MOLMO-SCENEGRAPH: ADDING STRUCTURE TO PRETRAINING YIELDS STRONG VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretraining on large, diverse collections of image-caption pairs is the standard method for training strong Vision-Language Models (VLMs). However, captions have several drawbacks due to their lack of consistent structure: they can be brief, ambiguous, and leave out information that annotators think is "obvious", leading to failures on downstream grounding and reasoning tasks. Motivated by human visual processing, we instead use a *structured* scene description: scene graphs. We curate PixMo-SceneGraph, a dataset of 704K image-scene-graph pairs derived solely from captions in the PixMo-Cap dataset, enabling a controlled comparison with unstructured caption data. We pretrain Molmo2 on PixMo-SceneGraph and finetune the resulting models on 8 standard academic datasets, finding that scene-graph pretraining improves performance over caption pretraining, particularly in generalizing to the grounding and reasoning tasks where caption-based models struggle. Combining scene graphs with captions yields further gains. Moreover, the structure of scene graphs also provides an interface for targeted data augmentation: we augment PixMo-SceneGraph with information typically omitted from captions—temporal statements, negations, affordances, and external knowledge—further improving performance. Our suite of models, Molmo-SceneGraph, outperform Qwen3-VL at 2B, 4B, and 8B scales across a broad range of academic, grounding, and reasoning benchmarks, despite being trained on 200–500$\times$ fewer tokens. Our results highlight the value of structure in pretraining data, and encourage researchers to rethink standard methods for VLM pretraining.

1 INTRODUCTION

Modern vision-language models (VLMs) (Radford et al., 2021; Liu et al., 2023b; Bai et al., 2023) are pretrained on large and diverse collections of image-caption pairs (Schuhmann et al., 2022). However, captions have fundamental limitations as a source of supervision: they are often brief, ambiguous, and systematically omit information that people consider "obvious", e.g., spatial relations and notions of causality (Gordon & Van Durme, 2013; Misra et al., 2016). As a result, important aspects of visual data are under-represented in pretraining corpora, limiting the ability of VLMs trained on them to perform important tasks such as grounding and reasoning (Kamath et al., 2026; Al-Tahan et al., 2024). This raises a question: *can we design supervision to encourage representations that (1) are dense and unambiguous; and (2) incorporate information omitted by captions?*

We posit that introducing *structure* into VLM pretraining data leads to better representations. We pretrain VLMs on a structured image representation: scene graphs. Scene graphs instantiate our desire for *unambiguity* by providing explicit and consistent dense supervision for grounded objects, attributes, and relations (Krishna et al., 2017; Park et al., 2025). Further, they more closely align with how humans think and perceive the world (Behrens et al., 2018; Hafri & Firestone, 2021).

We curate PixMo-SceneGraph, a dataset of 704K image-scene-graph pairs, by synthetically converting 704x captions from PixMo-Cap (Deitke et al., 2024) into scene graphs. To isolate the effect of structure, i.e., *unambiguous* information organization, the scene graphs are synthesized without access to the images. We serialize each graph as a JSON with explicit object identities, attributes, relations, and counts, and use these targets to pretrain the Molmo2 VLM (Clark et al., 2026). After

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

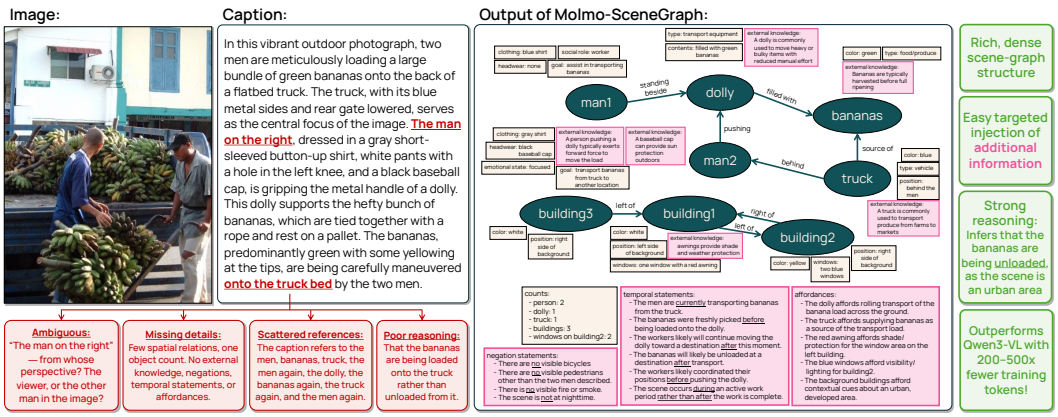


Figure 1: Molmo-SceneGraph is a powerful VLM pretrained on image-scene-graph pairs rather than image-caption pairs. The structured data allows for dense supervision and targeted data augmentation with information typically left out of captions, resulting in a model with strong reasoning capabilities that outperforms Qwen3-VL despite being trained on only 704K images.

pretraining, caption- and scene-graph-pretrained variants are post-trained with identical downstream data from 8 academic benchmarks. We call this suite of models Molmo-SceneGraph. Figure 1 shows an abbreviated scene graph output of Molmo-SceneGraph in comparison to a PixMo-style caption for the depicted image. The generated scene graph is dense, and while some relations in the caption are ambiguous ("the man on the right"—from whose perspective?), the scene graph is explicit ("building1 $\xrightarrow{\text{left of}}$ building2").

