FORMATTING INSTRUCTIONS FOR ICLR 2027 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) excel at visual understanding, but still struggle to reason reliably about spatial relationships. Models may capture relevant object cues while failing to bind them into a globally consistent spatial configuration, limiting their ability to reason about relations across multiple objects. To address this, we propose **BindOPSD**, a post-training framework that strengthens spatial reasoning through configurational binding. BindOPSD first constructs a privileged teacher context with a joint spatial view and correspondence-aware attention, which organizes local object evidence within a shared spatial frame. It then transfers this capability to the student through on-policy self-distillation, enabling the model to internalize the enhanced spatial reasoning while requiring only the original image at inference. To preserve general visual capabilities during spatial adaptation, we further introduce gradient decomposition to reduce interference between spatial learning and general capability retention. We also introduce **FineBind**, a fine-grained spatial reasoning benchmark designed to evaluate global configuration and object-to-position binding. Experiments across diverse spatial reasoning benchmarks demonstrate consistent improvements over strong baselines, with effective transfer beyond the training distribution and no additional visual inputs at inference.

1 INTRODUCTION

In recent years, Vision-Language Models (VLMs) have developed strong visual perception capabilities Bai et al. (2025a); Zhu et al. (2025); Huang et al. (2026), enabling them to connect language to visual evidence and answer questions about complex scenes. Yet their spatial reasoning remains challenging Jia et al. (2026), motivating recent efforts to enhance spatial reasoning through spatially focused training and the incorporation of region-level and geometric cues Chen et al. (2024); Cheng et al. (2024). Although these approaches can improve performance, they often rely on specialized supervision or additional visual components that provide useful local cues but may hinder the model's ability to integrate competing evidence across the scene, ultimately leading to unreliable spatial judgments Qi et al. (2026). Spatial reasoning therefore requires an integrated view of the relevant scene that preserves both local details and the spatial relationships that connect them. Human vision offers a useful perspective, gathering information from successive fixations while maintaining its relation to the broader scene Lu & Golomb (2024); Chiu & Golomb (2025). This suggests that effective spatial reasoning depends not only on perceiving individual spatial cues, but also on integrating them into a coherent scene-level representation.

This limitation does not necessarily mean that VLMs lack spatial knowledge, but may reflect their difficulty applying it when visual evidence reaches the reasoning stage as disconnected pieces Yuan et al. (2026). Bringing these pieces together inside the model would require information to flow between local cues and the wider scene. Yet this process unfolds through latent states that are difficult to observe or supervise directly, motivating us to intervene earlier, before visual encoding. In current VLMs, however, this organization is constrained by visual encoding, which compresses the full image into a limited set of tokens. Background content can obscure subtle object cues, while salient regions may dominate the representation Chen et al. (2025). To address this bottleneck, which makes spatial knowledge harder to use Hannan et al. (2026), a visual representation must keep local details clear and preserve their links to a shared spatial layout. A representation can support spatial reasoning only if it preserves both, a condition we define as configurational sufficiency.

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.

To address this, we propose BindOPSD, a method designed to strengthen spatial binding while allowing the final model to reason from the original image alone. We find that providing a privileged visual view that jointly covers all referred objects can substantially strengthen spatial reasoning. To construct this view, JOINT extracts and rescales the smallest region of the original image covering all referred objects, providing stronger evidence for these objects while preserving their relative layout. Because a single joint view may miss details that are clearer in individual crops, we retain those crops as additional evidence. This creates a correspondence problem across views, as the model must match each crop to the correct position in the JOINT view. We address this with Correspondence-Prior Attention (CPA), which uses the known mapping from each crop to JOINT to guide the match through attention logits. Together, JOINT and CPA provide the teacher with privileged spatial evidence that strengthens its spatial reasoning without updating its parameters. We then transfer this advantage to a model that reasons from the original image alone using On-Policy Self-Distillation (OPSD) Zhao et al. (2026). The student generates its own trajectory from the original overview, while the teacher uses the privileged visual information to provide a next-token distribution at each state. This supervision distills the teacher's enhanced spatial reasoning capability into the student. To preserve general abilities during spatial post-training Phan et al. (2026), we introduce gradient decomposition. It separates aligned and conflicting directions between spatial learning and general capability retention, then preserves aligned updates and suppresses conflicting ones. To train and evaluate the binding capability targeted by BindOPSD, we introduce FineBind, a fine-grained spatial reasoning benchmark with a focused training split. Training BindOPSD on this split alone improves performance across diverse external spatial reasoning tasks, suggesting that FineBind captures a spatial binding weakness that extends beyond a single benchmark.

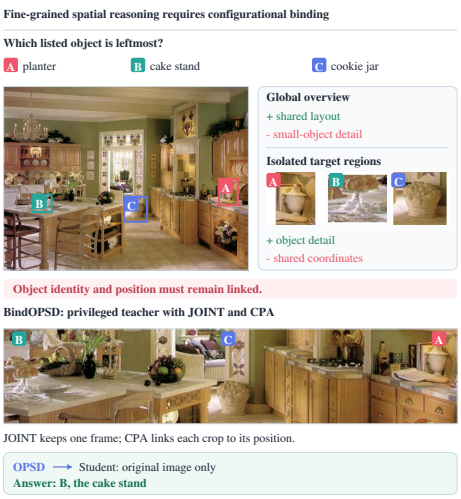


Figure 1: **Why configurational binding matters for spatial reasoning.**

Our main contributions are summarized as follows:

1. We identify a binding bottleneck in multi-object spatial reasoning and show that training on such scenes is crucial. To address this bottleneck, we introduce the first multi-object on-policy distillation method for VLM post-training, achieving strong results across spatial reasoning benchmarks.
2. We propose BindOPSD, which uses JOINT and CPA to construct a privileged teacher with substantially stronger spatial reasoning. It then transfers this capability to a student VLM through OPSD. The student achieves strong performance without requiring any privileged modules during inference, reducing inference time and cost. We also use gradient decomposition to improve generalization.
3. We further introduce FineBind, a fine-grained spatial reasoning benchmark built around complex scenes with multiple objects. It tests the binding capabilities of models and includes a focused training split. BindOPSD achieves state-of-the-art results on FineBind and other spatial reasoning benchmarks.

2 Related Work

2.1 Vision-Language Models and Spatial Evidence

Image–text training gave VLMs broad visual perception Radford et al. (2021); Jia et al. (2021). Flamingo Alayrac et al. (2022), BLIP-2 Li et al. (2023), and LLaVA Liu et al. (2023) connected visual encoders to language models to generate answers from images. Qwen2-VL Wang et al. (2024),

Qwen2.5-VL Bai et al. (2025b), and Qwen3-VL Bai et al. (2025a) handle varied visual inputs. InternVL3 Zhu et al. (2025) and InternVL3.5 Wang et al. (2025) also strengthen visual perception and reasoning. These gains do not fully resolve spatial reasoning errors. An attention study Chen et al. (2025) finds that errors can coincide with attention to irrelevant objects. Beyond Semantics Qi et al. (2026) shows that semantic signals can obscure positional cues, while Seeing Beyond Redundancy Hannan et al. (2026) links visual token compression to task complexity. Penguin-VL Zhang et al. (2026a) preserves fine spatial cues, while Phi-4-reasoning-vision Aneja et al. (2026) and Fine-R1 He et al. (2026) strengthen perception and visual reasoning. Other methods add spatial supervision or geometric information. SpatialVLM Chen et al. (2024) uses 3D spatial data, and RoboSpatial Song et al. (2025) provides spatial training data for robotics. SpatialRGPT Cheng et al. (2024) adds depth and region cues. SpatialGeo Guo et al. (2025) combines geometric and visual features, and GR3D Cheng et al. (2026) adds 3D grounding. Self-Evolving Spatial Reasoning Liu et al. (2026a) encourages consistent answers across geometric transformations. BindOPSD focuses on how evidence for several objects is linked to their positions within one shared scene.

