EVALUATING THE GAP BETWEEN VLMS AND SPECIALIZED VISION MODELS AT VISUAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite impressive improvements in general visual question-answering capabilities, Vision-Language Models (VLMs) continue to struggle at complex spatial reasoning tasks (e.g., comparing the relative positions of objects in a scene or identifying the direction of camera motion given two successive frames in a video) compared to specialized computer vision models (e.g., object detection/segmentation models). We perform a rigorous empirical evaluation to study this performance gap, focusing on two kinds of structural representations: (1) involving a single image or multiple images of unrelated scenes (e.g., bounding boxes), and (2) those involving multiple images of the same or similar scenes (e.g., correspondence points). The key challenge is that specialized models cannot directly solve realistic visual reasoning tasks. We propose to incorporate structural representations predicted by specialized models into the VLM's Chains-of-Thought (CoT), compared to incorporating the VLM's own predictions of the same structure. We find that using specialized models substantially outperforms using the VLM's own predictions at downstream reasoning tasks, rigorously establishing the limited spatial understanding of VLMs. Furthermore, we consider using supervised finetuning or reinforcement learning to improve the VLM's visual understanding capabilities; we find that in some cases these strategies can improve performance, but the gap compared to specialized models persists. Our experiments also provide insights into which kinds of structural representations are useful for improving VLM reasoning.

1 INTRODUCTION

A remarkable consequence of scaling deep learning for natural language processing (NLP) was that the resulting representations encode traditional NLP tasks (Tenney et al., 2019). With the advent of large language models (LLMs), these tasks could be solved through prompting alone without any finetuning (Radford et al., 2019; Brown et al., 2020; Qin et al., 2023; Laskar et al., 2023). With scaling, their capabilities at these tasks exceed those of any specialized model trained using finetuning, demonstrating that unsupervised pretraining alone serves as a strong foundation for NLP.

A natural question is whether the same is true for computer vision. We perform a systematic comparison of the capabilities of vision-language models (VLMs) to models trained to solve specialized computer vision tasks, including single-image tasks such as object detection and depth prediction as well as multi-image task such as camera change estimation and correspondence point prediction.

A key challenge is that traditional computer vision models typically do not directly solve realistic visual understanding problems (e.g., image question answering), and instead focus on intermediate tasks such as object detection/segmentation and depth prediction. To address this issue, we leverage the prompting capabilities of VLMs: we include predicted structure (e.g., bounding boxes, correspondence points, etc.) in the VLM prompt and ask the VLM to solve a downstream task. We then compare predicted structure from specialized models with using the VLM's own predictions. We find a substantial performance gap—even for frontier models and models targeting embodied intelligence, predictions from specialized models results in significantly stronger performance.

While prompting alone is not enough to elicit good task performance, it could be the case that the VLM's representations are such that a small amount of finetuning on task-specific data would be enough to bridge the gap. To answer this question, we apply both supervised finetuning as well as recent reinforcement learning based finetuning methods (Guo et al., 2025) to the VLM on task-specific

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.

data. We find that while this strategy can bridge performance to some degree, there continues to be a large gap between the VLM's capabilities and the specialized models.

Collectively, our results provide convincing evidence that current VLMs do not have broad capabilities in the same way as LLMs, and that significant improvements in either pre- or post-training are required to bridge this gap. Taking a broader view, our results are also consistent with Moravec's paradox (which predicts that computer vision is more challenging than NLP) (Moravec, 1988): despite the early success of deep learning at computer vision, achieving foundational computer vision capabilities turns out to be more challenging than foundational NLP capabilities.

2 Related Work

Representation-aware models. The question of whether visual representations other than standard RGB images improve model performance has previously been explored in robotics settings (Zhou et al., 2019; Chen et al., 2020), where policies have been trained end-to-end using visual features to model the perception state. However, it is difficult to utilize such intermediate representations for VLM reasoning, since even large VLMs struggle to outperform expert computer vision models on the traditional computer vision tasks best-suited to specific representations (Ramachandran et al., 2025). Thus, several approaches modify the underlying model architecture and pretraining loss functions to accommodate non-RGB visual representations. Examples include SpatialRGPT (region feature extractors that include depth information) (Cheng et al., 2024), SpatialBot (depth encodings) (Cai et al., 2025), and SpatialLLM (mixed vision encoders alongside 3D feature alignment loss functions) (Ma et al., 2025b). Alternative approaches have opted to modify tokenization approaches to include perception tokens (Bigverdi et al., 2025). However, all of these approaches train on much smaller dataset subsets than existing large VLMs and often require spending extensive amounts of time to ensure that the VLM is trained to reliably produce intermediate representations while retaining enough language capacity to solve the evaluation tasks.

Post-training methods. Numerous approaches have attempted to use SFT or reinforcement learning for learning intermediate representations to improve reasoning chains-of-thought. These techniques have been used to improve VLM performance on tasks including spatial question-answering (Sarch et al., 2025; Rui Yang, 2025; Sarch et al., 2026; Ma et al., 2025a; Hu et al., 2023), counting (Tan et al., 2025; Qi et al., 2024), STEM question-answering (Sarch et al., 2026), and OCR (Yu et al., 2025; Qi et al., 2024). However, several of these methods either 1) require extensive curation of expert reasoning traces for SFT prior to reinforcement-learning or 2) focus primarily on bounding boxes or 2D/3D object coordinates as the intermediate representation in reinforcement-learning only settings.

