# EMBODIMENTSEMANTIC: A SPATIAL SCENE-GRAPH DATASET AND BENCHMARK FOR VISION-LANGUAGE MODELS ON EMBODIED MANIPULATION TRAJECTORIES

**Anonymous authors**
Paper under double-blind review

Code Data

## ABSTRACT

Spatial grounding remains a key limitation of vision-language-action (VLA) systems for robotic manipulation. While current models can recognize objects and follow language instructions, they often lack an explicit representation of how objects are arranged in space, including support, containment, ordering, occlusion, and depth-sensitive relations. We introduce EmbodimentSemantic, a spatial scene-graph dataset and benchmark for evaluating relational grounding in embodied manipulation. EmbodimentSemantic represents scenes as directed object-relation-object triplets, where each triplet specifies a spatial relation between an ordered pair of objects using a fixed set of relations. This representation enables direct evaluation of object binding, relation prediction, and spatial consistency. The dataset includes real-world manipulation observations collected with the low-cost SO101 robot arm, together with generated scene graphs for studying spatial grounding in practical robotic settings. To provide controlled validation, we also introduce a simulator-grounded LIBERO benchmark with over 60K paired manipulation timesteps and more than 120K camera-specific scene graphs across paired third-person and wrist views, where ground-truth relations are derived automatically from MuJoCo geometry, world coordinates, camera projections, and visibility constraints. Experiments across open-source and commercial VLMs show that current models often predict plausible relations but struggle with exact depth-aware and viewpoint-dependent spatial structure. EmbodimentSemantic provides a unified framework for diagnosing spatial grounding in VLM perception.

## 1 INTRODUCTION

Since the development of the transformer architecture Vaswani et al. (2017), robot learning has seen several major advances Ichter et al. (2023); Black et al. (2025). Recent vision–language–action (VLA) models extend this trend by adapting internet-scale vision–language models (VLMs) to robot control, typically through fine-tuning on robot demonstrations Zitkovich et al. (2023); Kim et al. (2025). Despite their scale, however, these models remain weakly grounded in scene geometry. Several works show that VLA policies can learn shortcut associations between actions and task-irrelevant visual or semantic context, such as background, texture, viewpoint, familiar objects, or linguistic priors, rather than reliably representing spatial relations such as distance, relative pose, object size, and viewpoint Xing et al. (2025); Zhang et al. (2025); Fang et al. (2025).

Recent work has begun to address this gap by injecting spatial supervision and structured spatial representations into VLMs and VLAs, including large-scale metric spatial VQA data, robot-relevant annotations from 3D scans, and explicit egocentric spatial encodings or action grids Chen et al. (2024); Song et al. (2025); Qu et al. (2025). However, these methods still largely rely on learned spatial priors and do not remove the fundamental ambiguity of inferring metric geometry from RGB appearance alone. For example, if one object has a larger image footprint than another, it may be



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.

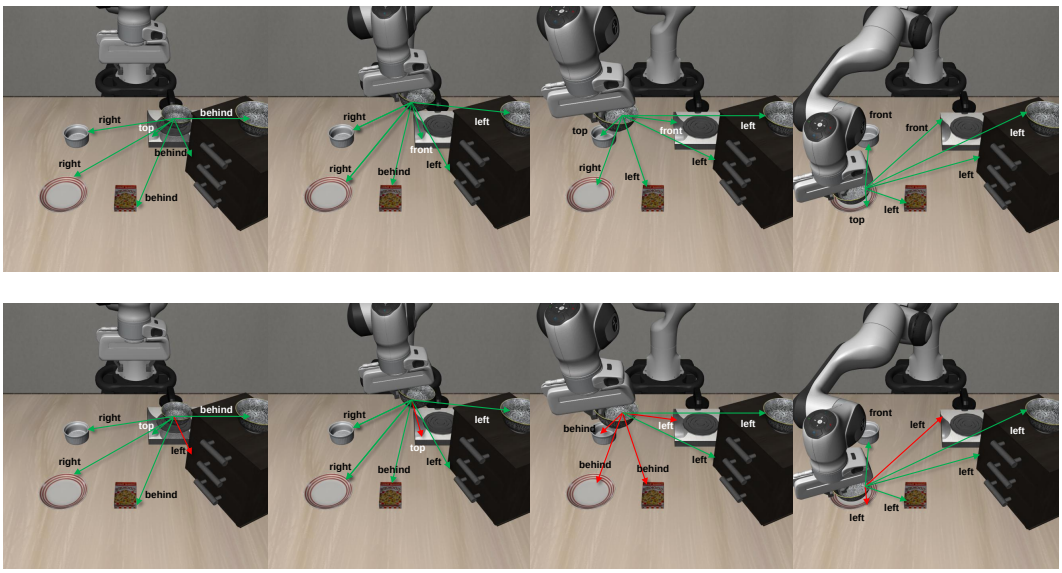


Figure 1: LIBERO scene-graph comparison. The top row shows simulator-grounded ground-truth relations, and the bottom row shows `gemini-3.1-pro` predictions for the same frames. Green edges indicate correct triplets and red edges indicate errors, highlighting both successful relation recovery and remaining depth-sensitive failures.

physically larger, closer to the camera, or both. This size–distance ambiguity means that appearance alone does not uniquely determine metric scale, depth, or pose without additional geometric cues such as depth, stereo, motion, camera/object pose, or object-size priors Battaglia et al. (2011).

Earlier robotic perception work addressed geometry more explicitly through object-pose estimation from RGB or RGB-D observations Xiang et al. (2018); Tremblay et al. (2018); Hodan et al. (2018), and such estimates have long supported manipulation through visual servoing Hutchinson et al. (1996); Chaumette & Hutchinson (2006). More recent manipulation policies similarly explore explicit spatial structure by augmenting RGB-language inputs with representations such as 3D keypoints, predicted end-effector poses, point clouds, or object-centric trajectories Shridhar et al. (2023); Goyal et al. (2023); Li et al. (2024; 2025a); Hsu et al. (2025); Li et al. (2025b). While these representations make geometry more directly available to the policy, they often shift the burden onto the perception system, requiring reliable depth sensing, calibration, segmentation, pose tracking, or related infrastructure, which can limit generalization.