2.2 On-Policy Distillation for Visual Reasoning

Knowledge distillation transfers teacher predictions to a student Hinton et al. (2015). LLAVADI Xu et al. (2024), LLaVA-KD Cai et al. (2025), and LLaVA-MoD Shu et al. (2025) extend this idea to multimodal models. VisPer-LM Jain et al. (2025) distills visual embeddings, while CompoDistill Kim et al. (2026) aligns visual attention. Training on fixed teacher outputs can leave a gap between training and student inference. On-Policy Distillation (OPD) Agarwal et al. (2024) addresses this gap by supervising trajectories generated by the student. On-Policy Self-Distillation (OPSD) Zhao et al. (2026) uses different contexts to form a teacher and student within one model. Vision-OPD Yuan et al. (2026) gives the teacher a detailed crop. VA-OPD Liu et al. (2026b) selects tokens that depend on visual evidence, while DOPD Yoon et al. (2026) steers learning toward visual grounding. VCSD Liang et al. (2026) and S^2VOPD Li et al. (2026) use altered visual evidence to create training signals. VAD Zhang et al. (2026b) isolates visual corrections, and OPD-V Aniri et al. (2026) uses modality balance to select tokens. These methods strengthen visual supervision, but a relation among several objects also requires each local cue to be matched to its position in the same scene. BindOPSD distills this shared spatial configuration into a student that operates solely on the original image. Reasoning post-training can also weaken general capabilities Phan et al. (2026); gradient decomposition aims to limit conflicting updates.

2.3 Spatial Reasoning Benchmarks

Spatial benchmarks differ in the evidence and viewpoint they require. What'sUp Kamath et al. (2023) holds object identity fixed while changing relations in controlled photographs. COCO-Spatial and GQA-Spatial from the same study use natural images, which contain more objects and familiar layout biases. Spatial-MM Shiri et al. (2024) tests object relations and questions with several reasoning steps, including questions from camera and human perspectives.Recent benchmarks probe further cases. CLEVR-3D-DeRef Martin et al. (2026) asks models to resolve references using depth and relational context. SpatiaLQA Xie et al. (2026) tests spatial logic, while InfiniBench Wang et al. (2026) varies scene complexity. OmniSpatial Jia et al. (2026) covers a broad range of spatial tasks. SpinBench Zhang et al. (2026c) tests perspective and rotation, and CityCube Xu et al. (2026) studies changes in viewpoint in urban scenes. OST-Bench Lin et al. (2025) evaluates spatial understanding during online exploration. MMSI-Bench Yang et al. (2026) asks models to combine evidence across several images. These settings vary the evidence across views and time. FineBind focuses on several objects within one complex image. It is designed to evaluate global configuration and object-to-position binding.

3 Methods

3.1 Problem Setting and Framework

We study spatial reasoning for the configurational sufficiency. An instance from our target domain is represented as

$$(I^o, \mathcal{C}, q, y), \qquad \mathcal{C} = (I^c_1, \ldots, I^c_K), \tag{1}$$

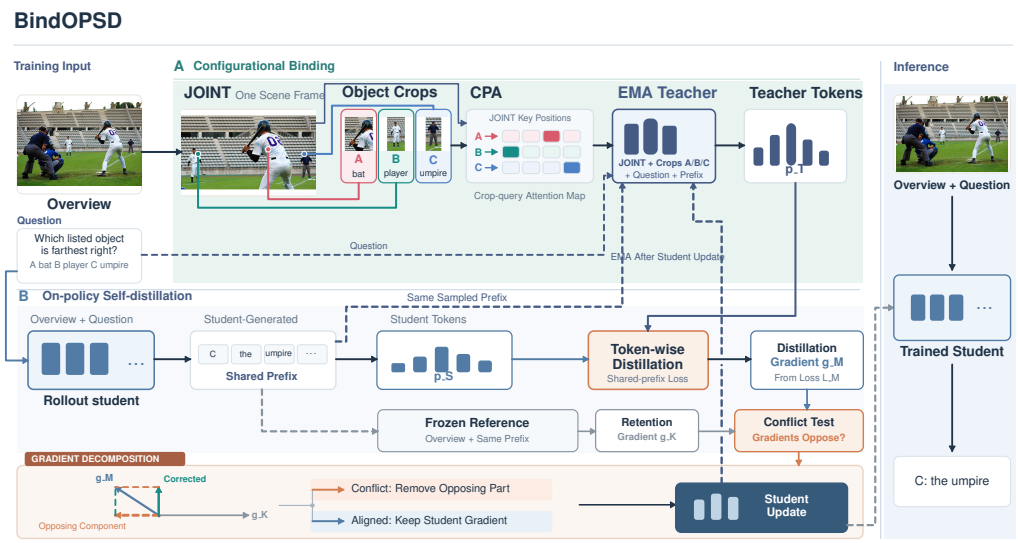


Figure 2: BindOPSD overview. JOINT builds a shared scene view, and CPA links crop evidence to locations in that view. The EMA teacher supervises the overview-only student on its sampled prefixes. A frozen reference guides gradient correction; inference uses the trained student.

where I^o is an overview image, $\mathcal{C}$ contains K object crops, q is a multi-object question, and y is its annotated answer. The student observes only $x_S = (I^o, q)$, whereas the Dense privileged teacher receives a direct sequence-level concatenation of the overview and all object crops, $x_T^D = [I^o; I_1^c; \cdots; I_K^c; q]$. Semicolons denote multimodal sequence order. Bounding boxes are used only to prepare the crops: neither box coordinates, y, nor the target relation are supplied to the model, and our objective contains no supervised cross-entropy on y. At inference, only (I^o, q) is available.

Our framework, BindOPSD, equips the privileged teacher with JOINT and correspondence-prior attention (CPA). JOINT restores a shared coordinate frame for the relevant objects; CPA binds crop-level identity evidence to the corresponding locations in that frame. The resulting teacher supervises an overview-only student on student-generated prefixes. A frozen overview-only reference additionally detects updates that conflict with the pretrained policy and corrects the distillation gradient. We maintain four model states: the trainable student θ, a rollout snapshot θ^-, an exponential-moving-average (EMA) teacher $\bar{\theta}$, and a frozen reference θ_{ref}.

Figure 2 summarizes the training and inference paths.

