

UCLA

UCLA Electronic Theses and Dissertations

Title

Grounded-Knowledge-Enhanced Instruction Understanding for Multimodal Assistant Applications

Permalink

<https://escholarship.org/uc/item/5rh3w127>

Author

Wu, Te-Lin

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Grounded-Knowledge-Enhanced Instruction Understanding for Multimodal Assistant
Applications

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

by

Te-Lin Wu

2024

© Copyright by
Te-Lin Wu
2024

ABSTRACT OF THE DISSERTATION

Grounded-Knowledge-Enhanced Instruction Understanding for Multimodal Assistant
Applications

by

Te-Lin Wu

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2024

Professor Nanyun Peng, Chair

With the recent advancements in artificial intelligence (AI), researchers are making endeavours towards building an AI that can understand humans, collaborate with humans, and help or guide them to accomplish certain everyday chores. The actualization of such an assistant AI can pose several challenges including planning (on certain events), comprehending human instructions, multimodal understanding, and grounded conversational ability.

Imagine a scenario that one wishes to perform a task, such as “making a plate of fried rice”, or “purchasing a suitable sofa bed”, which can require multiple steps of actions and manipulation of certain objects. How would an assistant AI collaborate with humans to accomplish such desired tasks? One crucial aspect of the system is to understand *how* and *when* to take **a certain action**, which is often learned from interpreting and following a **guidance**, a piece of resource that encompasses knowledge about accomplishing the task and potentially the events that will occur during task completions. The guidance can come from human verbal interactions (*e.g.*, in the form of a conversation or a question) or static written instructional manuals.

In the first part of this thesis, I will decompose the proposed system framework into three foundational components: (1) **task-step sequencing/planning**, where the AI needs to understand the appropriate sequential procedure of performing each sub-task to accomplish the whole task, especially when the task knowledge is learned from instructional resources online

that can be many and do not always come consolidated with proper ordering; (2) **action-dependencies understanding**, where an agent should be able to infer dependencies of performing an action and the outcomes after executing a particular action, in order to examine the situations and adjust the plan of accomplishing tasks; (3) **multimodal grounding and active perception**, that we equip the AI with the ability to actively ground the visually perceived surroundings to the textual instructions (or verbal interactions) and perform reasoning over multimodal information along the task completions.

In the second part of this thesis, I will introduce two newly curated resources that foresee the next-phase challenges towards building a strong and helpful assistive AI. One such resource focuses on **counterfactual reasoning**, a type of reasoning capability humans frequently rely on when performing complex decision making processes; while the other presents a **comprehensive suite of multimodal capabilities** of an assistive AI to function in a virtually created world.

Combining the two parts, the foundational components as well as the established novel challenging benchmarks, this thesis aims at providing a comprehensive research road map for the research direction of next-generation (multimodal) AI assistants.

The dissertation of Te-Lin Wu is approved.

Wei Wang

Yizhou Sun

Aditya Grover

Nanyun Peng, Committee Chair

University of California, Los Angeles

2024

To those who support me and trust me.

Table of Contents

1	Towards Actualizing Virtual Assistant AI	1
1.1	The Roadmap	1
1.1.1	Temporal Action Dynamics: Sequencing Instruction Steps	2
1.1.2	Action-and-Condition Dependencies	2
1.1.3	Situated Action-Knowledge-Driven Conversational AI	3
1.1.4	Multimodal Task-Centric Counterfactual Reasoning	3
1.2	Contributions and Structure of the Thesis	4
1.3	Other Relevant Publications	5
I	The Foundations of Multimodal Assistive AI	6
2	Temporal Action Dynamics: Multimodal Instruction Sequencing	8
2.1	Introduction	8
2.2	Background and Related Work	10
2.3	Problem and Datasets	11
2.3.1	Problem Definition	11
2.3.2	Datasets	11
2.3.3	Human Annotation	12
2.4	Sequence-Aware Multimodal PreTraining	14
2.4.1	Technical Challenges	14
2.4.2	Input Encoders	14
2.4.3	Sequence-Aware Pretraining	15
2.4.4	Order Decoder – BERSON	18
2.5	Experiments	18
2.5.1	Evaluation Metrics	18

2.5.2	Implementation Details	19
2.5.3	Standard Benchmark Results	20
2.5.4	Evaluating with Alternative Orders	21
2.6	Summary	24
3	Learning Action Dependencies for Comprehending Task-Knowledge	25
3.1	Introduction	25
3.2	Background and Related Work	27
3.3	Terminologies and Problem Definition	28
3.4	Datasets and Human Annotations	29
3.4.1	Annotations and Task Specifications	30
3.5	Training With Weak Supervision	31
3.5.1	Linking Heuristics	31
3.5.2	Linking Algorithm	33
3.6	Models	35
3.6.1	Non-Contextualized Model	36
3.6.2	Contextualized Model	36
3.6.3	Learning	36
3.7	Experiments and Analysis	37
3.7.1	Training and Implementation Details	37
3.7.2	Experimental Setups	37
3.7.3	Experimental Results	38
3.8	Summary	41
4	Tracing Active Objects Throughout Tasks with Symbolic World Knowledge	42
4.1	Introduction	42
4.2	Background and Related Work	44
4.3	Tasks and Terminologies	45
4.3.1	Technical Challenges	45
4.3.2	Datasets	46
4.4	Method	48

4.4.1	Adapting GLIP	48
4.4.2	LLM for Action-Object Knowledge	50
4.4.3	Object-Centric Joint Inference	52
4.5	Experiments and Analysis	53
4.5.1	Ego4D SCOD	53
4.5.2	TREK-150	56
4.6	Summary	58
II	New Challenges for Multimodal Assistive AI	60
5	ACQUIRED: A Dataset for Answering Counterfactual Questions In Real-Life Videos	62
5.1	Introduction	62
5.2	Background and Related Work	65
5.3	The ACQUIRED Dataset	67
5.3.1	Dataset Design & Collection	67
5.3.2	Dataset Statistics	71
5.4	Benchmarking Models	71
5.5	Experiments and Analysis	74
5.5.1	Training and Implementation Details	74
5.5.2	Experimental Setup	75
5.5.3	Experimental Results	75
5.6	Summary	77
6	SIMMC-VR: A Task-oriented Multimodal Dialog Dataset with Situated and Immersive VR Streams	79
6.1	Introduction	79
6.2	Background and Related Work	81
6.3	SIMMC-VR Dataset	82
6.3.1	Multimodal Dialog Generation	83
6.3.2	SIMMC-VR Dataset Analysis	90
6.3.3	Novel Challenges to SIMMC 2.0	91

6.4	SIMMC-VR Task Formulation	92
6.4.1	Multimodal Dialog State Tracking	92
6.4.2	Multimodal Coreference Resolution	92
6.4.3	Failure-Mode Prediction	93
6.4.4	Dialog Response Generation	93
6.5	Modeling & Experimental Analysis	93
6.5.1	Experimental Results	95
6.6	Summary	96
III Conclusion		97
7	Conclusion and Future Directions	98
7.1	Summary of Contributions	98
7.2	Summary of Technical Limitations	99
7.3	The Next Steps	103
7.4	Future Research Directions	105

List of Figures

1.1	The proposed roadmap of the fundamentals of building an assistant AI.	2
2.1	Multimodal task procedure sequencing: The left column shows unordered instruction steps from the manual <i>How To Make Wood Signs</i> . Each step is a text description and its associated image. Without the complementary information from the visuals, a novice may have difficulty inferring the proper task order. Considering multimodal information, the proper order can be correctly inferred (right column).	9
2.2	Sequence-aware pretraining includes: (1) masked language modeling (MLM), (2) image-swapping prediction (ISP/PISP) which requires the model to predict if some images (image-patches) are swapped, and (3) sequential masked region modeling (SMRM) where models are asked to reconstruct masked regions in each image within the input sequence.	16
2.3	Top-3 and least-2 categories of human-model performance difference (in PMR): The selected categories have >10 samples. The difference bars on the multimodal model series are compared against the text-only model series.	22
3.1	The Action Condition Inference Task: We propose a task that probes models' ability to infer both <i>preconditions</i> and <i>postconditions</i> of an <i>action</i> from instructional manuals. It has wide applications to <i>e.g.</i> assistive AI and task-solving robots. *Original instructions are rephrased for simplicity in this illustration.	26
3.2	Terminologies: (Left) We show a few exemplar actionables (light yellow) with their associated preconditions (light blue) and postconditions (light green). Notice that an actionable can have multiple pre- or postconditions and they can span across different instruction steps. For simplicity we do not show an exhausted set of text segments of interests, <i>i.e.</i> in the actual dataset there might be more. (Right) we show one sample SRL extractions which correspond to one of the action-condition dependency linkages on the left.	28

3.3	Model architectures: (a) Non-contextualized pairwise model: The model only considers a pair of given text segments. (b) Contextualized model: The model takes the whole instruction paragraphs (<i>i.e.</i> contexts) and wrap each text segment with our special tokens (<a>), where each segment representation is obtained by taking an average over its token representations. The <i>ordered</i> concatenated segment representations will then be fed into an MLP to make the final predictions.	35
4.1	Active object grounding is the task of localizing the active objects undergoing state change (OUC). In this example action instruction "cut the pawpaw into half with the knife", the AI assistant is required to firstly infer the OUC (pawpaw) and the Tool (knife) from the instruction, and then localize them in the egocentric visual scenes throughout the action trajectories. Symbolic knowledge including pre/post conditions and object descriptions can bring additional information to facilitate the grounding.	43
4.2	Ego4D SCOD active grounding: Example object undergo change (OUC) due to the instructed actions and associated Tools, spanning: the pre-condition, point-of-no-return (PNR) and post-condition frames.	46
4.3	Overview of proposed framework that comprises a base multimodal phrase grounding model (GLIP), a frame-type predictor, a knowledge extractor leveraging LLMs (GPT), and predictions supervised by both bounding box regression of the objects and their ranked scores.	48
4.4	Model architecture (knowledge-enhanced grounding): On the left depicts the word-region alignment (contrastive) learning of the base GLIP architecture, where the model is trained to align the encoded latent word and image features with their dot-product logits being supervised by the positive and negative word-region pairs. On the right illustrates the enhanced object-knowledge grounding. During training we apply an object-type dependent mask to propagate the positive alignment supervisions; while during inference time the frame-type predictor (offline trained by the encoded textual and image features) acts as a combinator to fuse dot product-logit scores from both (extracted) object phrases and corresponding knowledge. (Note that for simplicity we do not fully split some phrases into individual words.)	49
4.5	The GPT knowledge extraction pipeline. Demonstrated through an example from the Ego4D SCOD Dataset.	50

4.6	Qualitative inspections , mainly on the effectiveness of the GPT generated symbolic knowledge. Bounding box color code: Ground truth boxes , Models with uses of symbolic knowledge (MD-1) , <i>i.e.</i> the GPT+Conds.+Desc. ; Models without uses of symbolic knowledge (MD-2) , <i>i.e.</i> , the vanilla GPT	57
5.1	The ACQUIRED dataset is a video question answering (QA) dataset that specifically focuses on <i>counterfactual reasoning</i> on diverse real-world events. Our dataset concerns three types of commonsense reasoning dimensions: physical, social, and temporal, and encompasses videos from both third-person (upper) and first-person (lower) viewpoints. Each question is curated with a correct and a distractor answer. Each answer is by itself individually judgeable, and hence our dataset can be approached in either binary True/False or multiple-choice setting.	63
5.2	Data collection workflow	69
5.3	Top-40 frequent word-types in the dataset.	73
6.1	SIMMC-VR is a Situated Interactive Multi-Modal Conversation dataset that features task-oriented user↔assistant dialogs <i>streamed immersively</i> in a virtual-reality (VR) environment. The dataset is created on programmed realistic shopping scenarios and actively-rendered photorealistic user visual observations, which brings new challenges for complex spatial-temporal reasoning on the multimodal interactions (visual cues and grounded-dialogs).	80
6.2	Dialog generation flow: (Upper half) a meta-agenda is firstly programmed to sample an <i>object-centric</i> flow (grounded in the environment), which is used by the goal generator to sample high-level dialog goals. These goals are then used by both user and assistant simulators to synthesize templated utterances, which are then manually paraphrased by linguistic experts for diversity and naturalness (lower half).	84

6.3	Multimodal dialog generation: (Right most) meta-agenda illustrates an exemplar shopping scenario that concerns user demanding <i>complementary</i> (<i>i.e.</i> <u>can go with</u>) types for the first two items (jacket ↔ skirt) and the <i>same</i> type between the 2nd and 3rd items. Colors and patterns are not constrained, while the scenario simulates longer traversal is required (<i>far</i>) between the first two items and the latter two are <i>close-by</i> . (Middle) Path planning: the navigational utterances will be grounded on the planned path (displacements and orientations) and the referential objects (left most) used to facilitate the guidance are sampled according to <i>softmax</i> scores on a ranking (via features <i>e.g.</i> eye-gaze, color-contrast) of most suitable landmarks.	85
6.4	Plots of: (a) utterance lengths in dialogs, (b) acts and activities, and (c) co-reference distance between object mentions.	91
6.5	Dialog act(s) transitions for the first four rounds of dialogs in the <u>fashion</u> domain. The acts and activities are denoted for brevity as ACT:ACTIVITY: [A U] [turn_index] , where U and A denote user and assistant, respectively. The shown branching and inter-connectivity justifies the diversity of the synthesized dialog flows.	91
6.6	Baseline models: The inner grey box (denoted “GPT-2/BERT”) is the language model either as (is) the MM-DST model or the language encoder of the video-language model (VIOLET adopts BERT). The video-language model predicts MM-Coref via dense object descriptors, while MM-DST model generates (via GPT-2) the flattened target strings. . .	94
7.1	The human-feedback improvable system: is designed to comprehend and elicit effective human feedback to improve the multimodal AI models.	104

List of Tables

2.1	General statistics of the two datasets: We provide the detailed component counts of the datasets used in this work, including the statistics of tokens and sentences from the instruction steps (lower half of the two tables).	12
2.2	Golden-test-set performance: Models which take multimodal inputs (for both Visual-BERT and CLIP-ViL encoders) consistently outperform the ones that only take unimodal inputs. Our proposed sequence-aware pretraining is shown consistently helpful throughout the three modality variants. Humans show larger performance gain when both modalities of inputs are provided, and are more robust to the local ordering as implied by the smaller gaps between L_q and L_r	20
2.3	Model ablation studies: We provide a performance breakdown for incremental combinations of the pretraining objectives, ablated on the best performing models (CLIP and CLIP-ViL) from Table 2.2 for each dataset and modality.	20
2.4	Multi-reference performance: (\dagger denotes human performance) Our golden-test-set can be decomposed into two subsets: Single where each instance in this subset only has one single originally authored ground truth, and Multi . where each instance features multiple ground truths from alternative orders. For the Multi . subset, two types of performance can be computed: single considers only the originally authored ground truth and multi computes the multi-reference performance. All denotes the entire test-set combining the results from Single and Multi . subsets. Results are reported on the two main competitors: multimodal and text-only using the best performing models from Table 2.2 in each modality. % of instances benefit w. multi-reference indicates that of what percentage of instances <i>in each multi-reference subset</i> humans and the models benefit (for each instance if its performance improves <i>in any of the metrics</i>) from alternative ground truth orders.	22
2.5	Multi-reference subset statistics: We report the count (cnt) of multi-reference instances in each dataset across the three modalities, and their basic statistics.	23

2.6	Top-5 mean alternative orders by categories: We list top-5 categories in WikiHow according to the number of average ground truth references in their multi-reference subset. We again only list the categories with total instance count >10.	24
3.1	Heuristics used for determining condition linkages between text segments, with sample use-cases and descriptions.	33
3.2	Keywords for deciding a potential linkage: If a keyword is at the beginning of a sentence, we use the (first) comma of that sentence to separate it to two segments and link them accordingly, while the keyword itself is used as the separator otherwise. The segments are then either refined with SRL or kept as they are if SRL does not detect a valid verb.	35
3.3	Annotated-test-set performance: The best performance is achieved by applying all of the proposed heuristics (heus.) and undergoing the two-stage training: finetuned on the annotated-train-set first and then perform the self-training . Note that for the Instructables, both <i>Finetuned</i> and <i>Self</i> are done on the WikiHow training sets and a zero-shot transfer is performed.	38
3.4	Heuristics ablations: The models used here are contextualized models without the second-stage self-training for both datasets, and "-" indicates exclusion (from using all). In general, each of the designed heuristics give incremental performance gain to both datasets, where the temporal component is particularly effective in postcondition predictions (compare to Table 3.3).	39
3.5	Varying annotated-train-set size: on WikiHow (test-set size is fixed at 30%). We use the (best) model trained with all the proposed heuristics and the self-training paradigm.	39
3.6	Exemplar model errors. The second row are from distant segments not link-able even via the keyword heuristic.	40
4.1	Qualitative Analysis of GPT Knowledge Extraction: Examples of cases where GPT-extracted symbolic knowledge are wrong or conflict with Ego4D annotations. Here the GPT-extracted or dataset-annotated knowledge are displayed in GREEN if they match human analysis and RED otherwise. Explanations for each example are provided on the right.	51
4.2	Automatic evaluation of GPT entity extraction. Abbreviations: EM: exact string matching; Overlap: The ratio of GPT extractions fully covering the ground truth phrases	51

4.3	Human evaluation of GPT symbolic knowledge extraction. Abbreviations: Textual : i.e. "textual correctness" "Based on text alone, does the GPT conds./desc. make sense?"; Visual : i.e. "visual correctness": "Does the GPT conds./desc. match what is shown in the image?"	51
4.4	Model performance on Ego4D SCOD. OD : pure object detection. Instr : grounding with instructions. We highlight best OUC performance in RED for and best Tool performance in GREEN	55
4.5	PNR to Post OUC tracking ablation study. Since tracking module only produce a single box for each frame, we report the top-1 performance of our grounding model. (Normally COCO API reports max 100 detection boxes.)	56
4.6	Model performance on TREK-150. OPE denotes One-Pass Evaluation Dunnhofer et al. (2022) and OPE-Det is a variant to OPE where each tracker is initialized with its corresponding object detector prediction on the first frame. Success Score (SS) and Normalized Precision Score (NPS) are standard tracking metrics.	58
5.1	Comparisons of different visual question answering datasets. ACQUIRED is the first to feature all the dimensions.	65
5.2	Sample data points of our dataset.	68
5.3	Annotation drop rate for the first 5 batches. Each video gives 3 pairs of question - correct/distractor answers.	71
5.4	General statistics of the two video domains.	72
5.5	Deployed model fooling rates during collection.	72
5.6	Model benchmarking performance on our ACQUIRED dataset.	76
6.1	Digital assets categories used in SIMMC-VR for both fashion and furniture domains.	84
6.2	Meta-Agenda Programs	87

6.3	Exemplar utterance template and paraphrases in SIMMC-VR. In each row under the second column, the upper terms are the goals and the lower terms are the dialog acts (consisting of acts and activities). We show a few representative dialog acts with their corresponding sample templates (each act may have multiple templates as options) and a sample paraphrase. In each template, the subscripts denote the type of the placeholders, where the contents are filled-in grounded by the multimodal contexts (<i>e.g.</i> , sampled objects, user eye-gazes) or sampled attributes (<i>e.g.</i> , types or colors of the desired item).	89
6.4	SIMMC-VR dataset statistics. On average there are 13.2 objects mentioned in a dialog and more than 20 visible in each video frame, making the video-grounded dialogs diverse and rich in contents. Each video roughly lasts 2 minutes, equating to a total of >130 hours long VR streams.	90
6.5	Baseline performances for Multimodal (1) Dialog State Tracking (DST), (2) Object Coreference (Coref.), (3) Response Generation (Gen.), and (4) Failure Mode Prediction (Fail.). In the lower half, we report the corresponding performance from SIMMC-2.0 with the MM-DST model.	95

Acknowledgements

First and foremost, I would like to thank my advisor, Nanyun Peng. A Ph.D. journey, instead of being a short-term sprint, is more like a long distance marathon that undergoes ups and downs, where numerous obstacles are in between the roads towards the seemingly never-ending finishing line. It would be rather impossible to endure such a long and challenging journey, without the help from a bright and helpful mind. During my five Ph.D. years, Nanyun has given me countless useful and insightful pieces of advice, which not only inspire me on how to perform proper research, but also how to grow myself into a better independent researcher, leader, and even a better and more reliable person. Her prompt guidance always arrives at the time when I needed them the most, just as much like my thesis topic, a just-in-time guidance. I am really grateful and enjoy the time working with Nanyun, and I would like to hereby attribute my deepest appreciation to her, who has been the major reasons that made what I am today as a researcher.

I would like to thank all of the lab members of UCLA PlusLab and NLP Group. If it were not for their kind support, insightful research feedback, joyful late night brainstorming, and most importantly, their company, I am not so sure if I could have endured the long academic journey until its end. They are all not just my colleagues or labmates, I see all of them truly as my great friends. I thoroughly enjoyed the time we all spent together, and I hope I can carry on more exciting discussions, collaborations, and fun moments, even after my Ph.D. years have come to an end.

I would also like to thank all my internship mentors and peers for fruitful and unforgettable internship experiences: Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky from Google Research; Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, and Seungwhan Moon from Meta Reality Labs; Anjali Narayan-Chen, and Aishwarya Padmakumar during my times at Amazon AI; Rajiv Jain, Yufan Zhou, Puneet Mathur, and Vlad Morariu at Adobe Research. I have enjoyed very much the times that I spent and learned a lot from all of them during my research internships at each great labs and places.

During my Ph.D. journey, I am fortunate to have the chance to collaborate with many brilliant and nice people, including Caiqi Zhang, Alex Spangher, Hong Chen, Rujun Han, Yu Zhou, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Mingyu Ma, Shikhar Singh, Nuan Wen, Shao-Hua

Sun, Kuan Fang, Marjorie Freedman, Ralph Weischedel, and Kai-Wei Chang. Much of the collaborations with them contribute to the main parts of this thesis.

Special thanks to my long-time friend, De-An Huang, who was one of the main reasons that I set foot into this exciting and rewarding academic journey, and much appreciation for his company throughout the most down times of mine as not only my best friend, but also my mentor, and life coach. Thanks Cho-Ying Wu and Chin-Cheng Hsu who have accompanied me through countless panicking nights that I have questioned myself why in the first place I have chosen to pursue this journey. I am fortunate to have them as my friends and roommates during some of my hardest times and I thoroughly enjoyed much laughter that they had brought to me to cheer me up whenever I needed. Thanks Mingyu Ma, I-Hung Hsu, Jiao Sun, and Rujun Han as not only my labmates, but also my dear friends, whose kind support, in many details, enabled me to adopt a better life during my Ph.D. studies.

I would like to attribute my deepest appreciation to my parents and my family members, who are my strongest support throughout all my life's ups and downs, as my most impenetrable fortress where I could be shielded and take shelter from. Last but not least, I would like to thank my partner who has always been by my side to emotionally support me, to comfort me, and give me much strength for perseverance along the journey. Her and BB (*BB is our chihuahua*) have often been the reasons that I could come back from being emotionally down.

Finally, I thank those who always support me and trust me. You are the reasons why I am here, and where I will be, in the future.

Curriculum Vitae

- 2010-2014 B.S. in Electrical Engineering
National Tsing Hua University, Hsinchu, Taiwan
- 2015-2017 M.S. in Electrical Engineering
Stanford University, Stanford, California, USA
- 2020 Research Intern
Google Research, Mountain View, California, USA
- 2022 Research Scientist Intern
Meta Reality Labs, Redmond, Washington, USA
- 2022 Applied Scientist Intern
Amazon Alexa AI, Manhattan Beach, California, USA
- 2023 Research Scientist Intern
Adobe Research, San Jose, California, USA
- 2019-2024 Research Assistant, UCLA PlusLab
University of California, Los Angeles, California, USA

Publications

- Te-Lin Wu*, Yu Zhou*, and Nanyun Peng. Localizing active objects from egocentric vision with symbolic world knowledge. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023a.
- Te-Lin Wu*, Zi-Yi Dou*, Qingyuan Hu*, Yu Hou, Nischal Reddy Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Acquired: A dataset for answering counterfactual questions in real-life videos. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.
- Te-Lin Wu, Caiqi Zhang, Qingyuan Hu, Alex Spangher, and Nanyun Peng. Learning action conditions from instructional manuals for instruction understanding. In *Proceedings of the Conference of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023a.
- Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Nanyun Peng, Babak Damavandi, and Seungwhan Moon. Simmc-vr: A task-oriented multimodal dialog dataset with situated and immersive vr streams. In *Proceedings of the Conference of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023b.
- Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. Character-centric story visualization via visual planning and token alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

- Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the Conference of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.
- Mingyu Derek Ma, Muhao Chen, Te-Lin Wu, and Nanyun Peng. Hyperexpan: Taxonomy expansion with hyperbolic representation learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2021.
- Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. Com2sense: A commonsense reasoning benchmark with complementary sentences. In *Proceedings of Findings of the Conference of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-Findings)*, 2021.
- Te-Lin Wu, Shikhar Singh, Sayan Paul, Gully Burns, and Nanyun Peng. Melinda: A multimodal dataset for biomedical experiment method classification. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021a.
- Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Mike Bendersky. Lampret: Layout-aware multimodal pretraining for document understanding. In *VIGIL-NAACL21 (2021)*, 2021b.
- Shao-Hua Sun*, Te-Lin Wu*, and Joseph J Lim. Program guided agent. In *International Conference on Learning Representations (ICLR)*, 2019.
- Kuan Fang*, Te-Lin Wu*, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2139–2147, 2018.
- Amir R Zamir*, Te-Lin Wu*, Lin Sun, William B Shen, Bertram E Shi, Jitendra Malik, and Silvio Savarese. Feedback networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1308–1317, 2017.
- Arianna Yuan, Te-Lin Wu, and James L McClelland. Emergence of euclidean geometrical intuitions in hierarchical generative models. In *CogSci*, 2016.

CHAPTER 1

Towards Actualizing Virtual Assistant AI

1.1 The Roadmap

With its recent rapid advancements, AI technology is becoming more and more prevalent in our daily lives. In the near future, humans are able to gain access to powerful AI assistants that can aid our needs in a just-in-time fashion, just very much like the AI named J.A.R.V.I.S.¹, in the well-known Iron Man movie series. One can easily imagine that, we can get help from the AI assistant by conversing with it during any of our daily activities, or when learning to do certain complex new tasks.

For the notion of assistance, particularly for some complex real-world physical tasks, it often involves the processing of relevant instructions, *i.e.*, the *guidelines* that teach us when to perform what actions. In order for an AI assistant to guide us humans throughout the accomplishment of certain tasks, the AI requires to be equipped with three main capabilities that center around learning and leveraging action-centric knowledge within these instructions: (1) **Temporal action dynamics**: As most of the everyday tasks consist of multiple finer-grained steps, the AI should learn to sequence task-steps and plan the procedure dynamically during taking actions at each sub-task. (2) **Action-dependency knowledge**: The AI should be able to consolidate more detailed instructions that are relevant to the human commands, and structure the instructions into a systematic plan to follow. (3) **Active perception and multimodal grounded interactibility**: The AI agent will eventually be deployed to the real-world, and in many cases, the surroundings are to be actively perceived that the verbal interactions with the humans should be grounded with such an active

¹<https://en.wikipedia.org/wiki/J.A.R.V.I.S.>

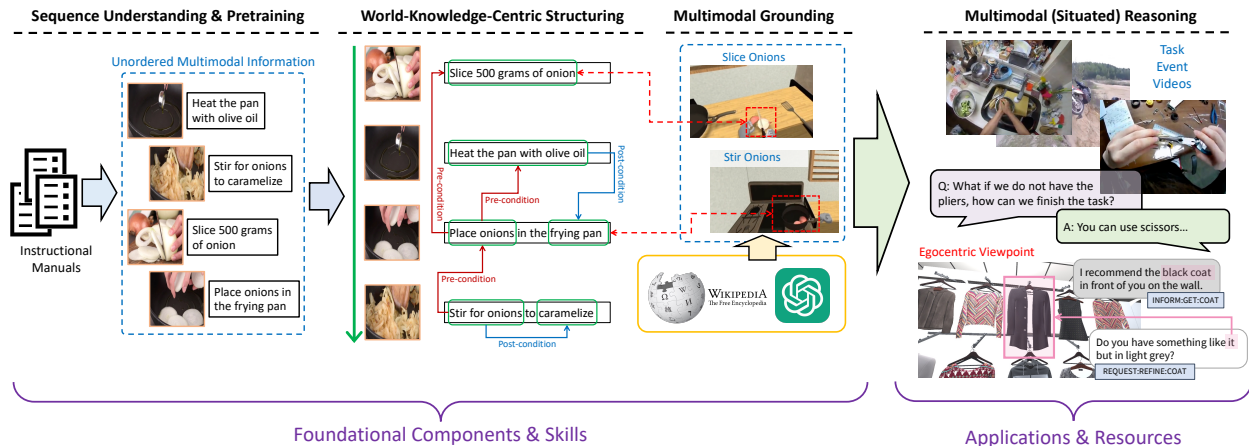


Figure 1.1: The proposed roadmap of the fundamentals of building an assistant AI.

perception. In this chapter of roadmap, I will discuss the overview of these essential aspects and illustrate the designed roadmap that this thesis is based on, as depicted in Figure 1.1.

1.1.1 Temporal Action Dynamics: Sequencing Instruction Steps

As previously mentioned, *guidelines* are common media that are processed and utilized when facing a novel and complex tasks. Fortunately, real-world knowledge of accomplishing certain tasks are often communicated through a set of human written procedural instructions, and hence these *guidelines* are ubiquitous for one to obtain.

However, as much useful as instructions are, they may not always come in a proper sequential order, for example, when we obtain task instructions from multiple different sources or when each sub-task can lead to another series of multi-step instructions. Therefore, *sequencing unordered task-steps* is crucial for comprehending and inferring task procedures, which very much requires temporal and causal common sense reasoning ability, as well as basic task-solving knowledge which often grounded by reality. This sequential nature aspect is illustrated on the leftmost of Figure 1.1.

1.1.2 Action-and-Condition Dependencies

Another essential aspect for comprehending instructions is to infer the dependencies of executing the instructed actions, including their *preconditions*, the prerequisites to be met prior to executing an action, and *postconditions*, the effect caused after performing the action. It is crucial for an agent to understand satisfying which preconditions will allow one to proceed

to the next action, as well as what postconditions (effects) imply the success of performing an action. This is helpful for both of the assistant AI and any autonomous robots, where the former needs to be able to recognize the condition fulfillment for giving the user proper guidance, and the latter should examine the current state or situation to decide on whether it is on the right track towards accomplishing the desired task.

In Figure 1.1 mid-left, once the instructions are properly consolidated, the overall procedure guidance should be parsed into a more systematic and structural form, consisting of primary instructed actions and their aforementioned dependencies and supposed outcomes, for the AI to refer to while guiding or performing the tasks. Understanding the dependencies among actions can also help inferring certain missing task details that are omitted by manually written instructions due to their triviality to humans, which results in an overall more comprehensive and structured understanding of the real-world tasks.

1.1.3 Situated Action-Knowledge-Driven Conversational AI

In addition to the standard multimodal grounding between rather static visual cues to language, on the mid-right of Figure 1.1, one can also observe that certain grounding should be inferred by the AI when humans are actively interacting with the environment. For example, grounding the current visual states to the preconditions inferred from the instructions requires the agent to *actively* explore and examine the environment to produce the correct judgements. Furthermore, during the verbal interactions with the assistant AI, important ability such as multimodal co-reference resolution (*e.g. Which exact object is being referred to?*), visiolinguistic episodic memory retrieval (Grauman et al., 2022c) (*e.g. Where did I put the object?*), guided navigation and manipulation, and giving instructions due to spatiotemporal understanding of both the verbal interaction (conversation) history (*e.g. Can you check to your left whether the ingredients are ready?*) and the user active environmental trajectories.

1.1.4 Multimodal Task-Centric Counterfactual Reasoning

Alongside accomplishing a task, there might be certain detours that need to be made due to some unforeseeable circumstances, such as accidents or missing requirements. For example, imagine if a person is trying to mix some food ingredients, and the required tool instructed by the manual, an electric mixer, is missing. A question, such as *"What if I do not have a*

mixer?" may arouse, and the assistant AI is supposed to respond with: *"You could use a whisk instead."* Another example could be *"Could I avoid spilling the milk if I had chosen a larger bowl?"*.

These questions are of a specific type of reasoning called counterfactual reasoning (Qin et al., 2019), where it concerns a rationale over intervened facts. Such reasoning activity is rather prevalent and common in our everyday life where the *counterfactuality* often comes grounded by certain visually perceived events.

1.2 Contributions and Structure of the Thesis

This thesis mainly discusses our contributions in each of the aforementioned aspects towards actualizing the virtual assistant AI, that humans can converse to while performing their desired tasks. The thesis is divided into two major parts: (I) 2 the foundational components or skills of such an assistant AI, and (II) newly constructed resources to benchmark how AI models are capable of demonstrating the required foundational skills.

In the first part of this thesis, we discuss the importance of the procedural understanding and the utilization of multimodal complementary information in the instructional resources (Wu et al., 2022) to enhance the sequential awareness of AI models (Chapter 2). Followed by the organization of the instructional information, structurally comprehending the consolidated instructions will enable the assistant AI to provide more relevant just-in-time guidance. We particularly focus on inferring the action-and-condition dependencies in multimodal instructional sources (Wu et al., 2023a) (Chapter 3). The comprehensive understanding of the instructions and being able to infer additional knowledge, the assistant AI is then required to translate the knowledge to the visual world to track and interpret users' interactions for immediate assistance (Wu* et al., 2023a) (Chapter 4).

