MaRR: Map-Based Rotation Reasoning for Vision–Language Models

Anonymous authors
Paper under double-blind review

Abstract

Forward mental rotation requires predicting an object's unseen pose after a specified rotation, rather than recognizing a visible configuration. Our controlled diagnostics show that frozen vision–language model features contain decodable rotation information and support a supplied transformation, while inferring and reusing transformations across objects remains unreliable. Building on this insight, we introduce MaRR (Map-based Rotation Reasoning), a teacher-guided self-distillation framework that combines object-adaptive representations with a shared, object-independent transformation procedure. Using teacher traces, MaRR teaches VLMs to *map* object-specific features, *rotate* their spatial relations, and *match* the predicted configuration to candidate images. Fine-tuning Qwen3.5-4B with this supervision improves held-out CAD accuracy, with gains extending to shifted CAD renderings and block objects, although transfer to public benchmarks remains selective. A matched comparison shows that trace design introduces trade-offs: natural teacher responses yield stronger OOD transfer and physical-choice consistency, while standardized MaRR traces achieve higher accuracy under candidate permutation. Together, these findings suggest that structured supervision can help VLMs organize and use their latent spatial knowledge, while the way the procedure is expressed in teacher traces affects transfer and robustness.

1 Introduction

Vision–language models (VLMs) increasingly answer questions about spatial relations, depth, and three-dimensional scenes (Cheng et al., 2024; Cai et al., 2025; Wang et al., 2025). Forward mental rotation poses a distinct challenge: the relevant object state is not directly observed, but must be predicted from an image and a requested transformation (Hong et al., 2023; Zhang et al., 2026). We study a fixed-camera setting in which the object rotates in a fixed world frame and the model selects the resulting pose from four candidate images. Success requires more than recognizing a familiar shape or matching the reference to its closest visible counterpart.

A successful solution must identify object structure, interpret the rotation command, apply it consistently, preserve part correspondence through changes in visibility, and compare the predicted state with the candidates. Failure at any of these steps can produce the same incorrect final answer. This raises a central question: are failures primarily due to missing spatial information, an inability to transform available representations, or an inability to compose these ingredients reliably across different objects?

Our diagnostics reveal usable rotation information in frozen features and improved prediction under a supplied transformation, but unreliable cross-object inference and reuse. These findings narrow the failure modes without establishing a unique bottleneck (Section 2).

Coordinate-based supervision further shows that explicit rotation procedures are learnable, although matched-domain gains do not ensure transfer. This motivates a design that structures the transformation without forcing every object into the same exhaustive description.

We introduce MaRR (Map-based Rotation Reasoning) around the principle of *adaptive object representation plus shared transformation procedure*. MaRR organizes reasoning as *map–rotate–match*: identify a small set of object-specific features, transform their spatial relations under the requested rotation, and compare the predicted configuration with the candidates (Figure 1). The scaffold is fixed; the object-specific representation is not. Teacher-generated traces make this interface available for supervised fine-tuning (SFT).

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.

