RECOGNITION GAINS, COMPOSITIONAL GAPS: RETHINKING SPATIAL REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning underpins real-world applications of vision-language models, including robotics, autonomous driving, and navigation. Yet how advances in spatial recognition relate to models' ability to combine and reuse spatial information remains insufficiently studied. To examine this relationship, we introduce SpaReCo-Bench (**Sp**atial **Re**cognition and **Co**mposition), which organizes 15 tasks across 4 families: state, identity, relation, and composition. Centered on object orientation and spatial transformations, the benchmark grounds basic judgments and compositional problems in shared visual contexts, making it possible to assess recognition and the use of recognized information together. Our evaluations show task-dependent transfer from spatial training to related decisions. SFT fine-tuning improves attribute recognition, while direct supervision improves attribute recombination. Identity-conditioned selection and transformation analogy exhibit distinct transfer patterns, showing that spatial training benefits depend on the operation required by the target task. SpaReCo-Bench provides a controlled foundation for distinguishing recognition gains from progress in compositional reasoning and for examining how spatial information supports decisions across related tasks. The benchmark, dataset, and code will be made publicly available.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have strengthened visual recognition and multimodal understanding (Hurst et al., 2024; Bai et al., 2025). Spatial reasoning extends these capabilities to judgments about object orientation, spatial relations, and changes across viewpoints (Kamath et al., 2023; Tasnim et al., 2025; Jung et al., 2025). In embodied applications such as robotic manipulation (Driess et al., 2023), these judgments must often be combined to select a target satisfying multiple conditions or apply a transformation to a new object, as illustrated in Figure 1. Yet it remains unclear when improvements in individual spatial judgments translate into more reliable compositional decisions.

Figure 1 illustrates how we compare recognition and compositional use. Source-task evaluation measures learning of component judgments, while linked target evaluation measures transfer to joint selection or transformation use. Direct target supervision tests whether the target task can be learned under task-specific training.

Existing studies have made spatial reasoning increasingly measurable. Some isolate individual judgments such as orientation, identity, or spatial relations, while others examine multi-view transformations and compositional scene understanding (Zhang et al., 2025; Gong et al., 2026; Huang et al., 2026; Xiao et al., 2026). These benchmarks provide task-specific accuracy profiles. SpaReCo links component judgments to compositional decisions over shared object views, allowing source learning, cross-task transfer, and direct target learning to be compared within a common evaluation framework.

We introduce SpaReCo-Bench, which organizes state, identity, relation, and composition questions around shared object views and explicit spatial annotations. Its linked questions support a direct study of how spatial information moves from individual judgments to joint selection and transformation decisions. We use targeted training as a measurement of this relationship: spatial training can strengthen individual abilities and benefit related compositional tasks, but the benefit is task dependent

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.

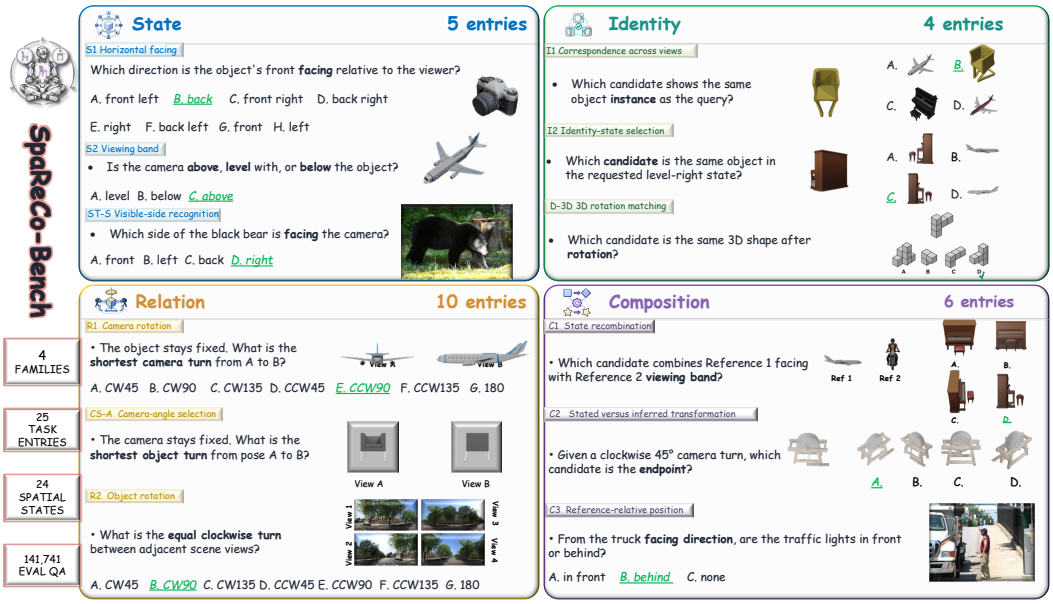


Figure 1: SpaReCo-Bench studies whether recognized spatial information can be reused in linked compositional decisions. The benchmark connects component judgments about attributes and transformations to target questions that require jointly using those judgments over shared visual evidence.

and sensitive to how visual evidence is presented. The benchmark and its linked evaluation structure make these differences measurable, with task families serving as measurement groups.