Another challenge is evaluation: success on spatial reasoning benchmarks does not necessarily imply recovery of a structured scene representation. Existing multimodal benchmarks primarily evaluate spatial understanding through captioning, visual question answering, relation classification, or image–text matching (Krishna et al., 2017; Yang et al., 2019; Liu et al., 2023b; Thrush et al., 2022; Wang et al., 2024; Chen et al., 2024). Although useful for diagnosing spatial and compositional failures, these evaluations typically produce free-form answers, binary judgments, or matching scores rather than explicit relational states. As a result, they do not directly measure whether a model can recover the structured representations needed for robotic planning, monitoring, or action conditioning.

We introduce EMBODIMENTSEMANTIC, a spatial scene-graph dataset and benchmark for evaluating VLMs on embodied manipulation trajectories. Our primary contribution is a real-world dataset collected with the low-cost SO101 robot arm, pairing manipulation observations with generated scene graphs that encode object-centric spatial relations as directed triplets (`objectA`, `relation`, `objectB`). To validate scene-graph generation under controlled conditions, we also construct a simulator-grounded LIBERO Liu et al. (2023a) benchmark in which ground-truth scene graphs are generated from simulator state and compared against VLM-predicted graphs from robot camera observations. Using both real-world and simulated data, we benchmark




Figure 2: Real-world SO101 scene-graph prediction examples. Generated scene graphs are overlaid on external-camera observations from two physical robot scenes. These examples show that the same object-centric scene-graph representation used in LIBERO can be applied to real SO101 observations.

open-source and commercial VLMs on their ability to recover manipulation-relevant spatial relations. We additionally release SAMGRAPH, a segmentation-mask-based baseline that converts object masks into directed scene graphs for LIBERO external and wrist views and SO101 external-camera observations.

## 2 RELATED WORK

**Scene graphs and spatial reasoning benchmarks.** Scene graphs provide a structured way to represent objects and their relations, and have been widely used to study relational visual understanding. Visual Genome (Krishna et al., 2017) introduced large-scale object, attribute, and relationship annotations for natural images, while SpatialSense (Yang et al., 2019), Visual Spatial Reasoning (VSR) (Liu et al., 2023b), and Winoground (Thrush et al., 2022) evaluate spatial or compositional grounding through relation classification, image–text pairs, and caption matching. More recent VLM benchmarks further target spatial reasoning: SpatialEval (Wang et al., 2024) evaluates relationship reasoning, counting, navigation, and positional understanding; SpatialVLM (Chen et al., 2024) trains on large-scale metric spatial question-answering data; SpatialRGPT (Cheng et al., 2024) studies grounded 2D and 3D spatial reasoning; and EmbSpatial-Bench (Du et al., 2024) and MV-RoboBench (Feng et al., 2026) move toward embodied or robotic scenes. These benchmarks provide useful diagnostics for spatial reasoning, but they typically evaluate answers, classifications, or matching decisions rather than the full relational structure of a manipulation scene. EMBODIMENTSEMANTIC instead evaluates whether models can directly produce ordered scene-graph triplets (`objectA`, `relation`, `objectB`) from embodied manipulation observations, using a fixed set of spatial relations. This makes it possible to isolate object-binding errors, relation reversals, directional inconsistencies, and hallucinated relations that may be hidden in language-facing benchmarks but can directly change the action required for manipulation.

## 3 EMBODIMENTSEMANTIC BENCHMARK AND DATASET

EMBODIMENTSEMANTIC is built from LIBERO Spatial manipulation trajectories. Each task consists of demonstrations in which the robot picks up `akita_black_bowl_1` from a different initial spatial configuration and places it on `plate_1`. The benchmark converts these demonstrations into frame-level spatial scene graphs over a fixed object and relation ontology. Each demonstration contains synchronized RGB observations from a fixed third-person camera (`agentview`) and a wrist-mounted camera (`eye_in_hand`). For each selected frame and camera view, EmbodimentSemantic provides spatial relations as ordered triplets of the form (`objectA`, `relation`, `objectB`).

**Object and relation ontology.** EmbodimentSemantic fixes the object vocabulary and relation set so that model predictions can be evaluated by exact triplet matching. The object vocabulary contains seven canonical LIBERO objects: the manipulated bowl, a distractor bowl, the cookie box, ramekin, plate, wooden cabinet, and stove. The relation vocabulary contains eight directed predicates arranged in four inverse pairs: left/right, front/behind, on-top/below, and inside/contains. This fixed

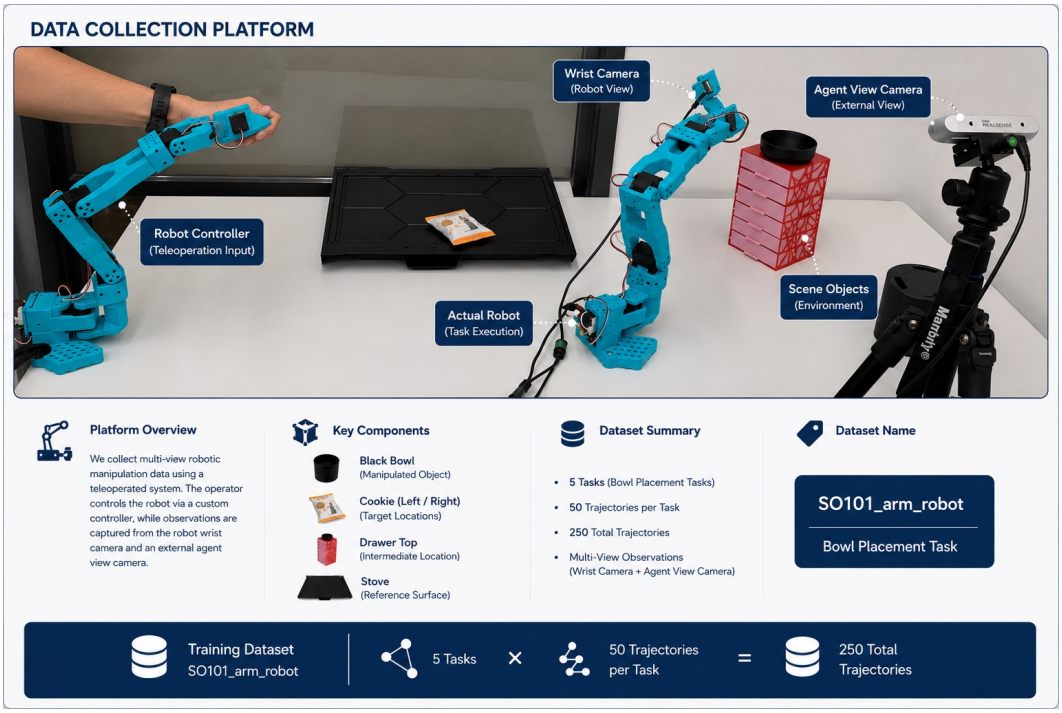


Figure 3: Teleoperation interface used to collect real-robot demonstrations for the bowl-placement tasks. The scene contains the SO101 robot arm, the manipulated black bowl, and task-relevant support and distractor objects.

ontology removes ambiguity in object naming and relation formatting, allowing direct comparison across models.