In the second part, we introduce the two newly curated resources that are aimed at evaluating if the AI models possess the fundamental capabilities and skills that we outlined in the previous sections. Chapter 5 discusses the counterfactual reasoning and how it connects to solving real world problems where such type of reasoning is required or beneficial. The proposed ACQUIRED dataset aims at justifying whether state-of-the-art video-language models excel at such important reasoning aspect (Wu* et al., 2023b). Chapter 6 presents a novel and interesting situated conversation dataset where a simulated user-AI interactions take place in

a virtually built shopping environment (Wu et al., 2023b), where this dataset examines a full suite of multimodal capabilities that are crucial to achieve a helpful and effective assistive AI.

Finally, we provide a conclusive summary of contributions made in this thesis, followed by a discussion on potential future research directions in Chapter 7.

1.3 Other Relevant Publications

During my Ph.D. years, I have also published several other relevant research work along the line of understanding the inherent multimodal world, and how to interact within it. In the Demo2Vec work (Fang* et al., 2018) we propose a model that can infer object affordance given the usage videos, whereas the Program-Guided Agent framework (Sun* et al., 2019) is able to comprehend structural instructions (such as programs) to control autonomous agent. While this thesis generally focuses on physical commonsense reasoning that is helpful for real world tasks, general commonsense reasoning (Singh et al., 2021) and concept learning (Ma et al., 2021; Zamir* et al., 2017) are also two important aspects to equip the assistant AI with strong world knowledge. Lastly, I have also conducted work in researching various fundamental multimodal capabilities, such as grounding and understanding complex documents (Wu et al., 2021a,b), geometry (Yuan et al., 2016) across modalities, and multimodal planning with clear story-lines (Chen et al., 2022).

Part I

The Foundations of Multimodal Assistive AI

As discussed in Chapter 1, there are three main components building the foundations of a helpful multimodal assistive AI, they are the capabilities of: (1) learning and modeling the sequential nature inherent in the instructed tasks, (2) structurally comprehending the organized instructional resources, and (3) actively grounding the relevant objects when giving the instructions. Combining the aforementioned three essential skills, the AI assistant is then equipped with the capability of providing just-in-time guidance while co-observing the users' visual viewpoints.

Chapter 2 begins with introducing a multimodal instruction sequencing task that is exactly inspired by how an AI is supposed to comprehend and sort unordered series of information that are useful for accomplishing a task. We evaluate several strong natural language processing (NLP) models as well as multimodal (vision-and-language) models, where the results suggest suboptimal utilization of multimodal information. We thus propose a sequence-aware pretraining technique to allow the vision-language models to more effectively leverage complementary information among different modalities.

Once an organized and consolidated instruction is obtained, it is crucial for the AI to be able to parse and interpret the instructions to give structural guidance. Particularly, Chapter 3 demonstrates the importance of inferring the relations between the (to-be-instructed) actions and their pre- and postconditions. The scarcity of relevant datasets lead us to propose a weakly supervised method to automatically curate training resources to harness the potential action-condition dependencies from (theoretically) unlimited amount of instructional data.

Chapter 4, followed by the previous component, aims at devising a framework that can actively track (and localize) objects involved (or supposed to be involved) in each of the instructed actions. We propose a method that can effectively utilize pre- and postcondition world knowledge to enhance the grounding capabilities of vanilla phrase grounding models, achieving significantly improved performance on tracing the key entities throughout the task accomplishments.

CHAPTER 2

Temporal Action Dynamics: Multimodal Instruction Sequencing

2.1 Introduction

Instructions are essential sources for agents to learn how to complete complex tasks composed of multiple steps (e.g., “making a wood sign from scratch”). However, instructions do not always come in a proper sequential order, for example, when instructions must be combined across sources (e.g., to accomplish a complex task there might be *multiple* useful resources for certain task-steps come out from a single Google search). Therefore, *sequencing unordered task-steps* is crucial for comprehending and inferring task procedures, which requires thorough understanding of event causal and temporal common sense.

Existing work has studied sequencing unordered *texts* from paper abstracts or short stories (Chen et al., 2016; Cui et al., 2018). However, real-life tasks are often complex, and multimodal information is usually provided to supplement textual descriptions to avoid ambiguity or illustrate details that are hard to narrate, as illustrated in Figure 2.1.

To investigate whether current AI techniques can efficiently leverage multimodal information to sequence unordered task instructions, we curate two datasets from *online instructional manuals* (Hadley et al.; Yagcioglu et al., 2018). We consider two representative instruction domains: cooking recipes and “How-To” instructions (WikiHow), and establish human performance for the sequencing task on a subset of each data resource. As certain steps to perform a task can potentially be interchangeable,¹ we also collect annotations of possible

¹For example, without special requirements, preparing certain ingredients of a dish, such as slicing carrots or cucumbers, does not necessarily need to follow a specific order.

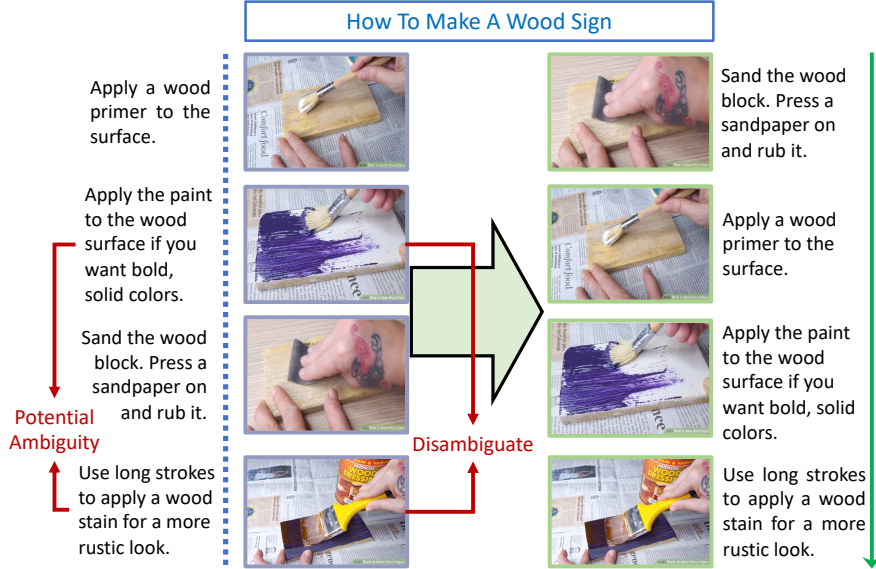


Figure 2.1: Multimodal task procedure sequencing: The left column shows unordered instruction steps from the manual *How To Make Wood Signs*. Each step is a text description and its associated image. Without the complementary information from the visuals, a novice may have difficulty inferring the proper task order. Considering multimodal information, the proper order can be correctly inferred (right column).

orders alternative to the originally authored ones to create *multiple references*.

We observe that multimodal information is consistently helpful for the sequencing task in some preliminary studies. However, compared to humans, current models are less efficient in utilizing multimodal information. We hypothesize that it is because the models do not effectively capture the sequential information in the vision modality nor the sequential alignment between multimodal contents. To address this, we propose to equip models with capabilities of performing **sequential aware multimodal grounding** by proposing several self-supervised objectives to pretrain the models before finetuning them on the downstream sequencing task.

The work of this chapter has the following key contributions: (1) We propose a multimodal sequencing task with two curated instructional manuals, and comprehensive human annotations. (2) We investigate model performance on sequencing unordered manuals, and propose sequence-aware pretraining techniques to more effectively use the multimodal information. We aim at providing insights on which task categories are most challenging for the state-of-the-art models. They also shed the light that more sophisticated sequential multimodal grounding are required to further improve the performance for the proposed

multimodal sequencing task.

2.2 Background and Related Work

The work introduced in this chapter is highly inspired by the *story sequencing test*, which is a popular way of examining children’s abilities on sequential reasoning that is shown evident for procedural understanding (Tomkins, 1952; Baron-Cohen et al., 1986; Loucks et al., 2017). In NLP, most existing works attempt the sequencing task as sorting a series of unordered sentences (Chen et al., 2016; Cui et al., 2018; Logeswaran et al., 2018; Oh et al., 2019; Lee et al., 2020; Calizzano et al., 2021) from paper abstracts or short paragraphs. While certain prior work also attempts to extend it to incorporate multimodality (Agrawal et al., 2016), the dataset used, Visual StoryTelling (Huang et al., 2016), features album images that were not intended to be procedural nor supply unstated details to complement the texts.² In computer vision, existing work leverages shuffle frame prediction for learning video representations (Lee et al., 2017; Xu et al., 2019; Wang et al., 2020; Li et al., 2020a). (Zellers et al., 2021b) also proposes a pairwise relative frame re-ordering objective to learn temporal common sense from scripted videos, however, as their downstream tasks mainly concern visual reasoning and ordering by frame-text-matching (also on Visual StoryTelling), the re-ordering objective is more focused on the visual modality (in contrast, our work’s focus is on both modalities).

In addition to the works introduced in Chapter ??, another prior work (Zhang et al., 2020) also considers WikiHow for learning event temporal ordering, but limited to only pairwise relations. Additionally, a recent work uses WikiHow to infer visual goals (Yang et al., 2021), which aligns with our goal to exploit the multimodal contents within this useful resource.

Another work attempts the original visual ordering task of RecipeQA (Liu et al., 2020) (also an multiple choice task). However, we argue that our task tackles a more complex task as the desired orders need to be directly derived and the event-wise complementary multimodal understanding is not an essential component in these existing works.

²The images come from photo albums and the annotators *create* stories based on their freely arranged image sequences.

2.3 Problem and Datasets

2.3.1 Problem Definition

Given a task procedure S consisting of N steps, where each step $S_i \in S$ can consist of two types of contents: a textual description T_i of tokens $\{T_{i,k}\}_{k=1}^{n_T}$ and/or image(s) $I_i = \{I_{i,k}\}_{k=1}^{n_I}$.³ A model is required to take as inputs a random permutation of S , *i.e.* $S_p = \{S_{p_1}, \dots, S_{p_N}\}$, where p is a permutation (S_{p_j} can take one of the following three modalities: T_{p_j} , I_{p_j} , and $\{T_{p_j}, I_{p_j}\}$), and predict the correct order of S_p , *i.e.* $\text{argsort}(S_p)$.

2.3.2 Datasets

There are three major features we require for the target datasets: (1) It is multimodal. (2) It consists of task procedures as sequences of steps. (3) Different modalities are used intentionally to complement each other. In light of these, we consider the following resources:

RecipeQA. We start from a popular as well as intuitive choice of instruction manuals, recipes, which fully fulfill the aforementioned criteria. RecipeQA (Yagcioglu et al., 2018) is a multimodal question answering dataset consisting of recipes scraped from *Instructables.com* (Yagcioglu et al., 2018). We utilize the recipes collected in RecipeQA and convert each unique recipe into sequential multimodal steps for our task.

WikiHow. To expand the types of instruction manuals for our task beyond recipes, we also consider a popular "How To ..." type of instructions, WikiHow (Hadley et al.), which is an online knowledge base that consists of human-created articles describing procedures to accomplish a desired task. Each article contains a high level goal of a task, a short summary of the task procedures, and several *multimodal* steps where each step consists of a description paired with one or a few corresponding images.

We scrape the entire WikiHow knowledge resource, containing more than 100k unique articles (mostly) with multimodal contents, as well as the hierarchically structured category for each article. Table 2.1 presents the essential statistics of the two datasets.

³For computational concerns, we set $n_I = 1$ in this work.

Type	Counts			
Total Unique Articles	109486			
Total Unique Images	1521909			
Train / Dev / Golden-Test	98268 / 11218 / 300			
Type-Token Ratio	216434 / 82396591 = 0.0026			
Type	Mean	Std	Min	Max
Tokens in a Step Text	52.95	26.25	0	5339
Sentences in a Step Text	3.36	1.3	0	50
Number of Steps of a Task	5.27	2.62	0	75

(a) WikiHow

Type	Counts			
Total Unique Articles	10063			
Total Unique Images	87840			
Train / Dev / Golden-Test	8032 / 2031 / 100			
Type-Token Ratio	91443 / 5324859 = 0.017			
Type	Mean	Std	Min	Max
Tokens in a Step Text	82.08	84.72	0	998
Sentences in a Step Text	4.19	4.22	0	73
Number of Steps of a Task	6.45	2.57	4	20

(b) RecipeQA

Table 2.1: General statistics of the two datasets: We provide the detailed component counts of the datasets used in this work, including the statistics of tokens and sentences from the instruction steps (lower half of the two tables).

2.3.3 Human Annotation

To ensure the validity of our proposed multimodal sequencing task, we establish the human performance via Amazon Mechanical Turk. Since our dataset is constructed from resources that are not directly designed for the sequencing task, the quality of random samples is unverified. Specifically, some articles in WikiHow may not have a notion of proper order among the steps.⁴ As a result, to construct a high quality test set particularly for WikiHow for establishing human performance, we first identify a set of categories which are more likely to feature proper order, *e.g.* *Home and Garden* and *Hobbies and Crafts*.⁵ A random

⁴No temporal or other dependencies among the task-steps, *e.g.* “How to be a good person”, where each step depicts a different aspect and tips of being a good person.

⁵Although the data used for training is not cleansed and thus can be noisy, we believe models can still learn to sequence from many of the articles designed to have proper order.

proportion is then sampled and the co-authors further downsample the subset to 300 samples with the aforementioned criteria via majority vote. For RecipeQA, we randomly sample 100 recipes from the dataset. And hence, the resulting two subsets serve as our **golden-test-set** for performance benchmarking.

Human Performance. Prompted with a task goal and a randomly scrambled sequence of the task-steps (can be one of the following modalities: multimodal or text/image-only), workers are asked to examine the contents and decide the proper performing order. Human performance are then computed against the original authored orders as the ground truths, averaged across the whole set.⁶

Alternative Orders. When performing a task, some steps can be interchangeable. To take the interchangeability into consideration in our benchmark task, we also collect possible alternative orders to the original ones to create multiple references. For each instance in our golden-test-set, given the instruction steps sequenced in their original order, we ask workers to annotate alternative orders if the presented task-steps can be performed following a different order.⁷

Although in this work we are mainly focusing on sequential instructions and hence the interchangeability is also gauged in a sequential manner, we want to point out that the nature of task-step interchangeability is also highly related to parallel (branching) steps of tasks (Sakaguchi et al., 2021a). We argue that the actions that can be performed interchangeably imply no direct dependencies are among these actions and thus can potentially be parallelized, and hence our alternative order formulation can help inferring these parallel actions.

⁶We design an algorithm to compute the inter-annotator agreements (IAAs). The IAAs for (*multimodal*, *text-only*, *image-only*) versions in WikiHow is: (0.84, 0.82, 0.69), and (0.92, 0.87, 0.81) in RecipeQA.

⁷The alternative order annotation IAAs for (*multimodal*, *text-only*, *image-only*) versions in WikiHow is: (0.73, 0.71, 0.78), and (0.79, 0.76, 0.79) in RecipeQA.

2.4 Sequence-Aware Multimodal PreTraining

2.4.1 Technical Challenges

To benchmark the proposed task, we construct models comprising: (1) an **encoder** which encodes multimodal or text/image-only inputs, and (2) an **order decoder** which utilizes the encoded representations to predict the orders. Our preliminary studies verify the effectiveness of multimodal information, however, we also observe that, compared to humans, the multimodality is much under exploited and properly utilized. We hypothetically attribute the reasons of the aforementioned issue to that the standard multimodal grounding techniques (Li et al., 2019; Lu et al., 2019; Su et al., 2020; Chen et al., 2020) do not explicitly concern the sequentiality of text and associated image sequences, and hence may fall short of effectively utilizing the sequential properties in multimodal inputs.

To address this and help models capture **sequentiality in task-steps** better as well as adapt to our target task domains, we pretrain the encoders with several self-supervised objectives on the instructions before integrating them with the decoder. Specifically, to encourage models to have better awareness of the sequential alignments in multimodal instruction steps, we propose to pretrain the encoders with the following self-supervised objectives: (1) masked language modeling (**MLM**), (2) (patch-based) image-swapping predictions (**ISP/PISP**), and (3) sequential masked region modeling (**SMRM**). Figure 2.2 illustrates an overview of the pretraining paradigm (and the model architectures).

2.4.2 Input Encoders

Text-Only Encoders. We use *RoBERTa* (Liu et al., 2019) for text-only inputs. Although the next-sentence prediction in BERT (Devlin et al., 2019) can potentially be exploited for sequencing, we empirically find that RoBERTa performs better.

Multimodal Encoders. We consider the following two V&L models mainly due to their easy adaptation to our proposed sequencing task:

VisualBERT (Li et al., 2019) grounds object detected image regions (*e.g.* by Faster-RCNN (Ren et al., 2016)) to language with a single transformer model (Vaswani et al.,

2017). VisualBERT is pretrained with: (1) multimodal masked language modeling (MLM)⁸, and (2) image-text matching prediction (ITM), where the image in an image-caption pair is randomly replaced with another one to create misalignment, and the model is required to predict whether the current pair is aligned.

CLIP-ViL (Shen et al., 2021) is also a single-stream V&L model similar to VisualBERT, while the visual encoder is replaced by a patch-based model inspired by the ViT (Dosovitskiy et al., 2021) in CLIP (Radford et al., 2021), where the image features are taken as *gridded-image-patches* as shown in Figure 2.2. The pretraining objectives remain the same as VisualBERT. Empirically, both (Shen et al., 2021) and this work find such patch-based model tends to yield better downstream performance.

Image-Only Encoders. We attempt to provide an image-only baseline on our sequencing task with two visual encoders: (1) *ResNet*-based (He et al., 2016) Faster-RCNN model (also the visual encoder in VisualBERT) where both the detected regional features and the whole-image-feature are used, and (2) the aforementioned patch-based *CLIP* model.⁹

2.4.3 Sequence-Aware Pretraining

For the proposed sequence-aware multimodal pretraining objectives, the inputs to the models are generally *ordered* instruction step sequences, which can be further sub-sampled to produce length-varying subsequences. Although we do not find this necessarily benefit the downstream performance, it is observed that the sub-sampling helps the model converge faster. While all of our proposed objectives can be applied to sequence with arbitrary length (≥ 2), without loss of generality and for simplicity, the following sections assume the sub-sampled sequence is of length 2. We will now describe them in details in the followings:

Masked Language Modeling:

The standard MLM (Devlin et al., 2019) is employed by the text-only models to adapt a pretrained language model to the target domain (task instructions). Following prior V&L works, we apply MLM to multimodal models. Specifically, we ensure that the textual de-

⁸RoBERTa is used to initialize VisualBERT and CLIP-ViL.

⁹Without confusion, throughout the paper we term the ViT- and CLIP-inspired visual encoder simply as CLIP.

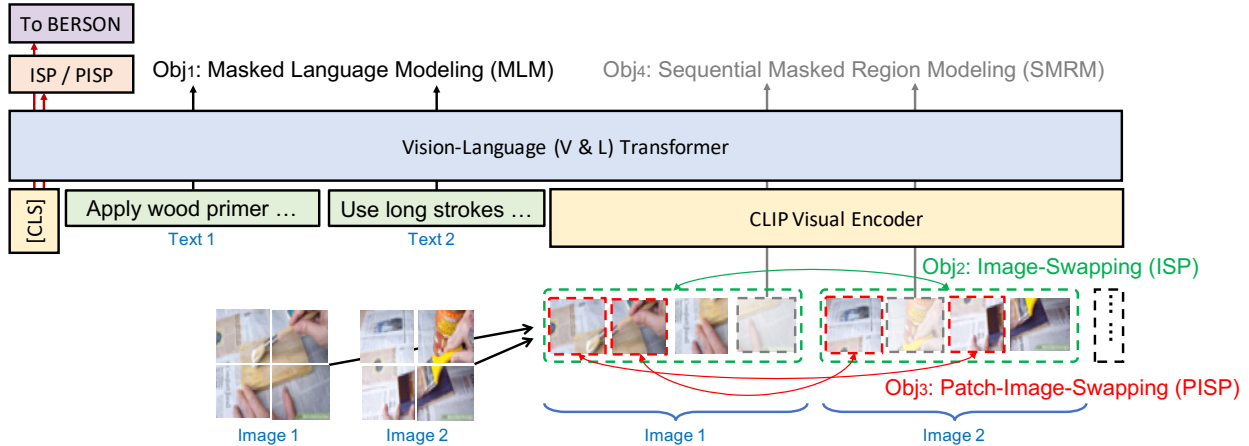


Figure 2.2: Sequence-aware pretraining includes: (1) masked language modeling (MLM), (2) image-swapping prediction (ISP/PISP) which requires the model to predict if some images (image-patches) are swapped, and (3) sequential masked region modeling (SMRM) where models are asked to reconstruct masked regions in each image within the input sequence.

scription of each step T_i gets similar amount of tokens being masked-out such that the models can potentially exploit the image sequences more.¹⁰

Swapping-Based Prediction:

This objective concerns, with certain probability, randomly swapping a pair of items in a sequence and asking the model to judge whether the resulting sequence is properly ordered or not (*i.e.* binary classification). We mainly perform the swapping in the image modality and hence it can be viewed as a sequence-aware version of ITM objective in most V&L models. As in ITM, the output representation at the [CLS] token is used to make the prediction.

Standard. For an ordered sequence S , we can randomly swap two¹¹ items of S , $\{S_i, S_j\}$, where $i < j$, to $\{S_j, S_i\}$, with a certain probability δ . Our preliminary studies find that swapping the textual contents does not necessarily help the downstream performance for either text-only or multimodal models, so we only perform the swapping on the images $\{I_i, I_j\}$ in both multimodal and image-only models. For patch-based image inputs (or regional features), the whole patches of an image are swapped with those of another one within the same sequence, as illustrated in **Obj**₂ in Figure 2.2.

¹⁰As higher chances that the complementary textual information is also masked out from different steps.

¹¹Two is our minimum number for a valid subsequence.

Patch-Based. We can perform the aforementioned swapping prediction with a finer granularity, directly on the image patches. Assuming each image I_i is cropped into w patches (or w detected regions), *i.e.* $\{\mathbf{i}_{i,k}\}_{k=1}^w = \{\mathbf{i}_{i,1}, \dots, \mathbf{i}_{i,w}\}$, we randomly select M (ranging from 1 to w) number of patches each from the two images I_i, I_j (*i.e.* $\{\mathbf{i}_{i,p}\}, \{\mathbf{i}_{j,q}\}, p, q \in M$ -sized sampled indices) to be swapped with probability δ . Specifically, for each image patch $\mathbf{i}_{i,m} \in I_i$, a randomly selected image patch $\mathbf{i}_{j,n} \in I_j$ is sampled to be swapped with. The sampled M -sized indices do not need to be the same set of integers for each image. The **Obj**₃ in Figure 2.2 illustrates the patch-based swapping prediction with $w = 4$ and $M = 2$.

Sequential Masked Region Modeling:

Prior works extend the masked learning to the visual modality, where the masked target is either a predefined discrete visual vocabulary (Sun et al., 2019; Bao et al., 2021) or (soft) object class labels (Lu et al., 2019; Su et al., 2020; Chen et al., 2020). In this work, we construct a feature-based target vocabulary dynamically in each training batch. We first randomly select the same amount of $X\%$ ($X = 15$) patches for **each image** to be masked out (replaced with 0-tensor), and then construct a target vocabulary from the original output representations (before masking) of these patches.

Concretely, denote the output representation of an input image-patch $\mathbf{i}_{i,m}$ as $h(\mathbf{i})_{i,m}$ and the masked positions of I_i as D_i , we can construct a candidate list from all the output representations of the patches at the masked positions of each image, *i.e.* $C = \{h(\mathbf{i})_{i,m}\} \cup \{h(\mathbf{i})_{j,n}\}, m, n \in D_i, D_j$. Denote the masked image patches (the gray-colored image patches in Figure 2.2) as $\mathbf{mask}(\mathbf{i})_{i,m}$, for each output masked representation $h(\mathbf{mask}(\mathbf{i}))_{i,m}$, we concatenate it with all the candidates, *i.e.* $h(\mathbf{mask}(\mathbf{i}))_{i,m} || h(\mathbf{i}'), \forall \mathbf{i}' \in C$, which results in $|C|$ concatenated representations for each masked position. A $|C|$ -way multi-class classification can then be performed by maximizing the probability of $p(\mathbf{i}_{i,m} | h(\mathbf{mask}(\mathbf{i}))_{i,m}; C)$. For robust training, we additionally: (1) shuffle the candidate set C for each masked position to prevent overfitting, and (2) ensure the overlapping of masked positions in each pair of images, $D_i \cap D_j$, is $< 50\%$, allowing the models to utilize information of similar regions from other images in the sequence.

Overall Training Objective:

As the mechanism in some objectives cannot guarantee mutually exclusive impacts (*e.g.* performing ISP and PISP simultaneously may create confusing swapped patches), we employ a turn-taking fashion, with uniform probability, one of the objectives (**Obj**) is sampled for each training mini-batch. The overall pretraining objective is defined as below:

$$L = L_{\text{MLM}} + L_{\text{Obj}}, \text{Obj} \sim \{\text{ISP}, \text{PISP}, \text{SMRM}\} \tag{2.1}$$

2.4.4 Order Decoder – BERTSON

BERTSON is a recently proposed state-of-the-art neural sentence ordering framework (Cui et al., 2020), where a pointer network (Vinyals et al., 2016) exploits both the local (relative pairwise order) and global (self-attentions on top of the entire input sequence) information of the inputs to decode the predicted order. BERTSON mainly exploits the [CLS] output representations for relational understanding, which aligns well with how our encoders are pretrained (Figure 2.2). We integrate our encoders (with or without sequence-aware pretraining) into BERTSON, replacing its original BERT encoder. The BERTSON-module-specific components are freshly initialized and then the entire integrated module is finetuned on our sequencing task.

2.5 Experiments

Our experiments seek to answer these questions: (1) How valid is the proposed task for humans to complete? (2) Is multimodality helpful? (3) Can the proposed sequence-aware pretraining utilize multimodality more effectively? (4) How would results differ when alternative orders are considered?

2.5.1 Evaluation Metrics

We mainly adopt evaluation metrics used in prior sentence ordering works (Chen et al., 2016; Cui et al., 2018, 2020):

Position-Based metrics concern the correctness of the absolute position of each item in a sequence, including: (1) **Accuracy (Acc)** which computes the ratio of absolute positions in the ground truth order that are correctly predicted; (2) **Perfect Match Ratio (PMR)** which measures the percentage of predicted orders exactly matching the ground truth orders;

and (3) **Distance (Dist.)** which measures the average distance¹² between the predicted and ground truth positions for each item.

Longest Common Subsequence computes the average longest subsequences in common (Gong et al., 2016) between the predicted and ground truth orders (\mathbf{L}_q). We also consider a stricter version, longest common substring, which requires the consecutiveness for the comparisons (\mathbf{L}_r).

Kendall’s Tau (τ) (Lapata, 2003) is defined as $1 - 2 \times (\# \text{ inversions}) / (\# \text{ pairs})$, where the inversion denotes that the predicted relative order of a pair of items is inverted compared to the corresponding ground truth relative order, and $\# \text{ pairs} = \binom{N}{2}$ for N -length sequence. Each metric focuses on different perspectives of the predictions, *i.e.* position metrics concern the absolute correctness, while common subsequence and τ metrics measure if general sequential tendency is preserved despite incorrect absolute positions.

2.5.2 Implementation Details

We use the original data splits for RecipeQA. For WikiHow, to prevent models’ exploiting knowledge from similar articles, we split the data so that certain (sub)categories do not overlap in each split. We use only the train splits in each dataset to perform their respective pretraining. Preliminary studies show that joint training with both RecipeQA and WikiHow data does not necessarily improve the downstream performance, thus the models evaluated in the two datasets are trained simply using their respective training sets for faster convergence.

We cap the overall sequence length at 5 and each step description with maximally 5 sentences for both models and humans. The maximum input length per step is 60 tokens (overall maximum length = 300) for training and GPU memory efficiency. $\delta = 0.5$ for both ISP and PISP. All images are resized to 224×224 , and 32×32 patch is used for CLIP-based models, resulting in $7 \times 7 = 49$ patches per image. Aside from standard positional embedding, we only supplement a modality token type embedding (text:=0, image:=1) to the multimodal models. Pretrained weights for each encoder is obtained either from their corresponding code bases or by running their codes on our setup.¹³

¹²Except for distance metric, higher scores are better.

¹³We initialize CLIP-ViL with our pretrained CLIP.

Modality	Encoders	Sequence-aware Pretraining	WikiHow Golden-Test-Set						RecipeQA Golden-Test-Set					
			Acc \uparrow	PMR \uparrow	L $_q$ \uparrow	L $_r$ \uparrow	τ \uparrow	Dist \downarrow	Acc \uparrow	PMR \uparrow	L $_q$ \uparrow	L $_r$ \uparrow	τ \uparrow	Dist \downarrow
Image-Only	ResNet	N	21.73	2.00	2.81	1.73	0.01	7.87	31.20	5.00	3.27	2.07	0.27	6.10
	CLIP	N	24.92	3.33	2.95	1.84	0.08	7.32	38.40	8.00	3.39	2.02	0.35	5.44
	CLIP	Y	28.24	5.00	3.09	1.96	0.16	6.80	47.20	16.00	3.68	2.40	0.52	4.12
	Human Performance		68.16	47.49	4.27	3.51	0.72	2.43	80.40	64.50	4.54	4.02	0.86	1.29
Text-Only	RoBERTa	N	74.75	56.67	4.47	3.78	0.82	1.71	74.00	52.00	4.45	3.68	0.83	1.64
	RoBERTa	Y	75.68	58.67	4.50	3.87	0.82	1.69	77.00	57.00	4.49	3.81	0.84	1.48
	Human Performance		83.35	66.91	4.63	4.11	0.89	1.06	88.92	78.56	4.76	4.41	0.93	0.70
Multimodal	VisualBERT	N	75.30	57.33	4.45	3.83	0.81	1.65	76.20	58.00	4.49	3.85	0.83	1.58
	VisualBERT	Y	77.30	59.67	4.50	3.86	0.83	1.58	78.20	60.00	4.56	3.91	0.85	1.44
	CLIP-ViL	N	76.15	59.00	4.49	3.87	0.82	1.68	79.20	60.00	4.57	3.93	0.85	1.29
	CLIP-ViL	Y	79.87	65.67	4.57	4.05	0.85	1.44	82.60	68.00	4.61	4.10	0.88	1.10
	Human Performance		91.03	79.61	4.78	4.46	0.94	0.52	92.12	83.13	4.82	4.53	0.95	0.45

Table 2.2: Golden-test-set performance: Models which take multimodal inputs (for both VisualBERT and CLIP-ViL encoders) consistently outperform the ones that only take unimodal inputs. Our proposed sequence-aware pretraining is shown consistently helpful throughout the three modality variants. Humans show larger performance gain when both modalities of inputs are provided, and are more robust to the local ordering as implied by the smaller gaps between L $_q$ and L $_r$.

2.5.3 Standard Benchmark Results

Modality	Pretrain	WikiHow Golden-Test-Set						RecipeQA Golden-Test-Set					
		Acc \uparrow	PMR \uparrow	L $_q$ \uparrow	L $_r$ \uparrow	τ \uparrow	Dist \downarrow	Acc \uparrow	PMR \uparrow	L $_q$ \uparrow	L $_r$ \uparrow	τ \uparrow	Dist \downarrow
Image-Only	ISP	27.31	4.00	3.02	1.82	0.12	7.00	43.20	9.00	3.49	2.05	0.47	4.46
	ISP + PISP	27.57	4.67	3.07	1.93	0.16	6.85	43.40	12.00	3.57	2.24	0.48	4.46
Multimodal	MLM	77.08	61.33	4.52	3.96	0.83	1.65	79.60	61.00	4.55	3.93	0.86	1.29
	MLM + ISP	77.61	62.00	4.54	3.97	0.83	1.60	80.00	61.00	4.56	3.93	0.86	1.26
	MLM + SMRM	77.94	62.33	4.54	3.98	0.84	1.60	80.00	59.00	4.53	3.89	0.87	1.26
	MLM + ISP + PISP	78.14	63.33	4.55	4.03	0.84	1.56	80.80	63.00	4.57	3.99	0.87	1.24
	MLM + ISP + SMRM	79.47	63.67	4.57	4.03	0.85	1.54	81.40	63.00	4.57	4.00	0.87	1.20

Table 2.3: Model ablation studies: We provide a performance breakdown for incremental combinations of the pretraining objectives, ablated on the best performing models (CLIP and CLIP-ViL) from Table 2.2 for each dataset and modality.

Table 2.2 summarizes both the human and model performance for each input modality evaluated using the original ground truth orders on the golden-test-set, whereas Table 2.3 summarizes a more detailed breakdown of the model performance when incrementing combinations of pretraining objectives.

As is shown, multimodal information is verified consistently helpful for humans. Compared under same scenario with or without the sequence-aware pretraining, the two multimodal models consistently outperform their text-only counterparts, where the proposed pretraining technique is shown particularly effective for the patch-based multimodal model (CLIP-ViL). However, our top-performing models still exhibit significant gaps below human

performance, especially in PMR.

Additionally, we observe a different trend in the two datasets where the multimodality benefits more in RecipeQA than WikiHow. The gap between the multimodal human and model performance is larger than the text-only counterparts in WikiHow, while a reversed trend is shown in RecipeQA. We hypothesize that recipes may contain more domain-specific language usages and/or less words for the pretrained language models and hence benefits more from the our in-domain sequence-aware pretraining. Humans, on the other hand, benefit more from the images in WikiHow as its texts are hypothesized to contain more ambiguities.

WikiHow Category Analysis. We are interested in knowing on which categories of WikiHow our models perform closer to humans, and on which the multimodal information is most efficiently utilized. In Figure 2.3 we select categories with the top and least performance gaps (with PMR metric, top=3, least=2) between the human and our best performing models. We observe that the categories on which multimodal models outperform the text-only ones the most are also the categories the models perform closest to humans, *e.g. Home and Garden*. We hypothesize that the images in these categories are well complementary to the texts and that our sequence-aware grounding performs effectively. In contrast, in categories such as *Arts and Entertainment* and *Hobbies and Crafts* where humans still enjoy benefits from multimodal information, our models have difficulty utilizing the multimodal information. We hypothesize that better visual understanding may alleviate the potentially suboptimal grounding as images of these categories can contain many objects that are not so commonly seen.