This work makes 3 contributions: First, it provides a benchmark that connects spatial recognition with compositional use across 15 task identifiers and 4 task families, reported as 25 evaluation entries when source-native stages are scored separately. Second, it preserves explicit links between shared visual evidence, component judgments, and target decisions, which supports controlled diagnosis beyond a single accuracy score. Third, it uses matched training and inference comparisons to study when spatial training gains transfer to related decisions and their task-specific limits.

2 Related Work

Spatial reasoning research has developed along 3 complementary directions: controlled recognition of spatial properties, composition of multiple spatial conditions, and training methods intended to improve transfer across these abilities. Table 1 places representative benchmarks within this landscape. The comparison uses task demands as an organizing device, with each category defining a measured evaluation scope.

Table 1 shows that the listed benchmarks typically emphasize recognition or one form of composition. SpaReCo-Bench provides explicit tasks in all four columns within one reporting suite, which enables source learning, target learning, and transfer to be compared under a common task inventory.

Recognition and spatial transformations. What's Up (Kamath et al., 2023) isolates basic spatial relations, EgoOrientBench (Jung et al., 2025) focuses on object-facing judgments, and DORI (Tasnim et al., 2025) examines orientation at multiple levels of granularity. RotBench (Niu et al., 2026) studies image-plane rotation, while SpinBench (Zhang et al., 2026) organizes identity, relation, translation, rotation, and perspective taking into a controlled diagnostic progression. These studies establish important ways to measure spatial information extraction and changes in visual reference frames.

Composition and diagnostic structure. SPHERE (Zhang et al., 2025) relates intermediate judgments to final reasoning questions and tests the effect of supplying intermediate facts. SpaCE-10 (Gong et al., 2026) combines atomic capabilities into multi-skill spatial questions, while KiVA (Yiu

Table 1: Comparison of spatial benchmarks across recognition and composition tasks. AR/AC and TR/TA denote attribute recognition/composition and transformation recognition/analogy, respectively.

Benchmark	Evaluation items	Domain	Source	Attributes		Transformations	
				AR	AC	TR	TA
BLINK (Fu et al., 2024)	3,807	Mixed	Real + Synthetic	✓	△	✓	✓
What's Up (Kamath et al., 2023)	820	Objects	Real	✓	✗	✗	✗
EgoOrientBench (Jung et al., 2025)	33,460	Objects	Real + Synthetic	✓	✗	✗	✗
DORI (Tasnim et al., 2025)	33,656	Objects	Real + Synthetic	✓	△	✓	✗
RotBench-Large (Niu et al., 2026)	1,200	General	Real	✓	✗	△	✗
3DSRBench (Ma et al., 2025a)	2,772	General	Real + Synthetic	✓	△	✗	✗
SPHERE (Zhang et al., 2025)	2,285	General	Real	✓	✓	✗	✗
SpaCE-10 (Gong et al., 2026)	5K+	General	Real	✓	✓	✗	✗
VSI-Bench (Yang et al., 2025a)	5K+	General	Real	✓	△	✗	✗
MMSI-Bench (Yang et al., 2026)	1,000	General	Real	✓	✓	✓	✗
SpinBench (Zhang et al., 2026)	2.7K	Mixed	Real + Synthetic	✓	△	✓	✗
KiVA / KiVA-adults (Yiu et al., 2025)	4,300	Objects	Real	△	✗	✓	✓
OmniSpatial (Jia et al., 2026)	1.5K	Mixed	Real + Synthetic	✓	△	✓	△
Spatial-DISE Bench (Huang et al., 2026)	559	Cognition	Real + Synthetic	✓	△	△	✗
Spatial-DISE-12K test (Huang et al., 2026)	1,854	Cognition	Real + Synthetic	✓	△	△	✗
SpatialTree (Xiao et al., 2026)	6,300	Mixed	Real + Synthetic	✓	△	✓	✗
SpatiaLQA (Xie et al., 2026)	9,605	General	Real	△	✓	✗	✗
SpaReCo-Bench (Ours)	141,741	Mixed	Real + Synthetic	✓	✓	✓	✓

✓ explicit task; △ related/partial; ✗ no explicit task.

et al., 2025) separates recognizing a transformation, describing its magnitude, and transferring it to a new object. MMSI-Bench (Yang et al., 2026) links multi-image questions with intermediate reasoning annotations. Spatial-DISE (Huang et al., 2026) organizes static and dynamic spatial reasoning into 4 quadrants and studies transfer to external tasks. Together, these efforts motivate connecting component judgments to the decisions they support. SpaReCo-Bench focuses on making that connection explicit for shared object views and linked target questions.