For aggregate analysis, the evaluator can also collapse inverse directions into four relation families: `is_left_of`, `is_in_front_of`, `is_on_top_of`, and `is_inside`. The primary benchmark, however, evaluates the full directed relation set.

**Dataset organization.** The benchmark contains ten task files, one per LIBERO Spatial task. Each file contains 50 demonstrations, giving 500 demonstrations in total. Demonstration lengths range from 75 to 197 timesteps, with a mean length of 124.5 timesteps. Across all demonstrations, EmbodimentSemantic contains 62,250 paired timesteps and 124,500 RGB frames across the two camera views. RGB observations are stored at $128 \times 128$ resolution.

Frame-level scene graphs are stored under camera-specific annotation keys. The `agentview` graph describes the third-person view of the scene, while the `eye_in_hand` graph is visibility-filtered for the wrist camera. During annotation generation, EmbodimentSemantic computes MuJoCo world coordinates and projected 2D bounding boxes for the relevant objects; the VLM benchmark evaluates the resulting spatial triplets.

The difference in graph density between the two views is intentional. The fixed `agentview` camera captures the global scene layout and therefore supports dense pairwise relation labels over the full object vocabulary. The wrist-mounted `eye_in_hand` camera observes a narrower and changing field of view, so its graph is filtered to visible objects. This paired-view structure allows EmbodimentSemantic to evaluate both global spatial ordering and egocentric spatial grounding.

### 3.1 Real-robot dataset

In addition to the simulator-grounded LIBERO benchmark, EmbodimentSemantic includes a real-robot bowl-placement dataset collected with a SO101 manipulation setup. The dataset contains 257 teleoperated episodes across five tabletop manipulation tasks. Each task requires the robot to

move a black bowl from a specified initial spatial context to a target support object, such as the stove or a cookie box.

The five real-robot tasks cover different starting relations for the manipulated bowl: on top of the drawer to the stove, left of the stove to the stove, right of the stove to the stove, on top of the drawer to the cookie box on the left, and on top of the drawer to the cookie box on the right. The demonstrations follow the LeRobot dataset format, with task-level metadata, trajectory data, and synchronized videos. The recorded video modalities include an external agent-view camera, a wrist camera, and an agent-view depth stream. Figure 3 shows the physical setup and teleoperation workflow used during data collection. An example of the generated scene graphs are shown in Figure 2.

## 4 SIMULATOR-GROUNDED GRAPH GENERATION

EmbodimentSemantic generates frame-level scene graphs deterministically from MuJoCo simulator state. Rather than relying on manual relation annotations, the pipeline extracts object geometry, world coordinates, and camera projections for each demonstration frame. Offline annotations are generated for both `agentview` and `robot0_eye_in_hand`.

### 4.1 OBJECT DISCOVERY, WORLD COORDINATES, AND BOUNDING BOXES

For each demonstration frame, the system identifies task-relevant objects from MuJoCo bodies and excludes robot and gripper components. For every object, EmbodimentSemantic extracts the world-space position and projects its geometry into camera views. Offline annotation uses both the fixed third-person `agentview` camera and the wrist-mounted `robot0_eye_in_hand` camera.

For each MuJoCo geometry primitive with world center $p \in \mathbb{R}^3$, rotation matrix $R$, and half-sizes $s = (s_x, s_y, s_z)$, the eight local geometry corners are:

$$c(\sigma_x, \sigma_y, \sigma_z) = \begin{bmatrix} \sigma_x s_x \\ \sigma_y s_y \\ \sigma_z s_z \end{bmatrix}, \qquad \sigma_x, \sigma_y, \sigma_z \in \{-1, +1\}. \tag{1}$$

Each corner is transformed into world coordinates as:

$$x_w = p + Rc. \tag{2}$$

Given camera intrinsics $K$ and inverse extrinsics $E^{-1}$, world points are transformed into the camera frame:

$$\tilde{x}_c = E^{-1} \begin{bmatrix} x_w \\ 1 \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix}. \tag{3}$$

Points with $z_c \leq 0.01$ are discarded, and the remaining points are projected using the standard pinhole camera model. The projected corners define an axis-aligned 2D bounding box, which is padded by 5 pixels and clamped to image bounds. Boxes that are implausibly large relative to the image size are discarded as projection artifacts.

This process produces, for each detected object, a world-space position and camera-specific projected bounding boxes. The world coordinates are used to assign spatial relations, while the projected boxes determine object visibility and support overlap-based vertical or containment checks.

### 4.2 SPATIAL RELATION ASSIGNMENT

Given a set of objects in a frame, EmbodimentSemantic constructs a directed scene graph by iterating over all ordered object pairs $(A, B)$, where $A \neq B$. Each ordered pair receives one relation from the fixed ontology. Relation assignment follows two stages: an overlap-based vertical or containment check, followed by dominant-axis world-frame ordering.