3.2 Teacher-Side JOINT and CPA

JOINT. A frozen pixel relocalizer matches each crop back to the overview, $(\widehat{B}_i, T_i, \rho_i) = \mathcal{M}(I_i^c, I^o)$, where T_i maps crop pixels to overview coordinates, $\widehat{B}_i$ is the matched region, and ρ_i is its reliability. JOINT and CPA are used only if every crop is localized reliably:

$$z = \prod_{i=1}^{K} \mathbf{1}\left[\rho_i \geq \eta_{\text{loc}} \ \wedge \ \text{Valid}(T_i, \widehat{B}_i)\right]. \tag{2}$$

The gate depends only on pixel correspondence, not on the answer or predicted relation. For $z = 1$, JOINT extracts a common scene window directly from the overview:

$$\begin{aligned} B^J &= \text{Clip}\left[\text{Expand}_m\left(\text{BBox}\left(\bigcup_i \widehat{B}_i\right)\right)\right], \\ I^J &= \mathcal{R}_J\big(\text{Crop}(I^o, B^J)\big). \end{aligned} \tag{3}$$

Here m is a fixed context margin and $\mathcal{R}_J$ is the actual resize-and-pad transform. If A_J maps overview coordinates through this transform, the crop-to-JOINT correspondence is $F_i^J = A_J \circ T_i$.

JOINT never pastes crop pixels: all targets remain in one scene-derived image and undergo the same transformation, preserving their relative layout while increasing their effective resolution.

When $z = 0$, JOINT is not constructed and the teacher falls back to the Dense input. Accordingly, the routed teacher input is

$$x_T(z) = \begin{cases} [I^J; I_1^c; \cdots; I_K^c; q], & z = 1, \\ x_T^D, & z = 0, \end{cases} \tag{4}$$

and CPA is disabled on this Dense fallback route.

Correspondence-prior attention. Let $\mathcal{U}_i$ and $\mathcal{P}_J$ be the content-token indices of crop i and JOINT, respectively. A frozen feature extractor gives a pooled crop descriptor $\bar{c}_i$ and JOINT token features $o^J(p)$. We combine their cosine similarity $s_i(p) = \cos(\bar{c}_i, o^J(p))$ with the recovered geometry. Specifically, let r_p be the normalized spatial center of JOINT token p, and let Ω_i^J be the normalized image of $\widehat{B}_i$ under A_J. CPA defines

$$G_i(p) = \frac{\exp\left[s_i(p)/\tau - d(r_p, \Omega_i^J)^2/(2\sigma_i^2)\right]}{\sum_{p' \in \mathcal{P}_J} \exp\left[s_i(p')/\tau - d(r_{p'}, \Omega_i^J)^2/(2\sigma_i^2)\right]}, \tag{5}$$

where d is the distance to the continuous matched region, τ is an appearance temperature, and σ_i is a spatial bandwidth.

At fixed layer–head pairs $\mathcal{A}_{\mathrm{CPA}}$, CPA changes only crop-query-to-JOINT-key logits:

$$\begin{aligned} \widetilde{A}_{ab} &= \mathrm{softmax}_b\left(\frac{Q_a K_b^\top}{\sqrt{d_h}} + M_{ab} + B_{ab}\right), \\ B_{ab} &= \mathbf{1}[a \in \mathcal{U}_i, b \in \mathcal{P}_J]\lambda_A \log(G_i(b) + \epsilon_A). \end{aligned} \tag{6}$$

The bias is zero outside $\mathcal{A}_{\mathrm{CPA}}$ and whenever $z = 0$; the original attention mask M is unchanged. Because JOINT precedes the crops in the serialized input, crop queries can retrieve scene evidence from their matched JOINT locations. Thus, JOINT supplies a relation-bearing configuration, while CPA associates crop-specific identity with positions inside it. Both modules are training-free and introduce no optimized parameters.

3.3 On-Policy Self-Distillation

The rollout snapshot samples an overview-only response

$$\hat{y} \sim \pi_{\theta^-}(\cdot \mid x_S). \tag{7}$$

At every sampled prefix $\hat{y}_{<t}$, student and teacher distributions are

$$p_{S,t} = \pi_\theta(\cdot \mid x_S, \hat{y}_{<t}), \qquad p_{T,t} = \mathrm{sg}\left[\pi_{\bar{\theta}}^B(\cdot \mid x_T(z), \hat{y}_{<t})\right], \tag{8}$$

where B denotes the CPA biases and sg stops gradients. Hence the teacher corrects states actually visited by the student rather than fixed expert prefixes.

For efficiency, we retain the student's stopped-gradient top-K_d support $\mathcal{S}_t$ and aggregate all remaining vocabulary mass into a shared tail symbol $\perp$: $\widetilde{p}_t(v) = p_t(v)$ for $v \in \mathcal{S}_t$ and $\widetilde{p}_t(\perp) = 1 - \sum_{v \in \mathcal{S}_t} p_t(v)$. With $m_t = \beta \widetilde{p}_{T,t} + (1-\beta)\widetilde{p}_{S,t}$, the token loss is

$$\mathcal{J}_t = \beta D_{\mathrm{KL}}(\widetilde{p}_{T,t} \| m_t) + (1-\beta) D_{\mathrm{KL}}(\widetilde{p}_{S,t} \| m_t). \tag{9}$$

For a logical batch $\mathcal{B}$, a fixed mask $w_{n,t}$ selects valid response tokens, $Z_n = \sum_t w_{n,t}$, and $\mathcal{B}_+ = \{n : Z_n > 0\}$. We optimize the trajectory-balanced objective

$$\mathcal{L}_M = \frac{1}{|\mathcal{B}_+|} \sum_{n \in \mathcal{B}_+} \frac{1}{Z_n} \sum_t w_{n,t} \mathcal{J}_{n,t}. \tag{10}$$

Sampling, support selection, and teacher predictions are held fixed within the update.

3.4 KL-GUIDED GRADIENT CORRECTION

Post-training can improve spatial reasoning while weakening performance on other vision-language tasks. To preserve the pretrained model's broader capabilities, we use a frozen reference to identify conflicting updates. The reference is evaluated on exactly the student's context: $p_{\mathrm{ref},n,t} = \mathrm{sg}[\pi_{\theta_{\mathrm{ref}}}(\cdot \mid x_{S,n}, \hat{y}_{n,<t})]$. The conceptual retention objective is the full-vocabulary

$$\mathcal{L}_K^{\mathrm{full}} = \frac{\lambda_{\mathrm{KL}}}{Z_K} \sum_{n,t} a_{n,t} D_{\mathrm{KL}}(p_{S,n,t} \| p_{\mathrm{ref},n,t}), \tag{11}$$

where $a_{n,t}$ selects retention positions. We obtain its practical gradient from the low-variance on-policy surrogate detailed in the appendix and denote the differentiated objective by $\mathcal{L}_K$.

Rather than adding $\mathcal{L}_K$ to the training loss, we use it only when the two gradients conflict. Let $g_M = \nabla_\theta \mathcal{L}_M$, $g_K = \nabla_\theta \mathcal{L}_K$, $\gamma = \langle g_M, g_K \rangle$, and $\kappa = \|g_K\|_2^2$. We apply the asymmetric correction

$$g_{\mathrm{final}} = g_M - \mathbf{1}[\gamma < 0] \frac{\gamma}{\kappa + \varepsilon_p} g_K. \tag{12}$$

When $\gamma < 0$, this removes the component of the distillation gradient that would increase the retention objective to first order; otherwise it leaves g_M unchanged. The correction is computed on globally accumulated, unscaled gradients before clipping, with distributed dot products reduced across parameter shards. It is a local gradient-conflict correction, not a hard constraint on the subsequent AdamW step.