Training and transfer. VST (Yang et al., 2025b) combines spatial perception supervision with reasoning-oriented training, and SpatialLadder (Li et al., 2025) studies staged improvements from spatial understanding toward reasoning. HiSpatial (Liang et al., 2026) analyzes dependencies across levels of 3D understanding, while SpatialReasoner (Ma et al., 2025b) introduces explicit representations for spatial inference. SpatialTree (Xiao et al., 2026) reports negative transfer among lower-level abilities and positive transfer toward higher-level abilities, showing why task-level transfer patterns characterize spatial learning. These studies motivate our central comparison: after a source judgment improves, how does the gain reach a related target decision? We address this question through linked recognition and composition tasks over shared visual evidence.

3 SpaReCo-Bench

SpaReCo-Bench studies how spatial attributes support joint selection and how demonstrated transformations support choices for a new object. Figure 2 summarizes the task inventory and shared visual evidence. Question counts and distinct-image counts measure different aspects of the benchmark: 1 view can support several questions with different information demands. The shared-image structure links these demands over common object views, enabling comparisons between component judgments and related target decisions.

3.1 Spatial tasks and benchmark scope

The benchmark contains 14 controlled tasks and 11 additional evaluation entries, giving 25 reporting columns across 4 families. Together with the retained C4 reference-relative task, these form 15 task identifiers; source-native stages are kept as separate reporting entries because they use their own answer spaces and scorers. State questions ask for horizontal facing (S1), viewing band (S2), their joint state (S3), and coarse facing (S4). Identity questions ask for cross-view object correspondence (I1) and selection by identity and state together (I2). Relation questions distinguish camera and object rotation (R1 and R2), measure directed offset (R3), identify changed components (R4), and ask for unsigned angular difference (R5). Composition questions recombine attributes from 2

references (C1), apply an explicitly stated transformation (C2), or infer a transformation from a demonstration and apply it to a new start (C3). State and identity provide attribute judgments, relation questions characterize changes, and composition questions test how that information is used. Family membership follows the question's primary demand.

SpaReCo-Bench combines retained object views and newly rendered controlled views from 6 object sources: OA-V1, an author-labelled object-view collection (Wang et al., 2024), Objaverse-OA (Deitke et al., 2023), ABO (Collins et al., 2022), HouseCat6D (Jung et al., 2024), CanoVerse (Jin et al., 2026), and Pix3D (Sun et al., 2018). These sources cover released object views, aligned 3D assets, product objects, household objects, canonicalized assets, and CAD furniture. Eligible assets are rendered with a controlled Blender Workbench pipeline under fixed azimuth and viewing-band settings, using source-specific alignment, materials, and textures where available. Evaluation questions extend the controlled tasks with distinct visual formats. SpatialTree (Xiao et al., 2026) supplies visible-side recognition (ST-S) and reference-relative position (C4). Spatial-DISE (Huang et al., 2026) supplies 3D and 2D rotation matching (D-3D, D-2D), while CoSpace (Zhu et al., 2025b) supplies camera-angle selection (CS-A). For KiVA (Yiu et al., 2025), K-T identifies the change type, K-M asks for its magnitude, and K-A tests analogy; KiVA-adults uses the corresponding KA-T, KA-M, and KA-A identifiers. Each KiVA stage is asked independently, with questions linked by source trial. Detailed source composition, rendering and selection rules, annotation pipelines, full prompts, and answer spaces appear in the Supplementary Material, Sections D, F, and G.

The core contains 485,914 training, 69,964 validation, and 139,295 test questions. 5 source datasets contribute 4,544 records: 2,098 Spatial-DISE training and validation questions and 2,446 evaluation questions. Each task retains its source-specific answer space and scorer.

Figure 2 separates question volume from visual coverage. Panel (a) shows that one image can support multiple questions, while panel (b) identifies where task entries reuse the same visual evidence. This reuse supports matched cross-task diagnosis while leaving the task definitions and answer spaces distinct; an edge denotes shared input, not a claimed prerequisite relation.

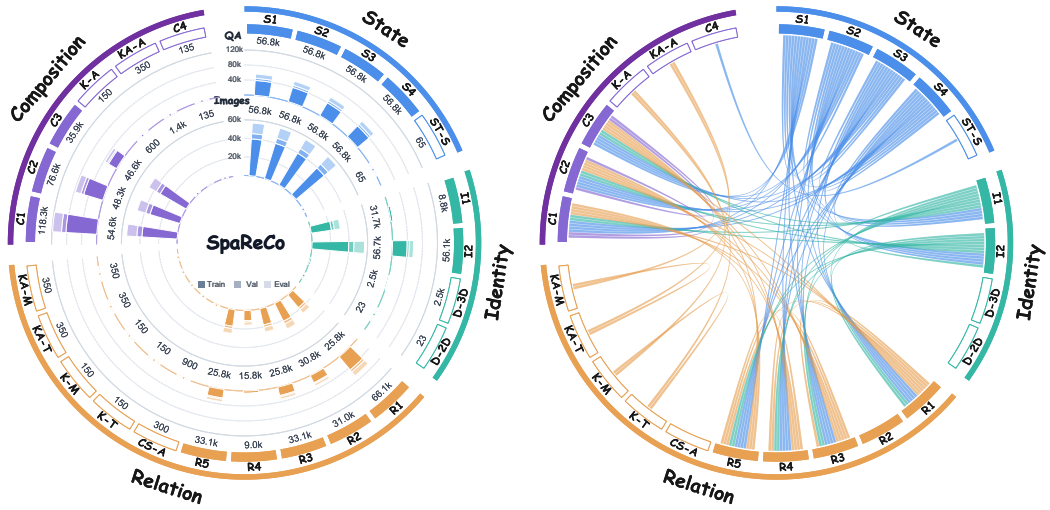
3.2 Linked questions for compositional diagnosis