2.5.4 Evaluating with Alternative Orders

For each instance where alternative ground truth orders exist, the performance is computed by the best each predicted order can obtain against all the ground truth orders¹⁴, denoted by **multi-reference performance**, and the subset containing these instances is denoted as the **multi-reference subset**.¹⁵

¹⁴Jointly considered from all the evaluation metrics.

¹⁵The overall average number of ground truth references becomes 1.19, 1.23, 1.09 for multimodal, text-only, and image-only versions in WikiHow; and 1.10, 1.17, 1.14 in RecipeQA.

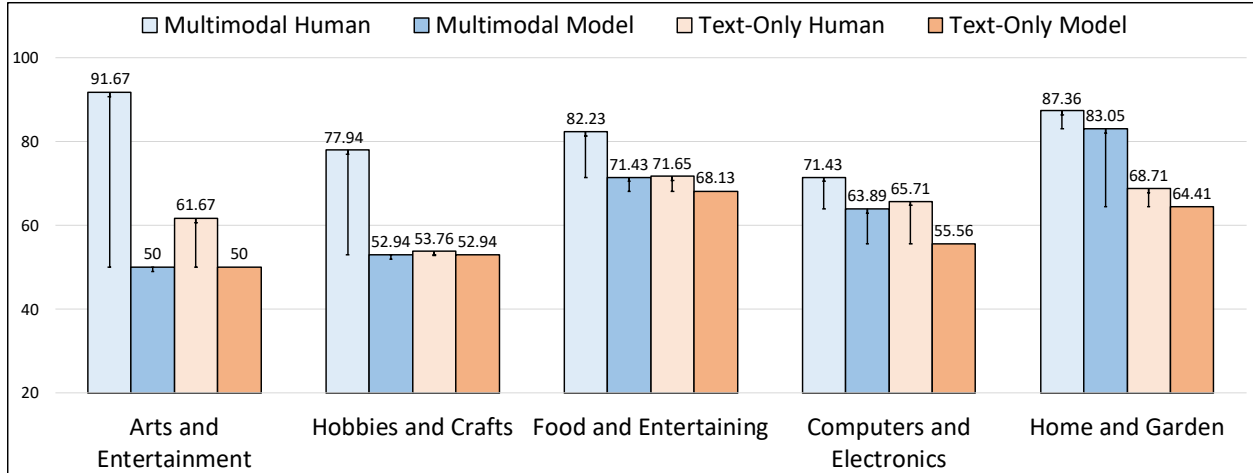


Figure 2.3: Top-3 and least-2 categories of human-model performance difference (in PMR): The selected categories have >10 samples. The difference bars on the multimodal model series are compared against the text-only model series.

Modality	Subset	WikiHow Golden-Test-Set (Size: 300)						RecipeQA Golden-Test-Set (Size: 100)					
		Acc \uparrow		PMR \uparrow		$L_r \uparrow$		Acc \uparrow		PMR \uparrow		$L_r \uparrow$	
		single	multi	single	multi	single	multi	single	multi	single	multi	single	multi
Text-Only	Single	77.30	—	61.75	—	3.98	—	79.32	—	60.23	—	3.90	—
	Multi.	67.35	80.00	40.82	59.18	3.35	3.86	60.00	75.00	33.33	58.33	3.17	3.92
	All	75.68	77.74	58.67	61.67	3.87	3.96	77.00	78.80	57.00	60.00	3.81	3.90
Text-Only	Single \dagger	85.57	—	71.41	—	4.24	—	90.27	—	80.41	—	4.47	—
	Multi. \dagger	72.03	85.51	43.84	71.38	3.46	4.14	79.00	87.00	65.00	80.00	3.95	4.40
	All \dagger	83.35	85.56	66.91	71.40	4.11	4.22	88.92	89.88	78.56	80.36	4.41	4.46
Multimodal	Single	81.68	—	69.90	—	4.15	—	83.71	—	69.07	—	4.12	—
	Multi.	70.98	78.82	47.05	61.22	3.59	3.90	46.67	60.00	33.33	33.33	3.67	3.78
	All	79.87	81.19	65.67	68.00	4.05	4.11	82.60	83.00	68.00	68.00	4.10	4.11
Multimodal	Single \dagger	92.86	—	83.67	—	4.56	—	91.88	—	82.61	—	4.52	—
	Multi. \dagger	82.09	92.22	59.80	83.33	3.99	4.54	100.00	100.00	100.00	100.00	5.00	5.00
	All \dagger	91.03	92.75	79.61	83.61	4.46	4.55	92.12	92.12	83.13	83.13	4.53	4.53

Multi.-subset-size (*text*, *multimodal*) are: (49, 51)/300 in WikiHow; (12, 3)/100 in RecipeQA.

Table 2.4: Multi-reference performance: (\dagger denotes human performance) Our golden-test-set can be decomposed into two subsets: **Single** where each instance in this subset only has one single originally authored ground truth, and **Multi.** where each instance features multiple ground truths from alternative orders. For the **Multi.** subset, two types of performance can be computed: **single** considers only the originally authored ground truth and **multi** computes the multi-reference performance. **All** denotes the entire test-set combining the results from **Single** and **Multi.** subsets. Results are reported on the two main competitors: multimodal and text-only using the best performing models from Table 2.2 in each modality. **% of instances benefit w. multi-reference** indicates that of what percentage of instances *in each multi-reference subset* humans and the models benefit (for each instance if its performance improves *in any of the metrics*) from alternative ground truth orders.

Modality	WikiHow (300)			RecipeQA (100)		
	Cnt	Min/Max	Avg/Std	Cnt	Min/Max	Avg/Std
Image-Only	24	2/4	2.1/1.4	13	2/3	2.1/0.3
Text-Only	49	2/6	2.4/0.9	12	2/6	2.4/1.1
Multimodal	51	2/4	2.1/0.5	3	2/6	4/1.6

Table 2.5: Multi-reference subset statistics: We report the count (cnt) of multi-reference instances in each dataset across the three modalities, and their basic statistics.

Statistics. Table 2.5 lists the essential statistics of the multi-reference subsets, including the counts of the multi-reference instance for each dataset and modality, as well as the per-instance statistics.

Multi-Reference Performance. The noticeable main competitors in Table 2.2 are multimodal and text-only models, and hence for conciseness, in Table 2.4 we mainly report the multi-reference version of their best performing variants with the selected metrics. Several trends still hold: (1) Multimodal models still outperform the text-only counterparts. (2) Human performance is still well above models’ even under multi-reference setups. Additionally, both humans and models perform significantly worse in the multi-reference subset when single (original) ground truth is enforced, implying the validity of our alternative order annotations.

We originally hypothesize that enforcing the original authored order to be the only ground truth would be unfair to the text-only models, as images can often better represent the detailed scene changes omitted by the texts, while in reality certain steps may not need to strictly follow the authored order. Judging from the number of instances that improve after evaluating with alternative orders, the text-only model indeed benefits more from the multi-reference setup. Examining the general trends in Table 2.4, one can conclude that the textual contents indeed possess certain levels of ambiguities where images can help to alleviate. However, as the performance gaps between multimodal and text-only models are still significant under the multi-reference settings, advantages of multimodality. Note that humans achieve perfect performance on the multi-reference subset in RecipeQA, though unlikely it may seem, it is mainly due to recipes tend to have rarer possible alternative orders.

WikiHow Categories. Table 2.6 lists the WikiHow categories with the most (top-5) an-

Categories	Mean Per-Instance Refs. (Cnt)		
	Multimodal	Text	Image
Home and Garden	2.00 (7)	2.14 (7)	2.00 (3)
Hobbies and Crafts	2.00 (5)	2.73 (11)	2.00 (2)
Food and Entertaining	2.20 (15)	2.22 (14)	2.17 (12)
Others	2.28 (7)	2.67 (5)	2.00 (4)
Personal Care and Style	2.33 (3)	2.00 (1)	2.00 (1)

Table 2.6: Top-5 mean alternative orders by categories: We list top-5 categories in WikiHow according to the number of average ground truth references in their multi-reference subset. We again only list the categories with total instance count >10 .

notated multi-reference ground truths. Note that the categories with more annotated alternative ground truths are also among the worse performance from both humans and models (refer to Figure 2.3).

2.6 Summary

In this work we present studies of language and multimodal models on procedure sequencing, leveraging popular online instructional manuals. Our experiments show that both multimodality and our proposed sequence-aware pretraining are helpful for multimodal sequencing, however, the results also highlight significant gaps below human performance ($\sim 15\%$ on PMR).

We provide insights as well as resources, such as the multi-reference annotations of the sequencing task, to spur future relevant research. We also anticipate that the alternative orders defined and annotated in our work can benefit more comprehensive task-procedure understanding. Future work such as predicting task steps which can be parallel or interchangeable, and understanding step dependencies can be explored.

CHAPTER 3

Learning Action Dependencies for Comprehending Task-Knowledge

3.1 Introduction

When performing complex tasks (*e.g. making a gourmet dish*), instructional manuals are often referred to as useful guidelines. To follow the instructed actions, it is crucial to understand the *preconditions*, *i.e.* prerequisites before taking a particular action, and the *postconditions*, *i.e.* the status supposed to be reached after performing the action. Knowledge of action-condition dependencies is prevalent and inferable in many instructional texts. For example, in Figure 3.1, before performing the action “*place onions*” in step 3, both *preconditions*: “*heat the pan*” (in step 2) and “*slice onions*” (in step 1) have to be successfully accomplished. Likewise, executing “*stir onions*” (in step 4), leads to its *postcondition*, “*caramelized*” (also in step 4).

For autonomous agents or assistant AI that aids humans to accomplish tasks, understanding the conditions provides a structured view of a task (Linden, 1994; Aeronautiques et al., 1998; Branavan et al., 2012b; Sharma and Kroemer, 2020) and helps the agent correctly judge whether to *proceed* to the next action and *evaluate* the action completions.

However, no prior work has systematically studied automatically extracting pre- and postconditions from prevalent data resources. To bridge this gap, we propose the *action condition inference task* on **real-world instructional manuals**, where a **dense dependency graph** is produced, as in Figure 3.1, to denote the pre- and postconditions of actions. Such a dependency graph provides a systematic task execution plan that agents can closely follow.

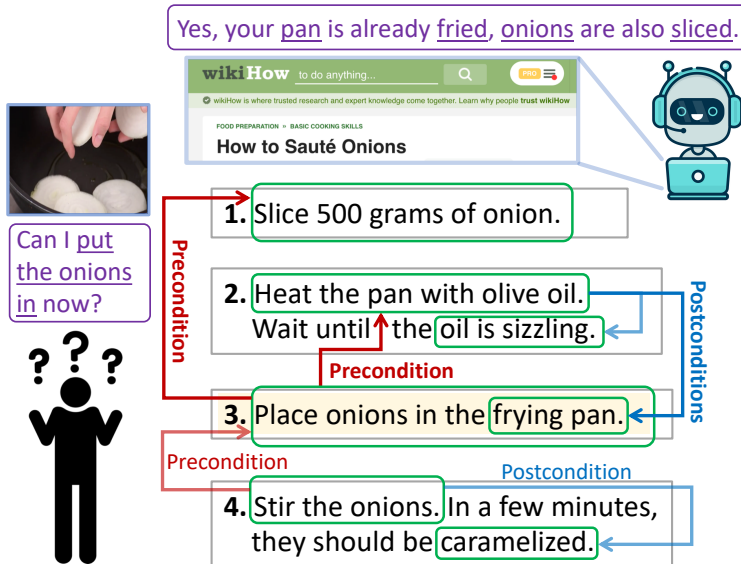


Figure 3.1: The Action Condition Inference Task: We propose a task that probes models’ ability to infer both *preconditions* and *postconditions* of an *action* from instructional manuals. It has wide applications to *e.g.* assistive AI and task-solving robots. *Original instructions are rephrased for simplicity in this illustration.

We consider two online instruction resources, *WikiHow* (Hadley et al.) and *Instructables.com* (ins), to study the current NLP models’ capabilities of performing the proposed task. As there is no densely annotated dataset on the desired action-condition-dependencies from real-world instructions, and annotating a comprehensive dependency structure of actions for long instruction contexts can be extremely expensive and laborious, we collect human annotations on a subset of totally 650 samples and benchmark models in either a **zero-shot** setting where no annotated data is used for training, or a **low-resource/shot** setting with limited amount of annotated training data.

We also design the following heuristics and show that they can effectively construct large-scale *weak supervisions*: (1) **Key entity tracing:** Key repetitive entity mentions (including **co-references**) across different instruction descriptions likely suggest a dependency. (2) **Keywords:** Certain keywords (*e.g.* the before in “do *X* before doing *Y*”) can often imply the condition dependencies. (3) **Temporal reasoning:** We adopt a temporal relation module (Han et al., 2021b) to alleviate the potential inconsistencies between the narrated orders of conditional events and their actual temporal orders to better utilize their temporally grounded nature (*e.g.* preconditions are *prior to* an action).

We benchmark two strong baselines based on pretrained language models with or without

instruction contexts on our annotated held-out test-set, where the models are asked to make predictions *exhaustively* on **every possible dependency**. We observe that contextualized information is essential ($> 20\%$ F1-score gain over non-contextualized counterparts), and that our proposed heuristics are able to augment an effective weakly-supervised training data to further improve the performance ($> 6\%$ F1-score gain) on the low-resource setting. However, the best results are still well below human performance ($> 20\%$ F1-score difference).

Our key contributions are three-fold: (1) We propose an action-condition inference task and create a densely human-annotated *evaluation dataset* to spur research on structural instruction comprehensions. (2) We design linguistic-centric heuristics utilizing entity tracing, keywords, and temporal reasoning to construct effective large-scale weak supervisions. (3) We benchmark models on the proposed task to shed lights on future research.

3.2 Background and Related Work

Two major lines of research work are closely related to the contribution in this chapter, understanding procedural texts (text pieces instructing a task procedure), and extracting event relations in contrived texts.

Procedural Text Understanding. Uncovering knowledge in texts that specifically features *procedural structure* has drawn many attentions, including aspects of tracking entity state changes (Branavan et al., 2012a; Bosselut et al., 2018; Mishra et al., 2018; Tandon et al., 2020), incorporating common sense or constraints (Tandon et al., 2018; Du et al., 2019), procedure-centric question answering (QA) (Tandon et al., 2019), and structural parsing or generations (Malmaud et al., 2014; Zellers et al., 2021a; ?). (Clark et al., 2018) leverages VerbNet (Schuler, 2005) with *if-then* constructed rules, one of the keywords we also utilize, to determine object-state postconditions for answering state-related reading comprehension questions. In addition, some prior works also specifically formulate precondition understanding as multiple choice QA for event triggers (verbs) (Kwon et al., 2020) and common sense phrases (Qasemi et al., 2021). We hope our work on inferring action-condition dependencies, an essential knowledge especially for understanding task-procedures, from long instruction texts, can help advancing the goal of more comprehensive procedural text understanding.

Drawing dependencies among procedure steps has been explored in (Dalvi et al., 2019; Sakaguchi et al., 2021b; Pal et al., 2021), however, their procedures are manually synthesized

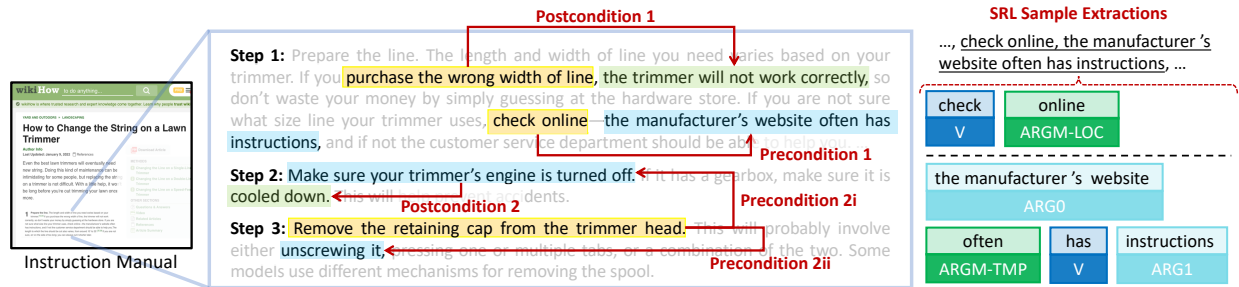


Figure 3.2: Terminologies: (Left) We show a few exemplar actionables (light yellow) with their associated preconditions (light blue) and postconditions (light green). Notice that an actionable can have multiple pre- or postconditions and they can span across different instruction steps. For simplicity we do not show an exhausted set of text segments of interests, *i.e.* in the actual dataset there might be more. **(Right)** we show one sample SRL extractions which correspond to one of the action-condition dependency linkages on the left.

short paragraphs. Our work, in contrast, aims at inferring diverse dependency knowledge directly from complex real-world and task-solving-oriented instructional manuals, enabling the condition dependencies to go beyond inter-step and narrative boundaries.

Event Relation Extraction. Our work is also inspired by document-level event relation extraction (Han et al., 2019, 2021a; Huang et al., 2021a; Ma et al., 2021). Specifically, certain works also adopt weak supervisions to learn event temporal relations (Zhou et al., 2020, 2021; Han et al., 2021b), while other relevant works aim at extracting causality relations (mainly cause-effect) automatically from texts (Cao et al., 2016; Altenberg, 1984; Stasaski et al., 2021). Our work combines multiple commonsensical heuristics tailored to the nature of the dependencies exhibited in actions and their conditions, in real-world instruction sources.

3.3 Terminologies and Problem Definition

Our goal is to learn to infer action-condition dependencies in real-world instructional manuals. We first describe essential terminologies in details:

Actionable refers to a phrase that a person can follow and execute *in the real world* (yellow colored phrases in Figure 3.2). We also consider negated actions (*e.g. do not ...*) or actions warned to avoid (*e.g. if you purchase the wrong...*) as they likely also carry useful knowledge regarding the tasks.¹

¹We ask workers to single out the actual *actionable* phrases, *e.g. purchase the wrong line* → *trimmer will not work*.

Precondition concerns the *prerequisites* to be met for an actionable to be executable, which can be a status, a condition, and/or another prior actionable (blue colored phrases in Figure 3.2). It is worth noting that humans can omit explicitly writing out certain condition statements because of their triviality as long as the actions inducing them are mentioned (*e.g.* heat the pan \rightarrow pan is heated, the latter can often be omitted). We thus generalize the conventional precondition formulation, *i.e.* sets of statements evaluated to true/false (Fikes and Nilsson, 1971), to a phrase that is either a passive condition statement or an *actionable that induces* the prerequisite conditions, as inspired by (Linden, 1994).

Postcondition is defined as the outcome caused by the execution of an actionable, which often involves status changes of certain objects (or the actor itself) or certain effects emerged to the surroundings or world state (green colored phrases in Figure 3.2).

Text segment in this paper refers to a textual segment of interest, which can be one of: {actionable, precondition, postcondition}, in an article.

In reality, a valid actionable should have both *pre-* and *postcondition* dependencies, however, we do not enforce this in this work as conditions can occasionally be omitted by human authors.

Problem Formulation. Given an input instructional manual and some text segments of interest extracted from it, a model is asked to predict the *directed* relation between a pair of segments, where the relation should be one of the followings: NULL (no relation), *precondition*, or *postcondition*.

3.4 Datasets and Human Annotations

As the condition-dependency knowledge we are interested in is prevalent in real-world instructions, we consider two popular online resources, **WikiHow** and **Instructables.com**, both consist of detailed multi-step task instructions, to support our investigation. For WikiHow, we use the provided dataset from (Wu et al., 2022); for Instructables, we scrape the contents directly from their website.

Since densely annotating large-scale instruction sources for the desired dependencies is extremely expensive and laborious, we mainly annotate a *test-set* and propose to train the models via weakly or self-supervised methods. We hence provide a small subset of the human-

annotated data to adapt models to the problem domain. To this end, we collect comprehensive human annotations on a selected subset in each dataset to serve as our **annotated-set**, and particularly the subsets used to evaluate the models as the **annotated-test-set**.² In total, our densely annotated-set has 500 samples in WikiHow and 150 samples in Instructables, spanning 7,191 distinct actions (defined by main predicate-object phrases) for diversity. In Chapter 5.5.2, we will describe how the annotated-set is split to facilitate the low-resource training. We also collect the human performance on the annotated-test-set to gauge the human upper bound of our proposed task.

3.4.1 Annotations and Task Specifications

Dataset Structure. The desired structure of the constructed data, as in Figure 3.2, features two main components: (1) **text segment** of interest (see Chapter 3.3), and (2) **condition linkage**, a *directed* and *relational* link connecting a pair of text segments.

Annotation Process. We conduct the annotated-set construction via Amazon Mechanical Turk (MTurk). Each worker is asked to carefully **read over thoroughly** a prompted complex multi-step instructional manual, where the annotation process consists of three main steps: **(1) Text segments highlighting:** To facilitate this step (and postulating the text segments for constructing weak-supervisions in Chapter 3.5), we *pre-highlight* several text segments extracted by *semantic role labelling* (SRL) for workers to choose from.³ They can also freely annotate (highlight by cursor) their more desirable segments. **(2) Linking:** We encourage the workers to annotate all the possible segments of interest, and then they are asked to connect certain pairs of segments that are likely to have dependencies with a directed edge. **(3) Labelling:** Finally, each directed edge drawn will need to be labelled as either a *pre-* or *postcondition* (NULL relations do not need to be explicitly annotated).

In general, for each article a worker is required to consider on average >500 pairwise relations with all associated article contexts (>300 tokens), which is a **decently laborious task**. Comparisons on the linkage annotations from different workers are as well made on *every* pair of *their respective annotated* text segments with the **actual candidate-**

²Following (Wu et al., 2022), we first choose from physical categories and then sample a manually inspected subset.

³SRL *V* and *ARGs* are connected alongside intermediate words to form contiguous segments.

consideration from the **entire** rest of article.

Since the agreements among workers on both text segments and condition linkages are sufficiently high⁴ given the complexity of the annotation task, our final human annotated-set retains the *majority voted* segments and linkages.

Variants of Tasks. Although proper machine extraction of the text segments of interest as a span-based prediction can be a valid and interesting task, we find that our automatic SRL extraction is already sufficiently reliable.⁵ In this paper, we thus mainly focus on the more essential linkage prediction (and their labels) task assuming that these text segments are given, and leave the possible end-to-end system with the (refined) text segment extraction, as the future work. Our proposed task and the associated annotated-set can be approached by a **zero-shot** or **low-resource** setting: the former involves no training on any of the annotated data and a heuristically constructed training set can be utilized (Chapter 3.5), while the latter allows models to be finetuned on a limited annotated-subset (Chapter 3.6.3). For the low-resource setting particularly, only 30% of the annotated data will be used for training.

3.5 Training With Weak Supervision

As mentioned in Chapter 3.4, our proposed task can be approached via a zero-shot setting, where the vast amount of **un-annotated instruction data** can be transformed into useful training resources (same dataset structure as described in Chapter 3.4.1). Moreover, it is proven that in many low-resource NLP tasks, constructing a much larger heuristic-based weakly supervised data can be beneficial (Plank and Agić, 2018; Nidhi et al., 2018).

3.5.1 Linking Heuristics

The goal of designing certain heuristics is to perform a rule-based determination of the linkage (its direction and the condition label). Our design intuition is to harness dependency

⁴The mean inter-annotator agreements (IAAs) per Fleiss Kappa for (segments, linkages) are (0.90, 0.57) and (0.88, 0.58) for WikiHow and Instructables. Note that the Kappa agreement measures the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly, so the agreement is high.

⁵~58% of the time SRL-proposed segments were directly used, with others mostly being few-word-span refinements.

knowledge by exploiting relations between actions and entities (*entity-level*), certain linguistic patterns (*phrase-level*), and *event-level* information, which should be widely applicable to all kinds of instructional data. Concretely, we design four types of heuristics: (1) **Keywords**: certain keywords are hypothesized to show strong implication of conditions such as *if, before, after*; (2) **Key entity tracing**: text segments that share the same key entities are likely indicating dependencies; (3) **Co-reference** resolution is adopted to supplement (2); (4) **Event temporal relation resolution** technique is incorporated to handle the inconsistencies between narrative order and the *actual* temporal order of the events.

SRL Extraction. Without access to human refinements (Chapter 3.4.1), we leverage SRL to postulate all the segments of interests to construct the weakly-supervised set. As SRL can detect multiple plausible ways to form the ARG frames with respect to the same *central* verb, we need to additionally determine the most desirable parses *for each action verb*. In this work, we simply select the most desirable SRL parses by choosing ones that maximize both: (1) the number of plausible segments (each centered around an action verb) *within a sentence*, where they do not overlap above a certain threshold (set to be 60% in this work), and (2) the number of ARGs in each of such segment.

Keywords:

Table 3.2 lists the major keywords that are considered in this work. Denote a text segment as a_i , keywords are utilized so as the text segments separated with respect to them, *i.e.* a_1 and a_2 , can be properly linked. Different keywords and their positions within sentences can lead to different *directions* of the linkages, *i.e.* $a_1 \rightleftarrows a_2$ (see second row of Table 3.1, note that here condition labels are not yet determined). For example, keywords before and after intuitively can lead to different directions if they are placed at non-beginning positions. We follow the rules listed in Table 3.2 to decide the directions.

Key Entity Tracing:

It is intuitive to assume that if the two text segments mention the same entity, a dependency between them likely exists, and hence a *trace* of the same mentioned entity can postulate potential linkages. As exemplified in the first row of Table 3.1, that *heating the pan* being a necessary precondition to *placing onions in the pan* can be inferred by the shared mention “pan”. We adopt two ways to propose the candidate entities: (1) We extract all the *noun*

standalone		
Heuristics	Examples	Descriptions
Entity-Tracing & Coref.		The shared entities are <u>pan</u> and <u>onions</u> (linked via co-references to <u>them</u>).
Keywords		Keywords are used to link the segments they separate. If the keyword is at the beginning (2nd example), the (1st) comma is used to segment the sentences.
Postcondition		Certain linguistic hints (e.g. SRL tags) are utilized to propose plausible (and likely) postcondition text segments.
Temporal		The action <u>prying</u> should occur prior to <u>stepping</u> , but these two segments are reversely narrated in the contexts.

Table 3.1: Heuristics used for determining condition linkages between text segments, with sample uses and descriptions.

phrases within the SRL segments (mostly ARG-tags), (2) Inspired by (Bosselut et al., 2018), a model is learned to predict potential entities involved that are not explicitly mentioned (e.g. *fry the chicken* may imply a *pan* is involved) in the context.

Co-References. Humans often use pronouns to refer to the same entity to alternate the mentions in articles, as exemplified by the mentions onions and them, in the first row of Table 3.1. Therefore, a straightforward augmentation to the aforementioned entity tracing is incorporating co-references of certain entities. We utilize a co-reference resolution model (Lee et al., 2018) to propose possible co-referred terms of extracted entities of each segment within the same step description (we do not consider cross-step co-references for simplicity).

3.5.2 Linking Algorithm

After applying the aforementioned linking heuristics, each text segment a_i , can have M linked segments: $\{a_1^{l_i}, \dots, a_M^{l_i}\}$. For linkages that are *traced* by entity mentions (and co-references), their directions always start from priorly narrated segments to the later ones, while linkages determined by the keywords follow Table 3.2 for deciding their directions. However, the

text segments that are narrated too much distant away from a_i are less likely to have direct dependencies. We therefore *truncate* the linked segments by ensuring any $a_j^{l_i}$ is narrated **no more than** “ S step” ahead of a_i , where S is empirically chosen to be 2 in this work.

Despite pruning the traces with the aforementioned design choice S can largely reduce *condition-irrelevant* segments, such heuristic indeed cannot guarantee the included text segments are always dependent with respect to an actionable. Our goal here is to exploit the generalization ability of language models to *recognize* segments that are most probable conditions by including as many heuristically proposed linkages as possible, where a better strategy on designing the maximum allowed step-wise distance is left as a future work.

Incorporating Temporal Relations:

As hinted in Chapter 3.3, the conditions with respect to an actionable imply their temporal relations. The direction of an entity-trace-induced linkage is naively determined by the narrated order of text segments within contexts, however, in some circumstances (*e.g.* fourth row in Table 3.1), the narrative order can be inconsistent with the actual temporal order of the events. To alleviate such inconsistency, we apply an event temporal relation prediction model (Han et al., 2021b) (trained on various temporal relation datasets such as *MATRES* (Ning et al., 2018)) to fix the linkage directions.⁶

We train the model on three different random seeds and make them produce a *consensus* prediction, *i.e.* unless all of the models jointly predict a specific relation (BEFORE or AFTER), otherwise the relation will be regarded as VAGUE. The model is then applied to predict temporal relations of each pair of event triggers (extracted by SRL, *i.e.* verbs/predicates), and then we invert the direction of an entity-trace-induced linkage, $a_j^{l_i} \rightarrow a_i$, if their predicted temporal relation is opposite to their narrated order (VAGUE is of course ignored).

Labelling The Linkages:

It is rather straightforward to label precondition linkages as a simple heuristic can be used: for a given segment, *any segments that linked to the current one that are either narrated or temporally prior to it* are plausible candidates for being preconditions. For determining postconditions, where they are mostly descriptions of status (changes), we therefore

⁶These do not include linkages decided by the keywords.

Keywords	Begin.	Within Sent.
<u>before</u> , <u>until</u> , <u>in order to</u> , <u>so</u>	$a_1 \longrightarrow a_2$	$a_1 \longleftarrow a_2$
<u>requires</u>	—	$a_1 \longrightarrow a_2$
<u>after</u> , <u>once</u> , <u>if</u>	$a_1 \longleftarrow a_2$	$a_1 \longrightarrow a_2$

Table 3.2: Keywords for deciding a potential linkage: If a keyword is at the beginning of a sentence, we use the (first) comma of that sentence to separate it to two segments and link them accordingly, while the keyword itself is used as the separator otherwise. The segments are then either refined with SRL or kept as they are if SRL does not detect a valid verb.

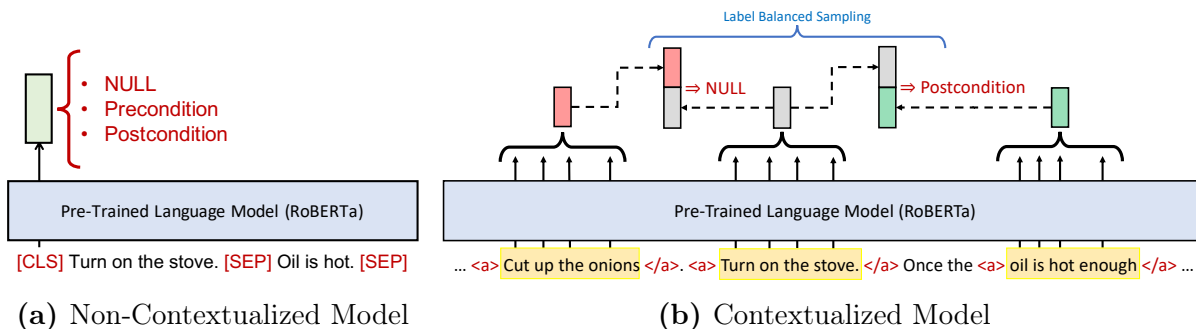


Figure 3.3: Model architectures: (a) **Non-contextualized pairwise model:** The model only considers a pair of given text segments. (b) **Contextualized model:** The model takes the whole instruction paragraphs (*i.e.* contexts) and wrap each text segment with our special tokens (<a>), where each segment representation is obtained by taking an average over its token representations. The *ordered* concatenated segment representations will then be fed into an MLP to make the final predictions.

make use of certain linguistic cues that likely indicate human written status, *e.g.* *the water will be frozen* and *the oil is sizzling*. Specifically, we consider: (1) *be-verbs* followed by present-progressive tenses if the subject **is an entity**, and (2) segments whose SRL tags start with ARGM as exemplified in Table 3.1.

3.6 Models

Our proposed heuristics do not assume specific model architecture to be applicable, and to benchmark the proposed task, we mainly consider two types of **base models**: (1) **Non-contextualized** model takes only the *two text segments* of interest at a time and make the *pairwise* trinary (directed) relation predictions, *i.e.* *NULL*, *precondition*, and *postcondition*; (2) **Contextualized** model also makes the relation predictions for every pair of input segments, but the inputs include the whole instruction article so the contexts are preserved. The two models are both based off pretrained language models (the non-contextualized model is

essentially a standard transformer-based language model finetuned for classification tasks), and the relation prediction modules are multi-layer perceptrons (MLPs) added on top of the language models’ outputs. Cross-entropy loss is used for training.

3.6.1 Non-Contextualized Model

The non-contextualized model takes two separately extracted text segments, a_i and a_j , as inputs and is trained similarly to the next sentence prediction in BERT (Devlin et al., 2019) (*i.e.* the order of the segments matters, which will be considered in determining their relations), as shown in Figure 3.3a.

3.6.2 Contextualized Model

The architecture of the contextualized model is as depicted in Figure 3.3b. Denote the tokens of the instruction text as $\{t_i\}$ and the tokens of i -th text segment of interest (either automatically extracted by SRL or annotated by humans) as $\{a_{ij}\}$. A special start and end of segment token, $\langle a \rangle$ and $\langle /a \rangle$, is wrapped around each text segment and hence the input tokens become: " $t_1, \dots, t_k, \langle a \rangle a_{i1}, a_{i2}, \dots, a_{iK} \langle /a \rangle, \dots$ ". The contextualized segment representation is then obtained by applying a mean pooling over the language model output representations of each of its tokens, *i.e.* denote the output representation of a_{ij} as $\mathbf{o}(a_{ij})$, the segment representation of $\mathbf{o}(a_i)$ is $AvgPool(\sum_{j=1}^K \mathbf{o}(a_{ij}))$. To determine the relation between segment i and j , we feed their *ordered* concatenated representation, $concat(\mathbf{o}(a_i), \mathbf{o}(a_j))$, to an MLP for the relation prediction.