After the student update, the teacher is refreshed by $\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta$. At inference, we discard JOINT, CPA, the EMA teacher, and the reference policy. The deployed model is therefore the original overview-only student and requires no crops, boxes, relocalization, or attention intervention.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We train on FineBind, which contains 2,500 training questions. Its 500-question test set was fixed before the student pilots. Each question concerns multiple objects in a single image and tests pairwise direction, extrema ordering, or binding between similar objects. We also evaluate on OST-Bench, COMFORT, MultiHopSpatial, Spatial-MM, SpatialMQA, SpatialQA, COCO-Spatial, GQA-Spatial, and SPINBench.

Baselines. We compare BindOPSD with its starting models, Qwen3.5-4B-Base and Qwen3.5-9B-Base, to assess the effect of training. The other evaluated models and their sizes are listed in Tables 1 and 2.

Implementation details. We train the 9B model on eight NVIDIA A100-SXM4-40GB GPUs using BF16 and a global batch size of 96. The learning rate is 2×10^{-6}, with 10 warmup steps followed by a constant schedule; weight decay is 0.01. The teacher uses JOINT with CPA when the correspondence gate passes and the Dense crop concatenation otherwise. CPA uses a bias strength of 0.05 at language layers 3 and 7. Distillation uses the top 100 tokens and a shared tail, together with an EMA teacher and KL-guided gradient correction.

Evaluation. At inference, BindOPSD receives only the original image and question. We report accuracy under each benchmark's answer rule.

4.2 MAIN RESULTS

Comparison with other models. Tables 1 and 2 show that BindOPSD achieves the best score on FineBind and eight of the nine spatial reasoning benchmarks. It ties Qwen3.5-9B-Base on SpatialMQA. Compared with the strongest competing result in each column, BindOPSD leads by 12.80 points on FineBind and 4.82 points on SpatialQA. Its margins on the other seven benchmarks it wins range from 0.48 to 1.60 points.

Comparison at the 9B scale. Compared with Qwen3.5-9B-Base, BindOPSD gains 19.20 points on FineBind, 19.02 points on SpatialQA, and 5.14 points on MultiHopSpatial. Its gains on the other six

Table 1: Main results on FineBind and nine spatial reasoning benchmarks.

Method	FineBind	OST-Bench	COMFORT	MultiHop Spatial	Spatial-MM
Qwen3.5 Backbones					
Qwen3.5-4B-Base	31.00	42.25	66.37	40.98	69.54
Qwen3.5-9B-Base	47.00	43.19	72.29	38.73	69.54
Other Open-source Backbones					
InternVL-2.5-8B	37.20	39.11	74.33	36.82	65.18
LLaVA-OV-7B	50.80	38.03	68.37	36.56	60.76
Spatial Reasoning Methods					
SpatialLadder-3B	45.40	36.88	73.78	29.89	59.58
GeoAnchor-2B	53.40	28.15	46.11	24.82	42.97
GeoThinker-7B	51.00	25.96	69.91	30.69	53.30
HATCH-3B	49.00	30.14	73.08	27.00	36.48
MindCube-RL-3B	43.80	35.82	74.22	33.89	35.68
ViewFusion-4B	50.00	42.12	62.69	38.42	62.03
VLM3R-7B	49.60	39.57	65.05	38.80	55.86
Ours					
BindOPSD-4B	**62.40**	**43.72**	**74.50**	**41.96**	**69.06**
BindOPSD-9B	**66.20**	**44.36**	**74.81**	**43.87**	**70.93**

Note. Model sizes denote the nominal evaluated versions and may exclude additional visual or geometric modules. GeoThinker uses the 7B Scaled+VGGT variant; VLM3R uses a 7B backbone with LoRA and CUT3R.

Table 2: Main results on five additional spatial reasoning benchmarks.

Method	SpatialMQA	SpatialQA	COCO-Spatial	GQA-Spatial	SPINBench
Qwen3.5 Backbones					
Qwen3.5-4B-Base	44.70	44.70	91.51	94.69	62.21
Qwen3.5-9B-Base	48.05	48.05	93.34	95.31	59.91
Other Open-source Backbones					
InternVL-2.5-8B	31.78	35.34	93.15	94.49	57.61
LLaVA-OV-7B	32.62	43.65	92.33	92.56	56.12
Spatial Reasoning Methods					
SpatialLadder-3B	29.74	45.67	84.96	85.04	50.42
GeoAnchor-2B	30.67	56.56	74.69	73.40	48.08
GeoThinker-7B	37.36	20.60	93.67	94.42	54.18
HATCH-3B	28.35	51.77	92.22	91.45	48.52
MindCube-RL-3B	20.07	46.91	88.72	89.94	49.18
ViewFusion-4B	35.78	62.25	92.97	94.69	58.38
VLM3R-7B	34.48	47.71	91.89	94.21	55.09
Ours					
BindOPSD-4b	**47.39**	**60.66**	**94.19**	**95.31**	**62.40**
BindOPSD-9B	**48.05**	**67.07**	**94.64**	**96.07**	**63.60**

spatial benchmarks range from 0.64 to 3.69 points, while its SpatialMQA score is unchanged. Both models have a nominal size of 9B.

4.3 Ablation Study

We examine the roles of JOINT and CPA at the 4B and 9B scales. Table 4 first evaluates the starting models with the original overview, JOINT, and JOINT+CPA. It then compares overview-only students distilled from JOINT and JOINT+CPA teachers. The students receive only the original image and question at inference.

Teacher-side visual configuration. JOINT without CPA produces mixed changes relative to the overview-only starting model. FineBind improves by 4.20 points at 4B and 0.40 points at 9B, while

Table 3: General mathematical and reasoning capabilities after spatial-oriented training. Results are reported on GSM8K, MATH-500, AIME2024, and Bamboogle. Higher scores are better. All scores are percentages. GSM8K and MATH-500 use greedy pass@1; AIME2024 uses avg@4 over four samples with temperature 0.6; Bamboogle uses greedy pass@1. Base models are included as general-capability references.

Method	GSM8K	MATH-500	AIME2024	Bamboogle
Qwen3.5-9B-Base	**96.51**	**91.40**	**41.67**	**47.20**
ViewFusion-4B	91.96	86.20	30.00	15.20
Qwen3.5-4B-Base	95.22	74.40	5.83	34.40
HATCH-3B	82.41	63.20	4.17	4.80
InternVL-2.5-8B	82.26	54.40	2.50	18.40
SpatialLadder-3B	77.63	52.60	0.83	6.40
LLaVA-OV-7B	75.51	45.80	0.83	11.20
VLM3R-7B	69.90	44.80	0.83	10.40
MindCube-RL-3B	73.01	44.60	0.00	6.40
GeoThinker-7B	39.42	25.00	0.00	8.80
GeoAnchor-2B	3.49	8.20	0.00	1.60
BindOPSD-4B	93.40	88.20	43.30	19.84
BindOPSD-9B	93.86	85.00	33.33	35.19

Table 4: Ablation of JOINT and CPA at the 4B and 9B scales. All results are accuracy (%). The upper block directly evaluates the starting models under different visual configurations. In the lower block, parentheses specify the teacher configuration used for distillation; every student is evaluated with the original overview only.