We then study the second part of our earlier question: once this structure is available, *can it serve as a structured interface for adding information that captions systematically omit?* We exploit the scene graph schema as an interface for targeted augmentation and add new fields to the pretraining data that enables Molmo-SceneGraph to generate, as seen in Figure 1, external knowledge ("Bananas are typically harvested before full ripening"), temporal inferences ("The bananas will be unloaded at a destination after transport"), negations ("There are no visible bicycles in the image"), and affordances ("The dolly affords rolling transport of the banana load across the ground")—topics that VLMs typically struggle with due to omission from pretraining corpora (Kamath et al., 2026). By systematically adding or removing these types of supervision, we can test not only whether structure helps, but which additional forms of information improve grounding and reasoning beyond the core graph: e.g., while the caption in Figure 1 incorrectly assumes that the men are loading bananas *onto* the truck, Molmo-SG's scene graph correctly reasons that the men are unloading bananas *from* the truck, as trucks are commonly used to transport produce from farms to markets (a generated external knowledge fact), and the scene takes place in an urban area (a generated affordance).

We evaluate our model on the 8 academic benchmarks it was finetuned on, 3 general-purpose benchmarks, and 4 benchmarks that test grounding and reasoning: spatial reasoning with VSR (Liu et al., 2023a), temporal inference with VisualCOMET (Park et al., 2020), Reporting Bias probes (Kamath et al., 2026), and object hallucination with POPE (Li et al., 2023). Relative to caption-only pretraining, scene-graph-only pretraining approximately preserves the standard average while slightly improving the out-of-domain average at the 2B and 4B scales. Combining graphs with captions improves these averages further, and targeted augmentation further benefits out-of-domain grounding and reasoning (Tables 1 and 2). This model exceeds the in-domain average of Qwen3-VL (Bai et al., 2025)—which uses the same text and vision backbones, but is trained on over 200x more tokens—at 2B, 4B, and 8B scales by 4.4 points, 2.9 points, and 1.7 points respectively, and its out-of-domain average at 2B and 4B by 2.0 points and 0.8 points; Qwen3-VL remains stronger out of domain at 8B by 0.6 points. Molmo-SceneGraph also outperforms LLaVA-OneVision-1.5 8B (An et al., 2025) and LLaVA-OneVision-2.0 8B (An et al., 2026), as well as Qwen3.5 9B (Team, 2026) in-domain. Further, our model outperforms the Molmo2 (Clark et al., 2026) finetuned model in-domain and out-of-domain, highlighting the value of our structured pretraining data.

Taken together, our results show that VLM pretraining benefits not only from *more* supervision, but from supervision that makes visual *structure* explicit. Structured targets preserve relationships that captions often leave implicit, provide a natural interface for adding new supervision, and improve grounding and reasoning without requiring more image-text data—outperforming models trained on substantially more data. These results suggest that the next step in VLM pretraining may be to rethink not just the scale of image-text corpora, but the form of the supervision itself. We will release all of our datasets and models in the spirit of open research.

2 Related Work

Vision-Language Model Training. Modern vision-language models (VLMs) commonly combine a pretrained vision encoder with a pretrained LLM through a learned connector, followed by multimodal instruction tuning (Liu et al., 2023b; Bai et al., 2023). The connector is typically trained to align visual representations with the LLM's language space using large-scale image-text data, often drawn from web-scale corpora such as LAION-5B (Schuhmann et al., 2022). After this, the model is finetuned on task-specific data, such as academic benchmarks and/or user inputs: e.g., LLaVA uses instruction tuning (Liu et al., 2023b); Qwen-VL uses a visual receptor and a multi-stage training pipeline with large-scale multimodal data (Bai et al., 2023); and Molmo pretrains on PixMo-Cap (Deitke et al., 2024) before finetuning on a combination of academic datasets, synthetic data, and pointing data. We use the Molmo2 architecture (Clark et al., 2026) to train Molmo-SceneGraph.

Structured Supervision for VLMs. Several works have explored incorporating structured supervision into vision-language learning to improve compositional and relational understanding. Yu et al. (2021) introduces object, attribute, and relationship prediction objectives derived from scene graphs, while Herzig et al. (2023) incorporate scene graphs into pretrained VLMs through an additional scene-graph component and fine-grained captions. More recently, Synthetic Visual Genome (Park et al., 2025) builds on Visual Genome (Krishna et al., 2017) by using language models to densely augment scene-graph annotations and train a multimodal model on the resulting data, demonstrating the value of dense relational supervision for visual reasoning. In contrast, we investigate scene graphs as the primary textual representation for VLM pretraining, rather than as auxiliary supervision or an augmentation of existing image-caption training.