3.6.3 Learning

Multi-Staged Training. For different variants of our task (Chapter 3.4.1), we can utilize different combinations of the heuristically constructed dataset and the annotated-train-set. For the low-resource setting, our models can thus be firstly trained on the constructed training set, and then finetuned on the annotated-set. Furthermore, following the **self-training** paradigm (Xie et al., 2020; Du et al., 2021), the previously obtained model predictions can be utilized to either *augment* (*i.e.* adding linkages) or *correct* (*i.e.* revising linkages) the original heuristically constructed data. And hence a second-stage finetuning can be conducted on this model-self-annotated data for improved performance.

Label Balancing. It is obvious that most of the relations between randomly sampled text

segment pairs will be NULL, and therefore the training labels are imbalanced. To alleviate this, we downsample the negative samples when training the models. Specifically, we fill each training mini-batch with equal amount of positive (relations are not NULL) and negative pairs, where the negatives are constructed by either *inverting* the positive pairs or *replacing* one of the segment with another randomly sampled *unrelated* segment within the same article.

3.7 Experiments and Analysis

Our experiments seek to answer these questions: (1) How well can the models and humans perform on the proposed task? (2) Is instructional context information useful? (3) Are the proposed heuristics and the second-stage self-training effective?

3.7.1 Training and Implementation Details

For both non-contextualized and contextualized models, we adopt the pretrained RoBERTa (-large) language model (Liu et al., 2019) as the base model. All the linguistic features, *i.e.* SRL (Shi and Lin, 2019), co-references, POS-tags, are extracted using models implemented by AllenNLP (Gardner et al., 2017). We truncate the input texts at maximum length of 500 while ensuring all the text segments within this length is preserved completely.

All the models in this work (*i.e.* both pretraining and finetuning) are trained on a single Nvidia A100 (40G RAM) GPU. The hyperparameters are manually tuned against different datasets, and the checkpoints used for testing are selected by the best performing ones on the held-out development sets.

3.7.2 Experimental Setups

Data Splits. The primary benchmark of WikiHow annotated-set is partitioned into **train (30%)**, **development (10%)**, and **test (60%)** set, respectively, resulting in 150, 50, and 300 data samples, for low-resource setting. We mainly consider the Instructables annotated-set in a **zero-shot setting** where we hypothesize the models trained on WikiHow can be well-transferred to it. For training conducted on the heuristically constructed data, including the second-stage self-training, we use respective held-out development sets to select the checkpoints around performance convergence for finetuning.

Evaluation Metrics. We ask the models to predict the relations on *every* pair of text

Model	Heus.	Finetuned/Self	WikiHow Annotated-Test-Set						Zero-Shot Transfer to Instructables					
			Precondition			Postcondition			Precondition			Postcondition		
			Prec.	Recall	F-1	Prec.	Recall	F-1	Prec.	Recall	F-1	Prec.	Recall	F-1
Prob. Random	—	N/N	3.55	4.42	3.54	0.61	0.86	0.68	2.94	3.88	3.04	0.46	0.46	0.42
Prompt. GPT-3	—	N/N	3.87	73.46	7.35	4.90	77.08	9.21	3.14	64.25	5.99	1.37	34.33	2.65
Adapt.-XPAD	—	Y/N	6.21	58.38	10.64	9.47	13.83	10.45	5.11	57.53	8.92	7.74	9.00	7.89
Non-Context.	Y	Y/N	8.21	79.52	14.32	15.43	44.99	20.56	6.49	65.05	11.31	13.64	43.50	18.65
	Y	Y/Y	8.56	81.19	14.91	26.53	65.95	34.31	6.64	67.13	11.54	24.53	61.93	31.78
	N	Y/N	34.01	58.33	39.27	34.44	43.15	36.79	26.93	53.43	32.92	32.16	41.39	34.42
	N	Y/Y	42.26	58.45	45.41	40.99	46.51	42.32	38.16	55.77	42.23	42.57	48.00	44.07
Context.	Y	N/N	10.69	34.79	15.05	10.34	11.88	10.49	10.34	16.17	11.42	4.52	4.15	4.15
	Y	Y/N	47.92	64.63	51.38	51.15	57.64	52.59	40.70	58.97	45.17	47.92	56.51	50.06
	Y	Y/Y	49.42	68.40	53.51	52.39	57.35	53.42	43.81	62.71	48.34	53.41	60.51	55.17
Human	—	—	83.91	83.86	83.55	77.39	84.81	78.81	84.74	81.32	82.78	71.90	82.51	75.53

Table 3.3: Annotated-test-set performance: The best performance is achieved by applying all of the proposed **heuristics (heus.)** and undergoing the two-stage training: **finetuned** on the annotated-train-set first and then perform the **self**-training. Note that for the Instructables, both *Finetuned* and *Self* are done on the WikiHow training sets and a **zero-shot** transfer is performed.

segments in a given instruction, and compute the average precision (Prec.), recall, and F-1 scores separately with respect to each (pre/post) condition labels.

Baselines. There is no immediate baseline we are aware of for the proposed action condition inference task. However, we note that (Dalvi et al., 2019)’s dependency graph prediction on scientific procedures (Mishra et al., 2018) shares high-level similarities to specifically our precondition inference task. Our non-contextualized model (without the second-stage self-training) with *only* the noun-phrase-based entity tracing heuristic resembles the KB-induced *prior dependency likelihood*, g_{kb} , in their proposed XPAD framework.⁷

Beside this **adapted-XPAD**, we also evaluate our task with (1) **probabilistic random-guess baseline** (random guesses proportional to the training-set label ratio), and (2) **zero-shot GPT-3** (Brown et al., 2020) where we prompt GPT-3 with exemplar data instances as the task definition (**contextualized**). These baselines help us to set up a benchmark and justify the challenges our task poses.

3.7.3 Experimental Results

Table 3.3 left half summarizes both the human and model performance on our standard split (30% train, 60% test) of WikiHow annotated-set. Contextualized model obviously

⁷With all entity-state-related components excluded (irrelevant to our task) and encoder replaced by RoBERTa model.

Heuristics.	WikiHow Annotated-Test-Set						Zero-Shot Transfer to Instructables					
	Precondition			Postcondition			Precondition			Postcondition		
	Prec.	Recall	F-1	Prec.	Recall	F-1	Prec.	Recall	F-1	Prec.	Recall	F-1
- temporal - coref. - keywords	45.60	61.22	48.59	43.71	47.56	44.35	39.35	57.03	43.49	38.45	42.96	39.39
- temporal - coref.	43.43	64.43	48.04	46.27	51.27	47.22	37.06	59.95	42.56	38.41	44.54	39.83
- temporal	45.83	62.48	49.17	47.72	52.70	48.81	39.39	59.53	44.23	46.81	52.15	48.23

Table 3.4: Heuristics ablations: The models used here are **contextualized** models without the second-stage self-training for both datasets, and "-" indicates exclusion (from using all). In general, each of the designed heuristics give incremental performance gain to both datasets, where the temporal component is particularly effective in postcondition predictions (compare to Table 3.3).

Train	Precondition			Postcondition		
	Prec.	Recall	F-1	Prec.	Recall	F-1
10%	41.34	61.71	46.06	45.24	55.56	47.95
20%	45.60	67.55	50.78	49.30	58.02	51.62
30%	57.38	64.46	57.53	50.49	54.57	51.09
40%	49.61	73.09	55.14	50.45	57.77	52.27
50%	54.27	70.89	57.84	51.35	55.85	52.23
60%	53.21	69.36	56.42	53.68	58.09	54.46

Table 3.5: Varying annotated-train-set size: on WikiHow (test-set size is fixed at 30%). We use the (best) model trained with all the proposed heuristics and the self-training paradigm.

outperforms the non-contextualized counterpart greatly, and all learned models perform well-above random baseline. Significant improvements on both pre- and postcondition inferences can be noticed when heuristically constructed data is utilized, especially when no second-stage self-training is involved. The best performance is achieved by **applying all the heuristics** we design, where further improvements are made by augmenting with second-stage pseudo supervisions. Similar performance trends can be observed in Table 3.3 right half where a zero-shot transfer from models trained on WikiHow data to Instructables is conducted.

Notice that the zero-shot GPT-3 performs quite poorly compared to our *best low-resource training setting*, and generally worse than our zero-shot contextualized model utilizing only the heuristically constructed data. We hypothetically attribute the poor performance to both the requirement of exhaustive search of the conditions across the whole manual, and its lacking of complex commonsense reasoning; justifying the effectiveness of our proposed training paradigm and the difficulty of our task. Nevertheless, there are still **large rooms** for improvement as the best model falls well-behind human performance (>20% F1-score gap).

Type	Example	Description
Heus. Overfit	<p>... use a sharp <u>blade</u> to cut ... Precondition ... look for a <u>blade</u> ... Precondition ✓</p>	Overfits on entity trace heuristic.
Lacking Causal Reason	<p>... body start leaning ... Precondition ... decrease pedal resistance ... Precondition ✓</p> <p>... can't completely dry ... Postcondition ... bacteria could form ... Postcondition ✓</p>	Knowledge-enhanced causal reasoning can be helpful.

Table 3.6: Exemplar model errors. The second row are from distant segments not link-able even via the keyword heuristic.

Heuristics Ablations. Table 3.4 features ablation studies on the designed heuristics. One can observe that keywords are mostly effective on inferring the postconditions, and co-references are significantly beneficial in the Instructables data, which can hypothetically be attributed to the writing style of the datasets (*i.e.* authors of Instructables might use co-referred terms more). Temporal relation resolution is consistently helpful across pre- and postconditions as well as datasets, suggesting only relying on narrated orders could degenerate the performance.

Error Analysis. While our (best) models perform well on linkages that exhibit similar concepts to the designed heuristics and generalize beyond their surface forms, we are interested in investigating under which situations they are more likely to err. We therefore sub-sample 10% of the annotated test-set for manual qualitative inspections and summarize our observations in Table 3.6. We find that our models can sometimes **overfit to certain heuristic** concepts as in Table 3.6 first row (within a food preparation context). Another improvement the models can enjoy is **better causal understanding**, which is currently not explicitly handled by our heuristics and can be an interesting future work (Table 3.6 second row, in a biking and cleaning contexts).

Humans, on the other hand, exhibit much superior performance than the models, tend to fail more often on two kinds of situations: (1) Missing preconditions (of an action) in those *much earlier paragraphs*, and (2) Sophisticated temporal ordering of the events (often not narrated sequentially in the texts). Especially, the first sentences of each task-step are often regarded as the starting actions, while in reality, they can be postconditions of the followed-

up detailed contexts. However, we think both aforementioned errors are rather remediable if the annotators are more careful and search more exhaustively for condition statements.

The Effect of Training Set Size:

Table 3.3 shows that with a little amount of data for training, our models can perform significantly better than the zero-shot setting. This arouses a question – how would the performance change with respect to the training set size, *i.e.* do we just need more data? To quantify the effect of training size on model performance, we conduct an experiment where we vary the sample size in the training set while fixing the development (10%) and test (30%) set for consistency consideration. We use the best settings in Table 3.3, *i.e.* with all the heuristics and self-training paradigm, for this study. We can observe, from Table 3.5, a plateau in performance when the training set size is approaching 60%, implying that simply keep adding more training samples does not necessarily yield significant improvements, and hypothesize that the discussed potential improvements are the keys to further effectively exploit the rich knowledge in large-scale instructional data.

3.8 Summary

In this work we propose a task on inferring action and (pre/post)condition dependencies on real-world online instructional manuals. We formulate the problem in both zero-shot and low-resource settings, where several heuristics are designed to construct an effective large-scale weakly supervised data. While the proposed heuristics and the two-staged training leads to significant performance improvements, the results still highlight significant gaps below human performance ($> 20\%$ F1-score).

We hope our studies and the collected resources can spur relevant research, and suggest two main future directions: (1) End-to-end propose (refined) actionables, conditions, and their dependencies, by fully exploiting our featured span-annotations of the text segments. (2) Inferred world states from the text descriptions as well as external knowledge of the entities and causal common sense can be factored into the heuristics for weak-supervisions.

CHAPTER 4

Tracing Active Objects Throughout Tasks with Symbolic World Knowledge

4.1 Introduction

Recent technological advancements in smart glasses (and headsets) from industry leaders such as Meta, Google, and Apple have attracted growing research in on-device AI that can provide *just-in-time* assistance to human wearers.¹ While giving (or receiving) instructions during task execution, the AI assistant should *co-observe* its wearer’s first-person (egocentric) viewpoint to comprehend the visual scenes and provide appropriate assistance. To accomplish this, it is crucial for AI to first be able to localize and track the objects that are undergoing significant state change according to the instruction and/or actions performed. For example in Figure 4.1, it can be inferred from the instruction that the object undergoing change which should be **actively grounded** and **tracked** is the *pawpaw*.

Existing works have focused on the visual modality alone for such state change object localization tasks, including recognizing hand-object interactions (Shan et al., 2020a) and object visual state changes (Alayrac et al., 2017). However, it remains under-explored whether the visual modality by itself is sufficient for providing signals to enable robust state change object localizing/tracking without enhanced signals from the textual modality. While utilizing a phrase grounding model (Liu et al., 2022a, 2023) is presumably a straightforward alternative, it leaves unanswered questions of which mentioned objects/entities in the instruction are supposedly the one(s) that undergo major state changes, *e.g.* , the *pawpaw* in Figure 4.1 instead of the *knife* is the correct target-object. Furthermore, how visual appearances of

¹Code at: <https://github.com/PlusLabNLP/ENVISION>

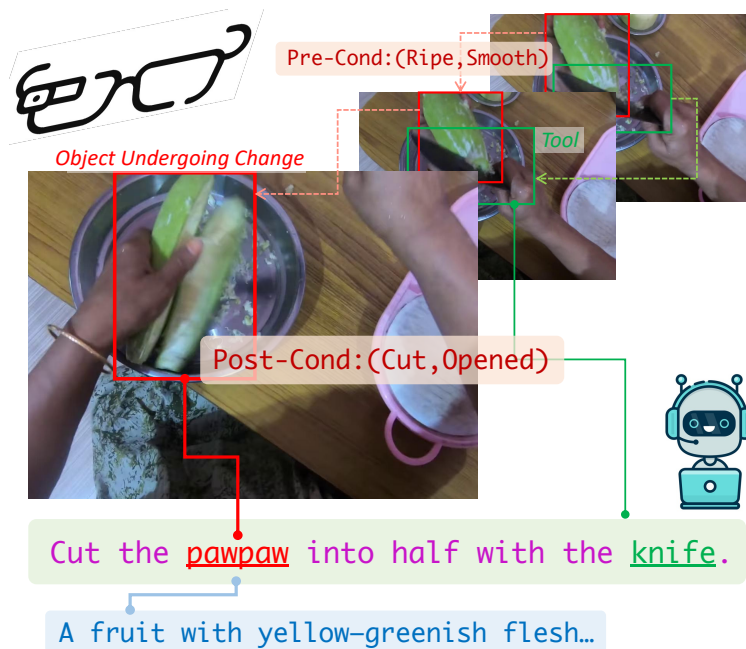


Figure 4.1: Active object grounding is the task of localizing the active objects undergoing state change (OUC). In this example action instruction "cut the pawpaw into half with the knife", the AI assistant is required to firstly infer the OUC (pawpaw) and the Tool (knife) from the instruction, and then localize them in the egocentric visual scenes throughout the action trajectories. Symbolic knowledge including pre/post conditions and object descriptions can bring additional information to facilitate the grounding.

the objects can help such multimodal grounding is yet to be further investigated.

In light of this, we tackle the active object grounding task by first extracting target object mentions from the instructions using large language models (ChatGPT (OpenAI, 2023a)) with a specifically designed prompting pipeline, and then finetuning an open-vocabulary detection model, GLIP (Li* et al., 2022), for visual grounding. We further hypothesize that additional action- and object-level symbolic knowledge could be helpful. As shown in Figure 4.1, state conditions *prior to* (*pre-conditions*: which indicate pre-action states) and *after* (*post-conditions*: which suggest at past state changes) the execution of the action are often considered when locating the objects, especially when the state changes are more visually significant. Furthermore, generic object knowledge including visual descriptions (*e.g.*, "yellow-greenish flesh"), are helpful for uncommon objects.² Based on this hypothesis, we prompt the LLM to obtain pre- and post-conditions on the extracted object mentions, along

²This should not contradict with the application where an assistive AI judges if the outcomes of the actions are desirable, as here we are only using general commonsensical conditions generated by an LLM, while in reality there can be more subtle and task-dependent conditions that need to be examined.

with a brief description focusing on specific object attributes.

To improve the grounding models by effectively using all the aforementioned action-object knowledge, we design an object-dependent mask to separately attribute the symbolic knowledge to its corresponding object mentions for training. During inference time, a pre-/post-condition dependent scoring mechanism is devised to aggregate the object and the corresponding knowledge logit scores to produce a joint inference prediction.

We evaluate our proposed framework on two narrated egocentric video datasets, Ego4D (Grauman et al., 2022a) and Epic-Kitchens (Damen et al., 2022) and demonstrate strong gains. Our main contributions are two folds: (1) We design a sophisticated prompting pipeline to extract useful symbolic knowledge for objects undergoing state change during an action from instructions. (2) We propose a joint inference framework with a per-object knowledge aggregation technique to effectively utilize the extracted knowledge for improving multimodal grounding models.

4.2 Background and Related Work

The contributions in this chapter are inspired heavily by three lines of research work, including egocentric visual perception, unveiling knowledge regarding action-object relations, and vision-and-language cross-modal grounding.

Egocentric Vision. Egocentric vision has recently attracted research attentions thanks to advancements in smart wearable devices and robotics. Datasets used in this work, Ego4D (Grauman et al., 2022a) and Epic-Kitchens (Damen et al., 2022, 2018; Dunnhofer et al., 2022) are two representative large-scale collections of egocentric videos recording tasks performed by the camera wearers. Other existing works have also investigated egocentric vision in audio-visual learning (Kazakos et al., 2019), object detection with EgoNet (Bertasius et al., 2017; Furnari et al., 2017), object segmentation with eye-gazes (Kirillov et al., 2023) and videos (Darkhalil et al., 2022).

Action-Object Knowledge. The knowledge of objects are often at the center of understanding human actions. Prior works in both NLP and vision communities, have studied problems such as tracking visual object state changes (Alayrac et al., 2017; Isola et al., 2015; Yang et al., 2022), understanding object manipulations and affordances (Shan et al., 2020a; Fang et al., 2018), tracking textual entity state changes (Branavan et al., 2012a;

Bosselut et al., 2018; Mishra et al., 2018; Tandon et al., 2020), and understanding textual pre-/post-conditions from action instructions (Wu et al., 2023a). While hand-object interactions (Shan et al., 2020a; Fu et al., 2022) are perhaps one of the most common object manipulation schemes, the objects undergoing change may not be directly in contact with the hands (see Figure 4.2). Here additional textual information can aid disambiguating the active object during localization and tracking. In this spirit, our work marries the merits from both modalities to tackle the active object grounding problem according to specific task instructions, and utilize action-object knowledge to further improve the models.

Multimodal Grounding. In this work, we adopt the GLIP model (Li* et al., 2022; Zhang* et al., 2022) for its compatibility with our settings and the joint inference framework, which indeed demonstrate significant improvements for the active object grounding task. There are many related works for multimodal grounding and/or leveraging language (LLMs) to help with vision tasks, including (but not limited to) Grounding-DINO (Liu et al., 2023), DQ-DETR (Liu et al., 2022a), ELEVATER (Li* et al., 2022), phrase segmentation (Zou* et al., 2022), visually-enhanced grounding (Yang et al., 2023), video-to-text grounding (Zhou et al., 2023), LLM-enhanced zero-shot novel object classification (Naeem et al., 2023), and multimodal object description generations (Li et al., 2022a, 2023b).

4.3 Tasks and Terminologies

4.3.1 Technical Challenges

The Active Object Grounding Task. For both robotics and assistant in virtual or augmented reality, the AI observes (or co-observes with the device wearer) the visual scene in the *first-person (egocentric)* point of view, while receiving (or giving) the task instructions of what actions to be performed next. To understand the context of the instructions as well as engage in assisting the task performer’s actions, it is crucial to closely follow the key objects/entities that are involved in the actions undergoing major state change.³ We term these actively involved objects as **objects undergoing change (OUC)**, and what facilitate such state change as **Tools**.

³State change can come from objects’ physical properties such as *composition, textures, and functionalities*; as well as attributes such as *sizes, shapes, and physical affordances*.

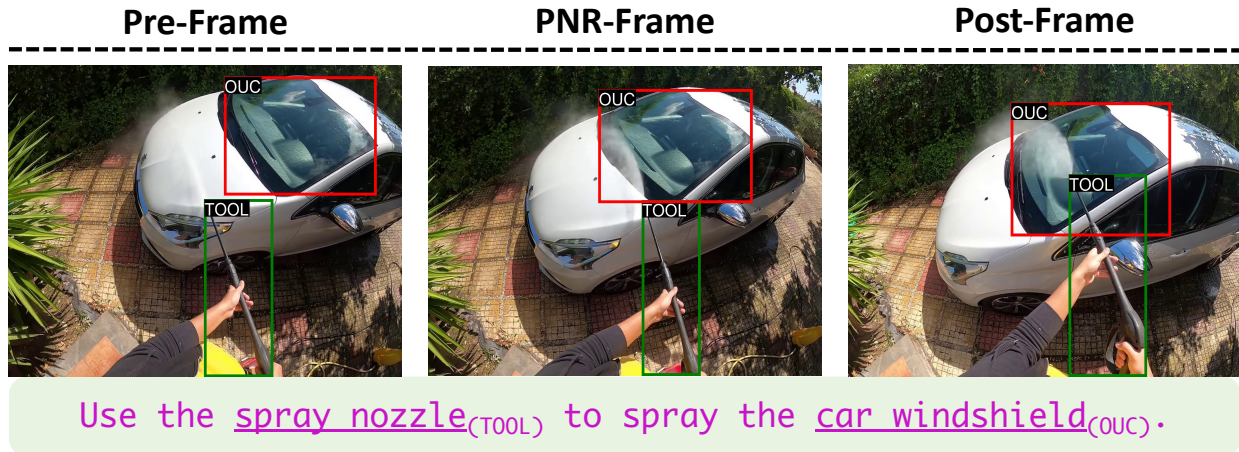


Figure 4.2: Ego4D SCOD active grounding: Example object undergo change (OUC) due to the instructed actions and associated Tools, spanning: the pre-condition, point-of-no-return (PNR) and post-condition frames.

Challenges. For a standard phrase grounding model, such as GLIP (Li* et al., 2022; Zhang* et al., 2022), it requires to know beforehand which entities and/or phrases in the texts to be grounded. Furthermore, localizing objects that undertake drastic visual changes, can often create out-of-distribution shifts in terms of recognizing the same objects (but under different states). This cause severe challenges to perform standard phrase grounding as it is hard to generalize to novel/unseen objects seamlessly. Our proposed framework will be essentially addressing such an issue, where the symbolic world knowledge extracted can provide more continuous grounding guidance. However, naively utilizing the symbolic knowledge could lead to noisy multimodal alignment, where the generic contrastive learning paradigm could fall short to capture subtle alignment between noisy textual descriptions with the visual (regions). Our proposed framework aims at coping with this challenge, by instructing the LLMs to produce concise and visually recognizable conditional symbolic knowledge to facilitate the underlying multimodal alignment with much reduced complexity while maintaining high diversity and principally decisive features.

4.3.2 Datasets

As there is not yet an existing resource that directly studies such active instruction grounding problem in real-world task-performing situations, we *re-purpose* two existing egocentric video datasets that can be seamlessly transformed into such a setting: **Ego4D** (Grauman et al., 2022a) and **Epic-Kitchens** (Damen et al., 2018). Both come with per-time-interval

annotated narrations transcribing the main actions occurred in the videos.⁴

Ego4D: SCOD. According to Ego4D’s definition, object state change can encapsulate both spatial and temporal aspects. There is a timestamp that the state change caused by certain actions start to occur, *i.e.*, the **point-of-no-return (PNR)**. Ego4D’s **state change object detection (SCOD)** subtask then defines, chronologically, three types of frames: the **pre-condition (Pre)**, the **PNR**, and the **post-condition (Post)** frames, during a performed action. Pre-frames capture the prior (visual) states where a particular action is allowed to take place, while post-frames depict the outcomes caused by the action, and hence also record the associated object state change. Each frame annotated with its corresponding frame-type is further annotated with bounding boxes of the OUC (and Tools, if applicable), that is required to be regressed by the models. Figure 4.2 shows an exemplar SCOD data point.

Our re-purposed active grounding task is thus as follows: *Given an instructed action and one of a Pre/PNR/Post-typed frames, localize (ground) both the OUC(s) and Tool(s) in the visuals.* While the official SCOD challenge only evaluates the PNR frame predictions, we consider all (Pre, PNR, and Post) frames for both training and inference.

Epic-Kitchens: TREK-150. TREK-150 object tracking challenge (Dunnhofer et al., 2022, 2021) enriches a subset of 150 videos from the Epic-Kitchens (Damen et al., 2018, 2022) dataset, with densely annotated per-frame bounding boxes for tracking a *target object*. Since the Epic-Kitchens also comprises egocentric videos capturing human performing (specifically kitchen) tasks, the target objects to track are exactly the OUCs per the terminology defined above. *Hence, given an instructed action, the model is required to ground and track the OUC in the egocentric visual scenes.*⁵

It is worth noting that some OUCs may occasionally go "in-and-out" of the egocentric point of view (PoV), resulting in partial occlusion and/or full occlusion frames where no ground truth annotations for the OUCs are provided. Such frames are excluded from the

⁴The narrations are paraphrased as imperative instructions.

⁵Unlike Ego4D SCOD task, TREK-150 does not contain any defined Pre/PNR/Post frames. Our proposed model is trained to perform joint inference and autonomously decide which of the pre- and post-conditions to weigh more based on the frame image and instructed action. And hence, in the TREK-150 task, frame-type information is not required.

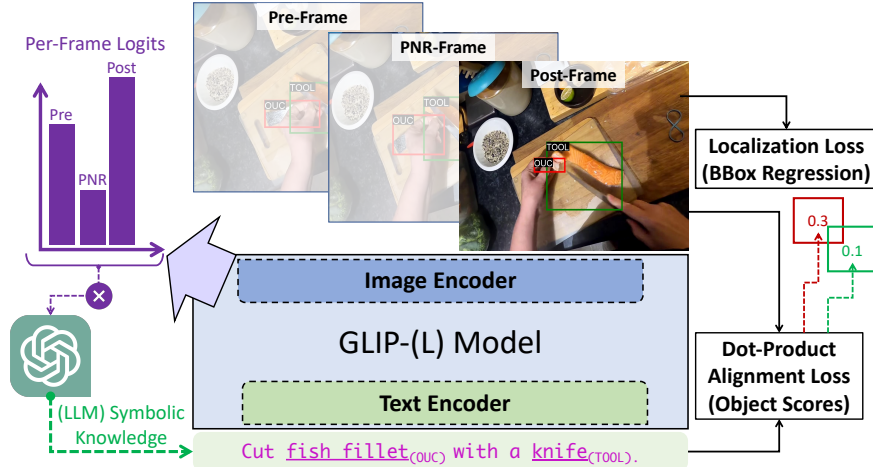


Figure 4.3: Overview of proposed framework that comprises a base multimodal phrase grounding model (GLIP), a frame-type predictor, a knowledge extractor leveraging LLMs (GPT), and predictions supervised by both bounding box regression of the objects and their ranked scores.

final evaluation. And in Chapter 4.5.2 we will show that our proposed model is very successful in predicting the objects when they come back due to the robustness of our symbolic joint inference grounding mechanism.

4.4 Method

Figure 4.3 overviews the proposed framework, consisted of: (1) A base **multimodal grounding architecture**, where we adopt a strong *open vocabulary object detection module*, GLIP (Li* et al., 2022). (2) A **frame-type prediction** sub-component which adds output projection layers on top of GLIP to utilize both image (frame) and text features to predict of which frame-type (Pre/PNR/Post) is currently observed. (Chapter 4.4.2) (3) A **prompting pipeline** that is engineered to extract useful action-object knowledge from an **LLM (GPT)**. (Chapter 4.4.2) (4) A **per-object knowledge aggregation** technique is applied to GLIP’s word-region alignment contrastive training. (Chapter 4.4.3)

4.4.1 Adapting GLIP

GLIP (Li* et al., 2022; Zhang* et al., 2022) achieves open vocabulary object detection by pretraining on a contrastive phrase grounding objective. Specifically, GLIP extends the text(caption)-to-image dot product matching objective from CLIP (Radford et al., 2021) to a **word-region**-level alignment objective. For some (tokenized) words of the textual

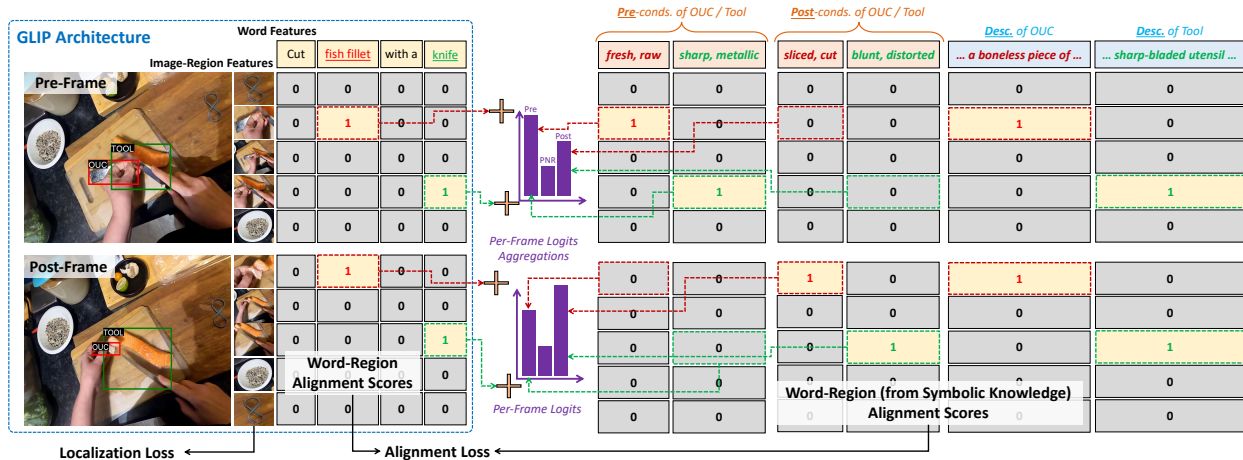


Figure 4.4: Model architecture (knowledge-enhanced grounding): On the left depicts the word-region alignment (contrastive) learning of the base GLIP architecture, where the model is trained to align the encoded latent word and image features with their dot-product logits being supervised by the positive and negative word-region pairs. On the right illustrates the enhanced object-knowledge grounding. During training we apply an object-type dependent mask to propagate the positive alignment supervisions; while during inference time the frame-type predictor (offline trained by the encoded textual and image features) acts as a combinator to fuse dot product-logit scores from both (extracted) object phrases and corresponding knowledge. (Note that for simplicity we do not fully split some phrases into individual words.)

description of an image, there are certain image region(s) that could be grounded to, while other regions are viewed as the negative samples for the CLIP-like alignment contrastive learning. During pretraining, GLIP utilizes both phrase grounding datasets (Ordonez et al., 2011; Plummer et al., 2015; Sharma et al., 2018) and object detection datasets (Krishna et al., 2017; Krasin et al., 2017; Shao et al., 2019).⁶

Contrastive Learning. We illustrate the GLIP training adapted to our task in Figure 4.4 left half. Notice that for simplicity we do not fully expand the tokenized word blocks, *e.g.*, "fish fillet" should span two words where each word ("fish" and "fillet") and its corresponding region is all regarded as the positive matching samples. The model is trained to align the encoded latent word and image features⁷ with their dot-product logits being supervised by the positive and negative word-region pairs. The alignment scores will then be used to score (and rank) the regressed bounding boxes produced by the image features, and each box will feature an object-class prediction score. Concretely, for j th regressed box, its grounding

⁶The descriptions of object detection are a simple concatenation of all the available object class labels.

⁷For details of GLIP's multimodal fusion technique, we refer the readers to Li* et al. (2022); Zhang* et al. (2022).

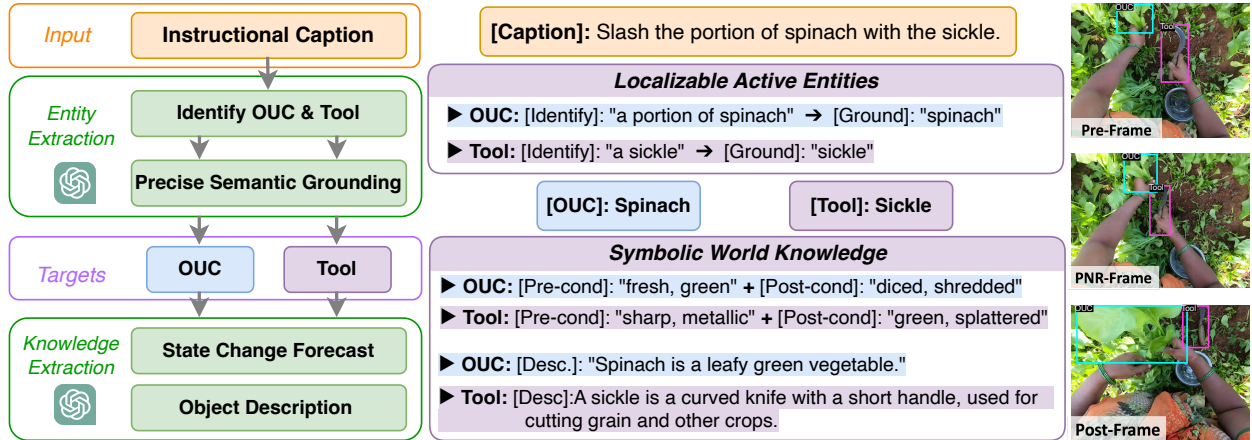


Figure 4.5: The GPT knowledge extraction pipeline. Demonstrated through an example from the Ego4D SCOD Dataset.

score to a phrase $W = \{w\}_{1:T}$ is a mean pooling of the dot-products between the j th region feature and all the word features that compose such a phrase: $\mathbf{S}_j^{box} = \frac{1}{T} \sum_i \mathbf{I}_j \cdot \mathbf{W}_i$. In this work, we mainly focus on the OUC and Tool object classes, *i.e.*, each textually-grounded region will further predict whether it is an OUC or Tool class.