An episode links a target question to existing component questions through view identities and semantic annotations. Each question retains its own input images and is answered independently. These links support comparisons between component accuracy and target selection; their coverage and input contexts determine which paired analyses are available.

State recombination. 2 reference objects supply different attributes: horizontal facing from the first and viewing band from the second. 4 candidate views of a third object cross 2 facing directions with 2 viewing bands. Either attribute alone leaves 2 candidates, while their conjunction identifies 1. The linked single-image questions ask for the relevant attribute of each reference and the joint state of each candidate. Their inputs differ from the six-image recombination question. 4-condition groups change the reference attributes while retaining the candidate set, testing whether selections follow both constraints. This design supports the component-response and reference-response analyses in Section 6.

Transformation analogy. A demonstration pair specifies a visual transformation, and a new start image defines the object to which it must be applied. The model selects the corresponding endpoint from a candidate set. A matched action-endpoint question supplies the same transformation explicitly while retaining the new start and candidates. These paired questions compare selection with a stated action against selection requiring its extraction from a demonstration. Together with independent state and relation queries, they extend the diagnosis from recombining attributes to transferring a transformation.

I2 provides a second conjunction: its candidate gallery crosses 2 object identities with 2 spatial states, leaving 1 candidate that satisfies both conditions. Identity alone matches 2 candidates, so an single-choice identity question cannot uniquely resolve this gallery. The model must combine identity with state to select the correct candidate. The construction isolates joint selection by identity and state independently of the C1 reference combination.



(a) Task and image distributions (b) Shared-image topology

Figure 2: SpaReCo-Bench task inventory and image sharing. (a) Training, validation, and evaluation question counts and distinct input-image counts by task ID. (b) Image sharing across task entries; connections trace shared visual evidence across tasks.

3.3 Construction, Splits, and Validation

The core construction aligns each object's views with semantic state labels, then forms pairs, reference groups, and candidate sets for each task. Semantic annotations link component judgments to compositional decisions, preserving image order and reference roles. Supplementary Material, Sections F and G, gives source-specific construction rules and task definitions.

The full inventory contains 699,717 question-answer (QA) records: 695,173 core questions across training, validation, and test, plus 4,544 questions from the 5 additional evaluation sources. Table 1 reports 141,741 evaluation questions: 139,295 core test and 2,446 additional evaluation questions. Core partitions are disjoint in object identity, image path, and image content, as checked for overlap. Original source partitions are retained where defined: Spatial-DISE preserves its official partitions, SpatialTree preserves its official test identity, and CoSpace and the KiVA series enter SpaReCo as evaluation tasks. KiVA contributes 3 independent requests per source trial, with trial grouping preserved for statistics and no previous answer passed between requests. All 135 C4 questions belong to the benchmark evaluation set. C2/C3 use nearest-45° horizontal changes within the same viewing band and allow at most 5° elevation difference.

Validation covers annotation consistency and visual answer validity. Annotation checks examine the agreement between answer options and spatial labels, the consistency of angular conventions, candidate uniqueness, and completeness of linked groups. Visual review additionally examines whether the requested facing or viewpoint is identifiable, including symmetry and occlusion cases. Review scope and task-specific eligibility are documented in Supplementary Material, Sections F.2 and G. The shared records support the full-task comparisons in Section 4 and the matched diagnostics in Section 6.

4 Benchmark Evaluation

We report task-wise benchmark performance and use controlled fine-tuning to test whether gains on a supervised source reach related held-out tasks. The linked episodes in Section 3.2 connect component judgments with target selections.

4.1 Experimental Setup

Benchmark models. Table 2 groups the selected models by architecture and spatial training. General-purpose models comprise Qwen2.5-VL-3B-Instruct (Bai et al., 2025), InternVL3-2B (Zhu et al., 2025a), LLaVA-OneVision-7B (Li et al., 2024), DeepSeek-VL2-Tiny-3B (Wu et al., 2024),

Table 2: Task-level accuracy (%) across SpaReCo-Bench. Columns retain task-specific answer spaces and scoring rules; dashes mark unreported measurements.