Agentic/tool-based approaches. Two types of methods involving utilizing tool calls for visual reasoning have shown promising results. In the first setting, the VLM is simply provided with a set of tools it can access while generating its chain-of-thought (Hu et al., 2024; Gupta and Kembhavi, 2023).In the second setting, the VLM is finetuned using reinforcement-learning on the end-to-end agent trajectories, thus learning to map question types to specific tool call trajectories (Wu et al., 2025; Wang et al., 2025a). However, these approaches primarily utilize segmentation, bounding box, cropping, and resizing tools, and do not discuss how to identify which representations should be used as tool calls while others can be distilled into the student VLM.

3 Methodology

In our work, we evaluate VLM performance under a variety of training regimes and evaluation setups. We first measure "base" VLM performance by evaluating the out-of-the-box VLM on our test set. Second, we finetune the VLM directly on a training set of question-answer pairs using Group Relative Policy Optimization (GRPO) Guo et al. (2025) with only an answer-scoring reward function R_{QA}. Thirdly, we obtain "teacher" representations from specialized computer vision models corresponding to each question, and utilize them in four ways: 1) through supervised finetuning on the training set, 2) by directly prompting the base VLM to answer the question with the representation in the prompt, 3) by training a copy of the VLM to produce the teacher representation using GRPO with a *representation reward* R_{rep} (so that it can serve as an expert teacher to a separate base copy), and 4) by training the VLM with GRPO to output both the representation *and* the answer to the question using a *mixed reward* R_{mix}.

We introduce the two representation families we study (per-image and scene-level relational representations) and how we construct their training sets in Section 3.1). Then, we describe the general reward formulation for the mixed-reward setting in Section 3.2. Finally, we introduce the representation rewards for per-image representations in Section 3.3 and scene-level relational representations in Section 3.4.

3.1 Representation Families

We distinguish between two families of learned representations. The first family consists of three unary **per-image representations** derived from MIMIC-IT (Li et al., 2025): *bounding boxes, 3D pose estimates, and object depth ordering*. These representations are computed independently for each image. Importantly, this does not restrict them to single-image tasks. For single-image QA, one image yields one structured representation. For paired-image or multi-view QA, each image yields its own structured representation, and the model may compare those per-image representations downstream in order to answer relational questions. For training our models on these representations, we construct a training set comprising 5000 samples from the MIMIC-IT dataset Li et al. (2025), with 3000 two-image question-answer pairs taken from the COCO General Difference subset, and 2000 single-image question-answer pairs taken from its LLaVA-Conversation subset. For answer scoring on MIMIC-IT tasks, we use a semantic GPT-5 judge directly:

$$R_{\mathrm{QA}} = J_{\mathrm{GPT}}(q, a, \hat{a}), \tag{1}$$

where a is the reference answer, $\hat{a}$ is the model answer, and $J_{\mathrm{GPT}} \in [0, 1]$ is a GPT-5 judge score for semantic equivalence.

The second family consists of three **scene-level relational representations**: *correspondence grids, camera pose, and semantic correspondence*. Instead of being attached to individual images, these representations encode relationships across views within the same scene, such as point correspondences, relative viewpoint change, or semantic alignment. A single image may therefore participate in structured relationships with one or many other images in the scene. For training our models on these representations, we construct a training set of 5000 samples containing frame pairs from 70 scenes from ScanNet (Dai et al., 2017), which are reformulated as question-answer pairs about relative motion between the frames. Our frame pair selection process involves selecting frames ≤ 240 indices and ≤ 4 seconds apart. These pairs are included only if a depth-reprojected anchor point moves 40–180 pixels with reprojection error ≤ 0.15m. For the ScanNet grid and semantic correspondence, the QA term is the left/right camera-motion label precomputed with VGGT (Wang et al., 2025b):

$$R_{\mathrm{QA}} = \mathbf{1}[\hat{m}_{\mathrm{lr}} = m_{\mathrm{lr}}]. \tag{2}$$

3.2 General Mixed Reward Formulation

In the mixed-reward setting, the objective consists of three parts: a small format-validity term, a representation-distillation term, and a question-answering term. The format term rewards valid task-specific output structure, the representation term rewards agreement with a teacher-produced intermediate representation, and the QA term rewards answering the original question correctly.

We train on six intermediate visual representations for each training regime: three on MIMIC-IT training dataset (bounding boxes, 3D pose, and object depth ordering) and three on the ScanNet training dataset (correspondence grids, camera motion, and semantic correspondence). Except for camera motion, each task uses the same top-level mixed reward decomposition:

$$R_{\mathrm{mix}}(x, y) = 0.10 R_{\mathrm{fmt}}(y) + R_{\mathrm{rep}}(x, y) + 0.25 R_{\mathrm{QA}}(x, y), \tag{3}$$

where x is the example, y is the model output, R_{fmt} measures whether the output follows the required schema, R_{rep} measures the quality of the intermediate representation, and R_{QA} measures final answer agreement. The representation contribution is itself a weighted sum of representation-specific subterms whose weights sum to 0.65. Equivalently, one can write $R_{\mathrm{rep}} = 0.65 \bar{R}_{\mathrm{rep}}$ for a normalized intermediate score $\bar{R}_{\mathrm{rep}} \in [0, 1]$. This shared weighting keeps the representation reward dominant while preserving pressure to answer the original question. Camera motion is the exception: its representation is already the answer to the left/right/up/down motion question, so we use direct per-axis classification accuracy instead of adding a separate QA term.

3.3 PER-IMAGE REPRESENTATION REWARDS

In this section, we introduce the representation rewards for our three per-image representations: bounding boxes, 3D pose estimates, and object depth ordering.