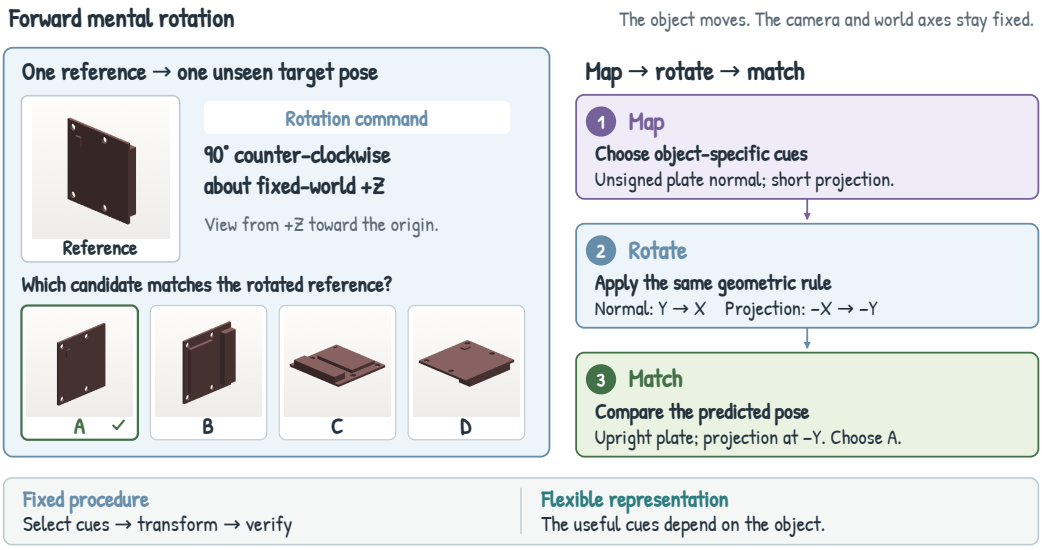


Figure 1: **Task and MaRR design principle.** The object rotates under a fixed camera and world frame; the green outline marks the correct candidate, A. MaRR maps object-specific cues, applies a shared rotation, and matches the predicted pose. Here the plate normal is an unsigned line. The diagram specifies a supervision interface, not separately executed student modules.

Our contributions are:

- controlled diagnostics separating readable rotation information, supplied transformation action, and inferred cross-object reuse;
- an object-adaptive map–rotate–match interface and a teacher-supervision pipeline for VLM training; and
- evidence that procedural supervision helps, while overly standardized traces can hurt transfer and physical-choice consistency, with tradeoffs in accuracy under candidate permutation.

2 Controlled Diagnostics: Why Isn't VLM Reasoning Enough?

We use controlled experiments on block objects to distinguish information present in frozen embeddings from the ability to transform and reuse it. These diagnostics later motivate the design of MaRR. Figure 2 summarizes the interventions and results, with protocol details in Appendix A.

Experiment A: Is rotation-relevant information present? We ask the model to choose the same object under rotation from four candidates given a reference. For each reference–candidate pair, we construct a feature vector by concatenating the elementwise sum, absolute difference, and product of the two normalized frozen image embeddings. A 4-layer multi-layer perceptron (MLP) is then trained on the feature vectors. The probing result yields above-chance accuracy, showing that useful task information is decodable from the selected visual features. This is aligned with emerging view that VLMs can be trained to represent and reason about 3D structure from 2D observations (Cai et al., 2026; Cheng et al., 2025).

Experiment B: Can the representation support a known transformation? In a separate experiment, a router receives the correct rotation ID and sends a query object's frozen feature to an expert netowrk that learns to predict the object's feature after transformation. Supplying this action improves prediction over a query-only baseline, whereas supplying a wrong rotation substantially reduces accuracy. A rotation is therefore learnable when the identity of the transformation is given.

Experiment C: Can the model infer and reuse the transformation reliably? We next replace the supplied action with a rotation inferred from a before–after pair of another object. A rotation reader must identify the change, based on which an operator learns the predicted feature. This pipeline