Information Content and Reporting Bias in VLM Data. Beyond the form of supervision, recent work has examined how the information content of VLM training data affects downstream capabilities. Doveh et al. (2023) show that both caption quality and density are important for compositional reasoning. Similarly, the PixMo-Cap dataset (Deitke et al., 2024) collects detailed human descriptions and demonstrates the effectiveness of high-quality, dense image-text data for VLM training. However, even detailed captions can systematically omit information that people consider obvious. Kamath et al. (2026) study this phenomenon as reporting bias (Gordon & Van Durme, 2013; Misra et al., 2016) and show that such omissions can limit downstream reasoning capabilities. Our work builds on these observations by asking whether structuring caption information as scene graphs and explicitly incorporating otherwise omitted information can improve VLM performance.

3 Adding Structure to VLM Pretraining

3.1 Why Structure, and Why Scene Graphs?

Natural language captions have two primary drawbacks in representing a visual scene. First, natural language is ambiguous: e.g., "left of the woman" in an image can refer to completely different positions depending on whether the woman's perspective is considered, or the viewer of the image (Zhang et al., 2025). Second, when writing captions, people (and models trained on human annotations) exhibit a reporting bias and systematically omit "obvious" or "irrelevant" information: e.g., spatial and temporal relations, negations, and counts (Kamath et al., 2026). Both of these drawbacks can be attributed to the unstructured nature of image captions: as there is no consistent slot for each type of information, what information to include is left to convention and inference (Von Stutterheim & Klein, 1989) rather than made explicit. The promise of structured representations is that a consistent, labeled format can both resolve ambiguity and guide information inclusion.

Structured representations separate content from structure, rather than compressing both into a single linear description. Following earlier work in this space (Krishna et al., 2017), we draw inspiration from how the brain processes images. Human visual representations are not just descriptions of an image; they preserve structured information that other cognitive systems such as memory, language, and motor control can access (Cavanagh, 2011), and they separate the processing of object identity from that of spatial location and action (Mishkin et al., 1983; Goodale & Milner, 1992). Perception also captures relations between objects, either automatically as part of seeing, or computed later through deliberate reasoning (Hafri & Firestone, 2021). Visual representations keep different kinds of information distinct rather than collapsing into a single description, similar to how a scene graph separates nodes from other nodes and edges.

The brain's memory system offers further reason to select scene graphs as a structured representation. Evidence suggests that the brain uses a relational encoding format, representing relationships among elements, not just the elements themselves (Biederman et al., 1982; Eichenbaum & Cohen, 2014; Constantinescu et al., 2016; Behrens et al., 2018). Computational modeling proposes that this format enables generalization to new environments because the relational graph is learned separately from sensory content (Whittington et al., 2020). In other words, factorizing structure from content enables transfer to new environments with entirely new observations.

These findings from neuroscience and cognitive psychology motivate representations that keep the content of a scene distinct from the relationships among its elements. We apply this principle to vision-language learning by representing scenes as graphs that explicitly encode objects and their relations: e.g., "cup $\xrightarrow{\text{on}}$ table" and "book $\xrightarrow{\text{on}}$ shelf" share relational structure despite differing in content. Unlike captions, where relations must be inferred from linguistic context, scene graphs make them explicit, giving the model direct supervision on what is in a scene and how its elements relate.

3.2 Curating PixMo-SceneGraph

Having selected scene graphs as embodying the type of structure desired in our pretraining data, we now curate PixMo-SceneGraph: a corpus of 704K synthetically-generated scene graphs for VLM pretraining. We source our images and captions from PixMo-Cap (Deitke et al., 2024), which contains 712K diverse, high-quality image-caption pairs annotated by humans.

Generating scene graphs. To enable a controlled comparison between caption-based and scene-graph-based pretraining, we generate each scene graph using only the corresponding PixMo-Cap caption, *not the image itself*. This ensures that the scene graph does not contain information about the image unavailable in the caption, and prevents the scene-graph generation process from effectively distilling information from a proprietary vision-language model. We provide GPT-5.4-nano with the original caption and a detailed prompt instructing it to convert the caption into a structured scene graph in the form of a JSON dictionary. The prompt specifies the schema of the scene graph and provides strict instructions to preserve all information contained in the original caption, and is provided in the Appendix. We discard PixMo-Cap examples for which the original image is no longer retrievable from its URL.

Each scene graph consists of four primary components: a dictionary of objects with an entry for each object alongside a unique object ID, object name, and associated attributes; a dictionary of relationships between objects, referring to objects by their unique IDs; a dictionary of object counts by mapping object types to their corresponding counts; and an additional notes field capturing information from the caption that does not naturally fit into the other fields. A scene graph abbreviated for space is shown in Figure 2, with the full scene graph provided in the Appendix.

Augmenting scene graphs. The structured representation also allows us to systematically augment the scene graphs with additional information. To avoid requiring separate data-generation passes, we prompt GPT-5.4-nano to generate several types of auxiliary information alongside the core scene graph, while restricting each type to a specifically designated dictionary key, which are used during postprocessing to construct separate versions of the training data with and without the additional information. We generate external knowledge about objects (two facts per object), temporal statements about the image, negation statements, and affordances. We additionally provide the option to generate information about agents in the caption and their potential goals, as well as external knowledge about relationships between objects (c.f. Fig. 2). Postprocessing details c.f. Appendix.