Bounding boxes. For bounding boxes, we use GPT-5.4 to first select one to five question-relevant objects per image. GroundingDINO (Liu et al., 2025) then provides teacher boxes for these selected objects. The student emits a set of object aliases and normalized boxes $\hat{\mathcal{B}} = \{(\hat{a}_j, \hat{b}_j, \hat{i}_j)\}$, where $\hat{a}_j$ is an object alias, $\hat{b}_j \in [0,1]^4$ is an $xyxy$ box, and $\hat{i}_j$ is the image index. The teacher set is $\mathcal{B} = \{(a_k, b_k, i_k)\}$.

We score box geometry with generalized IoU. For two boxes,

$$\mathrm{GIoU}(b, \hat{b}) = \mathrm{IoU}(b, \hat{b}) - \frac{|C \setminus (b \cup \hat{b})|}{|C|}, \qquad s_{\mathrm{giou}}(b, \hat{b}) = \mathrm{clip}\left(\frac{\mathrm{GIoU}(b, \hat{b}) + 1}{2}, 0, 1\right), \tag{4}$$

where C is the smallest enclosing box. Geometry is semantic-gated by alias agreement:

$$g(a, \hat{a}) = \mathrm{F1}_{\mathrm{tok}}(a, \hat{a}), \qquad s_{\mathrm{sem}}((a, b), (\hat{a}, \hat{b})) = g(a, \hat{a})\, s_{\mathrm{giou}}(b, \hat{b}). \tag{5}$$

Let M be the greedy one-to-one matching between teacher and student boxes within the same image. The bounding-box representation reward is

$$R_{\mathrm{rep}}^{\mathrm{bbox}} = 0.40 R_{\mathrm{geom}} + 0.25 R_{\mathrm{alias}}, \tag{6}$$

with the top-level 0.65 representation mass split into geometry and alias terms. The geometry term combines best semantic GIoU, coverage of teacher boxes, and a count penalty:

$$\begin{aligned} R_{\mathrm{geom}} = \mathrm{clip}\Big(& 0.75\, \overline{s}_{\mathrm{sem}} \\ & + 0.15\, \frac{|\{k : \exists j,\ (k,j) \in M,\ \mathrm{GIoU}(b_k, \hat{b}_j) \geq \tau\}|}{|\mathcal{B}|} \\ & + 0.10 \left(1 - \frac{||\mathcal{B}| - |\hat{\mathcal{B}}||}{|\mathcal{B}|}\right)_+ \\ & - \lambda_{\mathrm{extra}} (|\hat{\mathcal{B}}| - |\mathcal{B}|)_+ \Big), \end{aligned} \tag{7}$$

and

$$R_{\mathrm{alias}} = \frac{1}{|M_\tau|} \sum_{(k,j) \in M_\tau} \mathrm{F1}_{\mathrm{tok}}(a_k, \hat{a}_j), \tag{8}$$

where M_τ contains matched pairs whose raw GIoU exceeds the matching threshold.

3D pose estimates. For 3D pose, the target representation is an object-centric coarse 3D spatial state. We first identify question-relevant objects, localize them with GroundingDINO boxes, refine the object evidence with SAM 2 (Ravi et al., 2025) masks, and use the object-level depth ordering induced by Marigold (Ke et al., 2024) to assign coarse depth bins. Then, we obtain a coarse view/orientation label from a crop-level CLIP (Radford et al., 2021) orientation classifier.

For each teacher object, the target is

$$\mathbf{o}_k = (a_k, i_k, x_k, y_k, z_k, r_k), \tag{9}$$

where a_k is the object alias and i_k is the image index. The x coordinate is a coarse horizontal bin, $x_k \in \{\mathrm{left}, \mathrm{center}, \mathrm{right}\}$, obtained from the object box center. The y coordinate is a coarse vertical bin, $y_k \in \{\mathrm{upper}, \mathrm{middle}, \mathrm{lower}\}$, also obtained from the box center. The z coordinate is a coarse depth bin, $z_k \in \{\mathrm{foreground}, \mathrm{midground}, \mathrm{background}\}$, obtained by ranking the object-level depth estimates within the image. Finally, r_k is a coarse orientation label from the crop-level orientation classifier. The student is trained to emit the same tuple for each relevant object.

Student objects are matched to teacher objects in the same image using alias token F1. A match is valid when the alias score exceeds threshold τ_{alias}. Let M be the set of valid matches. Object-selection

quality is

$$R_{\text{sel}} = 0.50\frac{|M|}{|\mathcal{O}|} + 0.35\frac{1}{|M|}\sum_{(k,j)\in M}\text{F1}_{\text{tok}}(a_k, \hat{a}_j) + 0.15\left(1 - \frac{||\mathcal{O}| - |\hat{\mathcal{O}}||}{|\mathcal{O}|}\right)_+, \tag{10}$$

where $\mathcal{O}$ and $\hat{\mathcal{O}}$ are the teacher and student object sets. For matched objects, the discrete field accuracies are

$$R_x = \frac{1}{|M|}\sum_{(k,j)\in M}\mathbf{1}[x_k = \hat{x}_j], \qquad R_y = \frac{1}{|M|}\sum_{(k,j)\in M}\mathbf{1}[y_k = \hat{y}_j],$$
$$R_z = \frac{1}{|M|}\sum_{(k,j)\in M}\mathbf{1}[z_k = \hat{z}_j], \qquad R_r = \frac{1}{|M|}\sum_{(k,j)\in M}\mathbf{1}[r_k = \hat{r}_j]. \tag{11}$$

The 3D-pose representation reward is

$$R^{\text{3D}}_{\text{rep}} = 0.65\Big[0.40R_{\text{sel}} + 0.60\big(0.20R_x + 0.20R_y + 0.30R_z + 0.30R_r\big)\Big]. \tag{12}$$

Object depth ordering. For object depth ordering, we localize teacher-selected objects using GroundingDINO, refine masks using SAM 2, and generate a depth map using Marigold. For object o_k with mask M_k, we sample points $P_k \subset M_k$ and compute a scalar teacher depth

$$d_k = \text{median}_{p\in P_k} D(p). \tag{13}$$

Objects in each image are sorted by d_k to form a teacher near-to-far ordering π_i. The student emits relevant object aliases and an ordering $\hat{\pi}_i$.