	State					Identity				Relation										Composition					
Model	S1	S2	S3	S4	ST-S	I1	I2	D-3D	D-2D	R1	R2	R3	R4	R5	CS-A	K-T	K-M	KA-T	KA-M	C1	C2	C3	K-A	KA-A	C4
General-purpose VLMs																									
Qwen2.5-VL-3B-Instruct	15.97	42.04	4.91	38.66	30.77	32.90	25.43	27.42	21.74	13.62	14.58	13.72	34.07	21.55	50.00	82.00	50.00	82.29	39.43	26.63	25.13	25.15	27.33	29.14	43.70
InternVL3-2B	15.98	30.27	5.14	37.77	35.38	28.34	25.56	24.35	13.04	13.30	13.96	13.42	32.72	26.85	50.33	76.67	37.33	73.71	28.57	25.04	25.07	24.75	34.00	34.86	48.89
LLaVA-OneVision-7B	25.06	45.81	7.68	42.42	44.62	23.63	27.71	24.35	30.43	13.55	13.90	14.41	35.01	27.82	50.00	98.00	59.33	88.86	34.57	24.85	24.97	24.90	35.33	34.00	51.11
DeepSeek-VL2-Tiny-3B	13.83	30.29	4.28	37.26	40.00	23.15	25.59	36.64	21.74	13.05	14.46	13.73	33.33	26.38	50.00	36.00	42.67	34.29	46.00	25.00	24.96	25.58	34.00	34.00	51.11
GLM-Edge-V-5B	18.02	39.84	5.68	41.80	23.08	23.72	20.10	28.84	39.13	13.42	14.56	14.09	35.60	26.07	49.67	92.00	50.00	86.86	38.29	25.12	25.42	25.34	40.67	38.00	49.63
Spatially tuned VLMs																									
DORI-3B	22.74	39.54	7.97	37.48	35.38	23.03	24.49	29.79	34.78	15.38	14.85	15.11	34.17	27.15	47.00	70.00	56.67	60.86	42.57	25.03	24.82	25.54	31.33	26.00	54.07
EgoOrient-InternVL3-2B	16.06	29.92	6.79	38.37	35.38	20.47	23.72	24.35	13.04	14.30	14.40	13.14	32.40	27.08	50.33	76.67	37.33	73.71	28.57	24.99	24.65	25.14	34.00	34.86	48.89
VST-3B-SFT	15.39	32.93	7.31	36.64	40.00	9.15	19.10	28.84	26.09	13.93	13.65	15.20	38.46	29.10	50.00	88.67	52.00	84.00	46.00	25.67	25.23	25.11	38.67	37.14	51.85
VST-3B-RL	16.05	33.03	7.44	41.15	38.46	10.03	20.35	30.50	43.48	14.33	14.77	15.20	37.29	29.29	51.00	60.00	53.33	56.57	47.71	24.53	25.16	24.69	38.67	42.57	52.59
SpatialLadder-3B	13.09	40.27	4.90	37.36	29.23	32.38	27.31	30.97	26.09	13.84	14.35	13.96	35.95	27.91	50.00	38.00	53.33	26.57	49.43	25.09	24.64	24.83	34.00	29.43	42.96
Spatial-SSRL-3B	18.42	41.78	5.05	39.57	41.54	35.80	27.51	25.06	30.43	14.23	14.09	15.34	34.83	28.85	50.67	88.00	47.33	87.14	45.71	24.85	24.77	25.63	31.33	29.14	48.15

and GLM-Edge-V-5B. The spatially tuned comparisons comprise DORI-3B (Tasnim et al., 2025), VST-3B-SFT and VST-3B-RL (Yang et al., 2025b), SpatialLadder-3B (Li et al., 2025), Spatial-SSRL-3B (Liu et al., 2026), and our InternVL3-2B reproduction of EgoOrient training (Jung et al., 2025). Controlled transfer comparisons use the same Qwen base checkpoint; the broader model table describes performance across architectures and spatial training recipes.

Controlled fine-tuning. The controlled training arms start from Qwen2.5-VL-3B-Instruct. Task-specific supervision measures source learning and direct target learning, while evaluation on other tasks measures transfer. The main diagnostic arms use 12,000 unique training questions per source except I1 and R4, which use all 6,142 and 6,275 available questions, respectively. Epoch counts and training settings appear in Supplementary Material, Section A.2. Optimization steps and token budgets differ across tasks. Full-task training is evaluated separately.

Data and evaluation. We use the frozen benchmark partitions and the scoring rules in Supplementary Material, Sections A.3 and D. Task accuracy is the primary metric; class-balanced accuracy is a supplementary diagnostic for uneven semantic labels. A core macro-average weights the 14 core tasks equally; the additional evaluation entries are reported individually. Table 2 reports test accuracy by task. Controlled transfer conditions use identical held-out questions. We report task-wise accuracy and paired changes, resampling linked questions by owner or linked group. Source-level variation is reported separately. Complete-group accuracy counts an episode as correct only when all of its target questions are correct. Paired changes are reported in percentage points with 95% cluster-bootstrap intervals. For episodes sharing objects, connected owner groups define the clusters. Undefined estimates are marked NA; unmeasured results are marked by dashes.