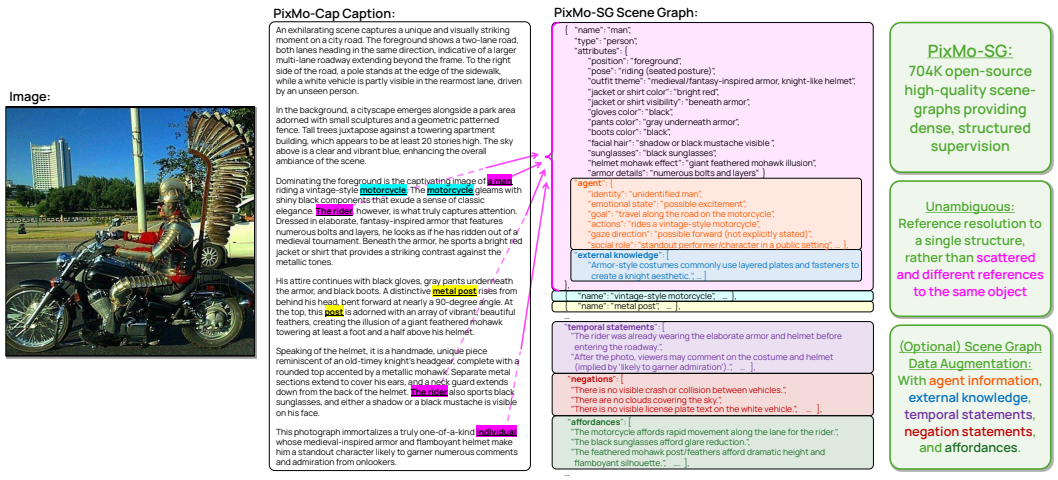


Figure 2: An example from the PixMo-SceneGraph pretraining corpus (abbreviated for space), alongside the PixMo-Cap caption. The scene graph structure allows for unambiguous reference resolution, as well as optional data augmentation of information typically omitted from captions.

3.3 PixMo-SceneGraph (PixMo-SG)

After filtering and postprocessing, we obtain 704K image-scene graph pairs: the PixMo-SceneGraph pretraining corpus. The resulting scene graphs contain an average of ∼1,200 tokens, or 2,200 tokens including augmentation, in comparison to PixMo-Cap captions of 200 tokens. Token-length distributions, including 1000 vision tokens per image, are provided in the Appendix.

4 Experiments

4.1 Model Training

Model. We follow the Molmo2 architecture (Clark et al., 2026), pairing a Qwen3 LLM (Yang et al., 2025) with a SigLIP2 So400m/14 vision encoder (Tschannen et al., 2025). Each image is encoded at 378x378 resolution as a resized global view along with up to 8 overlapping crops tiled over the image, which preserves fine-grained detail. For every image patch, we concatenate the encoder's third-to-last and ninth-to-last layer features. A connector turns these into visual tokens for the LLM. It first applies attention pooling over each 2x2 window of patches, using the window's mean feature as the query. It then projects each pooled feature into the language model's embedding space with a SwiGLU MLP. Visual tokens are inserted into the text sequence between image start and end tokens, with a separator token marking the end of each row. Training details in Appendix.

Pretraining. We pretrain on 704K images paired with scene graph (SG), captions (Cap), or both (SG+Cap), without and with augmentation (SG_{Aug}), for 4 epochs. Scene graphs are longer than captions, so to prevent them from dominating the token-level loss, we reweight each response's tokens by $2/\sqrt{n}$, where n is its number of target tokens. When an image has k annotations, each one is further downweighted by $1/\sqrt{k}$. For an image paired with both a caption and a scene graph, they are treated as independent data points rather than concatenated and conditioned on each other.

Finetuning. Each pretrained model is then finetuned identically on a mixture of academic vision-language datasets: TextVQA (Singh et al., 2019), ScienceQA (Lu et al., 2022), OKVQA (Marino et al., 2019), VQAv2 (Goyal et al., 2017), ChartQA (Masry et al., 2022), AI2-Diagram (Kembhavi et al., 2016), AOKVQA (multiple choice and direct answer) (Schwenk et al., 2022), and TallyQA (Acharya et al., 2019). Together, these cover general visual understanding, OCR, knowledge-based reasoning, counting, and chart and diagram understanding. We train for 14K steps at a batch size of 64, about 3 epochs per dataset on average. Following Molmo, each dataset is sampled in proportion to the square root of its size. TallyQA is capped at an effective size of 100K to prevent oversampling. Finetuning uses standard, unweighted token-level loss.

Table 1: **In-domain benchmark results across model scales.** Molmo-SceneGraph outperforms Qwen3-VL at all scales, LLaVA-OneVision-1.5 and -2.0 8B, Molmo2 8B, and Qwen3.5 9B.