The depth representation reward is

$$R^{\text{depth}}_{\text{rep}} = 0.25R_{\text{obj}} + 0.40R_{\text{ord}}, \tag{14}$$

again summing to the shared 0.65 representation mass. Object agreement combines alias quality, teacher-object coverage, and count agreement:

$$R_{\text{obj}} = 0.55\,\overline{\text{F1}}_{\text{alias}} + 0.35\,\text{coverage} + 0.10\left(1 - \frac{||\mathcal{O}| - |\hat{\mathcal{O}}||}{|\mathcal{O}|}\right)_+. \tag{15}$$

For ordering, predicted aliases are first mapped to teacher aliases by token F1. The pairwise order score for image i is Kendall-style:

$$K_i = \frac{\sum_{a\prec_{\pi_i} b}\mathbf{1}[a, b \in \hat{\pi}_i]\mathbf{1}[a \prec_{\hat{\pi}_i} b]}{\binom{|\pi_i|}{2}}, \qquad C_i = \frac{|\pi_i \cap \hat{\pi}_i|}{|\pi_i|}. \tag{16}$$

Then

$$R_{\text{ord}} = \frac{1}{N}\sum_i (0.75K_i + 0.25C_i). \tag{17}$$

3.4 Scene-Level Relational Representation Rewards

In this section, we introduce the representation rewards for our three scene-level relational representations: correspondence grids, camera pose, and semantic correspondence.

Correspondence grids. For correspondence grids, the model sees two ordered ScanNet frames and must choose one to three relevant objects on its own. For each selected object, it emits a triple

$$(\hat{a}_j, \hat{u}_j, \hat{v}_j), \qquad j = 1, \ldots, K, \quad K \in \{1, 2, 3\}, \tag{18}$$

where $\hat{a}_j$ is the selected object alias, $\hat{u}_j$ is a grid cell in frame 1 that lies on the object, and $\hat{v}_j$ is the predicted corresponding grid cell in frame 2. The grid is a fixed $G \times G$ partition of each image; in our

experiments $G = 6$. For reward computation, the teacher provides a set of salient object aliases and SAM 2 masks in frame 1, but the model output itself contains the selected objects. MAST3R (Leroy et al., 2025) is run offline on each image pair to map every frame-1 grid-cell center to a corresponding frame-2 grid cell, so no MAST3R call is needed during GRPO.

The representation reward averages the per-object scores over the K selected objects:

$$R_{\text{rep}}^{\text{grid}} = 0.20R_{\text{obj}} + 0.20R_{\text{mask}} + 0.25R_{\text{corr}}. \tag{19}$$

For each selected object, object agreement is the best token-F1 match to the teacher shortlist,

$$R_{\text{obj}} = \frac{1}{K}\sum_{j=1}^{K} \max_{a\in\mathcal{A}} \text{F1}_{\text{tok}}(\hat{a}_j, a). \tag{20}$$

After matching $\hat{a}_j$ to a teacher object, the frame-1 grounding term checks whether cell $\hat{u}_j$ lies on that object's mask. If the cell center is inside the mask the score is one; otherwise the score is a clipped mask-coverage credit:

$$R_{\text{mask}} = \frac{1}{K}\sum_{j=1}^{K} \begin{cases} 1, & \text{center}(\hat{u}_j) \in M_{\hat{a}_j}, \\ 0.75\,\text{clip}\left(\frac{|M_{\hat{a}_j}\cap\hat{u}_j|/|\hat{u}_j|}{\rho}, 0, 1\right), & \text{otherwise.} \end{cases} \tag{21}$$

The correspondence term compares the predicted frame-2 cell $\hat{v}_j$ to the offline MAST3R target cell $v^*(\hat{u}_j)$ for the chosen frame-1 cell:

$$R_{\text{corr}} = \frac{1}{K}\sum_{j=1}^{K} \text{clip}\left(1 - \frac{\|g(\hat{v}_j) - g(v^*(\hat{u}_j))\|_2}{\sqrt{2}\,(G-1)}, 0, 1\right), \tag{22}$$

where $g(\cdot)$ converts a grid-cell label to row-column coordinates. The QA term for this ScanNet task is the same precomputed left/right camera-motion label used by the semantic correspondence task.

Camera motion. For camera motion, the target is the egocentric motion from image 1 to image 2. VGGT estimates relative pose (R, t). Translation is discretized into forward/backward, left/right, and up/down labels, and rotation is discretized into yaw, pitch, and roll labels:

$$\ell = (\ell_{\text{fb}}, \ell_{\text{lr}}, \ell_{\text{ud}}, \ell_{\text{yaw}}, \ell_{\text{pitch}}, \ell_{\text{roll}}). \tag{23}$$

Each axis label is ternary, e.g. $\ell_{\text{lr}} \in \{\text{left}, \text{right}, \text{none}\}$. Translation labels are obtained from the relative translation vector in camera-1 coordinates; rotation labels are obtained from Euler-style yaw, pitch, and roll angles. Small components below fixed thresholds are assigned to `none`.