4.4.2 LLM for Action-Object Knowledge

Pipeline. As illustrated in Figure 4.5, we implement an LLM query pipeline to extract active entities and relevant symbolic knowledge from an instructional caption. To account for GPT’s verbose tendency, we forcibly instruct GPT to produce the active objects (OUC and/or Tool) following a specifically designed format and then apply heuristic-based post-processing to further refine the extractions. Conditioned on the extracted OUC (and Tool), two additional queries are made to generate: (1) the symbolic pre- and post-conditions of such objects induced by the actions, and (2) brief descriptions characterizing the objects and their attributes. Interestingly, we empirically find it beneficial to situate GPT with a role, *e.g.*, "From the first-person view."

GPT Intrinsic Evaluation. In Table 4.2, we automatically evaluate the OUC/Tool extraction of GPT against the labelled ground truth entities in both datasets. We report both exact (string) match and word overlapping ratio (as GPT often extracts complete clauses of entities), to quantify the robustness of our GPT active entity extractions.

Table 4.3 reports human evaluation results of GPT symbolic knowledge, including pre-

Examples	Explanations
GPT: Pick up some <u>green papers</u> _(OUC) from the table. [No Tool] Desc.: "Green papers are consultation documents issued by government."	Without visual knowledge input, GPT is not robust to phrase ambiguity, leading to undesirable definition.
GPT: Cut the <u>fish fillet</u> _(OUC) with a <u>knife</u> _(TOOL) OUC: [Pre-cond]: "fresh, raw" [Post-cond]: "sliced, cut" Tool: [Pre-cond]: "Sharp, metallic" [Post-cond]: "Blunt, distorted"	LLMs may hallucinate exaggerated state changes, in this case claiming the knife to be "blunt, distorted" after a single use, which is unreasonable.
GPT: Hold the <u>iron</u> _(OUC) on the ironing board with your hand. [No Tool] Ego4D gt-label: [OUC]: "pants" [Tool]: "iron"	"Pants" is not mentioned in the narration, GPT fails to capture OUC due to text narration reporting bias.
GPT: Spin the <u>mop</u> _(OUC) in the mop bucket <u>spinner</u> _(Tool) . Ego4D gt-label: [OUC]: "mop" [Tool]: "mop"	GPT prediction is more reasonable compared to the Ego4D ground truth label.

Table 4.1: Qualitative Analysis of GPT Knowledge Extraction: Examples of cases where GPT-extracted symbolic knowledge are wrong or conflict with Ego4D annotations. Here the GPT-extracted or dataset-annotated knowledge are displayed in **GREEN** if they match human analysis and **RED** otherwise. Explanations for each example are provided on the right.

Object Type	Ego4D SCOD		TREK-150	
	EM (%)	Overlap.	EM (%)	Overlap.
OUC	77.8	88.6	76.0	94.3
Tool	60.3	88.5	—	—

Table 4.2: Automatic evaluation of GPT entity extraction. Abbreviations: **EM**: exact string matching; **Overlap**: The ratio of GPT extractions fully covering the ground truth phrases

Knowledge Type	Ego4D SCOD		TREK-150	
	Textual	Visual	Textual	Visual
Pre-Cond.	86.5	81.6	83.0	79.9
Post-Cond.	75.2	70.3	76.6	73.5
Desc.	98.9	91.4	99.2	95.3

Table 4.3: Human evaluation of GPT symbolic knowledge extraction. Abbreviations: **Textual**: i.e. "textual correctness" "Based on text alone, does the GPT conds./desc. make sense?"; **Visual**: i.e. "visual correctness": "Does the GPT conds./desc. match what is shown in the image?"

/post-conditions and descriptions. Evaluation is based on two binary metrics, namely: (1). Textual Correctness: "Based on text alone, does the knowledge make sense?" and (2). Visual Correctness: "Does the conds./desc. match the image?" Despite impressive performance on

both intrinsic evaluations, we qualitatively analyze in Table 4.1 some representative cases where GPT mismatches with annotations or humans, including cases where GPT’s answer is actually more reasonable than the annotations.

Incorporating Knowledge:

Adding Knowledge. We use the following schema to enrich the instruction with the obtained knowledge: "*{instr.}* [SEP] object/tool (pre/post)-state is *{conds.}* [SEP] object/tool description is *{desc.}*", where [SEP] is the separation special token; *{conds.}* and *{desc.}* are the pre-/post-condition and object definition knowledge to be filled-in. Empirically, we find diffusing the post-condition knowledge to PNR frame yield better results. As Figure 4.4 illustrates (omitting some prefixes for simplicity), we propagate the positive matching labels to object/tool’s corresponding knowledge. In the same training mini-batch, we encourage the contrastiveness to focus on more detailed visual appearance changes grounded to the symbolic condition statements and/or descriptions, by sampling frames from the same video clips with higher probability.

Frame-Type Prediction. Using both the encoded textual and image features, we learn an additional layer to predict the types of frames conditioned on the associated language instruction. Note that the frame-type definition proposed in Ego4D should be generalizable outside of the specific task, *i.e.*, these frame types could be defined on any kinds of action videos. In addition to the annotated frames in SCOD, we randomly sub-sample nearby frames within 0.2 seconds (roughly 5-6 frames) to expand the training data. The frame-type prediction achieves a 64.38% accuracy on our SCOD test-set, which is then directly applied to the TREK-150 task for deciding the amount of pre- and post-condition knowledge to use given the multimodal inputs.

4.4.3 Object-Centric Joint Inference

Masking. As illustrated in Figure 4.4, a straightforward way to assign symbolic knowledge to its corresponding object type respectively is to construct a **per-object-type mask**. For example, an OUC mask \mathbf{M}_{OUC} will have 1s spanning the positions of the words from condition (*e.g.*, "*fresh,raw*" of the OUC "*fish fillet*" in Figure 4.4) and descriptive knowledge, and 0s everywhere else. We omit the knowledge prefixes in Chapter 4.4.2 (*e.g.*, the phrase "*object state is*") so that the models can concentrate on grounding the meaningful words.

Such mask for each object type can be deterministically constructed to serve as additional word-region alignment supervisions, and can generalize to object types outside of OUC and Tool (beyond the scope of this work) as the GPT extraction can clearly indicate the object-to-knowledge correspondences. In other words, we enrich the GLIP’s phrase grounding to additionally consider symbolic knowledge during the contrastive training. Note that the mask is frame-type dependent, *e.g.* , \mathbf{M}_{OUC}^{Pre} and \mathbf{M}_{OUC}^{Post} will focus on their corresponding conditional knowledge.

Aggregation. During the inference time, we combine the frame-type prediction scores \mathbf{S}^{fr} with the per-object mask to aggregate the dot-product logit scores for ranking the regressed boxes. Specifically, we have $\mathbf{S}_{OUC}^{box} = \sum_{fr} \mathbf{S}^{fr} * \mathbf{M}_{OUC}^{fr}$, where \mathbf{S} is a 3-way logit and $fr \in \{\text{Pre, PNR, Post}\}$.

4.5 Experiments and Analysis

We adopt the **GLIP-L** variant (and its pretrained weights) for all of our experiments, where its visual encoder uses the Swin-L transformer (Liu et al., 2021). We train the GLIP-L with our framework primarily on the SCOD dataset, and perform a zero-shot transfer to the TREK-150 task.

4.5.1 Ego4D SCOD

Experimental Setups:

Data Splits. We split the official SCOD train set following a 90-10 train-validation ratio and use the official validation set as our primary test set.⁸

Evaluation Metrics. Following the original SCOD task’s main settings, we adopt average precision (AP) as the main evaluation metric, and utilize the COCO API (Lin et al., 2014) for metric computation. Specifically, we report AP, AP50, (AP at $\text{IOU} \geq 0.5$) and AP75 (AP at $\text{IOU} \geq 0.75$).

⁸The official test-set only concerns the PNR frame, and deliberately excluded narrations to make a vision only localization task, which is not exactly suitable for our framework.

Baselines:

We evaluate three categories of baselines: (1) **Pure object detection** models, where the language instructions are not utilized. (2) **(Pseudo) referential grounding**, where certain linguistic heuristics are used to propose the key OUCs. (3) **GPT** with **symbolic knowledge**, where GPT is used to extract both the OUCs and Tools, with additional symbolic knowledge available to utilize.

Pure Object Detection (OD). We finetune the state-of-the-art model of the SCOD task from [Chen et al. \(2022\)](#) (**VidIntern**) on all types of frames (Pre, PNR, and Post) to serve as the pure object detection model baseline, which learns to localize the OUC from a strong hand-object-interaction prior in the training distribution. We also train an OD version of GLIP providing a generic instruction, "*Find the object of change.*", to quantify its ability to fit the training distribution of plausible OUCs.

Pseudo Grounding (GT/SRL). We experiment four types of models utilizing the instructions and certain linguistic patterns as heuristics: (1) We extract all the nouns using Spacy NLP tool ([Honnibal and Montani, 2017](#)) and randomly assign OUC to one of which (**Random Entity**). (2) A simple yet strong baseline is to ground the full sentence of the instruction if the only object class to be predicted is the OUC type (**Full-Instr.**). (3) Following (2), we hypothesize that the first argument type (**ARG1**) of the semantic-role-labelling (SRL) parses ([Shi and Lin, 2019](#); [Gardner et al., 2017](#)) of most simple instructions is likely regarded as the OUC (**(SRL-ARG1)**). (4) Lastly, to quantify a possible upper bound of simple grounding methods, we utilize the annotated ground truth object class labels from SCOD task and perform a simple pattern matching to extract the OUCs and Tools. For those ground truth words are not easily matched, we adopt the **ARG1** method from (3) (**GT-SRL-ARG1**).

GPT-based. For our main methods leveraging LLMs (GPT) and its generated action-object symbolic knowledge, we consider four types of combinations: (1) **GPT** with its extracted OUCs and Tools. (2) The model from (1) with additional utilization of object definitions (**GPT+Desc.**). (3) Similar to (2) but condition on generated pre- and post-conditions of the objects (**GPT+Conds.**). (4) Combining both (2) and (3) (**GPT+Conds.+Desc.**).

Base	Type	Method	Objects	Pre-Frame↑			PNR-Frame↑			Post-Frame↑		
				AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
VidIntern	OD	—	OUC	32.73	49.17	34.05	37.49	57.04	38.59	29.68	44.43	30.94
			Tool	16.39	23.43	17.25	16.53	24.51	17.14	14.03	21.70	14.44
	OD	—	OUC	26.91	42.83	27.86	29.74	47.70	30.47	24.13	38.74	24.71
		Zero-Shot on GTs	OUC	20.18	32.97	20.63	19.51	32.34	19.39	19.34	31.07	19.88
		Random Entity	OUC	25.90	42.17	26.20	26.85	44.21	26.80	24.45	39.17	25.10
		Full-Instr.	OUC	32.45	51.62	33.34	33.78	54.44	34.49	31.30	49.23	32.42
		SRL-ARG1	OUC	36.41	54.93	37.65	38.32	58.07	39.41	33.59	49.99	34.90
		GT-SRL-ARG1	OUC	37.87	56.35	39.55	39.64	59.41	40.73	34.97	51.34	36.69
			Tool	45.53	71.22	46.27	43.70	68.96	44.54	43.76	69.56	44.04
GLIP-L	Instr.	GPT	OUC	37.46	56.05	38.96	39.07	59.17	40.13	34.77	51.34	36.35
			Tool	38.41	60.66	39.33	37.64	60.26	39.29	37.67	59.73	38.24
		GPT+Desc.	OUC	36.97	56.16	38.35	38.49	59.38	39.41	34.09	51.18	35.56
			Tool	42.26	64.37	44.59	41.30	64.46	43.53	40.20	63.92	41.60
		GPT+Conds.	OUC	38.65	57.55	40.16	40.19	60.39	41.56	35.40	52.15	37.11
			Tool	43.48	65.78	45.58	42.37	64.97	44.77	41.08	63.26	42.07
		w/o obj.-mask	OUC	37.59	56.28	39.19	39.09	59.31	40.58	33.93	50.80	35.38
		GPT+Conds.+Desc.	OUC	38.27	57.79	39.65	39.96	60.91	41.35	35.37	52.82	36.95
			Tool	44.00	66.49	46.12	42.77	66.06	44.82	42.12	65.44	42.45

Table 4.4: Model performance on Ego4D SCOD. OD: pure object detection. **Instr:** grounding with instructions. We highlight best OUC performance in **RED** for and best Tool performance in **GREEN**.

Results:

Table 4.4 summarizes the overall model performance on Ego4D SCOD task. Even using the ground truth phrases, GLIP’s zero-shot performance is significantly worse than pure OD baselines, implying that many of the SCOD objects are uncommon to its original training distribution. Generally, the instruction grounded performance (**Instr.**) are all better than the pure OD baselines, even with using the whole instruction sentence as the grounding phrase. The significant performance gaps between our models and the VidIntern baseline verifies that visual-only models can be much benefit from incorporation of textual information (should they be available) for the active object grounding task.

Particularly for OUC, with vanilla GPT extractions we can almost match the performance using the ground truth phrases, where the both the conditional and definition symbolic knowledge further improve the performance. Notice that condition knowledge by itself is more useful than the definition, and would perform better when combined. We also ablate a row excluding the per-object aggregation mechanism so that the conditional knowledge is simply utilized as a contextualized suffix for an instruction, which indeed performs worse, especially for the post-frames. As implied in Table 4.4 , best performance on Tool is achieved

using the ground truth phrases, leaving room for improvement on more accurate extractions and search of better suited symbolic knowledge.

Method	Top-K	Post-Frame↑		OD Metric
		AP	AP50	AP75
Track from GT PNR	1	20.36	41.15	17.78
Track from Pred. PNR	1	10.21	21.27	8.63
GPT+Conds.+Desc.	1	29.85	43.53	31.45

Table 4.5: PNR to Post OUC tracking ablation study. Since tracking module only produce a single box for each frame, we report the top-1 performance of our grounding model. (Normally COCO API reports max 100 detection boxes.)

However, one may raise a natural question: if the OUC/Tool can be more robustly localized in the PNR frame, would a tracker improve the post-frame performance over our grounding framework? We thus conduct an ablation study using the tracker in Chapter 4.5.2 to track from PNR-frames using either the ground truth box and our model grounded box to the post-frames. Results in Table 4.5 contradicts this hypothesis, where we find that, due to viewpoint variations and appearance differences induced by the state change, our grounding model is significantly more robust than using tracking.

Qualitative Inspections:

Figure 4.6 shows six different examples for in-depth qualitative inspections. It mainly shows that, generally, when the models grounding with the symbolic knowledge outperforms the ones without, the provided symbolic knowledge, especially the conditional knowledge, plays an important role

4.5.2 TREK-150

Experimental Setups:

Protocols. TREK-150’s official evaluation protocol is **One-Pass Evaluation (OPE)**, where the tracker is initialized with the ground-truth bounding box of the target in the first frame; and then ran on every subsequent frame until the end of the video. Tracking predictions and the ground-truth bounding boxes are compared based on IOU and distance of box centers. However, as the premise of having ground-truth bounding box initialization can be generally impractical, a variant of **OPE-Det** is additionally conducted, where the

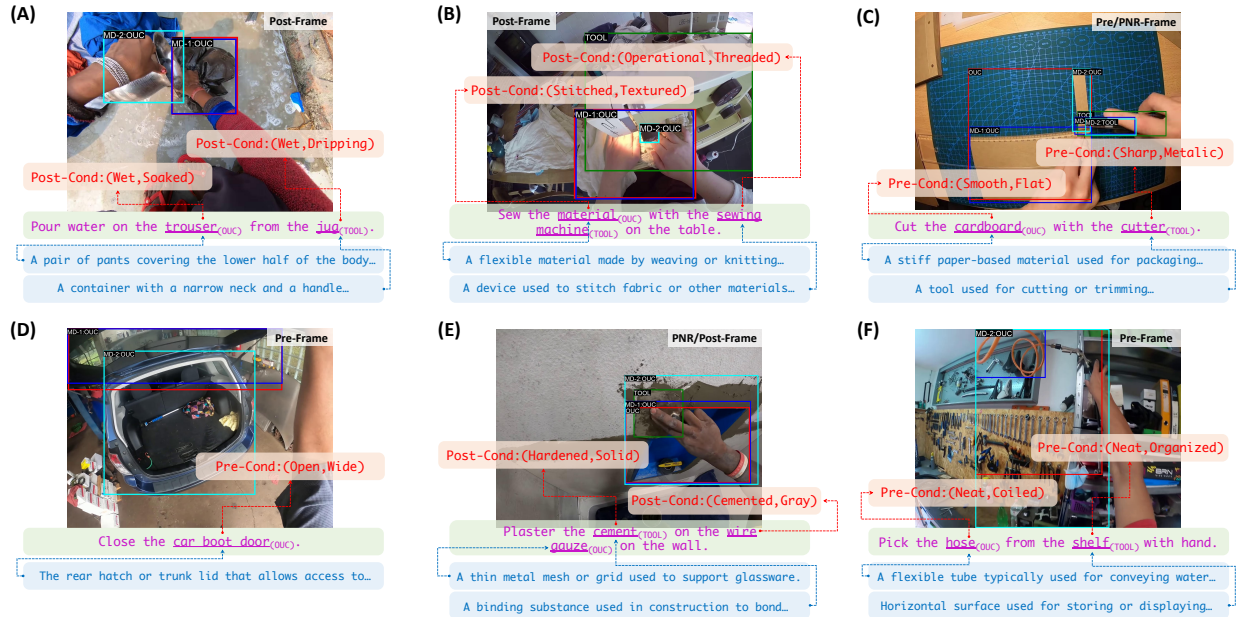


Figure 4.6: Qualitative inspections, mainly on the effectiveness of the GPT generated symbolic knowledge. Bounding box color code: **Ground truth boxes**, **Models with uses of symbolic knowledge (MD-1)**, *i.e.* the **GPT+Conds.+Desc.**; **Models without uses of symbolic knowledge (MD-2)**, *i.e.* , the vanilla **GPT**.

first-frame bounding box is directly predicted by our trained grounding model (grounded to the instructions).

Evaluation Metrics. Following Dunnhofer et al. (2022), we use common tracking metrics, *i.e.* , Success Score (SS) and Normalized Precision Score (NSP), as the primary evaluation metrics. In addition, we also report standard OD metric (APs) simply viewing each frame to be tracked as the localization task, as an alternative reference.

Baselines:

We adopt the best performing framework, the LTMU-H, in the original TREK-150 paper as the major baseline. LTMU-H integrates an fpv (first-person-view) object detector (HiC (Shan et al., 2020b)) into a generic object tracker (LTMU (Dai et al., 2020)), which is able to *re-focus* the tracker on the active objects after the (tracked) focus is lost (*i.e.* , identified by low tracker confidence scores).

Following the convention of utilizing object detection models to improve tracking (Feichtenhofer et al., 2017), we focus on improving object tracking performance by replacing the HiC-detector with our knowledge-enhanced GLIP models. We substitute the HiC-detector for all 8 GLIP-based models and the VideoIntern baseline trained on the SCOD task and

FPV OD	BBox Rank.	Method	OPE-Det \uparrow		OPE \uparrow		Std. OD Metric \uparrow		
			SS	NPS	SS	NPS	AP	AP50	AP75
HiC	MDNet	LTMU-H	0.267	0.261	0.505	0.520	—	—	—
HiC	MDNet	TbyD-H	0.047	0.018	0.433	0.455	—	—	—
Swin-L + DINO	VidIntern	—	0.341	0.340	0.526	0.541	29.49	41.47	30.85
GLIP-L	GLIP-L	Full-Instr.	0.355	0.361	0.521	0.537	38.51	60.06	40.17
		SRL-ARG1	0.373	0.377	0.528	0.544	40.00	60.52	40.96
		GT-SRL-ARG1	0.383	0.390	0.531	0.548	42.35	61.41	44.27
		GPT	0.379	0.389	0.529	0.545	41.85	61.89	43.46
		GPT+Desc.	0.402	0.409	0.528	0.543	41.40	60.70	43.17
		GPT+Conds.	0.412	0.422	0.541	0.557	45.90	67.26	47.94
		GPT+Desc.+Conds.	0.413	0.424	0.539	0.557	43.49	64.34	45.43

Table 4.6: Model performance on TREK-150. OPE denotes One-Pass Evaluation [Dunnhofer et al. \(2022\)](#) and OPE-Det is a variant to OPE where each tracker is initialized with its corresponding object detector prediction on the first frame. Success Score (SS) and Normalized Precision Score (NPS) are standard tracking metrics.

perform a zero-shot knowledge transfer (directly from Ego4D SCOD).⁹

Results:

Table 4.6 summarizes the performance on TREK-150. Our best GLIP model trained using GPT-extracted objects and symbolic knowledge outperforms the best HiC baseline by over 54% relative gains in the SS metric and over 62% relative gains in the NPS scores for the OPE-D task. It also outperforms the VideoIntern baseline by 16-18% relative gains in SS/NPS and even the GLIP-(GT-SRL-ARG1) model by 7-9% relative gains on both metrics. This demonstrates the transferability of our OUC grounding model in fpv-tracking. For the OPE task with ground-truth initializations, the gains provided by our GLIP-GPT models over LTMU-H narrow to 7-8% relative gains across both metrics while still maintaining a lead over all other methods. This shows that the model is still able to better help the tracker re-focus on the OUCs although the overall tracking performance is more empirically bounded by the tracking module.

4.6 Summary

In this work, we approach the active object grounding task leveraging two narrated egocentric video datasets, Ego4D and Epic-Kitchens. We propose a carefully designed prompting

⁹Mainly because: (1) The general bounding box annotations in Epic-Kitchens videos are *machine annotated*, and (2) we believe model learned from Ego4D’s more general visual domains should transfer well to kitchen activities.

scheme to obtain useful action-object knowledge from LLMs (GPT), with specific focuses on object pre-/post-conditions during an action and its attributional descriptions. Enriching the GLIP model with the aforementioned knowledge as well as the proposed per-object knowledge aggregation technique, our models outperforms various strong baselines in both active object localization and tracking tasks.

Part II

New Challenges for Multimodal Assistive AI

In the previous three chapters: we discuss the building blocks that are essential for a multi-modal assistive AI to be able to give effective instructions to humans and guide them towards the completion of the tasks. In the next two chapters, we will introduce two datasets that are particularly constructed to evaluate models' capabilities of utilizing multimodal and action information to reason over counterfactuality and converse back and forth to help solving users' problems.

In Chapter 5, we collect a counterfactual video question answering (video QA) dataset that aims at examining the models' commonsense reasoning capabilities to infer outcomes if certain realities are altered. This is crucial since humans perform such counterfactual judgements frequently during performing a task. For example, questions such as "*What if I do not have the spatula to stir the onions?*" or "*If I had stopped the engine earlier, would I be able to remove the trimmer's cap more easily?*", are commonly encountered as humans tend to infer alternative solutions or retrospectively search for better solutions, for accomplishing a desired task. We benchmark several strong video-language models on our collected dataset, and highlight important future endeavours to make for relevant research directions.

Chapter 6 introduces a novel and interesting dataset that simulates the conversational interactions between a user and an assistant AI within a (virtual) shopping environment. The virtual AI assistant can co-observe the users' egocentric viewpoints while conversing with the users to provide assistance such as recommending products, giving navigational guidance, and comparing items. The dataset is collected smartly in a two-phased pipeline where an algorithmically planned dialogue set is generated with dense dialogue attribute annotations, followed by human paraphrasing for more natural utterances. We propose four subtasks along the curation of the dataset that emphasizes specifically on multimodal grounding, tracing entities throughout the conversations, and action-condition dependencies mining.

CHAPTER 5

ACQUIRED: A Dataset for Answering Counterfactual Questions In Real-Life Videos

5.1 Introduction

Multimodal counterfactual reasoning refers to the ability to imagine and reason about what might have happened if certain conditions were different from what actually occurred based on vision and language inputs. It involves mentally simulating alternative scenarios and evaluating their potential outcomes. This cognitive process plays a crucial role in human intelligence, as it allows us to understand causality, make predictions, and learn from past experiences. For AI models, developing the capacity for counterfactual reasoning is a significant area of research and a challenging task. By enabling AI models to engage in counterfactual reasoning, we can enhance their understanding of causal relationships and their ability to assess the impact of interventions or changes in conditions.

However, despite the significance of counterfactual reasoning, it remains a relatively unexplored area of research. To assess the overall reasoning capabilities of models, several visual question answering datasets have been proposed on both images (Antol et al., 2015; Johnson et al., 2017) and videos (Yi et al., 2020; Xu et al., 2021). These datasets require reasoning skills such as commonsense reasoning, extracting human/object-to-object relations, and inferring physical properties.

One specific dataset in the realm of counterfactual reasoning is CLEVRER (Yi et al., 2020), which generates synthetic videos and associated questions in a controlled environment, featuring simulated object motion and rendered video frames. This dataset allows for evaluating models using descriptive, explanatory, predictive, and counterfactual ques-

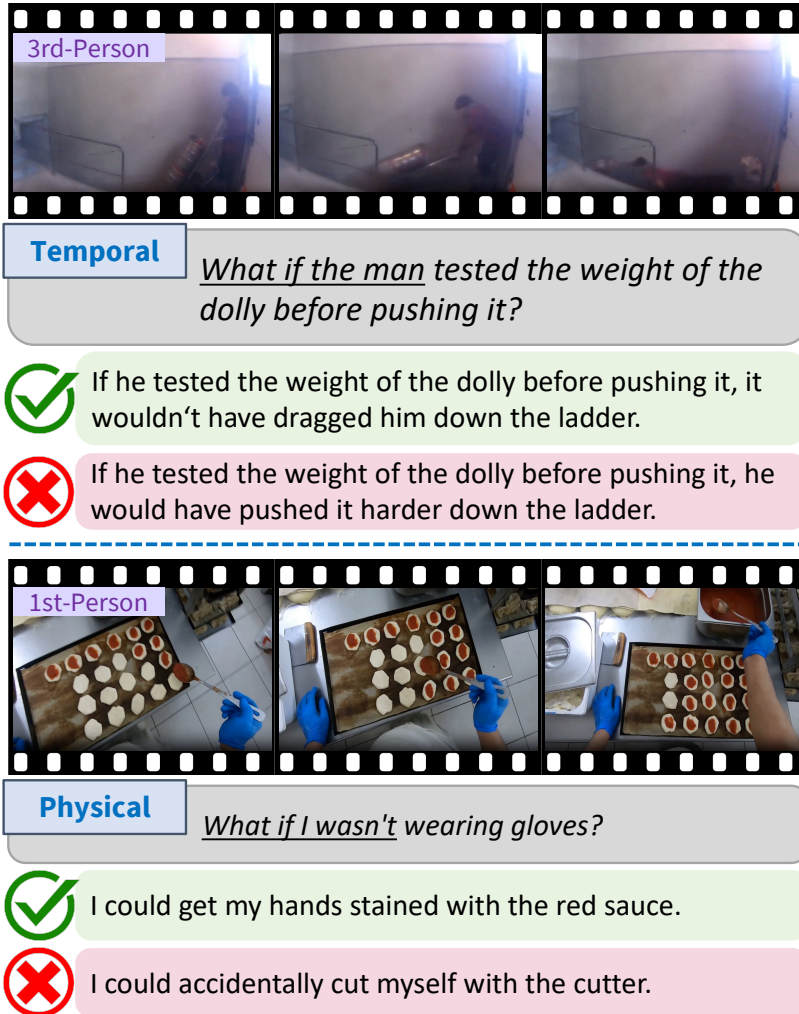


Figure 5.1: The **ACQUIRED** dataset is a video question answering (QA) dataset that specifically focuses on *counterfactual reasoning* on diverse real-world events. Our dataset concerns three types of commonsense reasoning dimensions: physical, social, and temporal, and encompasses videos from both third-person (upper) and first-person (lower) viewpoints. Each question is curated with a correct and a distractor answer. Each answer is by itself individually judgeable, and hence our dataset can be approached in either binary True/False or multiple-choice setting.

tions, covering a wide range of reasoning scenarios. However, the data generation process in CLEVRER is overly synthetic, limiting its effectiveness to assess models’ counterfactual reasoning abilities in realistic contexts. To address this limitation, TrafficQA (Xu et al., 2021) focuses on real-world traffic event cognition and reasoning in videos, specifically targeting scenarios like traffic accidents. It leverages crowdsourcing to gather diverse types of questions, including fundamental comprehension, counterfactual inference, and event forecasting. Nevertheless, because TrafficQA concentrates solely on traffic events, it fails to encompass

other real-life events, resulting in a substantial domain gap between TrafficQA and general video datasets such as Kinetics (Kay et al., 2017; Smaira et al., 2020) and YouTube (Abu-El-Haija et al., 2016; Zellers et al., 2022).

In this paper, we construct a benchmark that can evaluate the counterfactual reasoning abilities of visual models on various kinds of real-world events. We introduce **ACQUIRED**¹ that covers multiple dimensions of counterfactual reasoning and includes videos of both egocentric and exocentric views. Specifically, based on videos in both Oops (Epstein et al., 2020) and Ego4D (Grauman et al., 2022d), we crowd-source 11K questions over 3.7K videos targeting physical, temporal, and social counterfactual reasoning. Both the Oops and Ego4D datasets consist of human activities and interactions in numerous settings, making them ideal sources for curating video question answering datasets. In addition, many videos contain unintentional human actions (*e.g.*, the person accidentally falling down the ladder in Figure 5.1), which naturally enables people to come up with diverse *what-if* questions.

Inspired by Singh et al. (2021), we adopt a similar methodology for gathering counterfactual questions. Each question consists of a pair of answers, with one being the correct response and the other serving as a distractor. Importantly, the distractor answer represents a *minimal contrastive counterpart* to the correct answer. As we can see from examples in Figure 5.1, the design of using complementary pairs requires the model to understand the subtle differences between different options, which ensures that the model exhibits an intuitive grasp of counterfactual reasoning. In addition, having one distractor for each question allows for testing models in either True/False or multiple-choice setting.

We extensively evaluate numerous strong language models such as GPT-4, as well as state-of-the-art video-language models such as VALOR on our **ACQUIRED** dataset. The experimental results suggest that models struggle to effectively utilize the video contexts and perform counterfactual reasoning, with multimodal models achieving only comparable and sometimes inferior performance than language-only models. Moreover, the significant gap between the human and model (>13%) performance highlights the challenging nature of our task and room for improvements in visual counterfactual reasoning.

¹Abbreviation of: **A**nswering **C**ounterfactual **Q**uestions **I**n **R**eal-Life **V**ideos

Dataset	Visual Source	Question Source	Reasoning Domain			Counterfactual
			Physical	Temporal	Social	
<i>Image QA datasets</i>						
VQA Antol et al. (2015)	Diverse Real-world Event	Human	✓	✗	✗	✗
CLEVR Johnson et al. (2017)	Synthetic Object	Automatic	✓	✗	✗	✗
GQA Hudson and Manning (2019)	Diverse Real-world Event	Automatic	✓	✗	✗	✗
VCR Zellers et al. (2019)	Movie	Human	✓	✗	✓	✗
<i>Video QA datasets</i>						
CLEVRER Yi et al. (2020)	Synthetic Object Collision	Automatic	✓	✓	✗	✓
VLEP Lei et al. (2020)	TV & YouTube	Human	✓	✗	✗	✗
MovieQA Tapaswi et al. (2016)	Movie	Human	✓	✓	✓	✗
MSRVTT-QA Xu et al. (2017)	Diverse Real-world Event	Automatic	✓	✗	✗	✗
TGIF-QA Jang et al. (2017)	Tumblr GIF	Automatic & Human	✓	✓	✗	✗
MarioQA Mun et al. (2017)	Gameplay Video	Automatic	✓	✓	✗	✗
TVQA Lei et al. (2018)	TV	Human	✓	✓	✗	✗
Social-IQ Zadeh et al. (2019)	YouTube	Human	✗	✗	✓	✗
TrafficQA Xu et al. (2021)	Traffic Event	Human	✓	✓	✗	✓
NExT-QA Xiao et al. (2021)	Diverse Real-world Event	Human	✓	✓	✗	✗
Causal-VidQA Li et al. (2022b)	Diverse Real-world Event	Human	✓	✗	✗	✓
ACQUIRED	Diverse Real-world Event	Human	✓	✓	✓	✓

Table 5.1: Comparisons of different visual question answering datasets. ACQUIRED is the first to feature all the dimensions.

5.2 Background and Related Work

We will overview three lines of relevant research to this chapter: visual question answering, visual understanding models, and counterfactual reasoning.