	Text VQA	Science QA	OK VQA	VQAv2	Chart VQA	AI2D	A-OK VQA-MC	A-OK VQA-DA	TallyQA	Average
2B										
SG	87.7	95.8	73.0	90.9	75.0	91.5	85.9	69.0	80.5	83.2
SG+Cap	88.5	96.3	73.6	91.4	76.0	91.9	86.0	69.6	81.1	**83.8**
SG_{Aug}	86.9	95.7	74.2	90.7	75.0	91.3	85.2	69.1	80.7	83.2
SG_{Aug}+Cap	88.0	95.8	73.8	91.3	76.1	92.0	86.1	69.4	81.2	83.7
Cap	88.5	96.2	73.0	91.0	75.8	91.2	85.2	69.2	81.0	83.5
Qwen3-VL	88.3	87.3	63.6	87.7	80.0	88.0	83.2	58.6	77.8	79.4
4B										
SG	88.9	98.1	76.8	92.2	80.1	94.9	88.6	72.8	82.2	86.1
SG+Cap	89.9	98.0	76.9	92.6	81.0	95.3	87.7	72.5	82.6	**86.3**
SG_{Aug}	88.7	98.1	77.4	92.3	80.2	95.0	86.9	71.9	82.1	85.9
SG_{Aug}+Cap	90.0	98.0	77.5	92.7	80.8	94.7	87.6	72.4	82.8	**86.3**
Cap	89.8	98.3	76.5	92.6	80.6	94.5	88.1	72.3	82.7	86.2
Qwen3-VL	90.3	92.1	70.4	90.4	84.1	92.8	86.0	63.6	81.2	83.4
8B										
SG+Cap	90.3	98.3	77.8	92.9	82.0	95.3	87.8	73.6	83.2	86.8
SG_{Aug}+Cap	89.7	98.2	78.6	93.0	81.2	95.9	88.6	73.5	82.9	86.8
Cap	90.0	98.5	77.7	92.8	82.0	95.7	88.6	73.6	83.0	**86.9**
Qwen3-VL	90.9	94.5	74.0	91.0	85.4	94.1	88.3	65.7	81.8	85.1
Other strong VLMs										
LLAVA-OV-1.5 8B	87.7	98.7	76.0	90.5	86.2	94.2	88.9	68.1	80.2	85.6
LLAVA-OV-2 8B	86.3	98.2	73.7	89.8	86.0	94.3	88.2	63.9	79.2	84.4
Molmo2 8B	84.9	98.0	71.1	87.7	72.5	96.0	90.0	65.6	73.5	82.1
Qwen3.5 9B	91.8	95.5	73.1	91.5	86.3	94.5	90.2	67.5	82.7	85.9

4.2 Model evaluation

We evaluate Molmo-SceneGraph in two settings. The first is in-domain, on the academic datasets the models were finetuned on. The second is out-of-domain, on benchmarks unseen during training. The out-of-domain set includes general-purpose benchmarks: RealWorldQA (xAI, 2024); MM-Bench (Liu et al., 2024); and SEED-Bench (Li et al., 2024), as well as benchmarks that target grounding and reasoning: VSR (Liu et al., 2023a) tests spatial reasoning; VisualCOMET (Park et al., 2020) tests temporal reasoning; Reporting Bias (Kamath et al., 2026) has separate splits for spatial reasoning, negation, temporal reasoning, and counting; and POPE (Li et al., 2023) measures object hallucination with yes/no questions such as "Is there a banana in the image?". We report accuracy on all benchmarks except POPE, for which we report F1. We evaluate on test splits of benchmarks where available, and validation splits otherwise. Results are shown in Tables 1 and 2.

We modify two benchmarks before evaluation: for VisualCOMET, we convert the before/after event annotations into multiple choice questions; for Reporting Bias Temporal, we expand the benchmark from 200 to 568 examples to increase evaluation variance. Details are provided in the Appendix.

5 Findings

5.1 Structuring pretraining data improves generalization

Comparing the Molmo-SceneGraph models pretrained on unaugmented scene graphs alone with the models pretrained on captions alone, we find that structuring the pretraining data largely preserves in-domain performance while improving out-of-domain performance. In particular, scene-graph pretraining improves performance on previously unseen grounding and reasoning tasks, despite using no additional information beyond that contained in the original captions. The structuring of the same pretraining data may enable clearer compartmentalization of the model's skills, allowing it to better generalize to new skills; as discussed in the context of humans in Section 3.1.

Table 2: **Out-of-domain benchmark results across model scales.** Molmo-SceneGraph outperforms Qwen3-VL at 2B and 4B, particularly on grounding and reasoning benchmarks, and LLaVa-OneVision-1.5 8B and -2.0 8B and Molmo2 8B, but falls short of Qwen3-VL 8B and Qwen3.5 9B.