First, EmbodimentSemantic checks whether the projected 2D boxes of $A$ and $B$ overlap strongly. Let the intersection rectangle be defined by $(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$. We compute minimum-area overlap:

$$\mathrm{io}_{\min}(A, B) = \frac{(\hat{x}_2 - \hat{x}_1)(\hat{y}_2 - \hat{y}_1)}{\min\big((x_2^A - x_1^A)(y_2^A - y_1^A),\ (x_2^B - x_1^B)(y_2^B - y_1^B)\big)}. \tag{4}$$

If $\mathrm{io}_{\min}(A, B) > 0.8$ and at least one object is a bowl, the relation is assigned using the objects' world $z$-coordinates. For ordinary support cases, the higher object is labeled `is_on_top_of` and the lower object is labeled `is_below_of`. For the drawer task, we use a task-specific containment rule. For task T2, if `akita_black_bowl_1` and `wooden_cabinet_1` satisfy the overlap condition, the pair is assigned `is_inside` or `contains`. This hardcoded exception simplifies containment labeling for the only drawer-based task in EmbodimentSemantic.

If the vertical or containment condition does not apply, EmbodimentSemantic assigns a horizontal relation using world-frame object positions. Let:

$$\Delta x = x_A - x_B, \qquad \Delta y = y_A - y_B. \tag{5}$$

If $|\Delta x| \geq |\Delta y|$, the pair receives a depth relation, either `is_in_front_of` or `is_behind`. Otherwise, the pair receives a lateral relation, either `is_left_of` or `is_right_of`. The sign of the dominant offset determines the direction.

Importantly, EmbodimentSemantic defines left/right and front/behind relations in the MuJoCo world frame rather than directly in image-plane coordinates. This choice makes the labels consistent with the physical manipulation scene and simulator state, while camera projections determine which objects are visible in each view.

### 4.3 Camera-specific graph construction

EmbodimentSemantic augments the original LIBERO Spatial HDF5 files with additional semantic annotations; it does not regenerate or alter the original RGB observations. The stored RGB frames and projected bounding-box coordinates remain in the original LIBERO camera orientation. In particular, the released bounding boxes are defined with respect to the images as stored in the original HDF5 files.

Offline EmbodimentSemantic annotations are generated for both the fixed third-person `agentview` camera and the wrist-mounted `robot0_eye_in_hand` camera. Spatial directions such as `is_in_front_of`, `is_behind`, `is_left_of`, and `is_right_of` are assigned from the signs and dominant axes of MuJoCo world-coordinate differences, not from image-plane pixel coordinates. This makes direction assignment deterministic rather than random, although the semantic naming of front/behind and left/right depends on the chosen world-frame convention.

The graph-generation rule is shared across the two camera views. For the wrist view, objects are first visibility-filtered using projected `robot0_eye_in_hand` bounding boxes. In the current offline generation pipeline, relation assignment for the wrist graph then reuses the same world-frame relation logic used for `agentview`; wrist bounding boxes determine which objects are included, while MuJoCo world coordinates determine their spatial relations. Thus, the wrist graph is a visibility-filtered relational graph rather than a fully wrist-frame relation graph.

For visualization and human inspection, the `agentview` image may be rotated by $180^\circ$ to match a more natural viewing orientation. This rotation is not applied to the stored LIBERO images or to the released bounding-box coordinates.

The resulting annotations are stored under camera-specific keys:

- `agentview_scene_graph` for the fixed third-person camera;
- `robot0_eye_in_hand_scene_graph` for the wrist-mounted camera.

Each graph is represented as a list of ordered triplets:

$$(\texttt{objectA}, \texttt{relation}, \texttt{objectB}).$$

## 5 Experiments and results

Our experiments evaluate whether VLMs can recover object-centric spatial scene graphs from embodied manipulation observations. Given an RGB robot observation, the model predicts directed object–relation–object triplets over a fixed set of objects and relations. We then evaluate whether each predicted relation is assigned to the correct ordered object pair.

Table 1: Simulator-grounded VLM scene-graph prediction results. Mean per-task F1 is the main score. Macro F1 and micro F1 are computed as complementary aggregate scores but are not shown here for compactness.

| Model | agentview mF1 | eye_in_hand mF1 |
|---|---|---|
| gemini-3.1-pro | **0.5674** | **0.3773** |
| InternVL3-78B | <u>0.3519</u> | 0.2101 |
| Qwen3-VL-8B-Instruct | 0.2837 | <u>0.2108</u> |
| InternVL3_5-14B | 0.2561 | 0.1265 |
| gemma-4-E4B-it | 0.2209 | 0.1472 |
| Molmo2-8B | 0.1888 | 0.1459 |
| InternVL3_5-8B | 0.1769 | 0.1465 |
| nemotron-nano-12b-v2-vl | 0.1599 | 0.1270 |
| Qwen2.5-VL-7B-Instruct | 0.1405 | 0.1147 |
| MomaGraph-R1 | 0.1023 | 0.0926 |

This experiment is carried out in both simulated and real-world settings. In LIBERO-Spatial, the simulator state gives us ground-truth scene graphs for each selected frame, with observations from the simulated `agentview` and `wrist` cameras. The real-world SO101 dataset uses a matching external-camera and robot-camera setup with similar tabletop objects, allowing us to test the same scene-graph formulation on physical robot observations.

For LIBERO VLM evaluation, we sample every fifth frame, yielding 12,648 evaluated frames per camera and 25,296 camera-specific frames in total across the 500 demonstrations.

For each VLM, we use the same prompt and parsing protocol. The model receives the RGB observation, the task description, the object vocabulary, and the fixed set of relations and is asked to output one spatial triplet per line. We keep only valid triplets and compare the predictions against the reference graph for the corresponding task, episode, frame, and camera. Because the two black bowl instances are visually similar, evaluation treats predictions as invariant to swapping `akita_black_bowl_1` and `akita_black_bowl_2`. For each frame, we compute exact triplet F1 under the original prediction and under a swapped-bowl assignment, and score the higher-F1 assignment. This prevents penalizing models for visually ambiguous bowl-instance identity while still requiring exact relation and object-pair grounding for all other objects. We use mean per-task F1 as the main score for model comparison, since it gives equal weight to each manipulation task. In addition, we compute F1 in two complementary forms. Macro F1 first computes F1 independently for each evaluated frame and then averages these values, giving each frame equal weight. Micro F1 pools triplet-level true positives, false positives, and false negatives across all frames before computing the score. We report additional metrics including coverage, hallucination rate, reversal rate, direction consistency, per-relation performance, and per-object recall.

### 5.1 VLM BENCHMARKING

**Simulator-grounded evaluation.** We first evaluate VLMs on the simulator-grounded LIBERO benchmark. The controlled setting allows us to compare predicted scene graphs against reference graphs derived from simulator geometry.

