Verb Semantic Reasoning: a Semantics-Guided Approach for Improving Action Understanding in Vision–Language Models

Enjie Xu (xuenjiek15@gmail.com)[1], **Jingyi Wu**[1], **Ning Tang**[2], **Joshua K Hartshorne**[3], **Joshua B. Tenenbaum**[4], **Tao Gao (taogao@ucla.edu)**[1,2]

[1]Department of Communication, University of California, Los Angeles
[2]Department of Statistics & Data Science, University of California, Los Angeles
[3]Communication Sciences and Disorders, MGH Institute of Health Professions
[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Verbs are pivotal in human language and cognition. Psycholinguistic research has developed explicit, structured accounts of verb semantics, yet these theories have rarely been leveraged to improve models' visual action understanding. Meanwhile, current vision–language models (VLMs) often struggle with action semantics and exhibit unstable, weakly grounded judgments. Therefore, we introduce *Verb Semantic Reasoning* (VSR), a two-stage pipeline in which a Large Language Model (LLM) converts candidate action descriptions into structured event-semantic representations and a chain of semantic components questions; a VLM answers these questions over videos to select the best-supported action description. Results indicate that human performance was near ceiling, whereas both VLM baselines were substantially lower; VSR consistently improved accuracy across models and narrowed the human–model gap. These results suggest that explicit verb-semantic reasoning can significantly improve the accuracy of VLMs' action judgments, underscoring compositional, symbolic representations as an essential intermediate step for extracting semantics from the linguistic knowledge encoded in LLMs and leveraging it to guide VLMs' visual understanding.

Keywords: verb semantics; event structure; vision-language model; chain-of-thought reasoning;

Introduction

Verbs occupy a pivotal position in human language and cognition. In linguistics, a long tradition argues that verb meanings encode structured event representations (e.g., causation, motion, force dynamics) that systematically constrain argument realization and grammatical alternations (Gleitman, 1990; Hartshorne & Snedeker, 2026; R. S. Jackendoff, 2002; Levin, 1993; Pinker, 2007; Talmy, 2000). In vision, verbs are grounded in visual perceptual experience, shaped by objects' physical motion in the real world (H. Ji & Scholl, 2024; Y. Ji & Papafragou, 2020; Strickland & Scholl, 2015). Moreover, verbs may play an active role in human action planning: people can use verb semantics to anticipate how an event is likely to unfold and to evaluate candidate actions against these expectations (W. Liu et al., 2024).

In psycholinguistics, previous research has developed explicit, structured accounts of verb semantics, formalizing verb meaning as compositional event representations. In the conceptual-semantics tradition (R. Jackendoff, 1990; Pinker, 1989), verbs are not treated as unanalyzed labels. Instead, their meanings are represented as structured, compositional event descriptions built from a small inventory of primitives (e.g., `ACT`, `CAUSE`, `GO/MOVE`, `BECOME`, `CONTACT`). On this view, fine-grained action recognition is possible because viewers recover an underlying event structure from motion: they infer a limited inventory of conceptual primitives and relations (e.g., `GO` vs. `STAY`, `toward(to)` vs. `away-from`, `CAUSE` vs. `non-CAUSE`, `HAVE` change vs. no `HAVE` change) and compositionally combine them into action meanings sufficient for discrimination.

For instance, both "A pulls B" and "A pushes B" involve one entity causing another to move through contact, but the core semantics of *pull* specifies motion along a path toward the agent with distance decreasing, whereas *push* specifies motion along a path away from the agent with distance increasing. Similarly, both *hug* and *hit* involve physical contact, yet *hit* is a causal event that produces an immediate effect, whereas *hug* is a non-causal event whose central function is relation maintenance rather than outcome change.

However, most prior work has used these theories primarily to decompose linguistic semantics; while it is assumed these same structures are involved in general cognition, including the visual domain, non-linguistic theories to date have made little use of them (Dowty, 1991; Levin, 1993; Pustejovsky, 1995; Talmy, 2000). Therefore, this semantic framework has the potential to be applied directly to visual action understanding: by encouraging models to reason in terms of verb-semantic structure, we can build systems that translate perceptually available motion into discrete, composable semantic representations that can be identified, compared, and reasoned about. With this approach, many superficially similar events can be distinguished based on their underlying semantic differences.