Visual Question Answering Datasets. In Table 5.1, we list several representative visual QA datasets as well as their key features. The Visual Question Answering (VQA) dataset (Antol et al., 2015) is one of the pioneering works in this direction and has been a standard benchmark for evaluating the reasoning ability of image-language models (Goyal et al., 2017). Follow-up datasets such as CLEVR (Johnson et al., 2017) and GQA (Hudson and Manning, 2019) automatically construct compositional questions over real or synthetic images and perform the evaluation in a systematic way. To further evaluate the commonsense reasoning ability of models, VCR (Zellers et al., 2019) crowd-sources commonsense question-answer pairs associated with rationales over static images extracted from movies. Video question answering is more challenging than image question answering and is gaining increasing attention from the research community, leading to several video QA datasets being constructed (Lei et al., 2020; Tapaswi et al., 2016; Xu et al., 2017; Jang et al., 2017; Mun et al., 2017; Lei et al., 2018). Among them, CLEVRER (Yi et al., 2020) improves upon CLEVR and uses programmatically generated videos capturing collisions of synthetic objects

to evaluate the model reasoning abilities along multiple dimensions. Social-IQ (Zadeh et al., 2019) and TrafficQA (Xu et al., 2021) employ videos depicting real-world events, wherein Social-IQ primarily emphasizes human social interactions, while TrafficQA focuses on traffic events and accidents. To improve the diversity of the captured events, NExT-QA and Causal-VidQA collect videos from diverse domains and have human-annotated questions targeting different dimensions of reasoning.

As can be seen in Table 5.1, among all the visual QA datasets, there are only a few that attempt to evaluate the counterfactual reasoning abilities of models. In addition, the existing benchmarks are often limited in terms of the video sources and the question types, making it difficult to evaluate the model performance in a diverse real-world setting. ACQUIRED is the first dataset that can comprehensively evaluate the model counterfactual reasoning abilities spanning three distinct dimensions (*i.e.*, physical, social, and temporal) and cover videos that include a wide range of event types and from different viewpoints.

Visual Understanding Models. The creation of visual QA benchmarks allows for the development of visual understanding models. Many of the previous works have tried to solve these tasks using compositional approaches and scene graphs (Santoro et al., 2017; Hu et al., 2017; Hudson and Manning, 2018; Perez et al., 2018; Yi et al., 2018; Shi et al., 2019; Gao et al., 2020; Ding et al., 2021). For example, Hu et al. (2017) propose to train a modular network in an end-to-end manner to achieve both effectiveness and interpretability; Hudson and Manning (2018) utilize scene graphs and perform differentiable neural operations on the graphs to perform visual reasoning. Inspired by the success in pretraining on Internet-scale data (Devlin et al., 2019), pretraining models on large vision and vision-language tasks and then finetuning them on specific downstream tasks has become a standard in tackling visual understanding tasks (Sun et al., 2019; Li et al., 2020a; Zhu and Yang, 2020; Lei et al., 2021; Zellers et al., 2021b; Fu et al., 2021; Zellers et al., 2022; Wu et al., 2021, 2023b; Zhou et al., 2023). Existing works in this direction generally train models on large vision-language datasets with objectives such as masked language modeling and video-text matching. Despite the great progress in this direction, it is unclear if these models can perform counterfactual reasoning. To address this, we benchmark ACQUIRED against state-of-the-art models and systematically study their performance.

Causal and Counterfactual Reasoning. Humans can infer how an event would have

unfolded differently without experiencing this alternative reality and it has been a long-standing research topic in cognitive psychology (Van Hoeck et al., 2015). To empower such an important ability to artificial intelligence, researchers have tried to build learning models that can infer causal relations and perform reasoning in various fields (Qin et al., 2019; Yi et al., 2020; Baradel et al., 2020; Abbasnejad et al., 2020; Yue et al., 2021; Wang et al., 2021). Our constructed benchmark provides a valuable resource for developing and evaluating visual models with counterfactual reasoning abilities.

5.3 The ACQUIRED Dataset

5.3.1 Dataset Design & Collection

Problem Definition. As illustrated in Figure 5.1 and Table 5.2, each data point in ACQUIRED consists of a video and corresponding annotated question and answer pairs. We are inspired by prior works (Clark et al., 2019; Singh et al., 2021) to consider the surprisingly difficult nature of the T/F (yes/no) QA formats that could potentially exhibit less unintended biases/artifacts than curating data in the multiple choice (MCQ) settings. In light of this, for each question, we collect **one correct** and **one distractor** answer (which can be a slightly perturbed version of the correct one), where both of which are individually judge-able by themselves respectively. And hence, our dataset can be approached as a binary *True/False* (*T/F*) prediction task as well as a *multiple-choice* (*MCQ*) (2 choices in this case) question answering task.

It is worth noting that the distractors in our dataset are manually curated with certain twists towards the correct answers (examples in Table 5.2), forcing the models to truly understand the visual concepts involved in the counterfactual questions in order to answer correctly.

In Chapter 5.5.2, we will describe our adoption of a pairwise consistency metric that requires the model to answer correctly in both correct and distractor directions to be regarded as a success, in order to reduce the models’ exploiting surface-level heuristics to predict the answers.

Commonsense Dimensions. We adopt the commonsense knowledge categorization proposed in (Singh et al., 2021), which is inspired by the *Theory of Core Knowledge* (Spelke





Sub-sampled Key Video Frames	Question-Answer Pairs
	<p>(Temporal) Q: What if the two persons had swerved to their left before reaching the shore? Correct: They would not have had a beach landing. Wrong: They would have had a beach landing.</p>
	<p>(Social) Q: What if the skier was a stranger to the two people standing still? Correct: The skier does not throw the snowball. Wrong: The skier still throws the snowball.</p>
	<p>(Physical) Q: what if the wheel was in a bike? Correct: He would need to take out the screw before being able to set the wheel on the table Wrong: He would set the whole bike along with the wheel on the table.</p>
	<p>(Physical) Q: What if I let the cutting board lie on the counter? Correct: The cutting board would be dried slower. Wrong: The cutting board would be dried quicker as it occupies a larger area.</p>

Table 5.2: Sample data points of our dataset.

and Kinzler, 2007)², to collect QAs that focus on the following three dimensions: *physical*, *social*, and *temporal*. The **physical** dimension concerns the knowledge of objects involved in the events and their properties (*e.g.* , shape, size, functionalities, affordances), as well as the motion and location of the events. The **social** dimension looks at human social behaviors, particularly attributes such as personality, emotions, inner interests/intentions, and social activities.³ The **temporal** dimension regards the aspects of events/activities in their temporal orderings, duration, and frequency/speed of motions.

The three main dimensions are the building blocks towards a comprehensive commonsense reasoning, and helps systematically analyze in which aspects the models need to be improved upon more. Although some questions can be answered using more than one commonsense

²The capability of reasoning about physical objects, places, motions, and the social world.

³As most videos from Ego4D show tasks performed solely by the camera wearer without social interactions, we do not require the social dimension to be annotated.

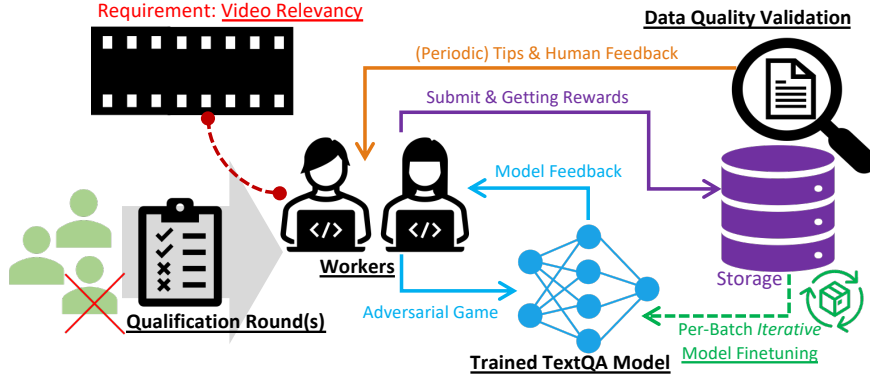


Figure 5.2: Data collection workflow.

dimension, we ask the annotators to label with the main one used.

Video Resources & Sampling. We utilize the Oops! (Epstein et al., 2020) dataset for third-person view videos and Ego4D (Grauman et al., 2022d) for first-person views, where both of which feature text descriptions of the video contents. Oops! concerns predicting the failing (oops) moment of an intended action in a video, and hence is event-rich and a good testbed for reasoning what could the outcomes turned out differently. Ego4D collects videos of humans performing daily activities in the first-person view, which adds a desirable task-knowledge layer on top of its event-richness.

As we are annotating subsets of videos from the aforementioned sources, we have the privilege to encourage a more balanced *key events* distribution from the videos to be annotated. Specifically, we (1) use NLP tools such as semantic role labeling (SRL) to extract key verbs (events) for each video description⁴, and group the videos accordingly, (2) each time sample an event group with a probability inverse proportional to the current launched key event distribution, (2) sample a video from the event group in (3), and repeat until reaching a desired number of videos (to be annotated).

The sampling strategy, combined with our pre-defined reasoning dimensions and video domains, is designed to improve the diversity of question-answer pairs.

Collection Workflow. We collect our dataset via Amazon Mechanical Turk (MTurk). Each MTurk worker is asked to carefully *watch a given video* for creating the QA pairs. As depicted in Figure 5.2, our dataset collection process comprises four main steps: (1) We

⁴We use the originally annotated narrations in Ego4D.

design a **qualification questionnaire** focusing on examining one’s understanding of the key concepts in our problem design, *i.e.* , the concept of counterfactual, the requirement of video relevancy, common sense reasoning dimensions, and what types of QA pairs are more desirable. (2) Once the workers pass the qualification test, they are directed to an interface where a **pretrained (text-only) QA model** is deployed in the loop of the QA creation process. Bonus monetary rewards are given if the deployed model fails to predict correctly the creations. (3) Internal members then conduct a **quality validation** on the created samples and provide customized tips and/or feedback to the workers for potential improvements. (4) Lastly, our deployed model is **iteratively finetuned** on the validated samples after each batch of annotations, which results in a constantly improved model to incentivize more challenging sample creations.

Integrating the model-in-the-loop protocol into the pipeline not only brings benefits in curating more challenging samples, but also helps diversify the answers as the models will not be easily fooled if there are similar patterns existed in the dataset or the questions can be simply guessed without visual inputs.

Quality Validation. In order to further ensure the sample quality as well as summarize common mistakes to provide custom human feedback to the annotators, our internal members conduct the second-phase manual sample validation in conjunction with the deployed model results. We cross-validate the annotations among our internal members in the ramping-up phase to ensure quality. We also accumulate detailed guidelines from our manual validation process for providing effective feedback. After scaling up, we continue to validate the annotations via uniform subsampling across each annotator. Our validation criteria are well aligned as can be seen in the high 0.85 Kappa score for commonsense dimension agreements; and 0.91 overlapping ratios for video relevancy.⁵

Validation Analysis. Table 5.3 reports the data drop-rates (majority voted to drop by all three validators) for the first 5 batches. We hope these rigorous safety checks can ensure a good data quality that also closely follows our guideline, and the validation should by no means introduce unnecessary biases as we indeed saw a decrease in the dropping rates in our

⁵We did not use Kappa score for video relevancy because there is an unbalanced "*agreed*" distribution of "yes" and "no" (22:1) in our validation results for this criteria, which would result in unfair Kappa score.

Batches	Annotation Drop-Rate (%)	Number of Videos
Batch-1	28	50
Batch-2	17.3	100
Batch-3	4.3	200
Batch-4	3.5	200
Batch-5	2.6	200

Table 5.3: Annotation drop rate for the first 5 batches. Each video gives 3 pairs of question - correct/distractor answers.

later collection batches.

5.3.2 Dataset Statistics

General Statistics. Table 5.4 summarizes the essential statistics of the collected dataset, where Table 5.4a is for videos obtained from the Oops! (Epstein et al., 2020) dataset whereas Table 5.4b is for videos from Ego4D (Grauman et al., 2022d). The frame-per-second rate (FPS) of videos from either source is mostly 30.

Key Annotated Events. We plot the distributions of most frequent key verbs (for main event types) and nouns (for entities involved in events) in Figure 5.3a and Figure 5.3b, respectively, to have a rough visual inspection of the diversity of the created samples. The key verbs/nouns are firstly determined by the SRL parses of the question and answer sentences (separately considered), and followed by lemmatization. Both plots are summaries of the two video sources.

Deployed Model. Table 5.5 reports the model fooling rates in our collected data across the two data sources. We encourage our annotators to develop QA pairs that can successfully fool our model by setting up monetary rewards and unlimited trials.

5.4 Benchmarking Models

We benchmark our dataset with both state-of-the-art language-only and vision-language models. Specifically, we perform experiments with DeBERTa (He et al., 2021), UnifiedQA (Khashabi et al., 2020), VIOLET (Fu et al., 2021), VALOR (Chen et al., 2023), and VL-Adapter (Sung et al., 2022) on our dataset.

Language-Only Models. While ACQUIRED is a multimodal dataset that has both vision and language inputs, previous works (Thomason et al., 2019) have pointed out that unimodal

Type	Counts
Total Unique Videos	2,664
Total Unique QA-Pairs	7,853
Type-Token Ratio	0.0158
Verb-Token Ratio (total # verb-types)	0.0341
Verb-Token Ratio (total # tokens)	0.0034
Noun-Token Ratio (total # noun-types)	0.0796
Noun-Token Ratio (total # tokens)	0.0063
Physical / Social / Temporal (%)	34 / 33 / 33

Type	Mean	Std	Max	Min
Tokens in a Question	11.5	3.5	39	5
Tokens in an Answer	8.2	6.6	60	5
Video Frames (Count)	296.8	207.3	3283	74
Video Duration (sec)	10.7	7.1	111.6	3.2

(a) Videos from Oops!

Type	Counts
Total Unique Videos	1,038
Total Unique QA-Pairs	2,695
Type-Token Ratio	0.0205
Verb-Token Ratio (total # video-types)	0.0586
Verb-Token Ratio (total # tokens)	0.0045
Noun-Token Ratio (total # noun-types)	0.1054
Noun-Token Ratio (total # tokens)	0.0081
Physical / Social / Temporal (%)	77 / 0 / 23

Type	Mean	Std	Max	Min
Tokens in a Question	11.1	3.3	32	6
Tokens in an Answer	9.4	6.0	41	5
Video Frames (Count)	399.1	54.8	572	240
Video Duration (sec)	13.3	1.8	19.3	8

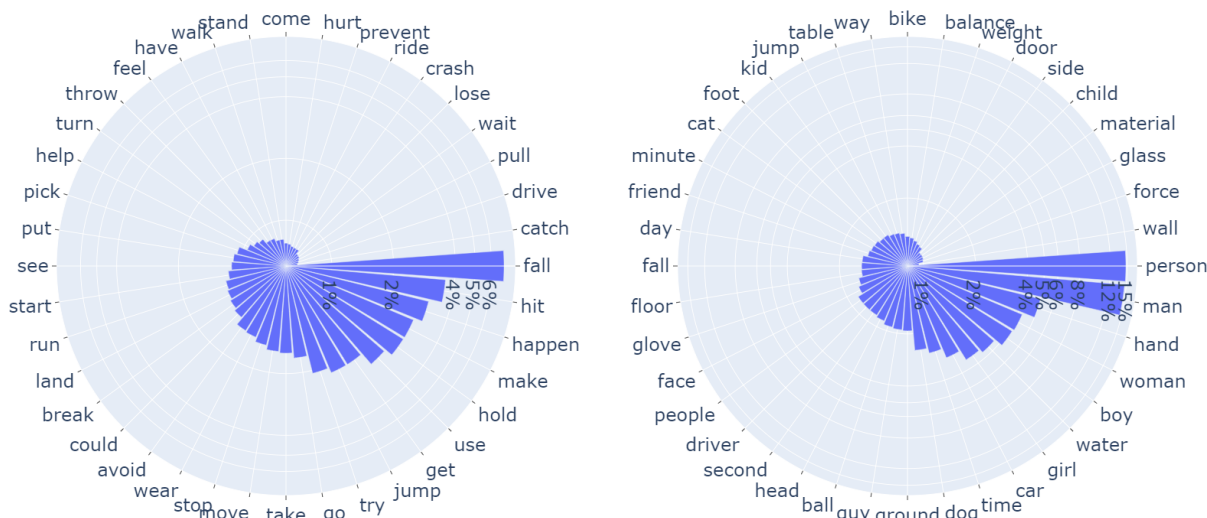
(b) Videos from Ego4D

Table 5.4: General statistics of the two video domains.

Videos From	Avg. Fool Rate (%)	Avg. Fool Accuracy
Oops!	57.69	42.31
Ego4D	51.43	48.57

Table 5.5: Deployed model fooling rates during collection.

models can sometimes achieve surprisingly strong performance because of the annotation bias. Therefore, we evaluate both DeBERTa-v3 (He et al., 2021) and the UnifiedQA model



(a) Verbs

(b) Nouns

Figure 5.3: Top-40 frequent word-types in the dataset.

family (Khashabi et al., 2020) (state-of-the-art question answering models based on the T5 architecture (Raffel et al., 2020)) on our dataset, which can reflect the dataset biases and provide an important reference point for multimodal models. The language-only models answer the textual questions without looking at the videos.

Inspired by the superior performance of the recent large language models, *i.e.*, the GPT model from OpenAI, we also evaluate its zero-shot performance on the textual parts of our dataset. Specifically, we consider both ChatGPT (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b). In addition, we further include a version of GPT models that can condition on pre-annotated descriptions describing the general contents of the videos, to serve as the pseudo visual (and situated) contexts of the questions.

VIOLET (Fu et al., 2021). VIOLET is a video-language model that has three components, including a video encoder (Swin Transformer-base (Liu et al., 2022b)), a language encoder (BERT-base (Devlin et al., 2019)), and a cross-modal transformer module that performs cross-modal fusion. The video and language encoders extract features from the video and language inputs respectively, and the extracted features are then fed into the cross-modal transformer for cross-modal interactions. VIOLET is pretrained on large-scale

video-text data with masked language modeling that predicts the original word tokens given the masked inputs, masked visual-token modeling (MVM) that recovers the masked video patches conditioned on the unmasked video and language inputs, visual-text matching that aims to align the paired video-text inputs between video and text modality.

VALOR (Chen et al., 2023). VALOR is a recently proposed multimodal model that can take video, language, as well as audio as inputs. Similar to VIOLET, VALOR also first encodes vision, audio, and text inputs separately, and the encoded features are then fed into a multimodal decoder for text generation. VALOR demonstrates strong performance across a wide range of tasks, including video retrieval, video captioning, video question answering, audio-visual captioning, text-to-audio retrieval, and audio captioning.

VL-Adapter (Sung et al., 2022). VL-Adapter uses a pretrained vision encoder (e.g. CLIP (Radford et al., 2021)) to extract vision features and feed the vision features as well as text tokens to a pretrained language model (e.g. T5 (Raffel et al., 2020)) so that the model can take both vision and language information. When adapting the model for downstream tasks, because it can be costly to finetune all the model parameters, VL-Adapter investigates different adapter-based parameter-efficient finetuning strategies and demonstrates that training the adapter allows them to only update a rather small portion (e.g. 4%) of total parameters and match the performance of finetuning the entire model. Because VL-Adapter supports different combinations of pretrained vision and language encoders, we employ different versions of CLIP-ViT-B/16 and UnifiedQA-Large as the vision and text encoders.

5.5 Experiments and Analysis

5.5.1 Training and Implementation Details

We obtain the pretrained weights of all the benchmarking models from their respective open-sourced releases and finetune them on our official training data split. The hyperparameters are manually tuned for each model, and the checkpoints used for testing are selected by their validation performance.

5.5.2 Experimental Setup

Data Splits. For our official (to-be-released) dataset, we follow a 45 – 5 – 50 ratio and randomly split the train-development-test datasets. The train split is mainly to adapt models to our QA task settings as well as the counterfactual reasoning style. We ensure that there are no overlaps between videos of different sets and the Oops! and Ego4D videos are equally distributed in each of the splits.

Evaluation Metrics. Models are evaluated by a simple accuracy metric, for both T/F and MCQ settings. We also further ablate the model performance along the commonsense dimensions and/or viewpoints, for a more detailed performance breakdown and analysis. We also include the pairwise accuracy in the T/F setting following Singh et al. (2021), where the model is considered correct if both individual judgments are correct in each pair.

Training Details. All the models in this work are trained on multi (at least 2-4) Nvidia A100 GPUs⁶ on a Ubuntu 20.04.2 operating system.

We train our models until performance convergence is observed on the training split (determined by the development set performance). All of the hyperparameters are manually tuned and searched, with multiple trials for better performance and training convergences.

5.5.3 Experimental Results

Table 5.6 reports benchmark performance. The best-performing multimodal model (VL-Adapter) performs slightly better than its text-only counterparts, UnifiedQA-large (*i.e.*, the language encoder of our VL-Adapter). While this shows that visual contexts and multimodality are effective, the performance gap is not substantial; therefore, there is room for improvement, and more effective methods of multimodal inputs are yet to be explored. While text-only UnifiedQA-3B achieves overall better performance in both T/F and MCQ settings, potentially due to its much larger learnable parameter space, its mediocre pairwise accuracy suggests that the model is still inept at robust counterfactual reasoning in the two facets of the same question.

In general, models perform better in the MCQ settings than the T/F ones. This is intuitive because in the MCQ settings, the model is aware that only one of the two given

⁶<https://www.nvidia.com/en-us/data-center/a100/>

Modality	Model	QA-Format	Viewpoints	Accuracy↑ (%)	Dimension Breakdowns		
					Physical	Social	Temporal
Text-Only	DeBERTa-V3	T/F	—	70.12	70.61	70.32	69.19
		MCQ	—	70.35	72.10	68.62	69.01
	UnifiedQA-base	T/F	—	68.93	70.22	69.32	66.33
		MCQ	—	67.63	68.53	69.01	65.13
	UnifiedQA-large	T/F	—	69.59	71.00	69.88	67.18
		MCQ	—	70.38	71.57	71.83	67.38
	UnifiedQA-3B	T/F	—	70.49	70.58	72.20	68.99
		T/F (Pair.)	—	54.91	55.31	56.21	53.26
	Vanilla ChatGPT	MCQ	—	73.40	73.36	75.80	71.60
		T/F	—	52.80	51.36	48.06	54.04
Desc.-ChatGPT	T/F	—	55.20	50.82	52.90	52.48	
	MCQ	—	42.40	36.96	43.22	47.83	
Vanilla GPT-4	T/F	—	53.80	53.89	53.16	54.32	
Desc.-GPT-4	T/F	—	56.20	55.00	58.23	55.56	
	MCQ	—	60.80	61.41	55.48	65.22	
Multimodal	VIOLET	T/F	All	66.15	70.20	64.45	60.24
		T/F (Pair.)	All	48.25	54.03	44.60	40.63
		MCQ	All	69.33	70.20	70.23	67.19
	VALOR	T/F	All	63.83	66.54	62.50	60.02
		T/F (Pair.)	All	43.00	46.51	42.46	37.26
		MCQ	All	55.06	58.28	51.76	51.69
	VL-Adapter	T/F (Pair.)	All	68.75	71.56	67.94	64.40
			3rd	66.32	66.01	67.90	65.07
			1st	72.63	75.49	—	62.82
	VL-Adapter	T/F (Pair.)	All	51.19	54.27	49.56	47.74
			3rd	47.82	47.60	49.50	46.40
			1st	60.40	62.23	-	53.44
		MCQ	All	71.53	72.70	70.39	70.25
			3rd	69.13	67.63	70.35	69.48
			1st	75.34	76.29	—	72.05
Human Performance	T/F	All	83.60	81.82	100	77.27	
	T/F (Pair.)	All	77.78	72.73	100	54.55	
	MCQ	All	92.59	90.91	100	90.91	

Table 5.6: Model benchmarking performance on our ACQUIRED dataset.

options is correct and only needs to compare them and select the more reasonable option. (Such a phenomenon is also studied/discovered in (Clark et al., 2019; Singh et al., 2021)) In the case of ChatGPT, its MCQ setting accuracy is lower than that of the T/F setting compared to others. We suspect that ChatGPT might have a weaker reasoning ability compared with GPT4. We observe that often ChatGPT refuses to give an answer in the MCQ settings because of insufficient conditions while it leans towards false when it was asked the same question in a T/F setting.

Perhaps surprisingly, despite the remarkable capabilities of the GPT series, they do not perform as impressively, even when provided with descriptions transcribing the major visual events in the videos. This suggests that the annotators in our curation task indeed closely examine many visual details in order to create more challenging samples.

Human Performance. We randomly sub-sample 500 videos to estimate human performance: these are reported in the last two rows of Table 5.6. The human performance highlights a significant gap above all the model results, especially for the MCQ settings. We hope future modeling endeavors can close the gap in visual counterfactual reasoning.

Commonsense Dimensions. The rightmost parts of Table 5.6 report the performance breakdown along commonsense reasoning dimensions. We observe a general trend: most of the models perform better in physical and social dimensions compared to the temporal dimension; the physical dimension generally exhibits the highest performance. That observation implies that, even after being finetuned on our dataset, the models still fall short of capturing temporal commonsense as opposed to the other two kinds of knowledge. This can also be hypothetically attributed to the fact that the pretraining data for the language models encapsulate more physical and/or human social knowledge.

Viewpoints. We take the best-performing multimodal model (VL-Adapter) and ablate its performance along different video viewpoints. We find that, despite being pretrained mostly on third-person viewpoint videos, the generalization ability of the models towards first-person viewpoints is sufficiently good. However, as the videos from Ego4D are not intended to explicitly contain failed actions from the camera wearers, it could be more challenging for our annotators to construct diverse and subtle counterfactual questions as compared to the videos from Oops!. Nevertheless, we argue that the counterfactual reasoning ability of the models should be equally crucial regardless of video viewpoint, and our dataset can inspire relevant research serving as a first-of-its-kind counterfactual video QA encapsulating videos from varying viewpoints.

5.6 Summary

In this work, we present a novel counterfactual-reasoning-focused video question answering dataset, named ACQUIRED. The dataset provides questions about counterfactual hypotheses over visual events (videos). We collect a correct and a distractor answer for three com-

nonsense reasoning dimensions: physical, social, and temporal. We benchmarked various state-of-the-art language models (including LLMs like GPT) and video-language models on the collected dataset, where the results demonstrate algorithm performance well below human performance ($>13\%$ accuracy). We hope our studies and the collected **ACQUIRED** dataset can spur relevant future research, specifically on testing multimodal models' capabilities in counterfactual reasoning, devising assistive AI for remedial and/or cause estimation of observed failures, and more sophisticated visual event understanding and reasoning.

CHAPTER 6

SIMMC-VR: A Task-oriented Multimodal Dialog Dataset with Situated and Immersive VR Streams

6.1 Introduction

With the growing popularity of smart glasses, studies on visually grounded conversational agents have gained significant interest. For instance, SIMMC-2.0 (Kottur et al., 2021) introduces an image-grounded, task-oriented dialog (TOD) dataset where an assistant agent co-observes the user’s egocentric viewpoint to aid with user requests. Many follow-up works (Huang et al., 2021b; Lee et al., 2022; Chiyah-Garcia et al., 2022) focus on challenges around dialog-image grounding, such as visual coreference resolution (*e.g.* ‘*the yellow dress behind the rack*’) of a static image.

However, several technical gaps still remain in applying prior work to build a real-world, *situated* multimodal assistant (Figure 6.1). For instance, a typical multimodal user-assistant scenario (with a video capturing capability) would include (1) spatial *and* temporal language references as grounding contexts (‘*the shirt I saw earlier when I entered the store*’), (2) actively perceived egocentric motions as part of conversation contexts (“*No – turn around the other way*”), (3) references to conversational memories from past sessions (‘*the one I bought earlier*’, the ‘*black coat*’ in Figure 6.1 being retroactively mentioned by both the assistant and the user), *etc.* While these scenarios are perceived as the expected capabilities of a next-generation multimodal assistant, our survey of datasets (Sec. 6.2) highlights that due to the static and constrained nature of the datasets’ grounding context, they lack sufficiently complex interactions.

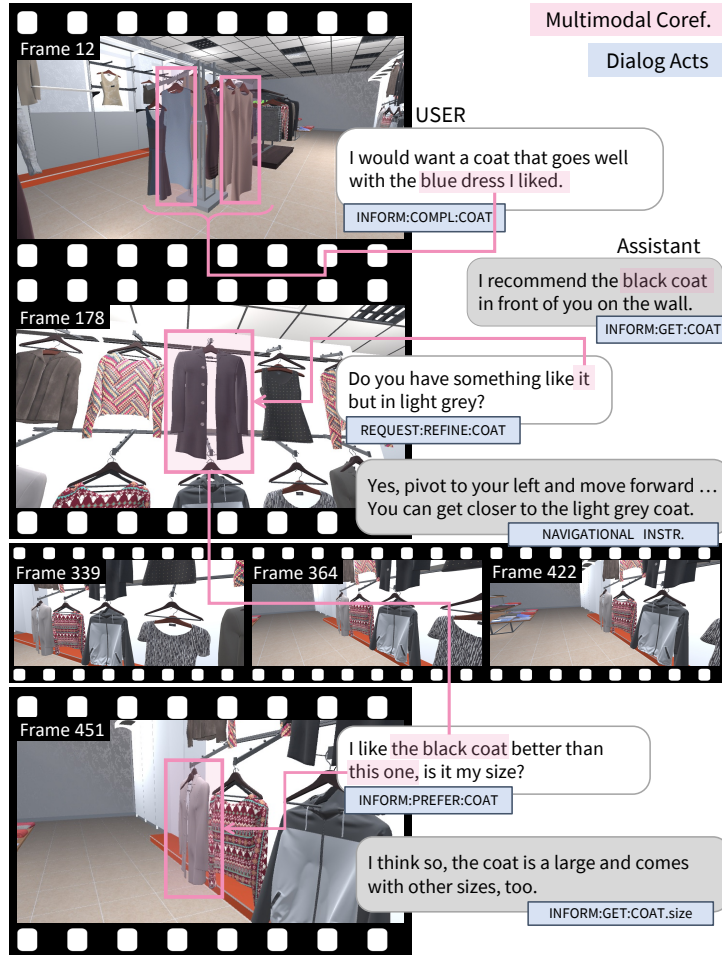


Figure 6.1: SIMMC-VR is a Situated Interactive Multi-Modal Conversation dataset that features task-oriented user↔assistant dialogs *streamed immersively* in a virtual-reality (VR) environment. The dataset is created on programmed realistic shopping scenarios and actively-rendered photorealistic user visual observations, which brings new challenges for complex spatial-temporal reasoning on the multimodal interactions (visual cues and grounded-dialogs).

To this end, we present SIMMC-VR, a video-grounded task-oriented dialog dataset comprising 4K user↔assistant task-oriented dialogs (95.3K utterances) grounded on diverse photorealistic VR video streams (4.8M frames). For data collection, we propose a novel two-stage approach with: (1) a multimodal interaction simulator that generates egocentric VR streams grounded on *object-centric* multimodal dialog flows, and (2) a manual paraphrasing step for naturalness and diversity while preserving multimodal dependencies between visual scenes and their grounding language. Our pipeline allows for flexible and cost-effective data collection, easily extendable to simulate any other domains given the availability of 3D virtual assets.

To measure progress towards real-world applicability, we propose four **SIMMC-VR** tasks that address new challenges in complex spatio-temporal dialog reasoning. We then extend state-of-the-art multimodal models to the **SIMMC-VR** tasks and discuss the limitations of current models.

Our contributions are as follows: (1) we present **SIMMC-VR**, a video-grounded task-oriented dialog dataset (95K utterances over 4.8M frames) targeted towards real-world applications for an assistant on smart glasses. (2) We propose the tasks with complex spatio-temporal conversational dependencies, and benchmark them by extending the state-of-the-art multimodal models. (3) Our data collection platform allows creation of a similar dataset in any target domains.

6.2 Background and Related Work

The proposed work in this chapter addresses unique requirements for a task-oriented assistant on smart glasses, making it a first-of-its-kind – while complementing other related works within multimodal NLP.

SIMMC (Moon et al., 2020; Kottur et al., 2021) is a class of research areas that the proposed work builds upon, which addresses using virtual environments to simulate a co-observing multimodal dialog agent. Moving away from the sanitized and static scenes that they concern for the limited use cases, **SIMMC-VR** introduces several additional challenges as summarized in Section 6.3.3.

Several models (Kung et al., 2021; Senese et al., 2021; Lee and Han, 2021; Huang et al., 2021c) are proposed for the **SIMMC** benchmark tasks – primarily focusing on grounding dialogs on visual objects from a single image. Taking inspirations from these works, we extend the models to accommodate temporal dependencies within frames.

Multimodal Dialog Datasets. Many of the existing literature in multimodal dialogs (Das et al., 2017b; Hori et al., 2018; Kottur et al., 2019; de Vries et al., 2017, 2018; Le et al., 2021) typically assume asymmetric visual information between two observers, *i.e.* *questioner* and *answerer*, where conversational goals are limited to reducing information asymmetry (similar to VQA). In contrast, we study task-oriented dialog scenarios – an assistant co-observes the same scene as a user does, thus focusing on serving user requests to achieve functional goals (*e.g.* giving recommendations).

The embodied AI dialog systems (Gao et al., 2022; Padmakumar et al., 2022), on the other hand, study the scenarios where a human participant *teaches* an AI agent a set of skills or gives navigational directions – hence posing an opposite role to an AI agent. While it is an important area to study, its distribution of utterance patterns is completely different and therefore not applicable for our target domain – building a situated AI *assistant*.

Egocentric Video Datasets. With the popularity of wearable devices, several datasets (Grauman et al., 2022b; Lv et al., 2022; Damen et al., 2021) are released to study the unique properties of egocentric videos. Our work also features similar visual properties, while adding conversational layers that showcase an assistant use case of such videos.

Task-Oriented Dialog Systems (Henderson et al., 2014; Rastogi et al., 2019; Budzianowski et al., 2018b; Eric et al., 2019) have long been studied to support various assistant scenarios (*e.g.* booking hotels). Our work takes its roots in this line of work – focusing on predicting user belief states and dialog acts to achieve functional goals – and extends it to a unique multimodal setting.

A popular thread in the task-oriented dialog system modeling is to fine-tune end-to-end causal LLMs (Hosseini-Asl et al., 2020; Peng et al., 2020; Chao and Lane, 2019; Gao et al., 2019; Crook et al., 2021). We extend this line of work and propose a multimodal extension to account for visual inputs.