The results show that exact spatial scene-graph prediction remains difficult across models. The proprietary `gemini-3.1-pro` model achieves the best performance on both views, with a mean per-task F1 of 0.5674 on `agentview` and 0.3773 on `eye_in_hand`. This sets the strongest reference point for the benchmark. Among open-source models, `InternVL3-78B` performs best on `agentview`, while `Qwen3-VL-8B-Instruct` performs best on `eye_in_hand`. Performance is lower on the wrist camera for most models, which is expected because the view is narrower, changes with robot motion, and contains partial observations of the fixed object set from which relations must be inferred.

The gap between the two camera views is important. The `agentview` camera provides a stable view of the full workspace with one object being moved. This tests whether VLMs update their predicted relations as the scene changes over time. While `eye_in_hand` requires the model to

Table 2: Main VLM diagnostics on `agentview`. Coverage is relation-type coverage, while F1 requires exact object–relation–object triplet matching. The large coverage–F1 gap shows that models predict plausible predicates but often bind them to the wrong ordered object pairs. Per-relation F1 is averaged across evaluated VLMs.

| Diagnostic | Value | Interpretation |
|---|---|---|
| Best coverage | 0.980 | Relation types are often mentioned |
| Best exact mF1 | 0.567 | Exact triplets remain difficult |
| Gemini coverage / F1 | 0.968 / 0.567 | Strongest model still has binding errors |
| Gemma coverage / F1 | 0.969 / 0.221 | High coverage, low exact grounding |
| MomaGraph coverage / F1 | 0.947 / 0.102 | Plausible predicates, poor binding |
| `is_left_of` F1 | 0.278 | Easiest relation family |
| `is_right_of` F1 | 0.276 | Easiest relation family |
| `is_in_front_of` F1 | 0.233 | Depth remains difficult |
| `is_behind` F1 | 0.222 | Depth remains difficult |
| `is_on_top_of` F1 | 0.196 | Support relation is harder |
| `is_inside` F1 | 0.191 | Containment is harder |
| `is_below_of` F1 | 0.189 | Vertical inverse is harder |
| `contains` F1 | 0.153 | Hardest relation |

reason from an egocentric and changing perspective. This makes the wrist-camera setting a stricter test of viewpoint-dependent spatial grounding. The lower wrist-view scores suggest that even the strongest open-source models still struggle when spatial layout must be inferred from partial and moving robot observations.

Mean per-task F1 gives the main comparison, while the broader metric set exposes related failure modes. Macro and micro F1 provide frame-level and pooled triplet-level views of the same exact-matching problem. Relation-type coverage measures the fraction of reference triplets whose predicate type appears anywhere in the model's prediction for the same frame, while hallucination rate measures over-generation. Reversal rate and direction consistency test whether the model preserves the ordered structure of relations. These diagnostics are useful because a model can predict plausible spatial predicates while still assigning them to the wrong object pair or reversing their direction.

Although several models achieve high relation-type coverage, their exact triplet F1 remains much lower, revealing a core failure mode that aggregate VQA-style spatial benchmarks can hide. For example, `gemini-3.1-pro` reaches 0.968 coverage on `agentview`, but only 0.567 mean per-task F1; similarly, `gemma-4-E4B-it` reaches 0.969 coverage but only 0.221 F1, and `MomaGraph-R1` reaches 0.947 coverage but only 0.102 F1. This gap shows that current VLMs often recover plausible spatial predicates while failing to bind them to the correct ordered object pairs. Per-relation results in Table 2 further show that failures are systematic: lateral relations are easier than depth, support, and containment relations, with `contains` performing worst overall. Thus, the dominant failure is not spatial vocabulary, but exact object-pair grounding under directed scene-graph evaluation.

The diagnostic metrics show that these errors are structured rather than random. Across complete `agentview` runs, the average hallucination rate is 0.690, indicating that VLMs frequently over-generate plausible but incorrect spatial triplets; at the same time, average direction consistency is only 0.205, showing that models rarely recover both directions of inverse relation pairs correctly. Reversal errors further indicate failures of asymmetric grounding: `qwen3-vl-8b` has the highest reversal rate at 0.288, while `InternVL3-78B`, `InternVL3_5-14B`, and `gemma-4-E4B-it` also show substantial reversal rates of 0.163, 0.149, and 0.133. Together with the relation-level gradient in Table 2, these results suggest that current VLMs often detect spatial co-occurrence and produce reasonable relation words, but struggle to resolve ordered, asymmetric spatial assignments from RGB observations, especially for depth, support, and containment relations.

**SamGraph baseline.** SAMGRAPH converts object segmentation masks into directed spatial scene graphs using mask geometry. On LIBERO `agentview`, evaluation over 500 task-episodes and 12,648 sampled frames yields 0.8209 precision, 0.8116 recall, 0.8162 micro F1, and 0.8196 mean per-task F1. Because the two black bowls are visually indistinguishable in a single RGB frame, we use the same permutation-invariant identity reconciliation as the VLM evaluator: the scorer com-

pares the original bowl-label assignment with a global exchange of the two bowl IDs and retains the higher-F1 assignment (3,980 frames). This changes only the two bowl names—never masks, relations, coverage, other objects, or ground truth—so the reported metric isolates relational grounding rather than temporal identity tracking (Appendix A.1). On the wrist view, evaluation over 100 task-episodes and 2,551 sampled frames yields 0.5321 precision, 0.3863 recall, 0.4476 micro F1, and 0.3017 mean per-task F1. The wrist evaluation is smaller and uses external-camera predictions to maintain object identity. For SO101, SAMGRAPH exports 16,268 predicted triplets from 835 sampled frames across 50 episodes; because verified triplet ground truth is unavailable, we report export coverage rather than accuracy.

**Real-world evaluation.** We also apply the same scene-graph prompting framework to the real-world SO101 dataset. This split is not used for quantitative scoring, since it does not currently include reference scene-graph annotations. Instead, we use the best-performing VLM from the previous experiment, `gemini-3.1-pro`, to generate scene graphs for physical robot observations.