For current vision–language models (VLMs) designed to understand visual semantics, understanding such abstract action semantics still remains a widely recognized challenge. Recent progress in action recognition increasingly leverages vision–language pretraining and multimodal alignment to help models extract semantics from visual information. A common recipe is to learn a joint embedding space for video and text via contrastive objectives (e.g., CLIP-style training; Radford et al., 2021), then cast action recognition as video–text matching or prompt-based classification (e.g., M. Wang et al., 2021), often combined with stronger temporal modeling (e.g., two-stream features, 3D CNNs, or transformer-based video

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

encoders) (Guo et al., 2025; Luo et al., 2020).

These approaches enable open-vocabulary recognition and improved transfer; however, strong benchmark accuracy does not necessarily imply robust and genuine semantic understanding. Beyond classification, recent diagnostic work suggests that VLMs can produce seemingly correct answers without stable, evidence-grounded understanding. For example, recent diagnostics show that VLMs' action judgments can be brittle: when the question or options are rephrased in semantically equivalent ways, their answers may become inconsistent (Shin et al., 2025). Moreover, their ability to determine what action or event occurred (e.g., chasing, pushing, pulling, leaving) remains limited and far from human performance (Saravanan et al., 2025). This suggests that simple video–text matching alone may be insufficient for inducing human-like, stable visual-semantic understanding in VLMs; instead, models may rely on statistical associations between superficial features in the input to make their judgments.

Given these limitations, we aim to build a system that can better understand action semantics from visual information. Our core idea is that if we can construct an action event's semantic structure by analyzing verb semantics and then use this structure to guide vision–language models' reasoning during action recognition, we may enable models to better understand events rather than merely match surface cues. In practice, we leverage the fact that modern Large Language Models (LLMs) already encode substantial knowledge relevant to verb and event semantics, making them suitable for producing structured event-semantic analyses from linguistic descriptions. Recent work at the intersection of psycholinguistics and LLMs further suggests that such models encode rich semantic information in their representations (Cai et al., 2024; T. Chen et al., 2021).

Figure 1 provides an overview of our theoretical framework. Given an action hypothesis, an LLM is prompted with instructions and examples to generate a structured semantic representation of the verb. This verb-semantic structure then serves as a scaffold: the VLM uses the resulting constraints to systematically evaluate whether the visual input supports the action hypothesis. To operationalize this idea, we first construct a lightweight verb-semantic structure inspired by the original Pinker/Jackendoff conceptual-semantics tradition, which (i) defines a small inventory of core predicates (e.g., `CAUSE`, `ACT`, `GO/MOVE`, `BECOME`, `BE`, `HAVE`, `CONTACT`) and semantic roles (Agent, Patient, Theme, Goal, Path), and (ii) provides reusable event templates together with a lexicon that maps candidate verbs to structured representations and prototypical effects.

Building on this structure, we introduce a **Verb Semantic Reasoning (VSR)** procedure that turns action discrimination into a staged, interpretable reasoning process (Figure 2). In our implementation, VSR is instantiated as a two-stage pipeline: (1) event semantic analysis, which uses the verb-semantic structure to derive a structured representation for each candidate action event and generates a chain

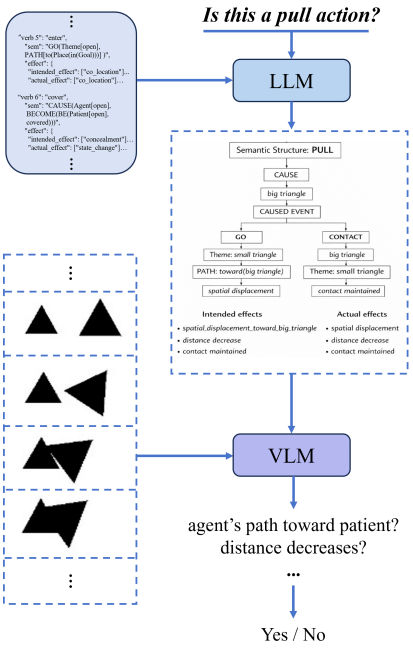

Figure 1: A semantics-guided framework for action understanding. Given an action hypothesis (e.g., *pull*), an LLM is prompted with instructions and examples to construct a verb-semantic structure. Conditioned on this structure, a VLM evaluates the visual input by checking diagnostic questions and outputs a decision (Yes/No). The semantic structure in the illustration shows an example we constructed based on Jackendoff/Pinker analyses of verb meaning.

of semantic components questions (perceptual → semantic → causal/effect → intent → decision); and (2) video reasoning, which applies the resulting question chain to videos to elicit evidence-grounded intermediate analyses and a final judgment. The key intuition is that many "similar-looking" motions for models become separable once we ask which fine-grained semantic components are licensed.