Method	FineBind	SpatialMQA	Spatial Lab	Spatial-MM
Direct evaluation of the starting model				
Qwen3.5-4B-Base	45.80	42.75	42.50	63.26
+ JOINT	50.00	42.94	45.00	63.37
+ JOINT + CPA	90.00	44.70	45.43	65.76
Qwen3.5-9B-Base	53.20	41.82	45.43	63.15
+ JOINT	53.60	43.59	44.29	63.26
+ JOINT + CPA	63.40	48.05	44.64	64.32
Overview-only students after distillation				
BindOPSD-4B (JOINT)	60.00	42.19	45.43	65.76
BindOPSD-4B (JOINT + CPA)	62.40	47.39	47.93	69.06
BindOPSD-9B (JOINT)	55.80	44.80	49.64	68.37
BindOPSD-9B (JOINT + CPA)	66.20	48.05	49.93	70.93

Spatial Lab decreases by 1.14 points at 9B. Holding the JOINT input fixed, adding CPA improves all four benchmarks at both scales. On FineBind, accuracy rises from 50.00 to 90.00 at 4B and from 53.60 to 63.40 at 9B. The corresponding SpatialMQA gains are 1.76 and 4.46 points. These results indicate that the shared scene view benefits from explicit correspondence between the object crops and their locations in that view. Because the Base-to-JOINT comparison also changes the visual input, the JOINT-to-JOINT+CPA comparison more directly isolates CPA.

Transfer to an overview-only student. Replacing a JOINT teacher with a JOINT+CPA teacher improves the student on every reported benchmark at both scales. At 4B, the gains are 2.40 points on FineBind, 5.20 on SpatialMQA, 2.50 on Spatial Lab, and 3.30 on Spatial-MM. At 9B, the corresponding gains are 10.40, 3.25, 0.29, and 2.56 points. The gains persist when the student no longer receives JOINT or object crops at inference, although their magnitude varies across benchmarks and model scales.

Inference cost. We measure the wall-clock cost of directly evaluating the starting models with Overview, JOINT, and JOINT+CPA. Table 5 projects the time for 1,000 serial questions.

The JOINT route increases projected time over Overview by 36.72 seconds at 4B and 51.10 seconds at 9B for 1,000 serial questions. Adding CPA to JOINT adds a further 0.94 seconds (0.4%) and 6.04

Table 5: Projected wall-clock time for 1,000 serial questions on one A100 GPU. Measurements exclude model loading and include image reading, processor preprocessing, and generation.

Starting model	Input	Time for 1,000 (s)	vs. Overview
Qwen3.5-4B-Base	Overview	195.16	–
	JOINT	231.88	+18.8%
	JOINT + CPA	232.82	+19.3%
Qwen3.5-9B-Base	Overview	205.76	–
	JOINT	256.86	+24.8%
	JOINT + CPA	262.90	+27.8%

seconds (2.4%), respectively. Thus, most of the observed runtime increase occurs on the JOINT route, while CPA has a smaller incremental cost in this implementation. These timings describe direct evaluation with the starting weights, rather than inference with the deployed overview-only student. We compare wall-clock times because the separately recorded CUDA timings do not cover the same preprocessing operations across routes.

4.4 Further Analysis

How does the teacher advantage transfer? The direct teacher gain does not predict the size of the student gain across model scales. On FineBind, CPA raises the 4B teacher from 50.00% to 90.00% with JOINT held fixed, while the corresponding overview-only student improves from 60.00% to 62.40%. At 9B, the teacher gain is smaller (53.60% to 63.40%), but the student gain is larger (55.80% to 66.20%). Thus, stronger teacher predictions alone do not guarantee a proportionate improvement after distillation. The ablation supports the value of CPA for both teachers and students, while leaving the factors that govern transfer efficiency open.

How far do the gains extend beyond FineBind? Training uses 2,500 FineBind questions. Relative to its 9B starting model, BindOPSD-9B improves on eight of the nine external spatial benchmarks and ties on SpatialMQA. The gains range from 19.02 points on SpatialQA and 5.14 on MultiHopSpatial to 1.30 on COCO-Spatial and 0.76 on GQA-Spatial. The latter two starting scores already exceed 93%, leaving less room for an absolute gain. Transfer also varies with model scale: BindOPSD-4B improves on eight external benchmarks but falls by 0.48 points on Spatial-MM. These results show transfer across different evaluations, without implying a uniform gain on every task.

What happens to general capabilities? The general capability results are mixed. Compared with Qwen3.5-9B-Base, BindOPSD-9B drops by 2.65 points on GSM8K, 6.40 on MATH-500, 8.34 on AIME2024, and 12.01 on Bamboogle. At 4B, MATH-500 and AIME2024 improve by 13.80 and 37.47 points, respectively, while GSM8K and Bamboogle decrease by 1.82 and 14.56 points. KL-guided gradient correction is designed to suppress conflicting updates, but these final scores cannot isolate its effect without an otherwise identical run that omits the correction. Spatial gains therefore should not be taken as evidence of uniform general capability retention.

5 Conclusion

We introduced BindOPSD to improve spatial reasoning by keeping object evidence and relative layout connected. JOINT builds a shared view of the referred objects, and CPA links individual crops to locations in that view. On-policy self-distillation transfers this privileged teacher signal to a student that receives only the original image and question at inference. We also introduced FineBind to test spatial binding in scenes with multiple objects. BindOPSD-9B gains 19.20 points on FineBind over its starting model, improves on eight of nine external spatial benchmarks, and ties on the remaining one. Ablations show that CPA adds value to JOINT and that this advantage can reach overview-only students. General capability changes remain mixed, and our evaluation focuses on single-image questions with object regions available during training. These limits motivate more direct tests of binding errors and capability retention.

AI USE STATEMENT

LLM-based tools assisted with writing. The authors are responsible for verifying AI-assisted material and for the final content of this work.

ETHICS STATEMENT

This work studies spatial reasoning in images and introduces FineBind for research evaluation. FineBind uses questions about multiple objects in a single image. Image collections may also cover some scenes and objects more than others, so benchmark accuracy does not establish equal performance across settings. Better spatial reasoning may support visual assistance, but errors that link an object to the wrong location can cause harm in navigation or other consequential uses.

REPRODUCIBILITY STATEMENT

Given resource constraints, we plan to release the training data, code, and model weights upon acceptance.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2204.14198.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15B technical report. *arXiv preprint arXiv:2603.03975*, 2026. URL https://arxiv.org/abs/2603.03975.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. OPD-V: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026. URL https://arxiv.org/abs/2608.05131.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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*, 2025b. URL https://arxiv.org/abs/2502.13923.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. LLaVA-KD: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 239–249, 2025. doi: 10.1109/ICCV51701.2025.00030. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cai_LLaVA-KD_A_Framework_of_Distilling_Multimodal_Large_Language_Models_ICCV_2025_paper.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024. doi: 10.1109/CVPR52733.2024.01370. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9910–9932. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25cr.html.

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*, volume 37, 2024. URL https://arxiv.org/abs/2406.01584.

An-Chieh Cheng, Yang Fu, Yatai Ji, Ligeng Zhu, Guanqi Zhan, Zhuoyang Zhang, Zhaojing Yang, Song Han, Yao Lu, Pavlo Molchanov, Vidya Nariyambut Murali, Jan Kautz, Xiaolong Wang, Hongxu Yin, and Sifei Liu. Grounded 3d-aware spatial vision-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16688–16700, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cheng_Grounded_3D-Aware_Spatial_Vision-Language_Modeling_CVPR_2026_paper.html.

Tzu-Yao Chiu and Julie D. Golomb. The influence of saccade target status on the reference frame of object-location binding. *Journal of Experimental Psychology: General*, 154(5):1183–1200, 2025. doi: 10.1037/xge0001718.