6.3 SIMMC-VR Dataset

SIMMC-VR is *actively* multimodal, where each data instance is a video from a user’s **egocentric** viewpoint recording all interactions within a virtual shopping environment, densely paired with dialog utterances and essential attributes. Each task-oriented dialog mimics real-world shopping scenarios where the assistant’s goal is to help the user make purchases and navigate through the environment. In each instance, the user walks around a virtual shop while the assistant provides product information or recommendations; as well as help the user locate and navigate to products of interest.

Dataset Collection Strategy. Multimodal or embodied dialogs (Das et al., 2017a; Padmakumar et al., 2022) are often constructed via a two-player game where participants interact with the *environment* and *converse* with each other (*i.e.* in a Wizard of Oz (WOZ) (Mrkšić et al., 2017; Budzianowski et al., 2018a) role-playing fashion). However, it can be overly

challenging to require annotators to role-play as the AI assistant in our complex and quite cluttered VR shop environments (>100 products). Furthermore, to match the potential retroactive reasoning shopping scenarios (*e.g.* concerning products priorly seen/mentioned), it could add much mental burden for annotators to memorize object attributes and their locations **while** composing *authentic* long dialogue interactions. Lastly, in conjunction with the aforementioned difficulties, it is rather unscalable and inextensible to manually annotate all the required labels (dialog acts, coreferences) cross-referencing complex moving scenes for a *task-oriented* dialog dataset.

We therefore collect the dataset through two phases: (1) **simulating multimodal dialog flows** with templated utterances – thereby programmatically generating fine-grained-scene-grounded annotations and systematically ensuring the diversity of the conversations, and (2) **manual paraphrasing**, which ensures the naturalness of utterances with a significantly less annotation overhead (Rastogi et al., 2020; Shah et al., 2018).

6.3.1 Multimodal Dialog Generation

Our pipeline for multimodal dialog generation simulates plausible and natural multimodal interactions in a virtual environment (Figure 6.2). The process is as follows: (1) Decide a **meta-agenda** based on object attributes and traversal routes. (2) Sample specific objects that fulfill the decided agenda as the **object-centric flow**. (3) Perform the user traversal **path planning** and video recording using the sampled objects as starting/ending points. (4) Synthesize the corresponding utterances via pre-written **templates** and the multimodal contexts. (5) Manually **paraphrase** the templated utterances.

We categorize a full dialog instance (generated through the previously described steps) into two phases: (a) **static phase** where the user *mostly* focuses on a specific viewpoint (with a small amount of randomness in movement or eye-gaze) when conversing with the assistant (Chapter 6.3.1), and (b) **active phase**, where the user navigates to another spot within the environment, at will or following assistant instructions, containing larger movements and actions (Chapter 6.3.1). The two phases interleave each other, creating a realistic shopping scenario (*e.g.* user walks into a shop, stopping by a few products, and wanders to other ones).

Virtual Environment. Following SIMMC-2.0, we use the same set of photorealistic VR shopping environments in Unity (Unity, 2020), where a set of seed scenes with pre-arranged

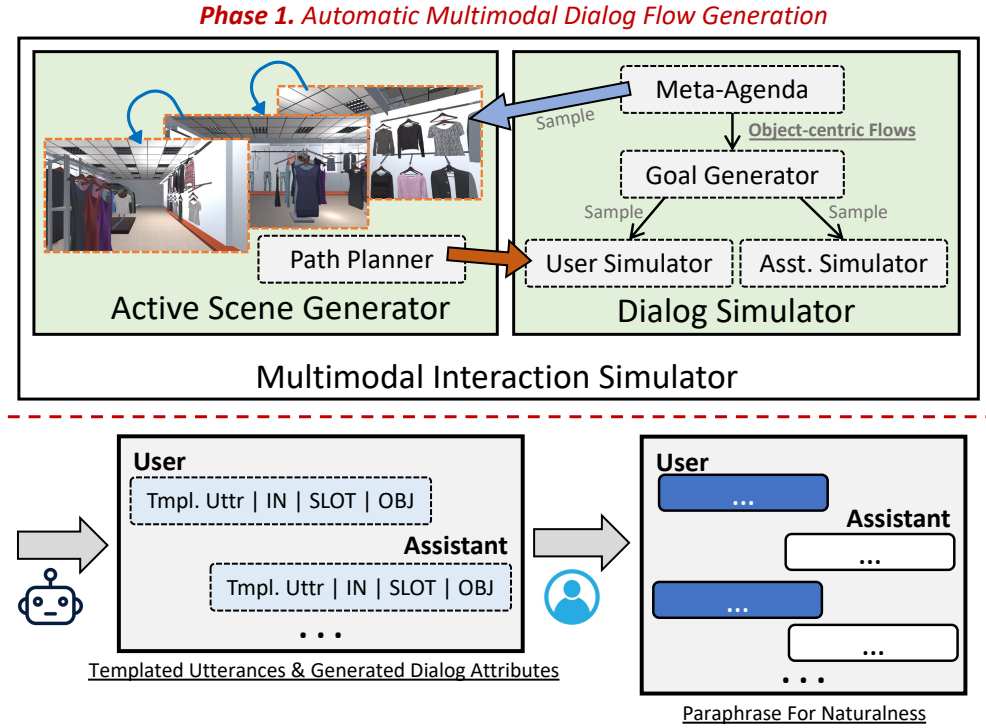


Figure 6.2: Dialog generation flow: (Upper half) a **meta-agenda** is firstly programmed to sample an *object-centric* flow (grounded in the environment), which is used by the goal generator to sample high-level dialog goals. These goals are then used by both user and assistant simulators to synthesize templated utterances, which are then manually paraphrased by linguistic experts for diversity and naturalness (**lower half**).

Fashion	hat, tshirt, jacket, hoodie, sweater, shirt, suit, vest, coat, trousers, jeans, joggers, skirt, blouse, tank top, dress, shoes
Furniture	area rug, bed, chair, couch chair, dining table, coffee table, end table, lamp, shelves, sofa

Table 6.1: Digital assets categories used in SIMMC-VR for both fashion and furniture domains.

digital assets (*e.g.* shirts, dresses for *fashion* domain and sofas, tables for *furniture* domain) are programmatically re-arranged into randomized larger sets of scenes.

Table 6.1 lists the asset (product item) categories used for constructing the SIMMC-VR dataset for both fashion and furniture domains.

Active Scene Simulation:

Figure 6.3ab illustrates the process of simulating visual observations of a user traversal, where a *path planning* is performed (connecting the start and end user position/orientation)

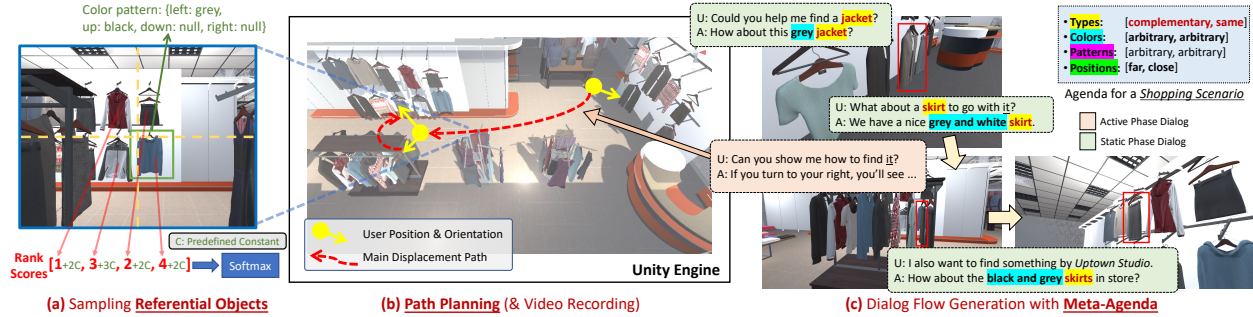


Figure 6.3: Multimodal dialog generation: (Right most) meta-agenda illustrates an exemplar shopping scenario that concerns user demanding *complementary* (*i.e.* can go with) types for the first two items (jacket \leftrightarrow skirt) and the *same* type between the 2nd and 3rd items. Colors and patterns are not constrained, while the scenario simulates longer traversal is required (*far*) between the first two items and the latter two are *close-by*. **(Middle) Path planning:** the navigational utterances will be grounded on the planned path (displacements and orientations) and the **referential objects (left most)** used to facilitate the guidance are sampled according to *softmax* scores on a ranking (via features *e.g.* eye-gaze, color-contrast) of most suitable landmarks.

in the environment, and the trajectories are rendered into egocentric videos.

Path Planning. Ideally, the navigational guidance should minimize the overall traversal distance (to a target spot), while taking the smoothness of movements into consideration. Given a start and end position in the extracted environment layout, we perform an A^* search to plan a trajectory simulating a user’s traversal within a shop. Additionally, we modify the standard A^* algorithm to minimize the amount of *turning* for smoother and more natural user movements¹, with random noises added to naturally jitter the planned path. We then augment the output path with rotation angles computed to account for the user orientation during the traversal. At each viewpoint on the planned path, a Unity camera snapshot is taken, and the traversal video is rendered by combining all the snapshots.

Referential Objects. Once the intended user-traversal video is planned and recorded, we define *key action points*, using the start/end viewpoints of user movements (*i.e.* displacement or turning actions). Inspired by the natural communication behavior, where we often refer to certain *landmarks* when giving navigational guidance, we derive a set of **referential objects** from objects placed across these viewpoints (*e.g.* “Turn left when you see the red shirt.”). Figure 6.3a illustrates the referential object sampling strategy: (1) Compute the cosine similarity between an egocentric viewpoint (3D) vector (gaze point at the center of **yellow**

¹ A^* ’s distance minimization may lead to excessive turns.

dotted lines) and a *look-at* vector to each of the objects within the scene – a higher similarity implies that it is closer to the eye-gaze line of sight, hence more probable to be referenced during conversations. (2) Augment the previously derived rankings with other plausible features such as stronger color contrast with neighboring objects. (3) Lastly, transform these rankings into sampling probabilities (via a *Softmax*) to sample object(s) for reference.

Scene Graphs & Disambiguation. When referring to an item in a cluttered environment, its surroundings often serve as good candidates to *disambiguate* items that may share similar attributes (often useful when users *under-specify* items). In light of this, for each object within the same scene, we build a **local scene-graph** to include the closest three objects to its *left, right, top, bottom* (four main directions). An object can then be referred to with its neighbors when further clarification is needed (*e.g. "Not that one, I mean the white hat below the red coat."*).

Scene Metadata. To facilitate templated utterances for paraphrasing (Chapter 6.3.1) and to formulate a modeling task with visual labels (Chapter 6.4), we compute 2D bounding boxes for all 3D assets in a particular viewpoint, where each object is cross-referenced across every frame. As the dense bounding box computation in a 3D environment is time-consuming (repeated for thousands of frames per dialog), we expedite this process via an approximate reconstruction. Specifically, we record the camera position and orientation for each video frame, and provide the mesh data for each asset and a function to reconstruct 2D bounding boxes on-the-fly.

Dialog Simulation:

In real-life shopping experiences, customers typically explore a shop with certain product attributes of interest in mind (*e.g. clothing colors, types*), thus shopping experiences are often **object-centric** (Yinyin, 2011). Inspired by this, we *program* several (extendable) *object-centric flows* that focus on certain objects within an environment to mimic how a user may wander (self-motivated or guided) around from one product to another.

Dialog Flows. To have full control over the diversity of dialog flows, and to encourage certain patterns of flows to emerge for more interesting user-AI conversations, we propose an *object-centric* generation pipeline. Specifically, to generate an *object-centric flow*, we (1)

Colors	same, arbitrary
Patterns	same, arbitrary
Types	same, arbitrary, alternative, complementary
Positions	far, close, come_back_to_X

Table 6.2: Meta-Agenda Programs

define a **meta-agenda**, a sequence of **meta-goals**² defined by certain object attributes that simulate a complete shopping experience (*e.g.* a customer looking for certain types or colors of clothing, or asking for a complementary item to match a previously purchased one) and (2) for each meta-goal, sample an object according to a planned traversal route (*e.g.* short or long travel distance, traveling back to a previously observed item) and a user-position/orientation to *look at* the object (where the path planning can perform on).³ The meta-agenda is either human-written or programmatically generated, and diversified while ensuring a balanced distribution of scenarios. The traversal route is engineered to ensure user’s navigation/orientation changes are necessary and natural.

For each of the sampled-objects, a **goal generator** will sample a high-level dialog *goal* to define the theme of a few turns of utterances (*e.g.* COMPARE → user requesting product comparisons). The **user simulator** then utilized both the sampled objects and goals to generate corresponding NLU labels following a probability distribution, consisting of user intents (*e.g.* INFORM:GET), request slots (*e.g.* color, brand) and object references. The **assistant simulator** then resolves the user requests, leveraging the multimodal context and the simulation API (*e.g.* for info lookup).⁴

Meta-Agenda. Table 6.2 lists the candidates that can be programmed into the *meta-agenda*. For *alternative* and *complementary* item mappings, we consider: (1) Relations in ConceptNet 5.0 (Speer et al., 2017) such as `distinct_terms` (*jacket* is `distinct_to` *coat*),

²We cap the max sequence length at 3, *i.e.* 3 *meta-goals*.

³Each flow is uniquely defined by the sampled object-sequence. We over-sample totally >1K object-centric flows evenly across 27 programmed meta-agenda (Figure 6.3c).

⁴In contrast, SIMMC-2.0 plans a dialog *only* by randomly sampling a sequence of abstract goals (*e.g.* BROWSE → GET_INFO → ...), often resulting in unrealistic scenarios.

`similar_terms` and/or `related_terms` (e.g. *sofa* is `related_to` *end-table*). And (2) Manual inspections and annotations, where we ask internal members to annotate the alternative and complementary items to a particular one of interest, and refine the annotated list with majority vote (e.g. *hat* is complementary to both *shirt* and *dress* as they can go in pairs, and *coat* is alternative to *jacket* as they share similar functionalities and thus can complement each other).

For the *positions* agenda, we pre-define a distance threshold to denote far or close depending on the environment room layout (differ in fashion and furniture domains). For the `come_back_to_X` program, we engineer that the user will traverse back to an item that is previously seen and indicated with interests, to simulate relevant shopping experiences in the real-world.

Templated Utterances. Grounded by the multimodal context, we pre-define a few utterance templates each associated with a specific dialog act, leaving the specific object-related information (e.g. object ids, modifiers, pronouns) as placeholders that are filled-in according to the visuals. This allows us to easily sample an utterance template that is suitable for a particular situation and the associated user or AI intention, determined by the dialog act. We list a few exemplar utterances and their paraphrases, and highlight the placeholders in Table 6.3. Notice that the local object scene-graphs (Chapter 6.3.1) are also useful for generating diverse reference expressions for the same object (second role of the Assistant examples in Table 6.3).

Manual Paraphrase. Next, we ask human annotators to paraphrase the templated utterances to better match the real-world natural language distribution. We design an interface that dynamically displays a multimodal scene that features either a still image (static dialog phase) or a user egocentric video (active dialog phase). When clicking on a specific turn of a dialog, the corresponding visual input is shown in the display panel to help annotators navigate through the entire dialog flow. We ask the annotators to pay attention to detailed and sophisticated spatial-temporal relations of objects and encourage writing interesting shopping experiences. The paraphrases are collected from more than 20 different linguistic experts for diverse language patterns/usages.

Once manual paraphrases are collected, we perform text-to-speech synthesis (TTS) on the utterances, and synchronize the speech with the relevant motion renders for improved

Role	Dialog Goal & Act	Example Templates & Paraphrases
User	BROWSE REQUEST:GET	Could you recommend something with {type:blouse} _[search-filter] ? ⇒ ‘I am looking for a <u>blouse</u> ; do you have anything to show me?’
	ALTERNATE_SEARCH INFORM:ALTERNATE	Do you have alternatives to [OID:34(hoodie,blue)] _[object] with {color:violet} _[search-filter] ? ⇒ ‘Any other options besides that? See if you have anything <u>violet</u> in store.’
	REFINE_SEARCH INFORM:REFINE	I would like to refine my search to include {type:skirt} _[search-filter] . ⇒ ‘I want to search more specifically for <u>skirts</u> . What are my options now?’
	ADD_TO_CART REQUEST:ADD_TO_CART	Please add to cart: [OID:50(hoodie,green), OID:50(hoodie, green)] _[object] . ⇒ ‘I like the <u>first hoodie</u> the best. Give me two of the <u>green</u> one.’
AI	ACTION INFORM:DIRECTION_STRAIGHT	Go {towards} _[direction] it. [OID:100(sweater,red)] _[object] will be on {far-left} _[relation] . ⇒ ‘Go straight forward until seeing a <u>red and white sweater</u> on your far left.’
	ACTION INFORM:DIRECTION_TURN	Turn {around} _[direction] and you will be able to see [OID:141(blouse,white)] _[object] , which is {on-right} _[relation] to [OID:154(jacket,black)] _[object] . ⇒ ‘Turn around and you will see that <u>white and black blouse</u> , on its left is a <u>black jacket</u> .’
	GET_INFO INFORM:GET	Here is the info on size: [OID:49(hat,green)] _[object] : {size:XS} _[slot-values] . ⇒ ‘That <u>green hat</u> you’re looking at is size XS.’
	COMPLEMENTARY_SEARCH INFORM:COMPLEMENTARY	How about these: [OID:77(skirt,brown)] _[object] ? They are {type:skirt} _[search-filter] . ⇒ ‘Yes we do. How about the <u>brown skirt</u> that is on the far right on the top row?’

* OID stands for object ID.

Table 6.3: Exemplar utterance template and paraphrases in SIMMC-VR. In each row under the second column, the upper terms are the goals and the lower terms are the dialog acts (consisting of acts and activities). We show a few representative dialog acts with their corresponding sample templates (each act may have multiple templates as options) and a sample paraphrase. In each template, the subscripts denote the type of the placeholders, where the contents are filled-in grounded by the multimodal contexts (*e.g.* , sampled objects, user eye-gazes) or sampled attributes (*e.g.* , types or colors of the desired item).

naturalness, making the rendered user shopping videos more realistic (and comprehensive). We use an open-sourced tool, *Coqui TTS* (Coqui.ai, 2022) to generate the spoken speech from the paraphrased utterances. This also helps computing the natural duration of each utterance when spoken so that we can interpolate certain number of video frames (under a fixed frame-rate) to fit such utterance would span.

Dialog Dataset Structures. Similar to other existing task-oriented dialog systems (Eric et al., 2019; Rastogi et al., 2020; Moon et al., 2020), each turn of SIMMC-VR’s dialog data consists of NLU (and NLG) intent and slot labels (*e.g.* “How do their prices compare?” → REQUEST:COMPARE, slots: price, objects: [1, 4]), as well as object references (a unique object ID across the same room environment) like SIMMC-2.0. In SIMMC-VR, due to the newly introduced *active dialog phase* and the richer dialog scenarios (*object-centric* flows), the list

Total # dialogs	4,075
Total # utterances	95,368
Avg # words per user turns	12.9
Avg # words per assistant turns	16.7
Avg # utterances per dialog	23.4
Avg # objects mentioned per dialog	13.2
Avg # objects in key video frames	24.6
Avg # objects per fashion environment	188.6
Avg # objects per furniture environment	62.0
Avg # frames (under fps = 10.0)	1197.7
Avg # seconds per TTS utterance	4.13

Table 6.4: SIMMC-VR dataset statistics. On average there are 13.2 objects mentioned in a dialog and more than 20 visible in each video frame, making the video-grounded dialogs diverse and rich in contents. Each video roughly lasts 2 minutes, equating to a total of >130 hours long VR streams.

of intents is expanded as compared to SIMMC-2.0.

6.3.2 SIMMC-VR Dataset Analysis

Table 6.4 shows the essential dataset statistics. In total, SIMMC-VR contains 4K dialogs with the corresponding videos (equating to 95.3K utterances).

Videos. We set the frame per second (fps) as 10.0, which roughly leads to an average of 1.2K frames per video (\sim 2 minutes length). On average there are 24.6 visible objects in the key video frames.

Dialog Acts & Flows. Each algorithmically generated flow, *i.e.* the **meta-agenda**-induced *object-centric flow* (Chapter 6.3.1), is capped to have at most 5 different dialogs with randomly sampled dialog goals and intents. The average number of utterances is 23.4, significantly larger than that in SIMMC-2.0 (10.4). Its length distribution over different turns is shown in Figure 6.4a. SIMMC-VR extends SIMMC-2.0’s annotation to a set of 5 dialog acts (*e.g.* INFORM, REQUEST) and 17 activities (*e.g.* REFINE, DIRECTION_TURN). Figure 6.4b shows their frequency breakdown. A visualization of dialog transition is shown in Figure 6.5 to illustrate the diversity and patterns of our generated dialog flows. Figure 6.4c plots the coreference distances according to how many utterances separate the mentions.

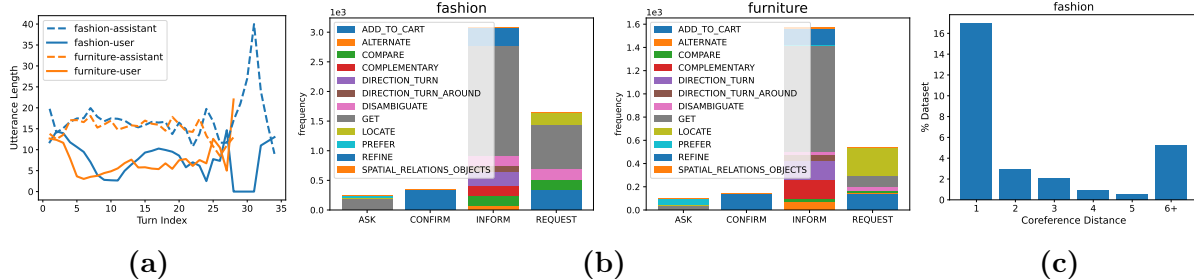


Figure 6.4: Plots of: (a) utterance lengths in dialogs, (b) acts and activities, and (c) co-reference distance between object mentions.

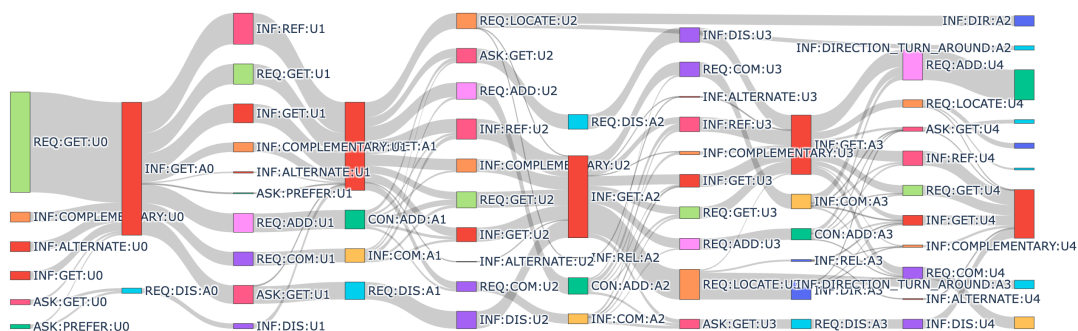


Figure 6.5: Dialog act(s) transitions for the first four rounds of dialogs in the *fashion* domain. The acts and activities are denoted for brevity as **ACT:ACTIVITY:[A|U] [turn_index]**, where **U** and **A** denote user and assistant, respectively. The shown branching and inter-connectivity justifies the diversity of the synthesized dialog flows.

6.3.3 Novel Challenges to SIMMC 2.0

SIMMC-2.0 shares the general goal of achieving multimodal task-oriented dialog systems for future real-world and VR applications. However, the active and rich multimodal contexts of SIMMC-VR introduce the following new challenges: (1) Anchoring *egocentric videos* as visual contexts, SIMMC-VR requires the spatial and the additional temporal multimodal reasoning, posing new categorical patterns of object coreferences and associated user/assistant utterances. (2) The novel dialog simulation pipeline allows for more diverse and realistic interactions (*e.g.* navigation and localization scenarios) with a number of transitory dialog actions and viewpoints, many of which have not been studied in the previous datasets. This results in the higher degree of complexities in conversational tasks – for instance, the coreference resolution task gets significantly harder with a much larger number of objects mentioned in a dialog (13.2 *vs.* 4.7 in 2.0), and with the increased average utterance counts (23.4 *vs.* 10.4 in 2.0). (3) SIMMC-VR requires that a perception model maintains object

correspondences across their variations from different angles and disjoint viewpoints over time, to ensure the correctness of their resolution. While this requirement poses a practical challenge for a real-world application, a robust solution has not been explored especially for its use in the context of the multimodal dialog management.

6.4 SIMMC-VR Task Formulation

The SIMMC-VR is created to help AI models cope with realistic shopping scenarios and assist human users in real-world applications in AR/VR. To investigate the (multimodal) conversational and assistive abilities of current AI systems in this immersive and situated environment, we propose four main benchmarking tasks leveraging the created dataset. Several tasks inherit from SIMMC-2.0 with additional challenges brought by the nature of active user scenes and expanded dataset annotations.

6.4.1 Multimodal Dialog State Tracking

Following SIMMC-2.0, in SIMMC-VR we retain the multimodal dialog state tracking (**MM-DST**) task, which aims at inferring structured information for understanding and planning out dialog policies/actions, with dialog utterances and/or multimodal contexts given. Each DST is required to resolve both the dialog intents (as a dialog *act*) and the user request slots, which is mainly evaluated by the F1 scores of the predicted slots and intents.

6.4.2 Multimodal Coreference Resolution

It is crucial for an assistant to be able to recognize objects that a user is referencing, either within the **current visual context**, or any **previously mentioned items**. Therefore, for each environment, a canonical ID is uniquely assigned to each object as the target for multimodal coreference (**MM-Coref**) resolutions, where the mentions can be resolved by both the dialog context (e.g. "Add the shirt I liked to the cart.") and the multimodal context (e.g. "How does the red shirt next to the jeans compared to the one before?"). Following SIMMC-2.0, we allow the models to take ground-truth bounding boxes as inputs to bypass the needs for perfect visual detectors. The evaluation metric is the F1 scores for the predicted object IDs. Note that as the multimodal contexts are videos, the models are implicitly conditioned to identify the frames that likely contain the target objects, leading to comprehensive multimodal spatial-temporal reasoning. Additionally, while there are no

explicit textual coreference annotations, the models are still implicitly required to perform textual coreference resolution for those utterances mentioning the same objects from prior dialogue turn(s).

6.4.3 Failure-Mode Prediction

SIMMC-VR features user failure-modes that simulate users accidentally failing to correctly follow the assistant guidance. In this task, given a dialog snippet (consisting of utterances in the *active phase*) and the video frames surrounding it, we ask the model to predict whether the current user actions correctly follow the instructions or not (*i.e.* binary classification evaluated by F1 scores). The task is highly multimodal as the model needs to understand the sophisticated active grounding of the visual and dialog contexts. During the training time, we pre-sample the same amount of negative samples to make the labels balanced.

6.4.4 Dialog Response Generation

This task requires a trained dialog agent to generate the assistant responses (measured in BLEU-4 (Papineni et al., 2002)), given user utterances as well as the *resolved* multimodal information (belief states and referred canonical object IDs). Note that even though the aforementioned information is given as ground-truths, the generation still needs to conform to natural language responses that do not contain flattened DSTs or object IDs (*e.g.* INFORM:COMPARE, (OBJ_ID: 5,9) \rightarrow "The white and blue shirts differ by ...").

6.5 Modeling & Experimental Analysis

In this section, we introduce the investigated baseline models to perform a preliminary benchmarking of the proposed dataset, where we hope to inspire more sophisticated and tailored modeling efforts from the community for future research.

Dataset Split. For the empirical modeling analysis and performance benchmarking, we randomly split the dataset into 3 sets: train (70%), dev (5%), and test (25%) sets, while ensuring both domains (fashion and furniture) have the same split distributions.

Baselines. To benchmark the dataset, we adopt:

(a) **MM-DST Model** is a 12-layered multi-task GPT-2 model (Radford et al., 2019; Kottur et al., 2021) trained with joint supervision signals from MM-Coref, MM-DST, and response

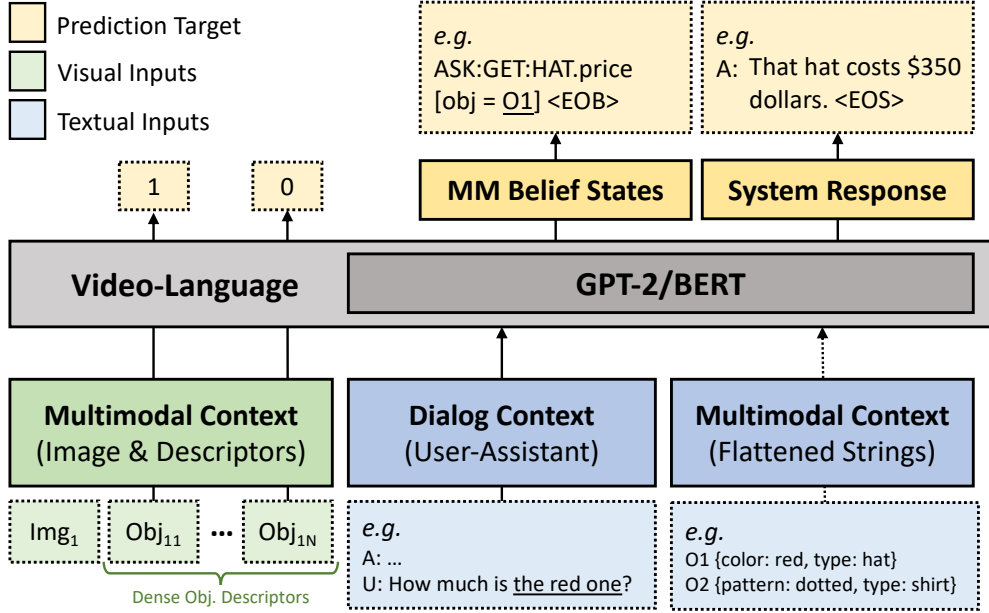


Figure 6.6: Baseline models: The inner grey box (denoted “GPT-2/BERT”) is the language model either as (is) the MM-DST model or the language encoder of the video-language model (VIOLET adopts BERT). The video-language model predicts MM-Coref via dense object descriptors, while MM-DST model generates (via GPT-2) the flattened target strings.

generation tasks, inspired by causal language modeling approach to dialog systems (Peng et al., 2020; Hosseini-Asl et al., 2020). The inputs to the model include both the dialog context (utterances) and the multimodal contexts flattened as structurally formatted text strings, where the outputs are the predicted DST labels. This baseline has two versions: one uses the ground-truth multimodal contexts provided from the scene generator (hence a soft oracle) to simulate the outputs from a robust object detector or from a controlled VR environment, whereas the other has to *infer* visual descriptors from raw videos, simulating real-world scenarios.

(b) Adapted-VIOLET Model is a multimodal video-language model based on VIOLET (Fu et al., 2021), adapted to fit our task structure (Figure 6.6). Due to computational limitations, we randomly sub-sample 10 – 15 video frames during training (while ensuring a proportion of these frames contain objects of ground-truth coreferences), and sweep through the entire video for test-time inference with a fixed window-size. In addition to the frame-level whole image feature, we feed the dense object descriptor features extracted in each ground-truth bounding boxes (assuming a perfect object detector) to the model for the

Model	DST			Coref	Fail.	Gen.
	Slot /	Int. /	Joint F1↑	F1↑	F1↑	BLEU↑
(Label Distribution)	19.4 /	9.39 /	8.73	0.66	34.1	—
MM-DST	72.4 /	78.6 /	33.9	17.1	—	0.117
MM-DST (no-gt.)	71.7 /	77.3 /	30.8	0.71	—	0.120
Adapt.-VIOLET	75.0 /	80.4 /	37.7	9.69	46.4	0.119

SIMMC-2.0 Performance (for comparison)						
MM-DST	89.6 /	94.5 /	44.6	36.6	—	0.192

Table 6.5: Baseline performances for Multimodal (1) Dialog State Tracking (DST), (2) Object Coreference (Coref.), (3) Response Generation (Gen.), and (4) Failure Mode Prediction (Fail.). In the lower half, we report the corresponding performance from SIMMC-2.0 with the MM-DST model.

MM-Coref task.⁵

All baseline models are trained for ten epochs, and the best model on the dev set is used for test.

6.5.1 Experimental Results

Table 6.5 summarizes the model performance and the probabilistic guess performance (proportional to training label distributions) for each sub-task.

Main Results. The baselines show strong overall performances especially in the DST task. The MM-Coref is understandably a very challenging task (resolving tens of items over moving frames), as evidenced in the relatively low scores – suggesting areas for future research. It is worth noting that without the ground truth multimodal contexts for assistant turns, the MM-DST model performs close to zero, indicating that the created dataset does not leak unintended artifacts for the object mentions (that language-only models can easily exploit without visual contexts). For the failure mode prediction, we prepare a test-set that focuses on the active scene utterances, where the random guess roughly equates to the amount of the failure probabilities (30%). We expect the future modeling efforts can better perceive discrepancies between the visual behaviors and the instructed guidance.

Effects of Temporal Grounding. We break down the MM-Coref performance by identi-

⁵Here to simplify the task, our dataset can also be approached without assuming any perfect vision modules.

fyng coref utterances with *temporal dependencies*. With the Adapted-VIOLET model, we get an F1 of 10.5 for utterances *without* temporal dependencies, and a significantly lower 2.81 for the others – suggesting the difficulty in encoding long-standing contexts.

Comparison with 2.0. We also include the MM-DST model performance for the SIMMC-2.0 dataset as a reference, to signify the new challenges that SIMMC-VR brings with the active VR-streams and the complex multimodal dialog flows.