	RealWorld QA	MM Bench	SEED-Bench	VSR	Visual Comet	RB-Spatial	RB-Negation	RB-Temporal+	RB-Counting	POPE	Avg.
2B											
SG	67.6	82.7	76.7	78.8	45.4	90.0	29.9	80.8	88.2	88.4	72.9
SG+Cap	67.8	83.4	76.6	78.4	44.7	94.5	35.2	78.6	90.0	88.7	73.8
SG_{Aug}	67.2	83.6	76.6	77.5	56.2	95.6	31.9	85.2	86.8	88.4	<u>74.9</u>
SG_{Aug}+Cap	67.9	83.6	76.9	79.4	52.5	98.1	32.4	83.5	89.2	88.3	**75.2**
Cap	66.7	83.6	76.1	74.9	44.9	91.5	34.7	76.0	89.8	88.6	72.7
Qwen3-VL	66.4	83.7	76.1	75.0	48.4	97.1	30.0	78.7	87.4	88.9	73.2
4B											
SG	71.2	86.5	77.9	80.7	64.6	99.0	54.7	86.4	88.6	88.8	79.9
SG+Cap	71.8	87.6	78.4	82.4	58.7	99.2	53.2	90.3	89.8	89.0	80.0
SG_{Aug}	69.8	86.5	77.8	81.3	67.0	99.3	51.4	90.5	89.0	89.0	<u>80.2</u>
SG_{Aug}+Cap	72.5	87.3	78.2	82.1	66.4	99.4	52.8	90.2	89.0	88.7	**80.7**
Cap	72.5	86.6	78.5	79.3	59.4	99.4	50.1	87.3	91.0	88.7	79.3
Qwen3-VL	72.7	89.3	79.0	83.7	66.6	99.3	48.4	86.8	84.2	89.1	79.9
8B											
SG+Cap	72.2	87.4	78.7	82.6	66.6	99.0	52.1	87.9	90.3	88.4	80.5
SG_{Aug}+Cap	71.2	87.9	78.7	81.7	69.8	99.3	53.0	92.4	90.5	88.4	<u>81.3</u>
Cap	72.5	87.0	78.5	80.4	67.0	98.8	51.9	90.1	89.9	88.8	80.5
Qwen3-VL	70.1	90.0	79.1	86.0	72.8	99.5	51.9	91.5	90.1	88.1	**81.9**
Other strong VLMs											
LLAVA-OV-1.5 8B	68.6	88.8	77.0	78.2	69.2	93.7	51.2	86.8	82.2	88.2	78.4
LLAVA-OV-2 8B	69.9	88.7	78.4	79.5	71.0	98.3	48.0	89.4	86.2	87.1	79.7
Molmo2 8B	77.6	87.8	75.1	80.0	60.8	99.3	41.5	85.0	89.2	88.3	78.5
Qwen3.5 9B	77.0	91.6	80.0	85.8	73.4	99.5	60.8	92.3	87.6	89.1	**83.7**

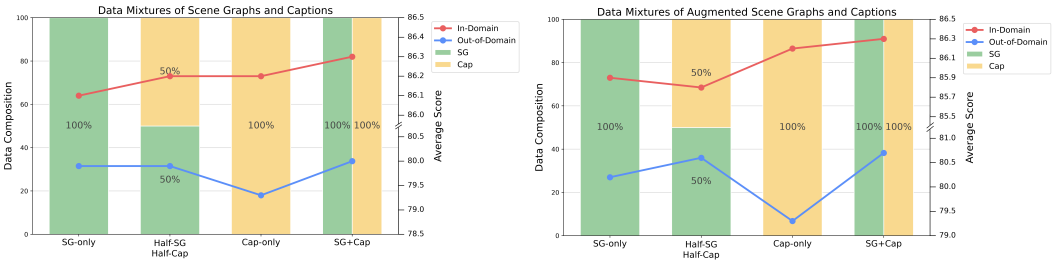
5.2 Combining structured and unstructured pretraining data improves performance

We next investigate the combination of structured and unstructured pretraining data. We find that combining the two forms of pretraining data—pairing each image with both its caption and its unaugmented scene graph—outperforms both caption-only and scene graph-only pretraining, both in-domain and out-of-domain. At 2B parameters, the combined model improves over caption-only pretraining by 0.4 points in-domain and 1.1 points out-of-domain; at 4B parameters, the corresponding improvements are 0.1 and 0.7 points. Potentially, having the captions at hand may allow the model to better parse the scene graph data, as the LLM is pretrained on natural language text, although we leave investigation of this mechanism to future work.

5.3 Augmenting the scene graph improves generalization

The structured representation also provides a convenient mechanism for incorporating information that tends to be absent in the original captions. As described in Section 3.2, we augment the scene graphs with temporal statements, negations, external knowledge, affordances, and agent information using dedicated fields in the scene-graph representation. We find that this additional information has little effect on in-domain performance, but substantially improves out-of-domain performance, particularly on grounding and reasoning tasks.

At 2B parameters, augmenting the scene graph improves the scene-graph-only model by 2.0 points overall out-of-domain and by 2.9 points on grounding and reasoning tasks. When captions and scene graphs are combined, the corresponding improvements are 1.4 and 1.9 points. At 4B parameters, the scene-graph only model improves on out-of-domain tasks after augmentation by 0.3 points overall, and 0.6 points on the grounding and reasoning tasks. The combined model improves by 0.6 and 0.9 points, respectively. At 8B parameters, where we train only the combined model, augmenting the scene graphs improves out-of-domain performance by 0.8 points overall and 1.2 points on grounding and reasoning tasks. Having open-sourced the method to generate synthetic captions, we leave scaling up the data to future work to potentially unlock greater gains at higher parameter scales.