Jiajie Guo, Qingpeng Zhu, Jin Zeng, Xiaolong Wu, Changyong He, and Weida Wang. SpatialGeo: Boosting spatial reasoning in multimodal LLMs via geometry-semantics fusion. *arXiv preprint arXiv:2511.17308*, 2025. URL https://arxiv.org/abs/2511.17308.

Darryl Hannan, John Cooper, Dylan White, and Yijing Watkins. Seeing beyond redundancy: Task complexity's role in vision token specialization in VLLMs. *arXiv preprint arXiv:2602.06914*, 2026. URL https://arxiv.org/abs/2602.06914.

Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-R1: Make multi-modal LLMs excel in fine-grained visual recognition by chain-of-thought reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.07605.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. STEP3-VL-10B technical report. *arXiv preprint arXiv:2601.09668*, 2026. URL https://arxiv.org/abs/2601.09668.

Jitesh Jain, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. Elevating visual perception in multimodal LLMs with visual embedding distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2743. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/75ed738c083f8941cfc990245e4f6265-Abstract-Conference.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

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.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. CompoDistill: Attention distillation for compositional reasoning in multimodal LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/74ee8b8a0b958199b440194fcf938bec-Abstract-Conference.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026. URL https://arxiv.org/abs/2608.14144.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026. URL https://arxiv.org/abs/2607.21556.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. OST-Bench: Evaluating the capabilities of MLLMs in online spatio-temporal scene understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0385. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/10d3c8a6046d56cf27470fa7b74ab432-Abstract-Datasets_and_Benchmarks_Track.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.08485.

Junming Liu, Yuqi Li, Yifei Sun, Maonan Wang, Piotr Koniusz, Yirong Chen, and Ding Wang. Self-evolving spatial reasoning in vision language models via geometric logic consistency. *arXiv preprint arXiv:2605.18162*, 2026a. URL https://arxiv.org/abs/2605.18162.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026b. URL https://arxiv.org/abs/2605.21924.

Zitong Lu and Julie D. Golomb. Dynamic saccade context triggers more stable object-location binding. *Journal of Experimental Psychology: General*, 153(4):873–888, 2024. doi: 10.1037/xge0001545.

Mary Lynn Martin, Martha Palmer, and Maria Leonor Pacheco. CLEVR-3D-DeRef. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 9490–9503. ELRA Language Resource Association, 2026. doi: 10.63317/4hw2eqvxuhuf. URL https://aclanthology.org/2026.lrec-1.745/.

Hoang Phan, Xianjun Yang, Yuanshun Yao, Jingyu Zhang, Shengjie Bi, Xiaocheng Tang, Madian Khabsa, Lijuan Liu, and Deren Lei. Beyond reasoning gains: Mitigating general-capability forgetting in large reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34368–34398. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1717. URL https://aclanthology.org/2026.findings-acl.1717/.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. In *The Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=AOJ9kbpM0N.

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

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1195. URL https://aclanthology.org/2024.emnlp-main.1195/.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, Siming Fu, Haoyuan Li, Zhelun Yu, Si Liu, Hongsheng Li, and Hao Jiang. LLaVA-MoD: Making LLaVA tiny via MoE-knowledge distillation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1a2131ebe25bd55e4fc734126ea583ed-Abstract-Conference.html.

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. doi: 10.1109/CVPR52734.2025.01470. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_RoboSpatial_Teaching_Spatial_Understanding_to_2D_and_3D_Vision-Language_Models_CVPR_2025_paper.html.

Haoming Wang, Qiyao Xue, and Wei Gao. InfiniBench: Infinite benchmarking for visual spatial reasoning with customizable scene complexity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21594–21605, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_InfiniBench_Infinite_Benchmarking_for_Visual_Spatial_Reasoning_with_Customizable_Scene_CVPR_2026_paper.html.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Zhu Hao, 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*, pp. 2645–2657, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xie_SpatiaLQA_A_Benchmark_for_Evaluating_Spatial_Logical_Reasoning_in_Vision-Language_CVPR_2026_paper.html.

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. CityCube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9190–9215. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.416. URL https://aclanthology.org/2026.acl-long.416/.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. LLAVADI: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024. URL https://arxiv.org/abs/2407.19409.

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://proceedings.iclr.cc/paper_files/paper/2026/hash/fe5d4bd3e2af823701b5d8f70a3c7602-Abstract-Conference.html.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.00564.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026. URL https://arxiv.org/abs/2605.18740.

Boqiang Zhang, Lei Ke, Ruihan Yang, Qi Gao, Tianyuan Qu, Rossell Chen, Dong Yu, and Leoweiliang. Penguin-VL: Exploring the efficiency limits of VLM with LLM-based vision encoders. *arXiv preprint arXiv:2603.06569*, 2026a. URL https://arxiv.org/abs/2603.06569.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026b. URL https://arxiv.org/abs/2607.28590.

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*, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.18734.

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

Table 6: Roles of the four model states in the 9B training algorithm.

State	Origin	Input and update
Student θ	9B base checkpoint	Overview; optimizer
Snapshot θ^-	Student snapshot	Overview; refreshed from student
Teacher $\bar{\theta}$	Student EMA state	Routed views; EMA update
Reference θ_{ref}	9B base checkpoint	Overview; frozen

A On-Policy Distillation: Algorithm and Objective

A.1 Model States and One Training Step

The trainable student has parameters θ. A rollout snapshot θ^- samples overview-only responses. The EMA state $\bar{\theta}$ forms the privileged teacher; the frozen reference θ_{ref} sees the same overview-only input as the student. The 9B run uses `Qwen3.5-9B-Base` as its base checkpoint. The teacher distributions are detached during the student update. At inference, only the overview-only student is used.

1. Construct $x_S = (I^o, q)$ and the routed teacher context $x_T(z)$ for a logical training batch.
2. Draw a response $\hat{y}$ from $\pi_{\theta^-}(\cdot \mid x_S)$ and hold its prefixes fixed for the update.
3. Evaluate the student on x_S, the detached teacher on $x_T(z)$, and the detached reference on x_S at the selected prefixes.
4. Coarsen teacher and student distributions to the selected 100 tokens and a shared tail, then differentiate $\mathcal{L}_M$.
5. Form the retention gradient g_K. Apply gradient correction to accumulated, unscaled gradients before clipping and the optimizer step.
6. Update the EMA state from the new student state, and refresh the rollout snapshot under the training schedule.

The configuration sets `teacher_update_rate=0.05`. This is the recorded update-rate field; its mapping to the EMA retention coefficient μ in Property 9 depends on the update convention. The snapshot is a copy of the student state used to generate responses, rather than an independently trained teacher.

A.2 Top-K_d Support and Shared Tail

For a fixed prefix, set $K_d = 100$, select S_t using the detached student distribution and define a deterministic coarsening C_t that keeps every token in S_t and maps every other vocabulary item to the shared symbol $\perp$. Let P and Q be the full teacher and student distributions; their coarsened versions are $\widetilde{P} = C_{t\#}P$ and $\widetilde{Q} = C_{t\#}Q$. The main-paper token objective is the weighted Jensen–Shannon divergence $\mathrm{JS}_\beta(\widetilde{P}, \widetilde{Q})$.