Because this representation is itself the answer, the reward is direct per-axis classification accuracy:

$$R_{\text{motion}} = \frac{1}{6}\sum_{a\in\mathcal{A}} \mathbf{1}[\hat{\ell}_a = \ell_a], \qquad \mathcal{A} = \{\text{fb}, \text{lr}, \text{ud}, \text{yaw}, \text{pitch}, \text{roll}\}. \tag{24}$$

Malformed or missing axis labels receive a small negative penalty rather than a separate format reward.

Semantic correspondence. For semantic correspondence, the model again selects one to three relevant objects on its own from the two ordered ScanNet frames. For each selected object, it emits a region-level box in frame 1 and a corresponding region-level box in frame 2:

$$\hat{y} = \{(\hat{a}_j, \hat{b}_j^{(1)}, \hat{b}_j^{(2)})\}_{j=1}^{K}, \qquad K \in \{1, 2, 3\}. \tag{25}$$

Here $\hat{a}_j$ is the model-selected object alias, $\hat{b}_j^{(1)}$ is the selected region in frame 1, and $\hat{b}_j^{(2)}$ is the predicted corresponding region in frame 2. For reward computation, the teacher provides salient object names and SAM 2 masks in frame 1, but these are not given as a fixed answer list; the model output itself determines which objects are selected. The image-2 target is not precomputed. Instead, during reward computation we compare the DINOv2 (Oquab et al., 2024) feature of the proposed frame-1 crop with the DINOv2 feature of the proposed frame-2 crop.

The representation reward is

$$R_{\text{rep}}^{\text{sem}} = 0.20R_{\text{obj}} + 0.20R_{\text{ground}} + 0.25R_{\text{DINO}}. \tag{26}$$

Object agreement is the average best token-F1 match between the model-selected aliases and the teacher object set:

$$R_{\text{obj}} = \frac{1}{K}\sum_{j=1}^{K} \max_{a\in\mathcal{A}} \text{F1}_{\text{tok}}(\hat{a}_j, a). \tag{27}$$

Matches below the alias threshold receive no grounding credit. The frame-1 grounding score compares the predicted box with the matched teacher SAM mask using an F1-style mask-box overlap:

$$P_j = \frac{|M_j \cap \hat{b}_j^{(1)}|}{|\hat{b}_j^{(1)}|}, \qquad Q_j = \frac{|M_j \cap \hat{b}_j^{(1)}|}{|M_j|}, \qquad R_{\text{ground}} = \frac{1}{K}\sum_j \frac{2P_jQ_j}{P_j+Q_j}. \tag{28}$$

The correspondence term uses DINOv2 crop embeddings:

$$f_j^{(1)} = \frac{\phi(I_1[\hat{b}_j^{(1)}])}{\|\phi(I_1[\hat{b}_j^{(1)}])\|_2}, \qquad f_j^{(2)} = \frac{\phi(I_2[\hat{b}_j^{(2)}])}{\|\phi(I_2[\hat{b}_j^{(2)}])\|_2}, \tag{29}$$

$$R_{\text{DINO}} = \frac{1}{K}\sum_j \text{clip}\left(\frac{1+\langle f_j^{(1)}, f_j^{(2)}\rangle}{2}, 0, 1\right). \tag{30}$$

We keep the same top-level $0.10/0.65/0.25$ weighting here. The DINO cosine is mapped from $[-1, 1]$ to $[0, 1]$ to be numerically compatible with the other reward terms.

4 Experiments

4.1 Experimental Setup

Evaluation Test Set. We evaluate models on 410 questions from three benchmarks for per-image representation evaluations: BLINK (Fu et al., 2025), ERQA (Team et al., 2025), and SAT (Ray et al., 2025). For BLINK, we use 120 stratified examples and restrict evaluation to questions with at most three images. For ERQA, we use 150 stratified examples across categories. We also evaluate on the full SAT holdout, which contains 150 intended examples.

For our scene-level relational representations, we evaluate on 300 two-image questions, 100 each from (a) BLINK's multi-view reasoning task subset (Fu et al., 2025), (b) Visual Spatial Tuning (VST) (Rui Yang, 2025), and (c) SPAR-Bench-Tiny from SPAR-7M (Zhang et al., 2025). We provide examples of evaluation questions in Figure 1 from each of the representation families.

Training and Prompting Regimes. We study a common set of training and evaluation regimes across models and representations. **Base** refers to off-the-shelf pretrained VLMs evaluated directly on the target question-answering tasks without any task-specific adaptation. **PureQA** refers to GRPO with a question-answering reward only, without any explicit intermediate representation reward. **SFT** refers to supervised fine-tuning on teacher-produced structured outputs. **Direct Teacher** is a prompted teacher baseline, where the teacher model's output is provided to the student VLM in a prompt. **Pure Distillation Reward** refers to a small model trained only on a representation-specific teacher reward before being used as a direct teacher model for an *untrained* base model. Finally, **Mixed Reward** combines structured representation supervision with the original QA objective. We provide our prompts in Appendix A.

4.2 Results and Discussion

We present results in Table 4.2 for all our training and prompting regimes. We evaluate three models on all six regimes: SmolVLM2-2.2B, Qwen3VL-4B, and Qwen3VL-8B. Additionally, to compare the differences in their regime performance to larger models, we evaluate four state-of-the-art VLMs: GPT-5.4, GPT-6 Astra, Gemini-3.8-Flash, and Gemini-ER-1.6 on the same tasks in the Base and Direct Teacher regimes.