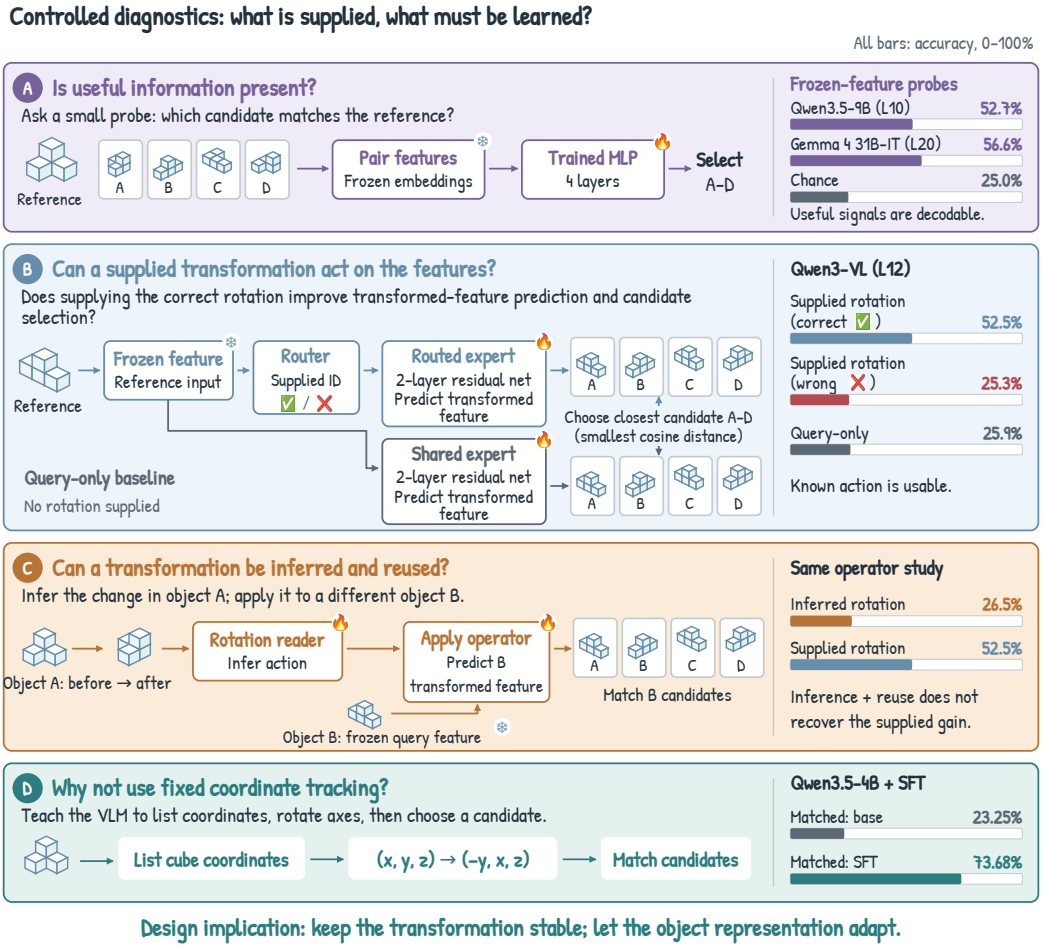


Figure 2: Controlled diagnostics. Experiment A to D show different interventions and their effects on accuracy. Parenthesized L labels identify the visual layer.

does not recover the accuracy gain as in Experiment B. The ability of reliable identification and cross-object reuse remains unresolved in the selected models.

Experiment D: Why not just use explicit coordinate tracking? We examine whether explicit coordinate tracking is a plausible approach to teach VLM rotation. Using SFT to teach the VLM to enumerate cube coordinates, transform and match the predicted pose to candidates, we observe strong gains on samples of block objects, manifesting that such a procedure is indeed learnable. This experiment motivates retaining an explicit transformation procedure.

Implications. Experiment A-C suggest that rotation reasoning requires more than access to useful visual features: models must identify the relevant transformation and apply it consistently across objects. Experiment D further demonstrate that an explicit rotation procedure can be learned. Motivated by these findings, MaRR structures supervision around a shared map–rotate–match sequence, connecting object representation, geometric transformation, and candidate comparison. Rather than implementing an exhaustive coordinate description, the representation is allowed to adapt to object geometry, and within a fixed procedure, we ensure consistent rotation learning through explicit tracking on the key object features.

3 Problem Setup and MaRR Framework

Task formulation. We study forward object rotation: an object rotates about its center while the camera, scene, and world coordinate frame remain fixed. Each question provides a reference image