Finally, we evaluate this approach on Heider–Simmel-style (Heider & Simmel, 1944) action events from a dataset introduced by Roemmele et al., 2016. We choose this dataset because it offers a clean diagnostic of abstract action semantics: the scenes contain only moving geometric shapes, substantially reducing shortcuts based on object identity, background context, or scene-specific cues. Such Heider–Simmel-style animation datasets are commonly used as benchmarks to test whether models can infer abstract action events and social-interaction semantics from motion trajectories alone (Chemburkar et al., 2024; Maslan et al., 2015; Yun et al., 2025).

In summary, our contributions are the following:

- We construct a lightweight verb-semantic structure inspired by the original Pinker/Jackendoff conceptual-semantics tradition, and show how an LLM can be used as a tool to automatically produce structured verb-semantic decompo-

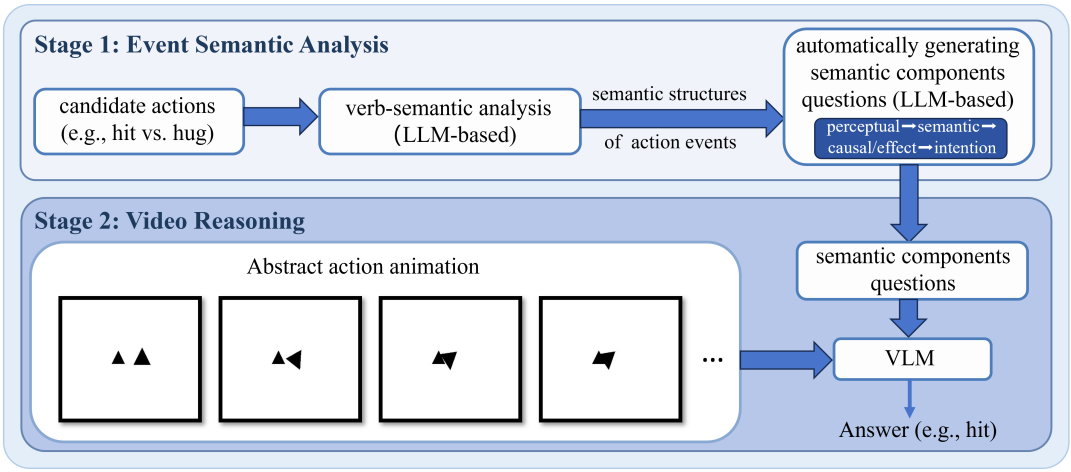

Figure 2: Verb Semantic Reasoning (VSR) pipeline. In Stage 1, an LLM (i) analyzes candidate action-event descriptions into verb-semantic structures and (ii) generates a level-wise question chain (perceptual → semantic → causal/effect → intent → decision). In Stage 2, a VLM answers the questions over abstract motion animations to select the best-supported description.

sitions by leveraging its rich semantic knowledge under our defined schema.

- Building on these decompositions, we use an LLM to automatically generate a chain of semantic components questions that operationalize verb-semantic constraints as verifiable, level-wise checks over visual evidence.
- We demonstrate that this semantics-guided question chain can effectively guide VLMs' action understanding, suggesting that Verb Semantic Reasoning (VSR) promotes deeper, verb-semantics-grounded interpretation of action events.

Related Work

Abstract concept understanding in machines Recent modeling work has increasingly sought to explore extracting explicit, structured representations for abstract concepts and reasoning. For example, *What Makes a Maze Look Like a Maze?* proposes Deep Schema Grounding (DSG), which decomposes an abstract concept into a dependency-graph schema and hierarchically grounds its components in images (Hsu et al., 2025). Related work in this vein includes logic-enhanced conceptual representations for grounding and compositional generalization (Hsu et al., 2024), structured conceptual scales for pragmatic meaning (Qiu et al., 2025), and situation-specific symbolic models for grounded theory-of-mind concepts (Ying et al., 2025). Collectively, these studies suggest that robust abstraction benefits from structured, compositional intermediate representations, rather than purely end-to-end mappings from inputs to labels. However, the work reviewed above primarily focuses on language or static images and rarely leverages psycholinguistic theories of event semantics when constructing representations; in contrast, we target abstract action semantics in dynamic videos, where models must infer theory-driven semantic structure from continuous low-level motion trajectories.