The SO101 tasks are designed to mirror the spatial structure of LIBERO Spatial with similar tabletop objects and external and wrist camera views. The model is prompted with the same relation ontology and a corresponding object vocabulary, producing directed object–relation–object triplets for sampled frames. These predictions are included as a real-world scene-graph resource rather than as ground truth.

We release both components of EMBODIMENTSEMANTIC: the real-world SO101 manipulation dataset and the simulator-grounded LIBERO annotations. The release includes physical robot observations, generated real-world scene graphs, and more than 120K camera-specific simulator-grounded scene graphs. Together, these components connect controlled simulator labels with real robot data and provide a resource for studying spatial grounding in embodied manipulation.

## 6 LIMITATIONS

EmbodimentSemantic is designed as a controlled diagnostic benchmark, and its scope is intentionally narrow. The current dataset focuses on one LIBERO-Spatial task family, seven canonical objects, and a fixed eight-predicate relations set. This controlled setting makes exact triplet-level evaluation possible, but it limits claims about broader open-vocabulary spatial grounding or real-world generalization.

The benchmark also relies on simulator-derived labels. This avoids manual annotation noise and provides dense frame-level supervision, but the resulting relations depend on the chosen MuJoCo world-frame convention, visibility filtering rules, and task-specific containment heuristics. In particular, left/right and front/behind are defined by world-frame axes rather than image-plane coordinates. This makes the labels consistent with the manipulation scene, but requires care when interpreting camera-specific model predictions.

EmbodimentSemantic evaluates spatial relation prediction under a fixed object vocabulary. Object names are provided to the model, so the benchmark isolates relational grounding rather than open-vocabulary object detection. This is intentional, but it means that strong performance on EmbodimentSemantic would not by itself imply robust object recognition in unconstrained robot scenes.

## 7 CONCLUSION

We introduced EmbodimentSemantic, a simulator-grounded benchmark for evaluating spatial scene-graph prediction in robot manipulation scenes. EmbodimentSemantic converts LIBERO Spatial demonstrations into frame-level object-relation-object triplets by extracting MuJoCo geometry, world coordinates, and projected bounding boxes from robot camera observations. This produces a controlled offline benchmark for VLM spatial grounding.

Our VLM results show that current models can achieve high relation coverage while still obtaining low exact triplet F1, indicating that plausible spatial language does not necessarily imply precise relational grounding. By evaluating object pairs, relation direction, hallucination, and viewpoint-specific structure directly, EmbodimentSemantic exposes spatial grounding as a measurable bottleneck for embodied foundation models.

The SAMGRAPH baseline complements prompted VLM evaluation with a segmentation-and-geometry pipeline that produces scene graphs for both simulated and real robot observations.

## AI USE STATEMENT

Generative AI tools were used solely to assist with language-level editing and information retrieval. For writing, they helped improve the grammar, clarity, concision, and formatting of text prepared by the authors. For retrieval and discovery, they helped locate potentially relevant literature and public documentation; the original sources were consulted before any information or citation was included. The scientific ideas, methodology, experiments, data, results, and conclusions were developed by the authors.

## REFERENCES


Peter W. Battaglia, Daniel Kersten, and Paul R. Schrater. How haptic size sensations improve distance perception. *PLoS Computational Biology*, 7(6):e1002080, 2011. doi: 10.1371/journal.pcbi.1002080.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi_0$: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010.

François Chaumette and Seth Hutchinson. Visual servo control. i. basic approaches. *IEEE Robotics & Automation Magazine*, 13(4):82–90, 2006. doi: 10.1109/MRA.2006.250573.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024. doi: 10.1109/CVPR52733.2024.01370.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, 2024.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers*, pp. 346–355, 2024.

Irving Fang, Juexiao Zhang, Shengbang Tong, and Chen Feng. From intention to execution: Probing the generalization boundaries of vision-language-action models. *arXiv preprint arXiv:2506.09930*, 2025. doi: 10.48550/arXiv.2506.09930.

ZhiYuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shi Shubin, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, Rushuai Yang, Ruichuan An, Leqi Zheng, Weijie Wang, Shuang Chen, Sicheng Xu, Yaobo Liang, Jiaolong Yang, and Baining Guo. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jXDZJAfRZB.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. RVT: Robotic view transformer for 3D object manipulation. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 694–710. PMLR, 2023.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. In *Proceedings of the European Conference on Computer Vision*, pp. 19–34, 2018.

Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se(3) pose trajectory diffusion for object-centric manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Seth Hutchinson, Gregory D. Hager, and Peter I. Corke. A tutorial on visual servo control. *IEEE Transactions on Robotics and Automation*, 12(5):651–670, 1996. doi: 10.1109/70.538972.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao,

Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

Hang Li, Qian Feng, Zhi Zheng, Jianxiang Feng, Zhaopeng Chen, and Alois Knoll. Language-guided object-centric diffusion policy for generalizable and collision-aware manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12834–12841, 2025a. doi: 10.1109/ICRA55743.2025.11127231.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024.

Xiaoqi Li, Liang Heng, Jiaming Liu, Yan Shen, Chenyang Gu, Zhuoyang Liu, Hao Chen, Nuowei Han, Renrui Zhang, Hao Tang, Shanghang Zhang, and Hao Dong. 3DS-VLA: A 3D spatial-aware vision language action model for robust multi-task manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2344–2359. PMLR, 2025b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8c3c666820ea055a77726d66fc7d447f-Paper-Datasets_and_Benchmarks.pdf.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 06 2023b. ISSN 2307-387X. doi: 10.1162/tacl_a_00566. URL https://doi.org/10.1162/tacl_a_00566.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.011.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 785–799. PMLR, 2023.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15768–15780, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 306–316. PMLR, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/89cc5e613d34f90de90c21e996e60b30-Paper-Conference.pdf.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems*, 2018. doi: 10.15607/RSS.2018.XIV.019.

Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3239–3266. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/xing25a.html.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Ji Zhang, Shihan Wu, Xu Luo, Hao Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. InSpire: Vision-language-action models with intrinsic spatial reasoning, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023.