4.2 Benchmark coverage and task-wise performance

Table 2 reveals an uneven performance profile across spatial tasks. On the evaluated questions, the reported accuracies for explicit transformation application (C2) and demonstration-based analogy (C3) cluster around 25%, the uniform-choice reference for their four candidates. This pattern occurs among both general-purpose and spatially tuned models. The state tasks also differ substantially: for example, LLaVA-OneVision obtains 45.81% on viewing band (S2) and 7.68% on joint state (S3).

Figure 3 summarizes the same results as four task-family macro-averages. The comparison includes general-purpose base models and spatially tuned expert models. The profile is heterogeneous rather than uniformly improved: some spatially tuned models obtain higher Relation or Composition averages, while State or Identity remains modest. Spatial-SSRL, for example, reaches 35.80% on I1, but its overall family profile does not dominate all four axes. This supports a task-dependent view of spatial training gains and shows that foundational recognition and downstream composition remain imperfect. Because the rows differ in architecture, training recipe, and checkpoint, the figure is descriptive rather than a causal comparison.

5 From Recognition Learning to Compositional Transfer

We compare independently trained Qwen2.5-VL-3B models on identical held-out questions. Training uses 12,000 unique questions per source, with the I1 and R4 exceptions described in Supplementary

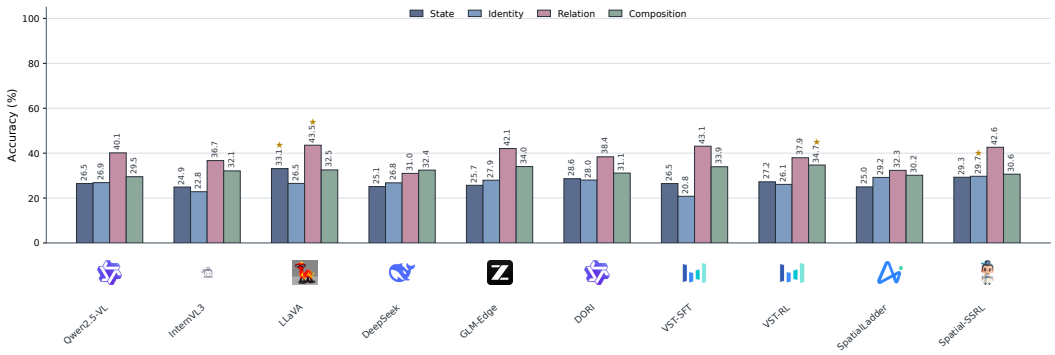


Figure 3: Task-family average accuracy across SpaReCo-Bench model families. Each bar is the unweighted macro-average of the task-level accuracies in Table 2; stars mark the highest available family average.

Table 3: Source learning and cross-task transfer. Entries are accuracy changes in percentage points relative to Qwen2.5-VL-3B.

Path	Source	Target	Source gain (pp)	Transfer gain (pp)	Direct target gain (pp)
Attribute path					
	S3	C1	+19.81 [18.30, 21.31]	-1.83 [-3.91, 0.00]	+14.42 [6.73, 18.24]
	S3	I2	+19.81 [18.30, 21.31]	-9.00 [-13.76, -5.56]	+39.50 [31.85, 47.89]
	I1	I2	+19.00 [15.06, 25.54]	-0.25 [-3.30, 4.52]	+39.50 [31.85, 47.89]
Transformation path					
	R3	C2	+7.76 [3.60, 12.09]	-0.83 [-4.52, 2.70]	+0.83 [-3.39, 5.13]
	R3	C3	+7.76 [3.60, 12.09]	+0.33 [-1.27, 1.95]	+0.50 [-4.30, 3.86]
	C2	C3	+0.83 [-3.39, 5.13]	+0.33 [-4.19, 2.34]	+0.50 [-4.30, 3.86]



Figure 4: Task-dependent transfer across ten source weights and six common target tasks. Cells report accuracy changes in percentage points on the shared diagnostic test; outlined cells mark source-task learning. Gray cells mark unavailable measurements.

Material, Section A.2. Table 3 separates learning the source, transfer to an untrained target, and learning the target through direct supervision. Each gain is measured against the base model on the same task. Supplementary Material, Section B.2, gives scores for six training sources; Section B.3 reports source-task accuracy for all ten.