Chain-of-Thought prompting Chain-of-Thought (CoT) prompting encourages a model to generate intermediate reasoning steps before producing a final answer, and has been shown to improve performance on tasks that require multi-step inference (Wei et al., 2022). In multimodal settings, CoT-style "thought chains" have also been used to structure VLM reasoning by decomposing a complex question into a sequence of simpler, grounded subquestions or explanations (Y. Chen et al., 2024; Lu et al., 2022). Nevertheless, much prior CoT work primarily targets logical, mathematical, or procedural reasoning. In contrast, our Verb Semantic Reasoning (VSR) is designed for *semantic* reasoning: it uses linguistically motivated verb-semantic structure to ask a staged set of questions, aiming to bridge visual evidence and action concepts. Although some recent multimodal CoT prompts incorporate coarse semantic cues, they are often shallow and not explicitly anchored in a principled linguistic theory of event semantics (Y. Chen et al., 2024; Lu et al., 2022); VSR aims to fill this gap.

Ground Language in Vision In computational accounts of event understanding, grounding language in perception has also long been a core theme. For example, Siskind's ABIGAIL system argued that verb meaning can be tied to visually recoverable event structure (Siskind, 1995). This line of work suggests that language and vision are complementary rather than isolated: structured, language-like representations provide an explicit scaffold for expressing event structure, while visual perception supplies the grounding needed to recover interaction dynamics from data (Andreas et al., 2016; Johnson et al., 2017; W. Liu et al., 2024; Mann et al., 1997; Siskind, 2001, 2003; Wong et al., 2023; Ying et al., 2025). Still, despite broad recognition that language and vision are complementary rather than isolated, prior work has rarely adopted explicit semantic structure as a formalism for constructing vi-

sual representations for VLMs, which remains an important missed opportunity. There is substantial linguistic and psycholinguistic evidence for Jackendoff/Pinker-style verb and event representations (Ambridge et al., 2018; Dowty, 1991; Hartshorne et al., 2014; R. Jackendoff, 1990; Levin, 1993; McRae et al., 1998; Pinker, 1989), yet these semantic structures have seldom been leveraged as an explicit framework for visual action events. Motivated by this gap, we propose VSR to integrate such verb-semantic structure with visual evidence for more faithful event understanding and description.

Method

To study action concept reasoning, we cast the problem in a visual question answering (VQA) setting. Each instance is represented as a tuple $\langle$v, q, a$\rangle$, where v denotes the visual input, q is a natural-language question, and a is the corresponding answer. Here, a can take the form of multiple-choice selection or free-form generation. In this work, we adopt a two-option VQA format: given a prompt such as `Which sentence best describes this event?`, the model chooses between two candidate event descriptions (Option A vs. Option B) that share the same argument structure but differ in the action predicate, e.g., `X [action_1] Y` versus `X [action_2] Y`. This two-alternative VQA-style formulation is commonly used to probe models' video–language inference capabilities (J. Liu et al., 2020; Saravanan et al., 2025).

Event semantic analysis

We facilitate the large language model (LLM) to reason with Pinker/Jackendoff-style conceptual semantics by first constructing a lightweight verb-semantic structure. Specifically, the structure provides:

- a small inventory of core semantic functions (e.g., `CAUSE`, `ACT`, `GO/MOVE`, `BE`, `BECOME`, `HAVE`, `CONTACT`, `PERCEIVE`) with informal definitions;
- a typed role set for event participants (e.g., Agent, Patient, Theme, Goal, Source, Path, Instrument, Experiencer, Stimulus, Location);
- path operators (e.g., `to`, `from`, `into`, `away-from`, `along`, `toward`) and composition rules for building nested structures;

To additionally capture the consequences that a verb may plausibly induce, we augment each lexical entry with an open-world effect specification. Effects are represented with:

- `intended_effect`: an outcome implied by an agentive intention (when applicable);
- `actual_effect`: an outcome that is supported by the visual evidence;

We adopt few-shot prompting to help the LLM interpret and analyze the semantic structure of each action event. Concretely, we write the verb-semantic structure as a concise instruction document and provide 34 manually constructed worked verb examples as reference. We then use prompt-based instruction to encourage the model to internalize this specification. Conditioned on the question stem and the two candidate options, the model composes the relevant elements in our semantic structure according to each candidate action and produces a structured event-semantic analysis of the described action. Regarding effects, because the effect inventory is treated as open-world, the model may propose novel effect labels grounded in the verb's lexical meaning when existing taxonomy examples are insufficient.

To illustrate how our two-option format is translated into event semantics, Figure 3 shows a worked example contrasting *hit* and *hug*. The analysis makes explicit which semantic dimensions distinguish the candidates—for example, whether a causative event is present (`CAUSE` vs. `ACT`), what kind of interaction is involved (e.g., contact), and how the candidates differ in their intended and actual effects (e.g., collision/sudden onset vs. contact maintained/affiliation). These dimensions provide a structured basis for generating diagnostic questions in the subsequent semantic components questions construction stage.