(a) Without augmentation (b) With augmentation

Figure 3: The impact of different pretraining mixtures on Molmo-SceneGraph 4B. Mixing half of each type of data falls behind combining all of both, but could be explored for data efficiency.

These results are consistent with the reporting-bias hypothesis (Gordon & Van Durme, 2013; Misra et al., 2016; Kamath et al., 2026): information that is omitted from captions can be important for downstream grounding and reasoning, even when it is implicit or "obvious" to human annotators. Explicitly providing this information during pretraining improves performance on tasks that require it, demonstrating that the structured representation can serve not only to reorganize existing supervision but also as a flexible interface for adding targeted information.

5.4 Our models outperform Qwen3VL and other strong VLMs

We compare Molmo-SceneGraph with Qwen3-VL as both use a Qwen3 language-model backbone and SigLIP2 vision encoder, and Qwen3-VL has models at 2B, 4B and 8B. Across model scales, our models achieve substantially stronger performance than Qwen3-VL on the in-domain benchmarks, despite being trained on only 704K images. Qwen3-VL is trained through a four-stage pipeline comprising more than 2.17T image and text tokens Bai et al. (2025). While exact training data composition is not public, the authors report using data spanning knowledge, OCR, document parsing, grounding and counting, spatial understanding, science, and multimodal reasoning. Thus, its training data encompasses capabilities evaluated by both our in-domain and out-of-domain benchmarks. In contrast, Molmo-SceneGraph variants are pretrained on 3.4–9.6B tokens and fine-tuned on an additional 0.75B tokens, for a total of 4.2–10.4B tokens. Our models therefore achieve stronger performance with orders of magnitude fewer training tokens and a much smaller image corpus.

Our strongest Molmo-SceneGraph model (SG_{Aug}+Cap) outperforms Qwen3-VL by 4.4, 2.9, and 1.8 points in-domain at the 2B, 4B, and 8B scales, respectively. On the out-of-domain benchmarks, our models outperform Qwen3-VL at the 2B and 4B scales by 2.0 and 0.8 points respectively, with particularly strong gains on grounding and reasoning tasks, on which they improve over Qwen3-VL by 2.5 and 1.5 points respectively. At 8B, however, Qwen3-VL performs better on the out-of-domain average by 0.6 points, highlighting that higher-parameter models may saturate our 704K image-text pretraining corpus, and suggesting future work with a larger pretraining corpus.

Molmo-SceneGraph outperforms LLaVA-OneVision-1.5 8B, LLaVA-OneVision-2.0, and Molmo2 8B in-domain and out-of-domain, and Qwen3.5 9B in-domain. These results demonstrate that high-quality, structured pretraining data can produce strong VLMs from a relatively small image corpus. This finding is consistent with observations from Molmo Deitke et al. (2024) that carefully curated data can support strong multimodal performance without relying solely on large-scale web data.

That Molmo-SceneGraph outperforms Molmo2 8B, despite having the same architecture, further highlights the value of structured pretraining data, and PixMo-SceneGraph as pretraining corpus.

6 Analysis

6.1 Mixtures of scene graph and caption data

To further explore different combinations of structured and unstructured pretraining data, we train variants of Molmo-SceneGraph on a different type of mixture of caption and scene graph data, where 50% of the images are paired with captions alone, and the other 50% with scene graphs alone. We

consider this setting both with and without augmentation of the scene graphs. Results averaged across the in-domain and out-of-domain benchmarks are shown in Figure 3.

Without augmentation, the model trained on an equal mixture of captions and scene graphs achieves the best of both worlds in-domain and out-of-domain, matching whichever of the full-caption and full-scene graph performances is higher. Further, its results are only slightly below that of the model trained completely on both captions and scene graph data.

With additional information, the equal mixture exhibits a different pattern. Its in-domain performance is slightly below that of both the full-caption and full-scene-graph models, but its out-of-domain performance actually exceeds both. Moreover, its out-of-domain performance nearly matches that of the model trained on the full combination of caption and scene graph data.

Together, these results suggest that mixtures of caption and scene graph data may offer additional opportunities for balancing in-domain and out-of-domain performance while reducing the number of training tokens, motivating further exploration of mixture ratios and data composition. Following our earlier hypothesis that the presence of captions may facilitate initial parsing of the scene graph by the text-pretrained LLM, curriculum training approaches may be worth exploring as well.

6.2 Emergent uncertainty estimation in Molmo-SceneGraph

PixMo-SceneGraph corpus was generated by an LLM from captions without access to the image. Given the discussed ambiguity of captions, there are times that the LLM is unable to resolve certain aspects of the caption into a scene graph (e.g., unclear coreference resolution); hence, we prompt the model to include an uncertainty estimate in its scene graph generation ("if an attribute is uncertain or inferred, mark it as 'possible' or 'likely'"). This culminated in an interesting emergent behavior in the Molmo-SceneGraph model trained on this data. Molmo-SceneGraph has access to the image, but in cases where it is uncertain, it uses the same language of uncertainty, or providing instantiation for its inferences. For example, in Figure 1, the full scene graph (c.f. Appendix) states that the dolly "likely" has wheels—upon closer inspection of the image, the base of the dolly is cropped out, explaining the model's stated uncertainty despite knowing that dollies often have wheels (generated in external knowledge). Molmo-SG's emergent uncertainty estimation calls for further study.