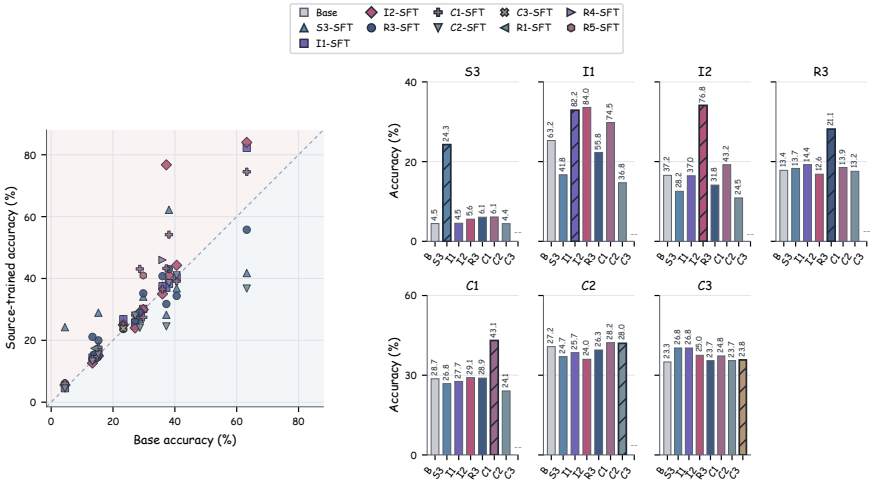
Source learning and target gains. S3 training improves S3 by 19.81 percentage points, yet changes C1 by -1.83 points and I2 by -9.00 points. Direct target supervision improves C1 by 14.42 points and I2 by 39.50 points. These contrasts show that the target tasks are learnable under direct supervision and that related-source gains have selective transfer patterns. Identity training provides another

asymmetric comparison: I1 improves by 19.00 points, while I2 changes by -0.25 points. In the reverse direction, I2 training improves I1 by 20.75 points.

Figure 4 compares ten source-trained models across six common target tasks: S3, I1, I2, R3, C1, and C2. Let $A_{s,t}$ be target-t accuracy after training on source s, and let $A_{0,t}$ be the corresponding Base accuracy. Each cell reports $\Delta_{s,t} = A_{s,t} - A_{0,t}$ in percentage points. Gray cells indicate unavailable measurements. Outlined cells show direct source-task gains when the training task is among the displayed targets; the remaining cells measure transfer to other tasks.

The common-target matrix summarizes direct learning and cross-task transfer. C1-SFT improves all six displayed targets, with a mean change of $+26.1$ points, whereas R3-SFT improves one target and has a mean change of -15.9 points. S3-SFT improves four of six targets (mean $+13.1$ points), while I2-SFT improves two and decreases four (mean -0.7 points). These averages include direct gains on the training task wherever it is displayed. Across State, Identity, Relation, and Composition, the sign and magnitude of each change depend on both the training source and the target task. The off-diagonal entries capture this directional variation in transfer between tasks.

Figure 5 summarizes the same ten-source, six-target evidence at two levels. The panels place direct gains alongside target-dependent off-diagonal changes, revealing selective transfer.



(a) All evaluated source-target pairs

(b) Target accuracy by training source

Figure 5: Source learning and cross-task transfer. (a) Base versus source-trained accuracy over the verified source–target pairs in the common-target matrix. Points above and below the diagonal show positive and negative transfer. (b) Absolute accuracy on the displayed targets for Base and the available source-trained weights. Outlined, hatched bars identify direct target supervision.

6 Compositional Answer Behavior

Linked questions reveal whether performance remains reliable across changing reference conditions and across related transformation tasks.

Condition following and transformation use. The four conditions in each original C1 group test whether choices respond consistently to changing references. Supplementary Material, Section B.6, reports the distribution of correctly answered conditions and the joint outcomes on paired C2/C3 questions. The latter share target and candidate views, while their five- and seven-image inputs distinguish explicit application from demonstration-based analogy. Their joint outcomes characterize task-level behavior, with direct-training controls provided by Table 3.

Complete success over all four reference conditions occurs in 6 of 300 groups (2.00%) after C1 training, compared with 0 groups for Base and S3-SFT. Reliable selection across all reference conditions thus remains rare even after direct C1 supervision. For paired application and analogy, both answers are correct on 58/600 pairs for Base, 52/600 for R3-SFT, and 59/600 for C2-SFT. Joint success on application and analogy remains low across these training conditions.

Table 4: External spatial evaluation of Base and S3-SFT.

	3DSRBench + BLINK							SAT + OmniSpatial	
Method	Orient.	View→Obj.	Parallel	Same Dir.	MV	Corr.	Rel. Depth	SAT Overall	Omni Overall
Base	32.5	11.4	11.4	46.8	42.9	25.3	64.1	51.3	40.3
S3-SFT	30.8 (↓ −1.7)	13.7 (↑ +2.3)	33.9 (↑ +22.5)	37.8 (↓ −9.0)	38.4 (↓ −4.5)	24.4 (↓ −0.9)	37.1 (↓ −27.0)	58.0 (↑ +6.7)	38.4 (↓ −1.9)

Corrections and newly introduced errors. A net gain can combine corrected answers with newly introduced errors. Separating these changes on identical target questions shows whether higher accuracy reflects consistent correction or a mixture of gains and regressions. Supplementary Material, Section C and Table S6, reports both components and uncertainty accounting for shared objects and groups.