Using specialized computer vision models as direct teacher models provides the largest performance increases on *per-image representation testing*, which usually cannot be recovered using reinforcement learning, distillation or SFT. The performance increases for most models evaluated

	Model	B. B.	3D Pose	Depth Ord.	Corr. Grid	Cam. Pose	Sem. Corr.
SmolVLM2-2.2B	Base	30.50%	30.50%	30.50%	23.33%	23.33%	**23.33%**
	PureQA	30.50%	30.50%	30.50%	15.00%	15.00%	15.00%
	SFT	33.90%	26.67%	30.98%	10.67%	23.00%	14.00%
	Direct Teacher	**33.66%**	35.12%	**39.75%**	**25.67%**	**23.67%**	16.33%
	Pure Dist. Reward	31.95%	35.37%	38.05%	11.67%	19.33%	21.33%
	Mixed Reward	32.93%	**37.07%**	37.32%	17.33%	21.33%	10.67%
Qwen3VL-4B	Base	52.68%	52.68%	52.68%	**30.33%**	**30.33%**	30.33%
	PureQA	57.80%	**57.80%**	57.80%	19.67%	19.67%	19.67%
	SFT	46.10%	50.00%	42.68%	29.00%	16.67%	29.33%
	Direct Teacher	**61.22%**	52.68%	**66.34%**	24.67%	27.67%	29.67%
	Pure Dist. Reward	54.39%	53.90%	53.17%	19.33%	19.67%	00.00%
	Mixed Reward	59.51%	36.34%	26.34%	22.67%	23.0%	**30.67%**
Qwen3VL-8B	Base	50.73%	50.73%	50.73%	33.00%	33.00%	33.00%
	PureQA	53.41%	53.41%	53.41%	22.33%	22.33%	22.33%
	SFT	53.90%	53.65%	53.65%	17.33%	18.67%	31.00%
	Direct Teacher	**63.41%**	**61.71%**	**63.66%**	**36.67%**	**46.00%**	27.67%
	Pure Dist. Reward	55.61%	52.12%	55.61%	18.33%	18.67%	31.33%
	Mixed Reward	57.56%	51.71%	55.12%	20.00%	22.67%	**34.00%**
GPT 5.4	Base	65.12%	65.12%	65.12%	30.00%	30.00%	30.00%
	Direct Teacher	**70.73%**	**71.71%**	**72.19%**	**32.67%**	**33.33%**	**33.00%**
GPT-6 Astra	Base	75.61%	75.61%	75.61%	**33.33%**	**33.33%**	**33.33%**
	Direct Teacher	**81.95%**	**80.00%**	**80.98%**	26.33%	26.67%	28.67%
Gemini-3.8-Flash	Base	48.78%	48.78%	48.78%	8.33%	8.33%	8.33%
	Direct Teacher	**60.24%**	**56.10%**	**58.05%**	**13.00%**	**10.00%**	**10.33%**
Gemini-ER-1.6	Base	70.73%	70.73%	70.73%	28.00%	**28.00%**	28.00%
	Direct Teacher	**77.07%**	**74.14%**	**74.14%**	**30.33%**	26.33%	**28.67%**

Table 1: Experimental results on all visual representations. We note that Base and PureQA numbers for each model are constant across all representations evaluated on the same test set, as those regimes are representation-agnostic. For models evaluated on more than two regimes, we mark the best results for each representation in **bold** and underline the second-best results.

using per-image representations is highest in the Direct Teacher setting (the only exceptions being SmolVLM2-2.2B and Qwen3VL-4B while using the 3D Pose Estimates representation) . While we observe an increase in performance on using pure distillation reward and mixed-reward reinforcement learning in several cases, they usually do not manage to recover the performance increases of the direct teacher setting. Thus, we can conclude that VLMs are capable of reasoning using these representations when provided, and thus augmenting VLM prompts with specialized computer vision models that produce these representations can help improve their reasoning performance.

Using specialized computer vision models as direct teacher models on *scene-level relational representation testing* results in similar or worse performance as the no-representation base setting for smaller models, but 8B and larger models can successfully leverage these representations to increase performance. Generally, we observe that for scene-level representations, it is difficult for smaller models to either use or learn to use these representations during evaluation. However, for 8B and larger models, we observe performance increases in most direct teacher settings, especially on using camera pose for Qwen3VL-8B.

Depth Ordering and Camera Pose produce the highest mean performance increases across six models when used as direct teachers compared to other per-image and scene-level relational representations respectively. For per-image representations, depth ordering and bounding boxes produce relatively higher mean performance improvements in the direct teacher setting compared to 3D pose estimates. On the other hand, all three scene-level relational representations produce relatively minor changes in mean performance compared to the per-scene representations.

The pure distillation reward setting is the best finetuning setting for improving performance for per-scene representations as it never degrades performance, although the mixed reward setting can lead to higher performance increases for certain representations. We observe that despite mixed-reward GRPO's strong performance improvements, especially for bounding box representations, it can degrade performance for other per-scene representations. Pure distillation,

on the other hand, achieves comparable or better performance in for all per-image representations compared to the base model, with second-strongest performance after Direct Teacher for several models using the Depth Ordering and 3D Pose Estimates representations.

Finetuning on scene-level relational representations does not generally lead to performance improvements. We observe that in most cases, finetuning on scene-level relational representations degrades model performance on reasoning tasks (with the exception of mixed-reward training on semantic correspondences, which maintains comparable performance to the base model). This suggests that while scene-level relational representations may help larger models in the direct teacher setting, these gains will likely not be transferable via finetuning.

5 Conclusions

In this paper, we have examined whether a VLM can reason using specialized expert computer vision model representations, as well as be finetuned to generate such representations in its own reasoning. We study VLM reasoning on a wide variety of tasks in several settings, focusing on two categories of representations: per-image and scene-level relational representations. Our results identify which visual representations lead to the highest performance gains when used in the prompt, and among those representations, which ones can be distilled into the VLM using finetuning approaches. We believe that our work presents a promising path forward for identifying representations that aid VLM reasoning, and lead to the design of better toolsets for improving VLM visual reasoning.