6.6 Summary

We present SIMMC-VR, a situated and interactive dialog dataset that features immersive VR streams as multimodal contexts, simulating realistic shopping scenarios along with user-assistant dialog interactions. The dataset consists of $4K$ user-egocentric videos paired with densely annotated dialog utterances. We build a novel meta-agenda generator for automatically synthesizing rich interactive dialogs grounded on active and diverse visual scenes, paraphrased manually for more natural speech. We propose four sub-tasks on SIMMC-VR which aims at inspiring future dialogue modeling endeavors on high-fidelity egocentric (user POV) environments; where the baseline performance highlights many challenges the dataset brings forth towards actualizing the real-world-ready VR/AR assistant. With rich annotations it provides, SIMMC-VR can as well expand beyond the proposed tasks to spur relevant future research, which includes (but not limited to): (1) augmented with speech-like spoken utterance interventions to enrich the naturalness of the dialogues, and (2) environments and room layouts beyond ones used under the scope of this paper.

Part III

Conclusion

CHAPTER 7

Conclusion and Future Directions

7.1 Summary of Contributions

In this dissertation, each of the presented contributions is closely tied to the roadmap designed in Chapter 1, from comprehending and consolidating task instructional resources, structurally interpreting the instructed actions and dependencies, to visually ground the learned instructed knowledge to the actual world. The proposed system-level outline should serve as the fundamental basis in actualizing helpful and effective assistant AI. In this chapter we summarize our contributions as follows.

In Chapter 2, we present a thorough study of language and multimodal models on the procedural understanding task, utilizing curated online instructional manuals from some popular sites such as WikiHow or recipes. We show that both multimodality and our proposed sequence-aware pretraining techniques are effective for models to learn to sequence and consolidate unorganized task steps. We also provide a multi-reference annotation in order to gauge the interchangeability of certain task steps to inspire future efforts along the line of research.

In Chapter 3, we introduce the essential instruction interpretation task on inferring action and pre-and-post-condition dependencies for structural understanding of the instructed information. The problem can be formulated as a low-resource learning setup, where we design several heuristics to automatically construct useful large-scale weakly supervised data, as well as a two-staged training methodology to improve the language models' capabilities of inferring the condition dependencies more robustly.

Followed the previous two chapters of comprehending instructions, Chapter 4 extends the learned procedural and structural dependency knowledge to actively ground the impor-

tant entities involved in the instructions to the visual world. We firstly propose a crafted prompting scheme to obtain useful action-object knowledge large language models (LLMs), followed by a per-object knowledge aggregation technique to improve vanilla phrase grounding modules on localizing and tracking state-changing key entities more accurately with longer tracking duration.

The aforementioned building blocks for actualizing the multimodal assistant AI leads us to curate novel and challenging resources to evaluate models of their relevant capabilities. In Chapter 5 and Chapter 6, we introduce two benchmarks we carefully collected, where one focuses on the counterfactual commonsense reasoning in diverse and action-rich event videos (Wu* et al., 2023b), while the other presents an interesting virtual shopping assistant that can converse to the users while giving guidance regarding the visual surroundings and tracing spoken items from the past conversations (Wu et al., 2023b). We hope these high-quality multimodal resources can shed lights on and inspire future endeavours of actualizing a better and improved assistant AI.

7.2 Summary of Technical Limitations

While every work in each chapter is motivated to improve the assistive capabilities of AI and aimed at either providing the essential building blocks for more advanced vision-language foundational components or useful and novel resources to learn from and evaluate upon, there are still some limitations in each of them. In this section, we will briefly discuss the technical limitations of each work, and how it may be improved for future advancements.

Task-Step Sequencing. While in this chapter (Chapter 2.3.3) we mentioned that alternative and interchangeable orders are annotated for certain task steps, there are still limitations yet to be addressed: **(1)** The current formulation assumes an *almost* sequential order of the task-steps, *i.e.*, if denoted the task as a graph, it would be a directed acyclic graph (DAG) which has one way traffic/flow of task orders. This chapter currently does not consider repeated steps and/or loops that could be mentioned during the task completion. **(2)** Following the previous point, the task steps could have underlying hierarchy as well, where certain higher level subtasks actually would lead to a few lower-level children task-steps. The most fundamental way of representing a task, is perhaps to extend our work into a representation of a finite state machine (FSM). This formulation would encompass both the

aforementioned loops as well as the inherent hierarchy, while obviously the learning complexity is much higher. **(3)** Currently our model is bounded by five steps, future work could consider extending our work to tackle task with much more steps and details.

Action-Condition Dependencies. There are three main limitations in this chapter: **(1)** Although our annotated dataset enables the possibility of learning an extractive model that can be trained to predict the span of the text segments of interest from scratch, we focus on the more essential action-condition dependency linkage inference task as we find that the SRL extraction heuristic currently applied sufficiently reliable. In the future, we look forward to actualizing such an extractive module and other relevant works that can either further refine the SRL-spans or directly propose the text segments we require. More specifically, the extractive module can be supervised and/or evaluated against with our human annotations on the text segment start-end positions of an article. **(2)** The current system is only trained on English instruction resources. Multilingual versions of our work could be as well an interesting future endeavors to make. **(3)** In this chapter, we mostly consider instructions from physical works. While certain conditions and actions can still be defined within more social domain of data (*e.g.* a precondition to *being a good person* might be *cultivating good habits*). As a result, we do not really guarantee the performance of our models when applied to data from these less physical-oriented domains.

Active Object Grounding. The two main limitations and potential future follow-ups are: **(1)** While we make our best endeavours to engineer comprehensive and appropriate prompts for obtaining essential symbolic action-object knowledge from large language models (LLMs) such as GPT, there are still few cases where the extracted objects are not ideal (see Table 4.1). Hence, our model performance could potentially be bounded by such limitation inherited from the LLM ability to fully and accurately comprehend the provided instructions. Future works can explore whether more sophisticated in-context learning (by providing examples that could be tricky to the LLM) would be able to alleviate this issue. Alternatively, we may utilize LLM-self-constructed datasets to finetune another strong language models (such as Alpaca (Taori et al., 2023)) for the object extraction task. **(2)** There is more object- and action-relevant knowledge that could be obtained from LLMs, such as spatial relations among the objects, size difference between the objects, and other subtle geometrical transitions of the objects. During experiments, we attempted to incorporate spatial and size information

to our models. However, experimental results on the given datasets did not show significant improvement. Thus we omitted them from this work. We hope to inspire future relevant research along this line to further exploit other potentially useful knowledge.

SIMMC-VR. While our SIMMC-VR dataset is constructed aimed at presenting a comprehensive suite of benchmarking tasks for multimodal assistive models, there are some improvements one can make: **(1)** The SIMMC-VR dataset, similar to the previous SIMMC families, focuses on shopping scenarios (clothing and furniture purchasing domains), one of the most common everyday activities that virtual reality could enable users to do from anywhere, anytime. We have not tested whether the models would generalize to domains outside of the shopping experiences, thus we cannot speak to the transferability of our results to environments with very different visual properties than what our virtual environments provide. **(2)** In this dataset, we hand-design several possible dialog acts that we assume are common for human buyers, as well as their associated scenarios. This may not exhaust all the possible interactions a shopper can do with the assistant.

ACQUIRED. The limitations of our ACQUIRED resource are: **(1)** ACQUIRED focuses on the three commonsense dimensions: physical, social, and temporal. While they likely span the most common types of the reasoning technique, there could be more, *e.g.* , numerical commonsense is not specifically dealt with in this work, nor is non common activities such as fantasies and fictions involved. **(2)** The videos used in this work are subsets of readily collected ones from both Oops! (Epstein et al., 2020) and Ego4D (Grauman et al., 2022d) mother sets, and hence the event distribution can be bounded by the activities they concern. While we argue that the dataset is, to our best knowledge, first of its kind video QA dataset in terms of diversity and dedication of counterfactual reasoning, the video resources spanning even more diversified situations can be further extended. **(3)** Unlike Oops!, there is not an obvious failed actions occurred in Ego4D, and hence the annotated questions could be confounded by more imagined situations. We argue that the required reasoning technique is essentially the same and the models learn on our dataset should generalize well to situations that actually involve failing actions from egocentric visual contexts. However, we encourage future research to extend the first-person viewpoint (egocentric) parts to encompass obvious failing actions to collect just-in-time assistive questions and their corresponding remedial responses.

Object-Centric v.s. Event-Centric. The essence of an action, often involves an act (*e.g.* , the predicate of a phrase) and the objects or environments involved. In Chapter 2 and Chapter 3, the sequential understanding of a task as well as the conditional dependencies, all draw the bases off the essence of actions, where how certain objects are being manipulated or acted on drives the relevant modeling and reasoning. To generalize such essence of action, the concept or notion of event, can come into place. Event can describe an action (Hsu et al., 2021; Parekh et al., 2023), a status, or generally any kinds of time-continuous occurrences. That said, both the actionables and conditions can be thought of as events in Chapter 3, where in Chapter 2 the model is essentially relying on the commonsense in ordering certain events.

In Chapter 4, the grounding is currently object-centric, that is, the notion of tracking and localizing key entities, even though driven by action-centric symbolic knowledge, concerns mainly the objects as the main goals. This may raise an interesting research question: *would grounding make sense to be generalized to action or event-centric?* Multimodal reasoning in events have recently drawn some attentions in visual event extraction and understanding in images (Li et al., 2020b, 2022) and/or videos (Chen et al., 2021), however, the actual underlying grounding (text-to-visual regions) is still mainly concerning object-level or object-centric concepts. Extending such grounding to temporal dimension in long untrimmed videos, with finer-granularity of spatiotemporal grounding as the focus, could be a very interesting and challenging next steps for the multimodal research community. Such capability can benefit many use cases, such as the spatio-temporal reasoning and/or query presented in Ego4D (Grauman et al., 2022a) tasks.

The event-centric understanding can also be seen in our ACQUIRED in Chapter 5, where event duration, ordering, and general understanding are essential for answering visual counterfactual reasoning questions. Furthermore, an extension of work in Chapter 4 generalizing the grounded regions to actually paired with generated textual descriptions on the recognized conditions, can push the work forward combining the benefits of object and even-centric grounding. Such grounding hierarchy can also be useful to answer questions in ACQUIRED, as many physical and temporal counterfactual questions regard the details of the objects and how they undergo event alternations.

7.3 The Next Steps

In order to keep further pushing research work along the line of multimodal agent/assistant directions, in this section I will outline two main possible next steps where one focuses on building the full-stack assistive AI agent and the other will shed light on how we can improve the agent continually through our human feedback.

Grounded Instruction Generation with Embodiment. This direction can be a suitable application for the work introduced in Chapter 3, as understanding the action-condition dependencies is able to help generating more current-situation-grounded instructions, *e.g.* agent knowing that the pan is not yet heated so it is less likely to utter the next action but to stay at the current state for such pre-condition to be met. Concretely, one can train a multimodal *perceiver* module to ground the inferred textual conditions to the current situations along the agent trajectories, which can later on be used as an engineered reward for training the agent and the instructor through reinforcement learning.

A possible framework can be built on top of a recent work (Dou and Peng, 2022) that extends the speaker-follower framework of (Fried et al., 2018; Shen et al., 2019) for vision-language navigation, where the authors have proposed a framework inspired by the back-translation in neural machine translation that can augment the *speaker* module to produce better instruction to guide the *follower* agent.

For potential suitable environments, in addition to navigation-centric tasks (as exemplified by SIMMC-VR), two recently released datasets, TEACH (Padmakumar et al., 2022) and BEHAVIOR-1K (Li et al., 2023a), where they both simulate embodied AI in a house-hold (or indoor) environments that perform a specifically given task and can utter back (*i.e.* chat) to the task instructor. Inspired by (Shridhar et al., 2021), one can consider building a text-based world simulator that is primarily based off instructions in *e.g.* WikiHow, where we abstract out the actual perception and grounding part but focuses on how world state changes can affect the generated instructions. The knowledge learned from the aforementioned approach can shed light on how we can transfer the ability to grounded multimodal settings.

Human-Feedback Improvable Multimodal AI. Recent trends of large-scale training on large language models, such as GPT-4 (Achiam et al., 2023), and large vision or multimodal

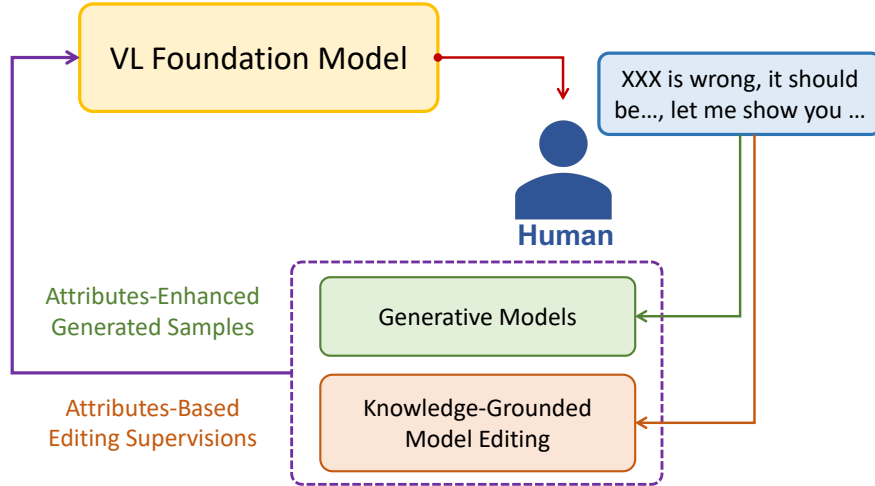


Figure 7.1: The human-feedback improvable system: is designed to comprehend and elicit effective human feedback to improve the multimodal AI models.

models (Black et al., 2024), have accumulated more and more attentions on aligning the pretrained models with human preferences to improve the model capabilities on numerous downstream tasks. This process is often referred to as the *post-training*, where unsupervised pretrained large models are tuned to align well with what humans want. Typically, the method where the models are post-trained by reinforcement learning from human feedback (RLHF) (Kaufmann et al., 2023) via proximity policy optimization algorithms (PPO) (Schulman et al., 2017), is predominant alongside several recently proposed preference optimization algorithms, such as direct preference optimization (DPO) (Rafailov et al., 2024) or Kahneman-Tversky optimization (KTO) as a human-centered loss functions (Ethayarajh et al., 2024). The reinforcement learning scheme can also be applied to training pixel-based generative models such as guided-image generations (Black et al., 2024) and large-scale multimodal models (Liu et al., 2024).

However, a notable caveat of using reinforcement learning is the curse of sample efficiency (Das et al., 2024), and furthermore, for knowledge-heavy or grounded concepts, directly editing the models (Mitchell et al., 2022) can enjoy both the efficiency as well as the effectiveness of improving the models. And hence, a promising future research direction is how to elicit and effectively utilize human feedback in more direct or sample efficiency ways. Figure 7.1 illustrates an exemplar possible instruction-tuned generative pipeline that could potentially improve a deployed vision-language-model on continually learning novel

concepts. When certain visual concepts are erroneously reasoned, human feedback can be elicited or prompted, followed by utilization of generative models or editing methodologies to correct or adapt the models with high-quality and controllable training samples (or supervisions). This is inspired by how humans learn novel concepts by mistakes and the *"show me more samples for this!"* fashion. In this way, the improvement is not only more interpretable and controllable, but also more efficient as we are not only relying on inefficient sample generations to seek useful human preference feedback.

7.4 Future Research Directions

At the time of writing this thesis, the AI research community has not been the same ever since the introduction of the large language models (LLMs). Industries, alongside some well-established research institutions have been radically competing in this space by training larger and more capable LLMs and tune them for many well-trained or emergent abilities and/or skills. While the scaling law might be true, that we could possibly always have stronger models when there is more ample of data as well as the computation powers, it still remains unclear whether what we are building, or the way we are training these models, is actually moving towards the true intelligence.

From this thesis's point of view, the human intelligence is powerful and advanced, not just because of it knows ample amount of knowledge, but owing to the fact that humans are highly adaptable when learning new tasks and knowledge, and can even create our own curriculum to teach us complex knowledge by taking one smaller step after another. Furthermore, the learning capability is not to be mixed up with generalizability, where the former is a procedure to pick up new skills possibly completely different or unseen during the training, which is also capable of absorbing external learning signals (like the feedback described previously); while the latter concerns more on the distribution shift of particular tasks, data, and/or scenarios. A future look-ahead research direction could be really unveiling how the actual learning procedure a model can take on, a true meta-learner that is able to follow or construct its own curriculum when noisy and unstructured learnable data is presented, and of course, able to supervise itself with external or intrinsic feedback.

Another futuristic direction going beyond the current trend of reliance on the parameter scaling law of multimodal learning, is to revisit how representations on individual modality

work, and how a fundamental mapping or alignment could be made in addition to simply relying on the large-scale pretraining stage. Language is fundamentally discrete (with symbols), while many other modalities are continuous, such as audio, visuals, and even sensor data streams. However, if certain key words or phrases, within a language context, can be mapped to or projected to a space where the alignment could be facilitated better. Such projection, inspired by revisiting some earlier work on aligning representations (Lample et al., 2018), can perhaps be integrated into the transformer-like to multi-layer neural models for stronger modality alignment while being trained to perform any multimodal skills. Furthermore, a pretrained symbolic-to-continuous alignment learning could possibly benefit a from-scratch trained multimodal LLMs, where no pretrained language or other modality encoder is utilized. Such from scratch pretraining perhaps is closer to how humans firstly develop our visual systems, followed-by a symbolic grounding/alignment, and lastly the full sapce of knowledge and skills are then largely scaled-up.

Lastly, multimodal learning with embodiment would make the real-life physical experiences groundable to the language and visuals, especially when it comes to more manipulative and affordance-related tasks. Performing a curricular sim-to-real transfer for the multimodal-embodied models, while directly being instruction-tuned or preference-based-learned, could be a very interesting future research agenda, in the next few years or beyond.

Bibliography

instructables.com. URL <https://www.instructables.com>.

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8M: A large-scale video classification benchmark. *arXiv preprint*, 2016.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. Pddl| the planning domain definition language. *Technical Report, Tech. Rep.*, 1998.

Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *International Conference on Computer Vision (ICCV)*, 2017.

Bengt Altenberg. Causal linking in spoken and written english. *Studia linguistica*, 38(1): 20–69, 1984.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations (ICLR)*, 2020.

- Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Mechanical, behavioural and intentional understanding of picture stories in autistic children. In *British Journal of developmental psychology*, volume 4, pages 113–125. Wiley Online Library, 1986.
- Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. First person action-object detection with egonet. In *Robotics: Science and Systems (RSS)*, 2017.
- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.
- Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. Simulating action dynamics with neural process networks. In *International Conference on Learning Representations (ICLR)*, 2018.
- S.R.K. Branavan, Nate Kushman, Tao Lei, and Regina Barzilay. Learning high-level planning from text. In *Association for Computational Linguistics (ACL)*, 2012a.
- SRK Branavan, Nate Kushman, Tao Lei, and Regina Barzilay. Learning high-level planning from text. In *Association for Computational Linguistics (ACL)*, 2012b.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018a.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018b.
- Rémi Calizzano, Malte Ostendorff, and Georg Rehm. Ordering sentences and paragraphs with pre-trained encoder-decoder transformers and pointer ensembles. In *Proceedings of the 21st ACM Symposium on Document Engineering*, pages 1–9, 2021.
- Mengyun Cao, Xiaoping Sun, and Hai Zhuge. The role of cause-effect link within scientific paper. In *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 32–39. IEEE, 2016.
- Guan-Lin Chao and Ian Lane. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

- Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. Joint multimedia event extraction from video and article. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
- Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022.
- Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. VALOR: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint*, 2023.
- Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*, 2016.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- Francisco Javier Chiyah-Garcia, Alessandro Suglia, José Lopes, Arash Eshghi, and Helen Hastie. Exploring multi-modal representations for ambiguity detection & coreference resolution in the simmc 2.0 challenge. *arXiv preprint arXiv:2202.12645*, 2022.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- Peter Clark, Bhavana Dalvi, and Niket Tandon. What happened? leveraging verbnet to predict the effects of actions in procedural text. *arXiv preprint arXiv:1804.05435*, 2018.
- Paul A. Crook, Satwik Kottur, Seungwhan Moon, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated interactive multimodal conversations (simmc) track at dstc9. *AAAI DSTC9 Workshop*, 2021.
- Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. Deep attentive sentence ordering network. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4340–4349, 2018.
- Baiyun Cui, Yingming Li, and Zhongfei Zhang. BERT-enhanced relational sentence ordering network. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6310–6320. Association for Computational Linguistics, November 2020.
- Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6297–6306, 2020.

- Bhavana Dalvi, Niket Tandon, Antoine Bosselut, Wen-tau Yih, and Peter Clark. Everything happens for a reason: Discovering the purpose of actions in procedural text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4496–4505, 2019.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.
- Dima Damen, Adriano Fragomeni, Jonathan Munro, Toby Perrett, Daniel Whettam, Michael Wray, Antonino Furnari, Giovanni Maria Farinella, and Davide Moltisanti. Epic-kitchens-100-2021 challenges report, 2021.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2022.
- Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335, 2017a.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017b.
- Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient rlhf via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.
- Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019.

- Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. *Neural Information Processing Systems (NeurIPS)*, 2021.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.
- Ziyi Dou and Nanyun Peng. A follower-aware speaker model for vision-and-language navigation. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), short*, 2022.
- Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2021.
- Xinya Du, Bhavana Dalvi Mishra, Niket Tandon, Antoine Bosselut, Wen-tau Yih, Peter Clark, and Claire Cardie. Be consistent! improving procedural text comprehension using label consistency. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.
- Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Is first person vision challenging for object tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2021.
- Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. In *International Journal of Computer Vision (IJCV)*, 2022.
- Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*, 2019.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.
- Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

- Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *International Conference on Computer Vision (ICCV)*, 2017.
- Richard E Fikes and Nils J Nilsson. Strips: A new approach to the application of theorem proving to problem solving. In *Artificial intelligence*, volume 2, pages 189–208. Elsevier, 1971.
- Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.
- Qichen Fu, Xingyu Liu, and Kris Kitani. Sequential voting with relational box fields for active object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET: End-to-end video-language transformers with masked visual-token modeling. 2021.
- Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. volume 49, pages 401–411. Elsevier, 2017.
- Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Shuyang Gao, Sanchit Agarwal Abhishek Seth and, Tagyoung Chun, and Dilek Hakkani-Ture. Dialog state tracking: A neural reading comprehension approach. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2019.
- Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *arXiv preprint arXiv:2202.13330*, 2022.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.
- Jingjing Gong, Xinchu Chen, Xipeng Qiu, and Xuanjing Huang. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*, 2016.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erappalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022c.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022d.

Chris Hadley, Katiana Uyemura, Kyle Hall, Kira Jan, Sean Volavong, and Natalie Harrington. Wikihow. URL <https://www.wikihow.com/Main-Page>.

Rujun Han, Qiang Ning, and Nanyun Peng. Joint event and temporal relation extraction with shared representations and structured prediction. In *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. Ester: A machine reading comprehension dataset for event semantic relation reasoning. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021a.

- Rujun Han, Xiang Ren, and Nanyun Peng. Econet: Effective continual pretraining of language models for event temporal reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021b.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021.
- Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272, 2014.
- Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
- Chiori Hori, Anoop Cherian, Tim K. Marks, and Florian Metze. Audio visual scene-aware dialog track in dstc8. *DSTC Track Proposal*, 2018.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*, 2020.
- I Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, et al. Degree: A data-efficient generation-based event extraction model. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2021.
- Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. Document-level entity-based extraction as template generation. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021a.
- Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2016.
- Xin Huang, Chor Seng Tan, Yan Bin Ng, Wei Shi, Kheng Hui Yeo, Ridong Jiang, and Jung Jae Kim. Joint generation and bi-encoder for situated interactive multimodal conversations. *AAAI 2021 DSTC9 Workshop*, 2021b.
- Yichen Huang, Yuchen Wang, and Yik-Cheung Tam. Uniter-based situated coreference resolution with rich multimodal input. *arXiv preprint arXiv:2112.03521*, 2021c.

- Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*, 2018.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint*, 2017.
- Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *International Conference on Computer Vision (ICCV)*, 2019.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UnifiedQA: Crossing format boundaries with a single qa system. In *Findings of EMNLP*, 2020.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
- Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*, 2019.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*, 2021.

- Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Open-images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*, 2017.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision (IJCV)*, 2017.
- Po-Nien Kung, Tse-Hsuan Yang, Chung-Cheng Chang, Hsin-Kai Hsu, Yu-Jia Liou, and Yun-Nung Chen. Multi-task learning for situated multi-domain end-to-end dialogue systems. *AAAI 2021 DSTC9 Workshop*, 2021.
- Heeyoung Kwon, Mahnaz Koupaee, Pratyush Singh, Gargi Sawhney, Anmol Shukla, Keerthi Kumar Kallur, Nathanael Chambers, and Niranjan Balasubramanian. Modeling preconditions in text with a crowd-sourced dataset. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.
- Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Association for Computational Linguistics (ACL)*, pages 545–552, 2003.
- Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. Dvd: A diagnostic dataset for multi-step reasoning in video grounded dialogue. *arXiv preprint arXiv:2101.00151*, 2021.
- Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, et al. Learning to embed multimodal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830, 2022.
- Haejun Lee, Drew A Hudson, Kangwook Lee, and Christopher D Manning. Slm: Learning a discourse language representation with sentence unshuffling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *International Conference on Computer Vision (ICCV)*, pages 667–676, 2017.
- Joosung Lee and Kijong Han. Multimodal interactions using pretrained unimodal models for simmc 2.0. *arXiv preprint arXiv:2112.05328*, 2021.

- Kenton Lee, Luheng He, and L. Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2018.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: ClipBERT for video-and-language learning via sparse sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023a.
- Chunyuan Li*, Haotian Liu*, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *Track on Datasets and Benchmarks, Neural Information Processing Systems (NeurIPS)*, 2022.
- Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022a.
- Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020a.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

- Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- 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 *Association for Computational Linguistics (ACL)*, 2020b.
- Manling Li, Ruochen Xu, Shuohang Wang, Luwei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16420–16429, 2022.
- 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. In *European Conference on Computer Vision (ECCV)*, 2014.
- Keith Vander Linden. Generating precondition expressions in instructional text. In *Association for Computational Linguistics (ACL)*, 1994.
- Ao Liu, Shuai Yuan, Chenbin Zhang, Congjian Luo, Yaqing Liao, Kun Bai, and Zenglin Xu. Multi-level multimodal transformer network for multimodal recipe comprehension. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1781–1784, 2020.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.
- Shilong Liu, Yaoyuan Liang, Feng Li, Shijia Huang, Hao Zhang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2022a.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.
- Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

- Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. Sentence ordering and coherence modeling using recurrent neural networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
- Jeff Loucks, Christina Mutschler, and Andrew N Meltzoff. Children’s representation and imitation of events: How goal organization influences 3-year-old children’s memory for action sequences. In *Cognitive Science*, volume 41, pages 1904–1933. Wiley Online Library, 2017.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019.
- Zhaoyang Lv, Edward Miller, Jeff Meissner, Luis Pesqueira, Chris Sweeney, Jing Dong, Lingni Ma, Pratik Patel, Pierre Moulon, Kiran Somasundaram, Omkar Parkhi, Yuyang Zou, Nikhil Raina, Steve Saarinen, Yusuf M Mansour, Po-Kang Huang, Zijian Wang, Anton Troynikov, Raul Mur Artal, Daniel DeTone, Daniel Barnes, Elizabeth Argall, Andrey Lobanovskiy, David Jaeyun Kim, Philippe Bouttefroy, Julian Straub, Jakob Julian Engel, Prince Gupta, Mingfei Yan, Renzo De Nardi, and Richard Newcombe. Aria pilot dataset, 2022.
- Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Ru-jun Han, and Nanyun Peng. Eventplus: A temporal event understanding pipeline. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Demonstrations Track*, 2021.
- Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. Cooking with semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 33–38, 2014.
- Bhavana Dalvi Mishra, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2018.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations (ICLR)*, 2022.
- Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*, 2020.
- Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural belief tracker: Data-driven dialogue state tracking. In *Association for Computational Linguistics (ACL)*, 2017.

- Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. MarioQA: Answering questions by watching gameplay videos. In *International Conference on Computer Vision (ICCV)*, 2017.
- Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Aldrian Obaja Muis Naoki Otani Nidhi, Vyas Ruochen Xu, and Yiming Yang Teruko Mitamura Eduard Hovy. Low-resource cross-lingual event type detection in documents via distant supervision with minimal effort. 2018.
- Qiang Ning, Hao Wu, and Dan Roth. A multi-axis annotation scheme for event temporal relations. In *Association for Computational Linguistics (ACL)*, 2018.
- Byungkook Oh, Seungmin Seo, Cheolheon Shin, Eunju Jo, and Kyong-Ho Lee. Topic-guided coherence modeling for sentence ordering by preserving global and local information. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2273–2283, 2019.
- OpenAI. Chatgpt, 2023a. URL <https://chat.openai.com>.
- OpenAI. Gpt-4 technical report, 2023b.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. 2011.
- Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 36, pages 2017–2025, 2022.
- Kuntal Kumar Pal, Kazuaki Kashihara, Pratyay Banerjee, Swaroop Mishra, Ruoyu Wang, and Chitta Baral. Constructing flow graphs from procedural cybersecurity texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://www.aclweb.org/anthology/P02-1040>.
- Tanmay Parekh, I Hsu, Kuan-Hao Huang, Kai-Wei Chang, Nanyun Peng, et al. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Association for Computational Linguistics (ACL)*, 2023.

- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 2020.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
- Barbara Plank and Željko Agić. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*, 2015.
- Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro Szekely. Corequisite: Circumstantial preconditions of common sense knowledge. In *West Coast NLP Summit (WeCNLP)*, 2021.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *2019 Conference on Empirical Methods in Natural Language Processing.*, Hongkong, China, nov 2019. Association for Computational Linguistics. URL <https://arxiv.org/pdf/1909.04076.pdf>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset.

- In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 34, pages 8689–8696, 2020.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 39(6):1137–1149, 2016.
- Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. proScript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2138–2149. Association for Computational Linguistics, 2021a.
- Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. proScript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021b.
- Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Neural Information Processing Systems (NeurIPS)*, 2017.
- Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Matteo Antonio Senese, Giuseppe Rizzo, Alberto Benincasa, and Barbara Caputo. A response retrieval approach for dialogue using a multi-attentive transformer. *AAAI 2021 DSTC9 Workshop*, 2021.
- Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*, 2018.
- Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.
- Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. pages 9866–9875, 2020b.
- Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *International Conference on Computer Vision (ICCV)*, 2019.
- Mohit Sharma and Oliver Kroemer. Relational learning for skill preconditions. In *Conference on Robot Learning (CoRL)*, 2020.

- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning. In *Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018.
- Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. Pragmatically informative text generation. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, June 2019.
- Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.
- Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255, 2019.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. Com2sense: A commonsense reasoning benchmark with complementary sentences. In *Association for Computational Linguistics (ACL)*, 2021.
- Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint*, 2020.
- Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2017. URL <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972>.
- Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.
- Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A Hearst. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, 2021.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020.

- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.
- Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Niket Tandon, Bhavana Dalvi Mishra, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. Reasoning about actions and state changes by injecting commonsense knowledge. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. Wiqa: A dataset for "what if..." reasoning over procedural text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. A dataset for tracking entities in open domain procedural text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, November 2020.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Coqui.ai. Coqui tts. <https://github.com/coqui-ai/TTS>, 2022.
- Unity. Unity. <https://unity.com/>, 2020.
- Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.
- Silvan S Tomkins. The tomkins-horn picture arrangement test. In *Transactions of the New York Academy of Sciences*, 1952.
- Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 2015.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

- Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *International Conference on Learning Representations (ICLR)*, 2016.
- J Wang, B Hu, Y Long, and Y Guan. Order matters: Shuffling sequence generation for video prediction. In *30th British Machine Vision Conference 2019, BMVC 2019*. Newcastle University, 2020.
- Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. 2021.
- Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Understanding procedural knowledge by sequencing multimodal instructional manuals. *arXiv preprint arXiv:2110.08486*, 2021.
- Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Association for Computational Linguistics (ACL)*, 2022.
- Te-Lin Wu, Caiqi Zhang, Qingyuan Hu, Alex Spangher, and Nanyun Peng. Learning action conditions from instructional manuals for instruction understanding. 2023a.
- Te-Lin Wu, Yu Zhou, and Nanyun Peng. Localizing active objects from egocentric vision with symbolic world knowledge. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.
- Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. 2017.
- Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10334–10343, 2019.
- Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9878–9888, 2021.
- Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

- Guang Yang, Manling Li, Jiajie Zhang, Xudong Lin, Shih-Fu Chang, and Heng Ji. Video event extraction via tracking visual states of arguments. In *AAAI Conference on Artificial Intelligence*, 2022. URL <https://api.semanticscholar.org/CorpusID:253264793>.
- Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikihow. *arXiv preprint arXiv:2104.05845*, 2021.
- Ziyan Yang, Kushal Kafle, Franck Deroncourt, and Vicente Ordonez. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.
- Wang Yinyin. Consumer behavior characteristics in fast fashion, 2011.
- Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQA: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. Piglet: Language grounding through neuro-symbolic interaction in a 3d world. In *Association for Computational Linguistics (ACL)*, 2021a.
- Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Neural Information Processing Systems (NeurIPS)*, 2021b.
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- Haotian Zhang*, Pengchuan Zhang*, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- Li Zhang, Qing Lyu, and Chris Callison-Burch. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, November 2020.
- Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. Temporal common sense acquisition with minimal supervision. In *Association for Computational Linguistics (ACL)*, 2020.
- Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2021.
- Yu Zhou, Sha Li, Li Manling, Lin Xudong, Shih-Fu Chang, Mohit Bansal, and Heng Ji. Non-sequential graph script induction via multimedia grounding. In *Proc. the 61th Annual Meeting of the Association for Computational Linguistics (ACL2023)*, 2023.
- Linchao Zhu and Yi Yang. ActBERT: Learning global-local video-text representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Xueyan Zou*, Zi-Yi Dou*, Jianwei Yang*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee*, and Jianfeng Gao*. Generalized decoding for pixel, image and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.