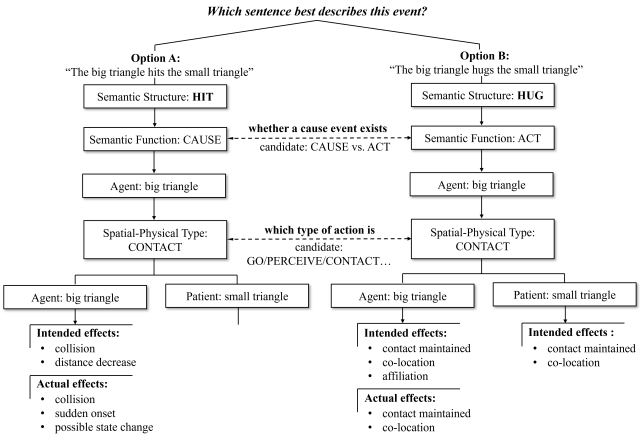

Figure 3: Example of event-semantic analysis for a two-option item (*hit* vs. *hug*). Each candidate description is decomposed into a verb-semantic structure (semantic function, participant roles) together with intended and actual effects. At the `CONTACT` level, to make clear that focus is on visual interaction properties, we represent this component as a spatial–physical type. Dashed annotations highlight key contrastive dimensions used for discrimination in this example.

Constructing the Semantic Components Questions After obtaining an event semantic representation, we further construct a chain of semantic components questions to guide how the downstream model infers the action from visual evidence. These questions are automatically generated by a prompted LLM, conditioned on the event-semantic representation produced in the previous stage. Figure 4 illustrates the intended structure: a five-level reasoning plan that will later be applied to videos, moving from *perceptual* (what is observable in the motion display) to *semantic* (which semantic components are instantiated), *causal/effect* (what causal re-

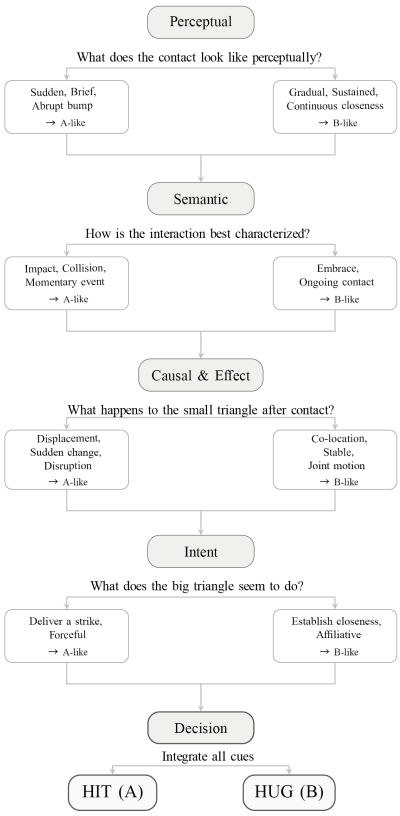

Figure 4: Simplified schematic of the semantic components questions used in our experiments for a two-option item (e.g., *hit* vs. *hug*). The question chain is automatically generated by a prompted LLM conditioned on the verb-semantic structures (Figure 3); different action-event structures yield different question chains, while the overall question-chain template (perceptual → semantic → causal/effect → intent → decision) is held fixed.

lations and outcomes are supported), *intent* (what intentions are plausible), and finally *decision* (which event description is best supported). At each level, the questions explicitly compare the two candidate events (Option A vs. Option B) along the corresponding dimensions, and the final decision stage aggregates evidence from earlier levels to select the most supported option. We emphasize that Figure 4 is a simplified schematic of the semantic components questions, capturing the core question structure and the key diagnostic features the model is prompted to analyze.

Video reasoning

We use a Heider–Simmel-style action dataset in which people create short animations of simple geometric shapes to depict a target action using a dedicated authoring interface (Roemmele et al., 2016). These animations depict interactions between two geometric shapes: a large triangle and a small triangle. From this dataset, we select 10 action types with high-quality, clearly interpretable animations (i.e., the action can be judged

Contact Actions				
Hit	Hug	Pull	Push	Throw
Pairs:				
Hit ↔ Hug				
Pull ↔ Push				
Pull ↔ Throw				
Hit ↔ Throw				
Hug ↔ Pull				
Hug ↔ Push				
No-Contact Actions				
Chase	Creep up on	Lead	Accompany	Leave
Pairs:				
Chase ↔ Creep up on				
Lead ↔ Accompany				
Accompany ↔ Leave				
Lead ↔ Leave				
Leave ↔ Creep up on				
Lead ↔ Chase				
Creep up on ↔ Accompany				

Figure 5: Selected action set and within-group action pairs used in the action-discrimination task (contact group vs. no-contact group).

directly from the display). We further categorize the selected actions into contact vs. no-contact events and pair actions within each group to be motion-similar yet semantically distinct (Figure 5). Each action type contributes four representative videos. The resulting 13 within-group pairs (contact group vs. no-contact group) define our two-option questions, and each question is evaluated on 8 videos (four depicting each action), yielding 104 trials in total.