Property 6: coarsening cannot increase divergence. For $0 < \beta < 1$,

$$\begin{aligned} 0 \le \mathrm{JS}_\beta(\widetilde{P}, \widetilde{Q}) &\le \mathrm{JS}_\beta(P, Q) \\ &\le -\beta \log \beta - (1-\beta)\log(1-\beta). \end{aligned} \tag{13}$$

Proof. Let $M = \beta P + (1-\beta)Q$. By the log-sum inequality,

$$D_{\mathrm{KL}}(C_{t\#}P \| C_{t\#}M) \le D_{\mathrm{KL}}(P \| M). \tag{14}$$

The same inequality holds for Q. Take their weighted sum. Weighted JS also equals the mutual information between a binary source label with prior $(\beta, 1-\beta)$ and a draw from the selected distribution. Mutual information does not exceed the entropy of that label. □

To identify the omitted information, write $T = \mathcal{V} \setminus S_t$ and denote the conditional distribution inside T by P_T, Q_T, or M_T. Applying the KL chain rule to the partition into singleton selected tokens and

one tail gives

$$\begin{aligned}&\mathrm{JS}_\beta(P,Q) - \mathrm{JS}_\beta(\widetilde{P},\widetilde{Q})\\ &\quad= \beta P(T) D_{\mathrm{KL}}(P_T \| M_T) \qquad (15)\\ &\qquad+ (1-\beta) Q(T) D_{\mathrm{KL}}(Q_T \| M_T). \qquad (16)\end{aligned}$$

Thus the approximation loses only information about how teacher and student distribute mass *within* the aggregated tail. It is not full-vocabulary JS. The run uses the selected top 100 tokens and one shared tail; the bound does not imply that this approximates full-vocabulary JS equally well on every prefix.

Property 7: student gradient of the token loss. Fix the prefix, the detached teacher $\widetilde{P}$, and the selected support S_t. Put $\widetilde{Q} = \widetilde{Q}_\theta$ and $\widetilde{M} = \beta\widetilde{P} + (1-\beta)\widetilde{Q}$. Differentiating the main-paper token loss gives

$$\nabla_\theta \,\mathrm{JS}_\beta(\widetilde{P},\widetilde{Q}_\theta) = (1-\beta) \sum_{u \in S_t \cup \{\perp\}} \nabla_\theta \widetilde{Q}_\theta(u) \log \frac{\widetilde{Q}_\theta(u)}{\widetilde{M}(u)}. \tag{17}$$

Proof. Differentiate $\beta \sum_u \widetilde{P}(u) \log(\widetilde{P}(u)/\widetilde{M}(u))$ and $(1-\beta)\sum_u \widetilde{Q}(u)\log(\widetilde{Q}(u)/\widetilde{M}(u))$. All terms containing $\nabla\widetilde{M}$ sum to $-\sum_u \nabla\widetilde{M}(u) = 0$. The additional derivative of $\widetilde{Q}\log\widetilde{Q}$ contributes $(1-\beta)\sum_u \nabla\widetilde{Q}(u) = 0$. The remaining term is Eq. equation 17. □

The tail probability remains differentiable: $\widetilde{Q}_\theta(\perp) = 1 - \sum_{v \in S_t} Q_\theta(v)$. The result requires the top-K_d indices to be held fixed during the update. It does not say that the EMA teacher stays fixed across training steps.

A.3 Trajectory Weighting and Effective Policy

The mask $w_{n,t}$ defines valid response positions, $Z_n = \sum_t w_{n,t}$, and $\mathcal{B}^+ = \{n : Z_n > 0\}$. Dividing by Z_n before averaging across records gives equal total weight to each nonempty response in a logical batch. This does not make tokens independent, and it does not turn a snapshot-generated trajectory into one sampled from the updated parameters. The actual rollout policy is θ^-; the lag relative to θ is controlled by its refresh schedule. The recorded rollout length limit is 64 response tokens. The training configuration selects token-level rollout importance sampling with threshold 2.0; this setting alone does not specify the snapshot refresh interval.

Property 8: response-balanced weights. For $n \in \mathcal{B}^+$, the total coefficient assigned to the response in Eq. (10) is

$$\sum_t \frac{w_{n,t}}{|\mathcal{B}^+| Z_n} = \frac{1}{|\mathcal{B}^+|}. \tag{18}$$

Proof. Substitute $Z_n = \sum_t w_{n,t}$ and cancel $Z_n > 0$. □ Hence longer valid responses do not receive more *total* weight solely because they contain more selected tokens. This property does not imply equal influence of examples on the parameter gradient, since token gradients can differ in magnitude.

Property 9: EMA lag under bounded student updates. Index the student after optimizer step s by θ_s, and update the teacher as $\bar{\theta}_s = \mu\bar{\theta}_{s-1} + (1-\mu)\theta_s$ for $0 \le \mu < 1$. Let $d_s = \theta_s - \bar{\theta}_s$. Then

$$\|d_s\| \le \mu^s \|d_0\| + \mu \sum_{j=1}^{s} \mu^{s-j} \|\theta_j - \theta_{j-1}\|. \tag{19}$$

Proof. Substitute the EMA update to obtain $d_s = \mu d_{s-1} + \mu(\theta_s - \theta_{s-1})$. Iterate this recursion and apply the triangle inequality. □ If every step has norm at most D and $d_0 = 0$, this reduces to $\|d_s\| \le \mu D(1-\mu^s)/(1-\mu)$. This is a parameter-distance bound conditional on update sizes, not an accuracy guarantee.

Property 10: a conditional bound on rollout lag. Fix an overview input and a common decoding scheme. Suppose the next-token kernels of θ^- and θ have total variation distance at most δ_j at generation position j, uniformly over every prefix at that position. Let $P^-_{<t}$ and $P^\theta_{<t}$ be their distributions on prefixes of length $t-1$, treating EOS as an absorbing token. Then

$$\mathrm{TV}(P^-_{<t}, P^\theta_{<t}) \le \min\left\{1, \sum_{j=1}^{t-1} \delta_j\right\}. \tag{20}$$

Proof. Couple the two token draws so they agree with maximal probability while their prefixes agree. Conditional on agreement so far, the next draw disagrees with probability at most δ_j. A union bound shows that their prefixes disagree by step $t-1$ with probability at most $\sum_{j<t} \delta_j$. Total variation is no larger than the disagreement probability of any coupling. □

The result explains why refreshing θ^- can matter; it does not assert that the required uniform δ_j values are small in the reported training run.

B Reference Retention and Gradient Correction

B.1 Gradient of the Full-Vocabulary Reference KL

Fix one sampled prefix. Let $p_\theta(v)$ be the student next-token distribution and $r(v)$ the detached reference distribution on the same overview-only input. If both distributions have positive support, then

$$D_{\mathrm{KL}}(p_\theta \| r) = \sum_v p_\theta(v) \log \frac{p_\theta(v)}{r(v)}, \tag{21}$$

$$\nabla_\theta D_{\mathrm{KL}}(p_\theta \| r) = \sum_v p_\theta(v) \log \frac{p_\theta(v)}{r(v)} \nabla_\theta \log p_\theta(v). \tag{22}$$

Proof. Direct differentiation yields $\sum_v (\log(p_\theta(v)/r(v)) + 1)\nabla_\theta p_\theta(v)$. The $+1$ contribution vanishes because $\sum_v \nabla_\theta p_\theta(v) = \nabla_\theta 1 = 0$. Substitute $\nabla p = p \nabla \log p$ in the remaining term. □

The identity concerns the full-vocabulary conditional objective on *fixed* prefixes. The sampled-token properties below require their own stated support and weighting assumptions; they do not follow merely from enabling a KL retention loss.