7 Generalization to External Benchmarks

Table 4 reports Base and S3-SFT on 3DSRBench (Ma et al., 2025a), BLINK (Fu et al., 2024), SAT (Ray et al., 2025), and OmniSpatial (Jia et al., 2026). Each benchmark retains its native split, prompt, and scorer.

At the benchmark level, full-data S3-SFT reaches 43.8% on 3DSRBench, 33.3% averaged over the three reported BLINK subtasks, 58.0% on SAT, and 38.4% on OmniSpatial. Relative to the reported Base references, these correspond to approximately -0.2, -10.8, $+6.7$, and -1.9 percentage points, respectively. S3 training therefore improves SAT overall performance but does not produce a uniform gain across external spatial benchmarks, supporting the task-dependent transfer pattern observed in the matched SpaReCo experiments. Table 4 also shows variation within benchmarks, including gains on parallelism and losses on relative depth. Together with Table 3 and Figures 4–5, these results show that the benefit of S3 supervision depends on the target task.

8 Conclusion

SpaReCo-Bench studies how vision-language models recognize spatial information and use it in linked decisions. It organizes 15 tasks across state, identity, relation, and composition, connecting basic spatial judgments with attribute selection and transformation use over shared visual evidence.

Our results show that spatial training produces selective gains. Models can improve on directly supervised spatial judgments and transfer those gains to some related tasks, while other tasks show little transfer or negative transfer. This task-dependent pattern separates recognition learning from compositional use and identifies which target decisions benefit from each training source. Linked questions and matched source–target evaluations make these relationships measurable at the task level.

Together, these results motivate evaluating spatial recognition and compositional use through linked tasks, with explicit visual evidence and task-specific answer spaces.

AI use statement

Generative AI tools were used only to assist with manuscript writing.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. ABO: Dataset and benchmarks for real-world 3D object understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21094–21104. IEEE, 2022.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. URL https://arxiv.org/abs/2303.03378.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73337-6_9. URL https://mlanthology.org/eccv/2024/fu2024eccv-blink/.

Ziyang Gong, Wenhao Li, Xianzheng Ma, Songyuan Li, Zhaokai Wang, Songze Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5b84864ff8474fd742c66f219b2eaac1-Abstract-Conference.html.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bMINsPQpME.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.03135.

Li Jin, Yuchen Yang, Weikai Chen, Yujie Wang, Dehao Hao, Tanghui Jia, Yingda Yin, Zeyu Hu, Runze Zhang, Keyang Luo, et al. CanoVerse: 3D object scalable canonicalization and dataset for generation and pose. *arXiv preprint arXiv:2603.07144*, 2026.

HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. HouseCat6D: A large-scale multi-modal category level 6D object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2024.

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14257–14267, 2025.

URL https://openaccess.thecvf.com/content/CVPR2025/html/Jung_Is_Right_Right_Enhancing_Object_Orientation_Understanding_in_Multimodal_Large_CVPR_2025_paper.html.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. URL https://arxiv.org/abs/2510.08531.

Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. HiSpatial: Taming hierarchical 3D spatial understanding in vision-language models. *arXiv preprint arXiv:2603.25411*, 2026. URL https://arxiv.org/abs/2603.25411.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-SSRL: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9581, 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_3DSRBench_A_Comprehensive_3D_Spatial_Reasoning_Benchmark_ICCV_2025_paper.html.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan L. Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-4703. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cdd63c6c813a6ec826051b6a49f11b5e-Abstract-Conference.html.

Tianyi Niu, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. RotBench: Evaluating multimodal large language models on identifying image rotation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2508.13968.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, and Kuo-Hao Zeng. SAT: Spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 2974–2983. IEEE, 2018.

Nazia Tasnim, Keanu Nichols, Yuting Yan, Nicholas Ikechukwu, Elva Zou, Deepti Ghadiyaram, and Bryan A. Plummer. Seeing isn't orienting: A cognitively informed hierarchical benchmark for object orientation in MLLMs, 2025. URL https://arxiv.org/abs/2505.21649v8.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. SpatialTree: How spatial intelligence branches out in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16701–16711, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xiao_SpatialTree_How_Spatial_Intelligence_Branches_Out_in_MLLMs_CVPR_2026_paper.html.

Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Hao Zhu, Rui Tang, Rong Wei, Mingli Song, Yuanyu Wan, and Jie Song. SpatiaLQA: A benchmark for evaluating spatial logical reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.20901.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025a. doi: 10.1109/CVPR52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b. URL https://arxiv.org/abs/2511.05491.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gHRoX4vXm3.

Eunice Yiu, Maan Qraitem, Anisa Noor Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. KiVA: Kid-inspired visual analogies for testing large multimodal models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vNATZfmY6R.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025. doi: 10.18653/v1/2025.acl-long.568. URL https://aclanthology.org/2025.acl-long.568/.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25390.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Chang Gao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a. URL https://arxiv.org/abs/2504.10479.

Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. CoSpace: Benchmarking continuous space perception ability for vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29569–29579. IEEE, 2025b.