Experiment

We used GPT-based vision–language models (GPT-4o and GPT-5) as our baseline VLMs, as they are widely used as strong baselines for evaluating chain-of-thought-style prompting and reasoning enhancements (Y. Chen et al., 2024; Hsu et al., 2025; Lu et al., 2022; X. Wang et al., 2022; Wei et al., 2022). To assess how well these baseline VLMs recognize actions in our Heider–Simmel-style dataset, we first benchmark GPT-4o and GPT-5 on our action-discrimination task and quantify their gap to human judgments. We then evaluate our VSR procedure and compare its accuracy against the baselines to test whether explicit verb-semantic reasoning reduces this gap and improves VLMs' understanding of abstract action meanings.

VLMs Implementation

We used a two-alternative choice task to assess action-discrimination ability in models. After viewing each animation, models were asked, "Which sentence best describes this event?" and chose between two options:

- Option A: "The big triangle [action 1] the small triangle."
- Option B: "The big triangle [action 2] the small triangle."

The correct answer was randomly assigned to positions A or B to control for position bias. Compared to GPT-4o, GPT-5 may engage additional deliberation via its built-in thinking

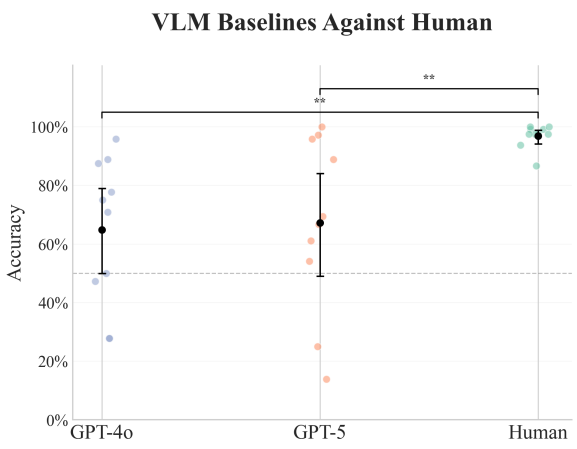

Figure 6: VLM baselines against humans in the action-discrimination task. Each dot shows accuracy for one action (10 actions total); black points and error bars summarize the mean across actions with 95% bootstrap confidence intervals.

mechanism, introducing potential confounds in baseline comparisons. To isolate the contribution of the explicit VSR reasoning process, we set GPT-5's reasoning_effort to minimal, reducing confounds from the model's internal reasoning by default. Separately from the baseline VLM evaluations, we used GPT-5 as the LLM in Stage 1 of VSR to generate verb-semantic structures and construct the chain of semantic components questions. All VLMs answer the same pre-generated questions over the same videos. During VSR inference, each level conditions on the previous level's question and answer, encouraging stepwise grounding of verb semantics before a final decision. For each model, we performed three independent runs and report results averaged across runs.

Participants Ten Zhejiang University students participated in an online experiment (Female: 5; Male: 5; mean age = 21.9). All participants received 25 RMB in compensation, and the study took approximately 50 minutes to complete. The experiment was programmed in HTML, JavaScript, CSS and displayed as a webpage with video size 640×480 pixels. All models and human participants are evaluated on the same 104 trials, using identical stimuli and the same (question, option-pair) prompts.

Results

Figure 6 summarizes performance of baseline VLMs and humans in the action-discrimination task. Humans performed near ceiling (96.92%), whereas both baseline VLMs were substantially lower (GPT-4o: 62.18%; GPT-5: 65.38%). Action-level analyses (10 actions; paired sign-flip permutation tests on per-action accuracies) showed no reliable difference between GPT-5 and GPT-4o (p = .702), but both models remained reliably below humans (GPT-4o vs. humans: p = .002; GPT-5 vs. humans: p = .006). In terms of stability across action types, human accuracy was tightly clustered near ceiling, whereas both VLMs showed markedly greater across-action variability, with several actions near or below chance (Figure 6). Overall, these results indicate a large and systematic human–model gap in action discrimination, establishing a strong baseline for testing whether VSR can reduce this gap.