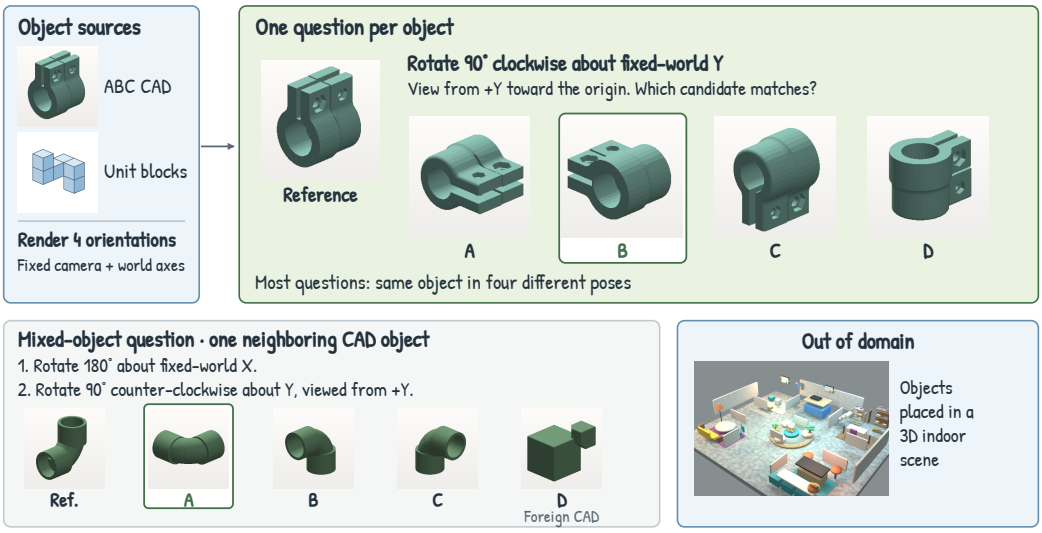


Figure 3: MaRR dataset construction. The original ABC CAD dataset provides only one pose per object. To create the MaRR questions, we render other poses of the same object with fixed camera and world coordinates.

I_0, four candidate images $\{I_k\}_{k\in\mathcal{K}}$, where $\mathcal{K} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$, and a specified rotation g. The input is

$$x = (I_0, I_\mathrm{A}, I_\mathrm{B}, I_\mathrm{C}, I_\mathrm{D}, g). \tag{1}$$

The model predicts an option $y \in \mathcal{K}$ identifying the object pose that results from applying g to the reference. In the standard setting, all candidates depict the same object under the same rendering conditions. The model must distinguish poses rather than object identities.

Map–rotate–match interface. MaRR organizes this prediction through a map–rotate–match interface:

$$m_0 = M(I_0), \qquad \widetilde{m} = R_g(m_0), \qquad \widehat{y} = \arg\max_{k\in\mathcal{K}} S\big(\widetilde{m}, M(I_k)\big). \tag{2}$$

Here M selects an object-adaptive map of recognizable components, their spatial relations and (or) their pointing directions. Action R_g transforms those geometric properties in the fixed world frame, and S compares the predicted configuration with each candidate. A map may track different features for different objects and a successful map need only retain enough information to discriminate the relevant poses.

Dataset construction. Figure 3 illustrates the construction settings. We draw geometrically varied objects from the ABC CAD dataset (Koch et al., 2019). CAD objects are digital 3D models of physical designed parts, such as brackets, plates, and mechaical components, whose varied shapes provide a diverse test of object rotation. We also create block objects as a complementary dataset. For each question, we render a reference image and four candidate poses under fixed camera and world coordinates. Most questions contain only one object across candidates, while mixed questions replace one or two incorrect candidates with other object(s).

For out-of-domain (OOD) dataset, we place the object in a separate indoor environment with varied viewpoint elevation, lighting, and appearance.

All training, validation, and test splits are strictly disjoint by object ID, with no object appearing in more than one split. Dataset composition and evaluation settings are detailed in Appendices B.1 and D.

Teacher-trace construction. We construct supervision teacher traces following the pipeline shown in Figure 4. Two candidate reasoning paths are generated in parallel and merged before the retained evidence is condensed into an SFT-ready trace.

In the staged path, the base VLM maps the reference image into a few features and expresses their directions or spatial relations in the fixed world frame. A deterministic geometry program rotates