# A APPENDIX

Table 3: LIBERO Spatial tasks used in EmbodimentSemantic. All tasks share the same goal, placing the black bowl on the plate, but differ in the bowl's initial spatial context.

| Task ID | Initial spatial context of `akita_black_bowl_1` |
|---|---|
| T0 | Between the plate and the ramekin |
| T1 | From table center |
| T2 | In the top drawer of the wooden cabinet |
| T3 | Next to the cookie box |
| T4 | Next to the plate |
| T5 | Next to the ramekin |
| T6 | On the cookie box |
| T7 | On the ramekin |
| T8 | On the stove |
| T9 | On the wooden cabinet |

Table 4: Object and relation ontology used by EmbodimentSemantic. Relations are represented as directed predicates arranged in inverse pairs.

| Type | Name | Description |
|---|---|---|
| Object | `akita_black_bowl_1` | Manipulated bowl instance |
| Object | `akita_black_bowl_2` | Secondary bowl instance / distractor |
| Object | `cookies_1` | Cookie box object |
| Object | `glazed_rim_porcelain_ramekin_1` | Ramekin reference object |
| Object | `plate_1` | Target placement object |
| Object | `wooden_cabinet_1` | Cabinet / drawer reference object |
| Object | `flat_stove_1` | Stove reference surface |
| Relation | `is_left_of` / `is_right_of` | Lateral world-frame ordering |
| Relation | `is_in_front_of` / `is_behind` | Depth world-frame ordering |
| Relation | `is_on_top_of` / `is_below_of` | Vertical support or stacking relation |
| Relation | `is_inside` / `contains` | Containment relation |

Table 5: EmbodimentSemantic dataset summary. A paired timestep contains one `agentview` frame and one `eye_in_hand` frame.

| Dataset attribute | Value |
|---|---|
| Tasks | 10 |
| Demonstrations | 500 |
| Demonstrations per task | 50 |
| Paired timesteps | 62,250 |
| Total RGB frames | 124,500 |
| Frames per demo | 75–197 |
| Mean frames per demo | 124.5 |
| Cameras | 2 |
| RGB resolution | $128 \times 128$ |
| Canonical objects | 7 |
| Directed relations | 8 |
| Mean triplets / frame (`agentview`) | 42.0 |
| Mean triplets / frame (`eye_in_hand`) | 16.73 |

### A.1 DUPLICATE-BOWL IDENTITY RECONCILIATION

LIBERO scenes contain two visually identical black-bowl instances, while the ground-truth scene graphs distinguish them by simulator-specific suffixes. Those suffixes are not observable from a single RGB frame. SamGraph therefore emits provisional identities for its two bowl tracks and maps them to the canonical bowl vocabulary only for scoring. For each sampled frame, the evaluator constructs exactly two assignments: the emitted assignment and the assignment obtained by exchanging `akita_black_bowl_1` with `akita_black_bowl_2` everywhere in the predicted triplets. It retains the assignment with the higher exact ordered-triplet F1. The exchange is selected on 3,980 of the 12,648 sampled `agentview` frames.

This permutation-invariant comparison changes no segmentation mask, predicted predicate, non-bowl identity, evaluated frame, or ground-truth triplet. It removes only the arbitrary naming penalty between two appearance-identical instances and is applied identically in the VLM and SamGraph evaluations. Consequently, the reported score measures spatial-relation recovery under indistinguishable instance labels; it should not be interpreted as a temporal identity-tracking metric.

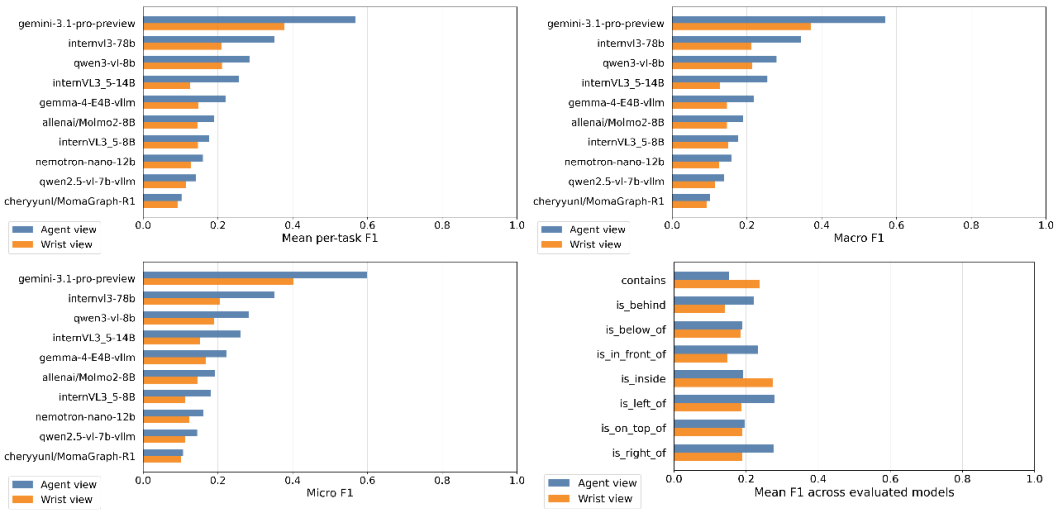


Figure 4: Additional aggregate F1 results for VLM scene-graph prediction.

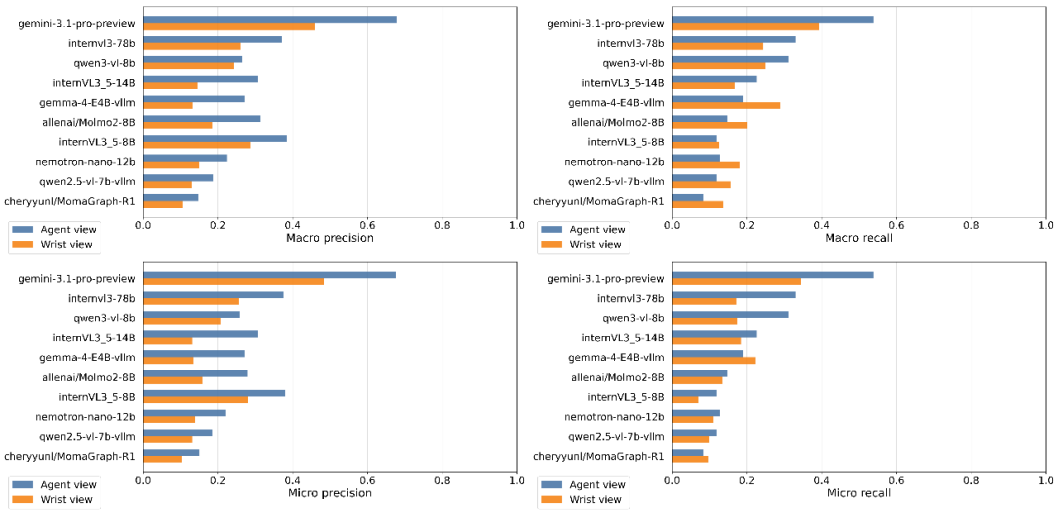


Figure 5: Additional precision and recall results for VLM scene-graph prediction.