Table 1 shows that VSR improves overall accuracy for both baseline models. Specifically, VSR yields absolute gains of 6.09% for GPT-4o and 6.41% for GPT-5, raising performance from 62.18% to 68.27% and from 65.38% to 71.79%, respectively. VSR also improves accuracy for both action categories: for GPT-4o, +1.39% on contact actions and +10.12% on no-contact actions; for GPT-5, +7.64% on contact actions and +5.35% on no-contact actions. Overall, VSR narrows the gap between baseline VLMs and human judgments, suggesting that explicitly guiding models to reason through verb semantics can improve action understanding.

Table 1: Overall and category accuracy by model. Parenthetical values indicate absolute improvements (Δ%) relative to the corresponding baseline model.

Model	Contact	No-Contact	Overall
GPT-4o	57.64%	66.07%	62.18%
GPT-4o+VSR	59.03% (↑ 1.39%)	76.19% (↑ 10.12%)	68.27% (↑ 6.09%)
GPT-5	65.28%	65.48%	65.38%
GPT-5+VSR	72.92% (↑ 7.64%)	70.83% (↑ 5.35%)	71.79% (↑ 6.41%)

Discussion

Contemporary LLMs possess substantial knowledge regarding verb meanings. Building on this foundation, we find that even with few-shot prompting, LLMs can reliably generate Jackendoff/Pinker-style verb-semantic structures by following a structure instruction and a handful of examples. However, our results show that VLMs do not automatically apply this knowledge to understanding action events when they are not engaged in externally guided reasoning. This indicates that training solely with standard video–text matching objectives does not guarantee that a model has formed a sufficiently semantic representation of visual motion. More broadly, it points to a limitation in current mechanisms for integrating vision and language: models may "know" verb semantics in their knowledge base, yet fail to use it as the basis for understanding visual events.

To address this limitation, our approach uses verb semantics as explicit scaffolding for action recognition. We use an LLM to automatically produce structured verb-semantic decompositions and to generate a chain of semantic components questions that operationalize these decompositions as verifiable checks over visual evidence; a VLM then answers the questions to arrive at a final action judgment. Crucially, we find that this semantics-guided question chain can reliably improve action discrimination, indicating that Verb Semantic Reasoning (VSR) effectively guides VLMs to interpret actions in terms of verb semantics rather than superficial visual cues.

Acknowledgments

This work was supported by ONR Vision Language Integration grant (PO# 951147:1) and DARPA ECOLE grant.

During the preparation of this manuscript, we used the AI-assisted tool GPT-5 to help refine and polish portions of the code used in our experimental implementation. All generated code was reviewed, edited, and validated by the authors, who take full responsibility for its correctness, integrity, and alignment with the research presented in this work.

References

Ambridge, B., Barak, L., Wonnacott, E., Bannard, C., & Sala, G. (2018). Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, *4*(1), 23. https://doi.org/10.1525/collabra.133

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 39–48. https://doi.org/10.1109/CVPR.2016.12

Cai, Z., Duan, X., Haslett, D., Wang, S., & Pickering, M. (2024). Do large language models resemble humans in language use? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 37–56.

Chemburkar, A., Feng, A., & Gordon, A. S. (2024). Evaluating vision-language models on the trianglecopa benchmark. *Preprint*.

Chen, T., O'Donnell, T. J., & Hartshorne, J. K. (2021). Verb-argument structure and semantics in contextual word embeddings [Abstract and slides]. *DUCOG: Annual Meeting of the Central European Cognitive Science Association*.

Chen, Y., Sikka, K., Cogswell, M., Ji, H., & Divakaran, A. (2024). Measuring and improving chain-of-thought reasoning in vision-language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619. https://doi.org/10.1353/lan.1991.0021

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Guo, Y., Gao, H., Yu, J., Ge, J., Han, M., & Ju, Z. (2025). Video action recognition meets vision-language models: Exploring human factors in scene interaction: A review. *Optoelectronics Letters*, *21*(10), 626–640. https://doi.org/10.1007/s11801-025-5058-9

Hartshorne, J. K., Bonial, C., & Palmer, M. (2014). The verb-corner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 397–402. https://doi.org/10.3115/v1/P14-2065

Hartshorne, J. K., & Snedeker, J. (2026). Conceptual nativism: The origins of language are in the structure of thought [PsyArXiv preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/ur3gj

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Hsu, J., Mao, J., Tenenbaum, J. B., Goodman, N. D., & Wu, J. (2025). What makes a maze look like a maze? [arXiv preprint arXiv:2409.08202]. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2409.08202

Hsu, J., Mao, J., Tenenbaum, J. B., & Wu, J. (2024). What's left? concept grounding with logic-enhanced foundation models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jackendoff, R. (1990). *Semantic structures*. MIT Press.

Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Ji, H., & Scholl, B. J. (2024). "visual verbs": Dynamic event types are extracted spontaneously during visual perception. *Journal of Experimental Psychology: General*, *153*(10), 2441–2453. https://doi.org/10.1037/xge0001636

Ji, Y., & Papafragou, A. (2020). Is there an end in sight? viewers' sensitivity to abstract event structure. *Cognition*, *197*. https://doi.org/10.1016/j.cognition.2020.104197

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Inferring and executing programs for visual reasoning. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2989–2998. https://openaccess.thecvf.com/content_iccv_2017/html/Johnson_Inferring_and_Executing_ICCV_2017_paper.html

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., & Liu, J. (2020). Violin: A large-scale dataset for video-and-language inference [arXiv preprint arXiv:2003.11618]. https://arxiv.org/abs/2003.11618

Liu, W., Chen, G., Hsu, J., Mao, J., & Wu, J. (2024). Learning planning abstractions from language [arXiv preprint arXiv:2405.03864]. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2405.03864

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, Ø., Clark, P., & Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems (NeurIPS)*.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., & Zhou, M. (2020). Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*. https://arxiv.org/abs/2002.06353

Mann, R., Jepson, A. D., & Siskind, J. M. (1997). The computational perception of scene dynamics. *Computer Vision*

and Image Understanding, *65*(2), 113–128. https://doi.org/10.1006/cviu.1996.0578

Maslan, N., Roemmele, M., & Gordon, A. S. (2015). One hundred challenge problems for logical formalizations of commonsense psychology. *Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning*, 107–113.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312. https://doi.org/10.1006/jmla.1997.2543

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT Press.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press. https://doi.org/10.7551/mitpress/3225.001.0001

Qiu, L., Zhang, C. E., Tenenbaum, J. B., Kim, Y., & Levy, R. P. (2025). On the same wavelength? evaluating pragmatic reasoning in language models across broad concepts. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 19913–19935. https://aclanthology.org/2025.emnlp-main.1008/

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (pp. 8748–8763, Vol. 139). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Roemmele, M., Gordon, A. S., Morgens, S.-M., & Morency, L.-P. (2016). Recognizing human actions in the motion trajectories of shapes. *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. https://doi.org/10.1145/2856767.2856793

Saravanan, D., Gupta, V., Singh, D., Khan, Z., Gandhi, V., & Tapaswi, M. (2025). Velociti: Benchmarking video-language compositional reasoning with strict entailment [arXiv preprint arXiv:2406.10889 (accepted to CVPR 2025)]. https://arxiv.org/abs/2406.10889

Shin, J., Lim, J., & Park, H. (2025). Do video language models really understand the video contexts? *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, 408–417.

Siskind, J. M. (1995). Grounding language in perception. *Artificial Intelligence Review*, *8*(5), 371–391.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, *15*, 31–90. https://doi.org/10.1613/jair.790

Siskind, J. M. (2003). Reconstructing force-dynamic models from video sequences. *Artificial Intelligence*, *151*(1–2), 91–154. https://doi.org/10.1016/S0004-3702(03)00112-7

Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, *144*, 570–580. https://doi.org/10.1037/a0037750

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 1). MIT Press. https://doi.org/10.7551/mitpress/6847.001.0001

Wang, M., Xing, J., & Liu, Y. (2021). Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*. https://arxiv.org/abs/2109.08472

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*. https://arxiv.org/abs/2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*. https://arxiv.org/abs/2201.11903

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*. https://arxiv.org/abs/2306.12672

Ying, L., Truong, R., Collins, K. M., Zhang, C. E., Wei, M., Brooke-Wilson, T., Tan, Z.-X., Wong, L., & Tenenbaum, J. B. (2025). Language-informed synthesis of rational agent models for grounded theory-of-mind reasoning on-the-fly. *arXiv preprint arXiv:2506.16755*. https://arxiv.org/abs/2506.16755

Yun, Y., Chen, Y.-C., Fu, S., & Lu, H. (2025). The emergence of latent force representation in human perception of social interactions. In D. Barner, N. R. Bramley, A. Ruggeri, & C. M. Walker (Eds.), *Proceedings of the 47th annual conference of the cognitive science society* (pp. 1793–1799). Cognitive Science Society. https://escholarship.org/uc/item/2hh240nk