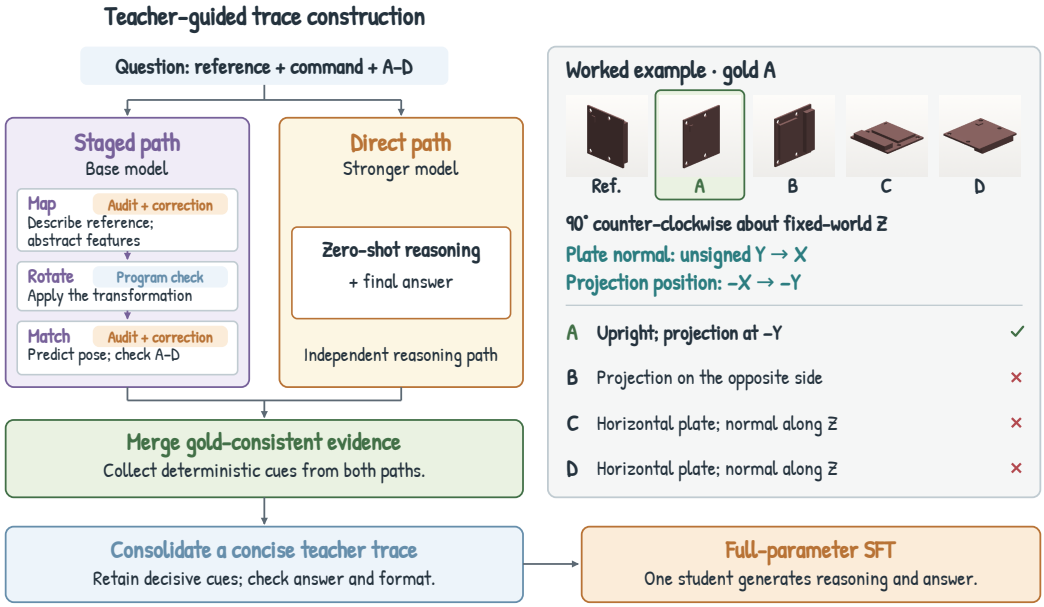


Figure 4: Teacher-trace construction and SFT. Staged and direct reasoning paths supply evidence for a concise training teacher-trace. Map and match receive audit and correction, while rotation receives a program check. In SFT, the student is trained on the complete response.

those properties. The base model then predicts the resulting pose and compares it with candidates A–D. A stronger VLM check and locally revise the visual descriptions whereas the rotation relationship is guaranteed to be correct by program check. In parallel, the stronger VLM solves the original question directly, providing an independent source of comparison evidence.

Both paths are saved before the consulting the gold answer. If neither path selects the correct candidate, the question is removed. If one succeeds, its evidence is used. If both succeed, the stronger-model comparison clues are preferred. The selected evidence is then trimmed to the decisive geometry, spatial relations, and visibility changes and rewritten as one concise map–rotate–match response. Of 2,350 construction questions, for 2185 at least one path succeeds, and 2,114 traces survive final validation. Detailed checkpoint counts and conditional retention rates are reported in Appendix B.2.

Method. The input x contains a reference image, candidates A–D, the fixed world-coordinate convention, and the requested rotation. The model generates a natural-language explanation followed by one boxed answer. The complete prompt is given in Appendix C.1.

We perform full-parameter SFT on the retained trace dataset $\mathcal{D}_{\mathrm{T}}$ using

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\frac{1}{|\mathcal{D}_{\mathrm{T}}|} \sum_{(x,\tau)\in\mathcal{D}_{\mathrm{T}}} \log \pi_\theta(\tau \mid x). \tag{3}$$

Here, τ is the retained teacher response, and $\pi_\theta(\tau \mid x)$ is the conditional probability assigned to that response by the student VLM with trainable parameters θ.

We teach map–rotate–match as one continuous respons. Appendix C specifies the sequence-loss implementation, including supervision of prompt tokens, and distinguishes the original selected checkpoint from the fixed-endpoint ablation runs.

4 Experiment

Training and model selection. We use Qwen3.5-4B as the base model on 2,114 retained CAD examples and GPT-5.6-sol as the stronger model for teacher-trace construction. A checkpoint is selected on the validation set evaluating both the accuracy rate and the robustness under deterministic

permutation. Dataset composition, optimization settings, and checkpoint selection are reported in Appendices B.1 and C.

Comparison models and inference. We compare with proprietary VLMs, specialized spatial-reasoning models, and general open-weight VLMs, using the models shown in Tables 1 and 2. External baselines are evaluated zero-shot with the same images and question text. Exact model identifiers and inference settings are given in the appendix.

Benchmarks. We prepare held-out test set from the ABC dataset, OOD CAD set, and block object set which test model generalization capacity to colored six-block and uniform eight-block objects. CAD and OOD CAD contain 300 questions each, and each block family contributes 300. Exact composition and the remaining human-review limitations are in Appendix D.

We additionally evaluate on SpinBench (218 samples) (Zhang et al., 2026), SRBench (400) (Hong et al., 2023), and OmniSpatial (28) (Jia et al., 2025). We retain only the fixed-camera object-rotation tasks. Details of the subsampling process are given in Appendix D.2.