Limitations. While our work evaluates multiple VLMs on diverse visual reasoning tasks, due to computational constraints, we can only evaluate models up to 8B parameters in size across all settings. Additionally, while we evaluate performance using a diverse set of representations, we leave the impact of combining multiple representations during training and inference to future work.

AI use statement

LLMs were used to generate and explain the code for executing some of the experiments, which was reviewed and checked by the authors.

References

Bigverdi, M., Luo, Z., Hsieh, C.-Y., Shen, E., Chen, D., Shapiro, L. G., and Krishna, R. (2025). Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3836–3845.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., and Zhao, B. (2025). Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498.

Chen, B., Sax, A., Lewis, G., Armeni, I., Savarese, S., Zamir, A., Malik, J., and Pinto, L. (2020). Robust policies via mid-level visual representations: An experimental study in manipulation and navigation. *arXiv preprint arXiv:2011.06698*.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. (2024). Spatialrgpt: Grounded spatial reasoning in vision-language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 135062–135093. Curran Associates, Inc.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N. A., Ma, W.-C., and Krishna, R. (2025). Blink: Multimodal large language models can see but not perceive. In Leonardis, A.,

Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G., editors, *Computer Vision – ECCV 2024*, pages 148–166, Cham. Springer Nature Switzerland.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gupta, T. and Kembhavi, A. (2023). Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962.

Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Krishna, R. (2024). Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379.

Hu, Y., Stretcu, O., Lu, C.-T., Viswanathan, K., Hata, K., Luo, E., Krishna, R., and Fuxman, A. (2023). Visual program distillation: Distilling tools and programmatic reasoning into vision-language models.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., and Schindler, K. (2024). Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502.

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., and Huang, J. (2023). A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

Leroy, V., Cabon, Y., and Revaud, J. (2025). Grounding image matching in 3d with mast3r. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G., editors, *Computer Vision – ECCV 2024*, pages 71–91, Cham. Springer Nature Switzerland.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Cahyono, J. A., Yang, J., Li, C., and Liu, Z. (2025). Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. (2025). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G., editors, *Computer Vision – ECCV 2024*, pages 38–55, Cham. Springer Nature Switzerland.

Ma, W., Chou, Y.-C., Liu, Q., Wang, X., de Melo, C., Xie, J., and Yuille, A. (2025a). Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*.

Ma, W., Ye, L., de Melo, C. M., Yuille, A., and Chen, J. (2025b). Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17249–17260.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Featured Certification.

Qi, J., Ding, M., Wang, W., Bai, Y., Lv, Q., Hong, W., Xu, B., Hou, L., Li, J., Dong, Y., et al. (2024). Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 1339–1384.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ramachandran, R., Garjani, A., Bachmann, R., Atanov, A., Kar, O. F., and Zamir, A. (2025). How well does gpt-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollar, P., and Feichtenhofer, C. (2025). Sam 2: Segment anything in images and videos. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R., editors, *International Conference on Learning Representations*, volume 2025, pages 28085–28128.

Ray, A., Duan, J., II, E. L. B., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B. A., Krishna, R., Zeng, K.-H., and Saenko, K. (2025). SAT: Dynamic spatial aptitude training for multimodal language models. In *Second Conference on Language Modeling*.

Rui Yang, Ziyu Zhu, Y. L. J. H. S. Y. S. Z. Z. L. X. L. S. L. W. W. Y. L. G. S. H. Z. (2025). Visual spatial tuning. *arXiv preprint arXiv:2511.05491*.

Sarch, G., Cai, L., Wang, Q., Wu, H., Chen, D., and Liu, Z. (2026). Vero: An open rl recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*.

Sarch, G., Saha, S., Khandelwal, N., Jain, A., Tarr, M. J., Kumar, A., and Fragkiadaki, K. (2025). Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*.

Tan, H., Ji, Y., Hao, X., Chen, X., Wang, P., Wang, Z., and Zhang, S. (2025). Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., Bohez, S., Bousmalis, K., Brohan, A., Buschmann, T., Byravan, A., Cabi, S., Caluwaerts, K., Casarini, F., Chang, O., Chen, J. E., Chen, X., Chiang, H.-T. L., Choromanski, K., D'Ambrosio, D., Dasari, S., Davchev, T., Devin, C., Palo, N. D., Ding, T., Dostmohamed, A., Driess, D., Du, Y., Dwibedi, D., Elabd, M., Fantacci, C., Fong, C., Frey, E., Fu, C., Giustina, M., Gopalakrishnan, K., Graesser, L., Hasenclever, L., Heess, N., Hernaez, B., Herzog, A., Hofer, R. A., Humplik, J., Iscen, A., Jacob, M. G., Jain, D., Julian, R., Kalashnikov, D., Karagozler, M. E., Karp, S., Kew, C., Kirkland, J., Kirmani, S., Kuang, Y., Lampe, T., Laurens, A., Leal, I., Lee, A. X., Lee, T.-W. E., Liang, J., Lin, Y., Maddineni, S., Majumdar, A., Michaely, A. H., Moreno, R., Neunert, M., Nori, F., Parada, C., Parisotto, E., Pastor, P., Pooley, A., Rao, K., Reymann, K., Sadigh, D., Saliceti, S., Sanketi, P., Sermanet, P., Shah, D., Sharma, M., Shea, K., Shu, C., Sindhwani, V., Singh, S., Soricut, R., Springenberg, J. T., Sterneck, R., Surdulescu, R., Tan, J., Tompson, J., Vanhoucke, V., Varley, J., Vesom, G., Vezzani, G., Vinyals, O., Wahid, A., Welker, S., Wohlhart, P., Xia, F., Xiao, T., Xie, A., Xie, J., Xu, P., Xu, S., Xu, Y., Xu, Z., Yang, Y., Yao, R., Yaroshenko, S., Yu, W., Yuan, W., Zhang, J., Zhang, T., Zhou, A., and Zhou, Y. (2025). Gemini robotics: Bringing ai into the physical world.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Wang, H., Su, A., Ren, W., Lin, F., and Chen, W. (2025a). Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. (2025b). Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306.

Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., and Nahrstedt, K. (2025). Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*.

Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., Peng, Y., Wei, H., Sun, J., Han, C., Ge, Z., et al. (2025). Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Zhang, J., Chen, Y., Xu, Y., Huang, Z., Mei, J., Chen, C., Zhou, Y., Yuan, Y.-J., Cai, X., Huang, G., Quan, X., Xu, H., and Zhang, L. (2025). From flatland to space: Teaching vision-language models to perceive and reason in 3d. In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N., editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

Zhou, B., Krähenbühl, P., and Koltun, V. (2019). Does computer vision matter for action? *Science Robotics*, 4(30):eaaw6661.

A Evaluation Prompts

A.1 Prompt Template

Prompt Template

```
{Task Prompt}
\n
\n
"Return ONLY one JSON object with exactly these keys: process,
    confidence, final\_answer.\n"
   "- process: a list with at most two very short observable
       phrases; do not include hidden chain-of-thought.\n"
   "- confidence: a number from 0 to 1.\n"
   "- final\_answer: concise answer string; when choices are
       provided, prefer the exact answer choice text.\n"
   "- No markdown, no code fences, no extra text."
\n
\n
Question:\n{question}
\n
\n
Answer choices:\n{choices}
```

A.2 Task Prompts

Base Prompt

```
"Task: Answer using only the images, question, and answer choices.\
    n\n"
```

Bounding Box Prompt

```
"Task: Answer using the images and teacher bounding-box evidence. "
"Coordinates are normalized [x1,y1,x2,y2] in [0,1].\n\n"
"Teacher bounding boxes: {bounding_boxes}"
```

3D Pose Estimates Prompt

```
"Task: Answer using the images and teacher coarse 3D pose estimates.
    "
"x is left/center/right, y is upper/middle/lower, z is foreground/
    midground/background/unknown.\n\n"
"Teacher 3D pose estimates: {pose_estimates}"
```

Depth Ordering Prompt

```
"Task: Answer using the images and teacher near-to-far depth
    ordering.\n\n"
"Teacher depth ordering: {depth_ordering}"
```

Correspondence Grid Prompt

```
"Task: Answer using the two images with a teacher correspondence
    grid overlaid on each image. "
"Grid cells are columns A-F and rows 1-6; use the grid to reason
    about corresponding points or motion.\n\n"
```

Camera Pose Prompt

```
"Task: Answer using camera-pose reasoning between image 1 and image
    2. "
"Estimate egocentric translation and rotation cues, then choose the
    answer.\n\n"
```

Semantic Correspondences Prompt

```
"Task: Answer using semantic correspondence reasoning. Match
    salient objects or regions across the two images, "
"then choose the answer.\n\n""
```

B EXAMPLE VISUAL REASONING QUESTIONS

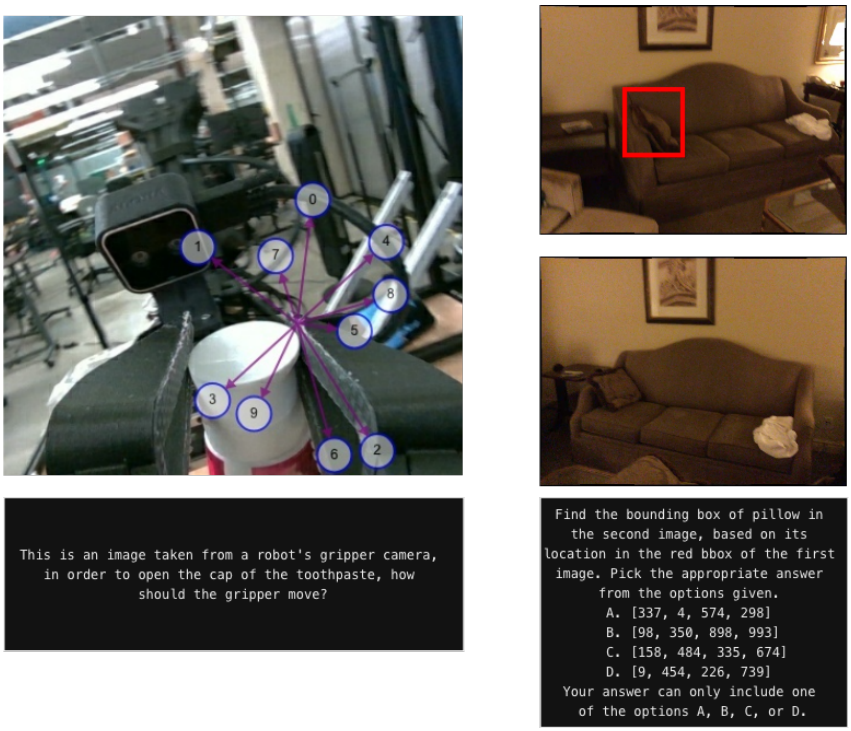


Figure 1: Comparison of evaluation question types for the two representation families. Left question is for per-image representations, taken from ERQA (Team et al., 2025). Right question is for scene-level relational representations, taken from SPAR-Bench-Tiny (Zhang et al., 2025).