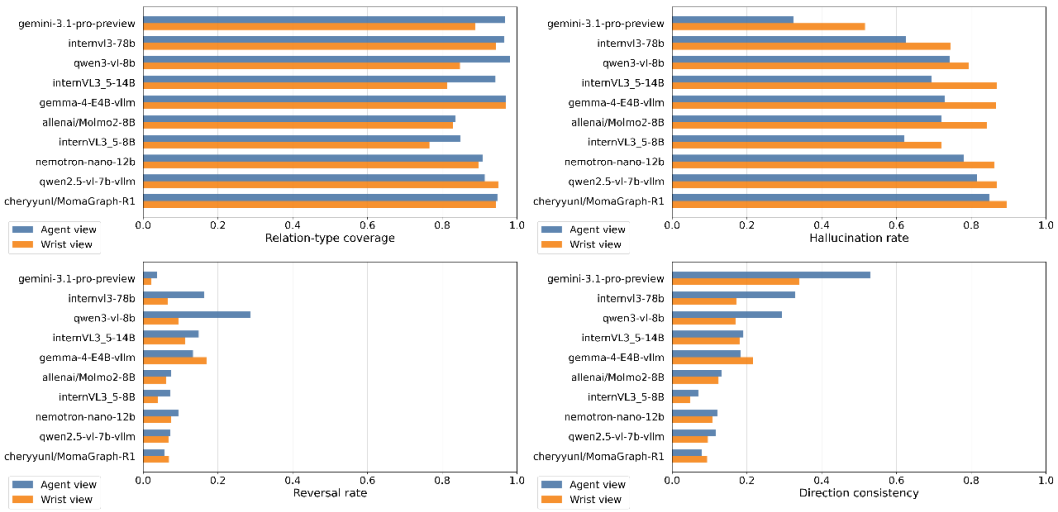


Figure 6: Diagnostic metrics for VLM scene-graph prediction.

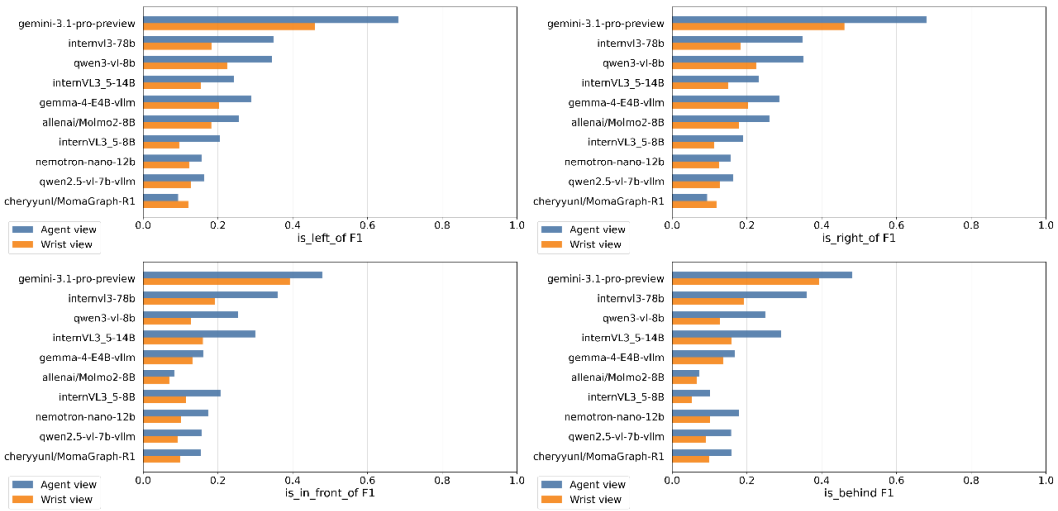


Figure 7: Per-relation F1 for lateral and depth relations.

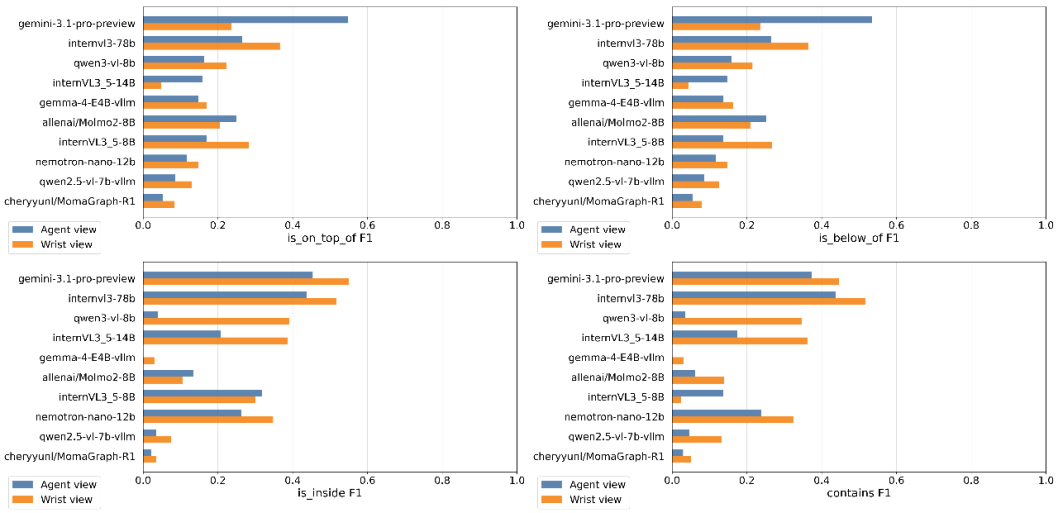


Figure 8: Per-relation F1 for support and containment relations.