarcone, Marten van Schijndel, Jorrig Vogels, and Vera Demberg. Saliency and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, 7:844, 2016.

- [262] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.
- [263] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2018.
- [264] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods. In *Proceedings of the Ninth International Conference on Multimodal Interfaces*, 2007.
- [265] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LI*, page 310–325, Berlin, Heidelberg, 2024. Springer-Verlag.
- [266] Elizabeth Qing Zhang, Edward Ruoyang Shi, and Lluís Barceló-Coblijn. Categorical perception and language evolution: a comparative and neurological perspective. *Frontiers in Psychology*, 14:1110730, 2023.
- [267] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Huang, Xiaogang Wang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5907–5915, 2017.
- [268] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [269] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

- [270] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [271] Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. Focus! relevant and sufficient context selection for news image captioning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6078–6088, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [272] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [273] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multi-modal learning: A survey. *arXiv preprint arXiv:2304.01008*, 2023.
- [274] Amir Zur, Elisa Kreiss, Karel D’Oosterlinck, Christopher Potts, and Atticus Geiger. Updating clip to prefer descriptions over captions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20178–20187, 2024.
- [275] Rolf A Zwaan. Embodiment and language comprehension: Reframing the discussion. *Trends in cognitive sciences*, 18(5):229–234, 2014.

- [276] Leyla Roskan Çağlar and Stephen José Hanson. Back to the future: The return of cognitive functionalism. *The Behavioral and brain sciences*, 40:e257, 2017.
- [277] Şeyda Özçalışkan, Susan Goldin-Meadow, Dedre Gentner, and Carolyn Mylander. Does language about similarity play a role in fostering similarity comparison in children? *Cognition*, 112(2):217–228, 2009.