7 Discussion and Future Work

We show that structuring VLM pretraining data can reduce the ambiguity inherent to captions, as well as serve as an interface to augment the corpus with data critical to grounding and reasoning that is typically left out of captions. We curate PixMo-SceneGraph, a dataset with 704K image-scene-graph pairs, and train Molmo-SceneGraph models, which achieve strong performance relative to Qwen3-VL and other popular models across model scales despite being trained on only a small fraction of the tokens. Our results suggest that the structure and information content of VLM pretraining data can be as important as its scale, and motivate further investigation into alternatives to the standard image-caption pretraining paradigm.

Our work raises several questions for future investigation. First, while we show that structured representations can improve VLM pretraining, it remains unclear which types of structure are most beneficial for different downstream tasks. For example, incorporating synthetic images such as screenshots into the pretraining data, together with structured descriptions of their contents, could potentially improve performance on web-agent tasks after fine-tuning on datasets such as MolmoWeb (Gupta et al., 2026), as the structured pretraining would align with the structured nature of web data. Incorporating more detailed affordance information could improve performance on embodied tasks such as ALFRED (Shridhar et al., 2020). Second, we discuss the uncertainty estimation that emerges from our model's generated scene graphs as a function of its pretraining. Making this uncertainty explicit and calibrated would be an interesting avenue for future work. Third, our scene graphs do not incorporate explicit grounding supervision, such as pointing annotations from PixMo-Points, bounding box data from Visual Genome, segmentation information from SAM3 (Carion et al., 2026), depth information, 2D/3D geometry, and tracking. Such information could be readily incorporated into the scene-graph representation by leveraging its natural flexibility to augment different types of supervision, and has strong potential to further improve VLM grounding performance.

AI USE STATEMENT

In this work, we used AI to (1) generate the scene graph training data (GPT), which we meticulously postprocessed and sampled to verify quality (samples provided in Appendix); (2) add evaluations to our codebase (Claude), which the first author verified manually; (3) polish sections of our paper after we wrote it (GPT), which the authors then took multiple passes through; (4) find any related work we missed (GPT), which the authors then searched for and read. We have NOT used AI for: (1) actual ideation; (2) writing the paper; (3) model training code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8076–8084, 2019.

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Timothy EJ Behrens, Timothy H Muller, James CR Whittington, Shirley Mark, Alon B Baram, Kimberly L Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Patrick Cavanagh. Visual cognition. *Vision research*, 51(13):1538–1551, 2011.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Alexandra O Constantinescu, Jill X O'Reilly, and Timothy EJ Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024. URL https://arxiv.org/abs/2409.17146.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.19595.

Howard Eichenbaum and Neal J Cohen. Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron*, 83(4):764–770, 2014.

Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pp. 25–30. ACM, 2013. doi: 10.1145/2509558.2509563.

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*, pp. 6904–6913, 2017.

Tanmay Gupta, Piper Wolters, Zixian Ma, Peter Sushko, Rock Yuren Pang, Yue Yang, Taira Anderson, Zhongzheng Ren, Harsh Trivedi, Winson Han, et al. Molmoweb: Open visual web agent and open data for the open web. In *European Conference on Computer Vision*, pp. 275–292. Springer, 2026.

Alon Hafri and Chaz Firestone. The perception of relations. *Trends in Cognitive Sciences*, 25(6): 475–492, 2021.

Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14077–14098, 2023. doi: 10.18653/v1/2023.emnlp-main.870. URL https://aclanthology.org/2023.emnlp-main.870/.

Amita Kamath, Jack Hessel, Khyathi Chandu, Jena D. Hwang, Kai-Wei Chang, and Ranjay Krishna. Scale can't overcome pragmatics: The impact of reporting bias on vision-language reasoning. *arXiv preprint arXiv:2602.23351*, 2026. URL https://arxiv.org/abs/2602.23351.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://link.springer.com/article/10.1007/s11263-016-0981-7.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308. IEEE, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3190–3199. IEEE, 2019.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2939, 2016.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. VisualCOMET: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3684_ECCV_2020_paper.php.

Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, Yejin Choi, and Ranjay Krishna. Synthetic Visual Genome. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2506.07643.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2210.08402.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10737–10746. IEEE, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Christiane Von Stutterheim and Wolfgang Klein. Referential movement in descriptive and narrative discourse. In *North-Holland Linguistic Series: Linguistic Variations*, volume 54, pp. 39–76. Elsevier, 1989.

James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3208–3216, 2021. doi: 10.1609/aaai.v35i4.16431. URL https://ojs.aaai.org/index.php/AAAI/article/view/16431.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025.