Property 11: an unbiased conditional sampled-token gradient. The following result characterizes an idealized, full-support sampled estimator. Fix the prefix and let $q(v)$ be the *actual* token sampling distribution, including any temperature or truncation. Assume $q(v) > 0$ whenever $p_\theta(v) > 0$, and define $R(v) = p_\theta(v)/q(v)$. If b does not depend on the sampled token, then

$$\begin{aligned} &\mathbb{E}_{v\sim q}\left[R(v)\left(\log \frac{p_\theta(v)}{r(v)} - b\right)\nabla_\theta \log p_\theta(v)\right] \\ &\quad = \nabla_\theta D_{\mathrm{KL}}(p_\theta \| r). \end{aligned} \tag{23}$$

Proof. Substitute $q(v)R(v) = p_\theta(v)$ into the expectation. The part without b equals Eq. equation 22. The baseline term is $-b\sum_v p_\theta(v)\nabla_\theta \log p_\theta(v) = -b\nabla_\theta \sum_v p_\theta(v) = 0$. □

For a fixed batch of prefixes, multiplying each sampled term by the main-paper position mask and normalizer gives the corresponding gradient estimator for the conditional batch objective. A differentiated surrogate can multiply $\log p_\theta(v)$ by a *detached* coefficient $R(v)[\log(p_\theta(v)/r(v)) - b]$. This stop-gradient construction implements Eq. equation 23; differentiating that coefficient as well would compute another gradient. If rollout sampling truncates the vocabulary, the support assumption fails for excluded tokens. Clipped ratios or self-normalized weights also change the expectation and generally introduce bias. The configured token-level importance-sampling threshold of 2.0 does not meet the unclipped assumption automatically, so this identity is not an assertion that the recorded training gradient is unbiased.

Property 12: the baseline minimizing conditional second moment. Set $f(v) = \log(p_\theta(v)/r(v))$ and $s(v) = \nabla_\theta \log p_\theta(v)$. The token-independent scalar baseline that minimizes the conditional second moment of Eq. equation 23 is

$$b^* = \frac{\mathbb{E}_q[R(v)^2 f(v)\|s(v)\|^2]}{\mathbb{E}_q[R(v)^2\|s(v)\|^2]}, \tag{24}$$

when the denominator is positive and both moments exist. *Proof.* The second moment equals

$$\mathbb{E}_q\left[R(v)^2(f(v)-b)^2\|s(v)\|^2\right]. \tag{25}$$

Differentiate this quadratic in b and set the derivative to zero. The estimator mean is independent of b by Property 11. The variance has the same minimizer. □

This is a theoretical reference, not evidence that the actual implementation uses b^* or has low variance. For stale snapshots, importance ratios can be large even when the estimator is unbiased.

B.2 Exact Projection and the Stabilized Rule in the Main Paper

Let $g_M = \nabla_\theta \mathcal{L}_M$, $g_K = \nabla_\theta \mathcal{L}_K$, $\gamma = \langle g_M, g_K\rangle$, and $\kappa = \|g_K\|_2^2$. For $\gamma < 0$ and $\kappa > 0$, the exact Euclidean projection of g_M onto the half-space $\{g : \langle g, g_K\rangle \geq 0\}$ is

$$g_{\text{proj}} = g_M - \frac{\gamma}{\kappa} g_K. \tag{26}$$

Indeed, $\langle g_{\text{proj}}, g_K\rangle = 0$. The correction is parallel to the normal of the half-space and has minimum Euclidean norm among feasible corrections. By Cauchy–Schwarz, $\langle g_{\text{proj}}, g_M\rangle = \|g_M\|^2 - \gamma^2/\kappa \geq 0$. Consequently, a sufficiently small plain-gradient step along $-g_{\text{proj}}$ does not increase either differentiable objective to first order. This statement is local and says nothing about finite steps or AdamW preconditioning.

Projection proof. If $\gamma \geq 0$, g_M already lies in the feasible half-space and is its own projection. If $\gamma < 0$, write any correction as $h - g_M = tg_K + u$, where $u \perp g_K$. Feasibility requires $\gamma + t\kappa \geq 0$, hence $t \geq -\gamma/\kappa$. Its squared length is $t^2\kappa + \|u\|^2$, which is minimized at $t = -\gamma/\kappa$ and $u = 0$. This yields Eq. equation 26. If $g_K = 0$, the constraint is vacuous and no correction is needed. □

The manuscript uses $\kappa + \epsilon_p$ in place of κ. In that *stabilized* case, when $\gamma < 0$,

$$\left\langle g_M - \frac{\gamma}{\kappa+\epsilon_p} g_K, g_K \right\rangle = \frac{\gamma\epsilon_p}{\kappa+\epsilon_p} < 0 \quad (\epsilon_p > 0). \tag{27}$$

Thus stabilization reduces the conflict but does not cancel it exactly. The recorded configuration sets $\epsilon_p = 10^{-8}$. The dot product and both norms in the correction are defined over the same trainable parameters after accumulation and before the optimizer step.

Property 13: retention change for a finite plain-gradient step. Assume the differentiated retention objective $\mathcal{L}_K$ has an L-Lipschitz gradient in a neighborhood of θ. For a plain step $\theta^+ = \theta - \alpha g$, the descent lemma gives

$$\mathcal{L}_K(\theta^+) - \mathcal{L}_K(\theta) \leq -\alpha\langle g_K, g\rangle + \frac{L\alpha^2}{2}\|g\|^2. \tag{28}$$

Proof. Smoothness implies $\mathcal{L}_K(\theta + d) \leq \mathcal{L}_K(\theta) + \langle g_K, d\rangle + L\|d\|^2/2$. Set $d = -\alpha g$. □

With exact projection in a conflicting batch, $\langle g_K, g_{\text{proj}}\rangle = 0$, so Eq. equation 28 permits an increase of order α^2. For the stabilized direction in Eq. equation 27, the first-order contribution $-\alpha\langle g_K, g\rangle$ is positive when $\gamma < 0$ and $\epsilon_p > 0$. Neither rule proves non-increase under AdamW, whose preconditioner and weight decay change the parameter step. The actual reference score remains an empirical question.

Property 14: sensitivity to an estimated retention gradient. Suppose a corrected direction g is orthogonal to an estimate $\widehat{g}_K$ of g_K. Then

$$|\langle g, g_K\rangle| = |\langle g, g_K - \widehat{g}_K\rangle| \leq \|g\|\|g_K - \widehat{g}_K\|. \tag{29}$$

Proof. Use $\langle g, \widehat{g}_K\rangle = 0$ and Cauchy–Schwarz. □

Thus even an exact projection onto a sampled gradient need not be orthogonal to the full-vocabulary reference gradient. Equation equation 29 makes the estimator error explicit; it does not provide its numerical size.

B.3 Diagnostics for the Two Objectives

The projection identities describe the correction to the parameter gradient, while the actual update also depends on the optimizer. Retention therefore has to be evaluated on the overview-only model rather than inferred from the geometry of one update. No numerical gradient-cosine curve is used to support the theoretical statements above.

C FineBind Benchmark

FineBind contains 2,500 training questions and a 500-question test set that was fixed before the student pilots. Each question uses a single image with multiple objects. The questions cover pairwise direction, extrema ordering, and binding between visually similar objects. Together, these cases test whether a model assigns each referred object to its correct location within a shared scene before judging spatial relations. At inference, BindOPSD receives only the original image and question.