Evaluation protocol. We report pass@1 accuracy using frozen task-specific parsers. Malformed or ambiguous responses are incorrect; relaxed CAD reparsing is reported only when explicitly marked as a format diagnostic. Public benchmarks are parsed in their native answer spaces, and unrelated tasks are not pooled into a single score. Parser specifications are provided in the appendix.

5 Results & Analysis

MaRR-SFT substantially improves its base model on unseen CAD objects and outperforms the evaluated baselines in general (Table 1). Mixed-object questions are generally easier than same-object questions, consistent with object identity providing an additional elimination cue. Gains also extend to both block object families despite training on CAD objects. This pattern supports transfer across object geometries within the forward-rotation setting, despite that the remaining errors reveal performance unreliability.

OOD generalization and spatial-benchmark transfer. The advantage over the base model persists under OOD dataset (Table 2), suggesting that the gains are not confined to the training environment. Transfer to public benchmarks is less consistent. MaRR improves on SRBench Easy and OmniSpatial but underperforms its base on SpinBench forward rotation and SRBench Hard. The OmniSpatial improvement is also based on a small test set and should be interpreted cautiously.

	(A) CAD object task			(B) Block object tasks		
Model	Same	Mixed	Overall	Colored n06	Generic n08	Overall
Proprietary models						
GPT–5.1	39.5	51.1	43.0‡	**31.0**	23.3	27.2‡
DeepSeek V4.1 Flash	5.7	13.3	8.0‡	0.0	0.0	0.0‡
Specialized spatial-reasoning models						
MultiHopSpatial	6.2	10.0	7.3‡	1.0	1.7	1.3‡
3DThinker-MindCube	19.5†	23.3†	20.7†	0.0	0.0	0.0
SpatialStack	20.0†	43.3†	27.0†	0.0	0.0	0.0
SpaceOm	26.7	44.4	32.0	19.0	23.0	21.0
SpaceQwen2.5-VL-3B	29.5	45.6	34.3	22.7	21.7	22.2
General open-weight models						
Qwen3.5-4B	12.9	35.6	19.7‡	13.7	14.3	14.0
Molmo2-4B	25.7	35.6	28.7	26.3	22.0	24.2
MiniCPM-V-4	27.1	41.1	31.3	24.3	24.0	24.2
InternVL3.5-4B	14.8	13.3	14.3	11.0	17.0	14.0
MaRR-SFT (ours)	**42.9**	**65.6**	**49.7**‡	30.3	**26.3**	**28.3**

Table 1: Primary rotation results. Pass@1 accuracy (%) on CAD objects and block objects. †Relaxed format diagnostic for models with no strict-format-valid CAD responses. ‡Strict score with at least one length-truncated response. Parsing and coverage details are given in Appendix D.3; a dash denotes an unavailable result.

	(A) OOD	(B) Other spatial benchmarks			
Model	CAD objects	SpinBench Fwd.	SRBench Easy	SRBench Hard	OmniSpatial
Specialized spatial-reasoning models					
3DThinker-MindCube	22.0†	18.3	26.0	**30.5**	21.4
SpatialStack Qwen3.5-4B	28.3†	37.2	27.0	25.5	21.4
General open-weight models					
Qwen3.5-4B base	19.3	**52.3**	28.0	23.5	25.0
Qwen3.5-9B	24.7	43.1	26.5	25.5	17.9
Qwen3-VL-8B-Instruct	23.0	36.2	20.5	13.5	7.1
InternVL3.5-4B	21.7	30.3	**31.5**	29.5	**35.7**
MaRR-SFT (ours)	**42.7**	29.4	31.0	20.0	28.6

Table 2: OOD generalization and public-benchmark transfer. Pass@1 accuracy (%). †Relaxed CAD format diagnostic for models with no strict-format-valid responses. Public benchmarks are reported separately without an aggregate score.

Breaking down. Table 3 summarizes improvements by question type. First, the gains are distributed across questions with different steps of rotation rather than concentrated in the shortest transformations. On OOD CAD, MaRR maintains similar accuracy across distances, while its base performs worst at the largest distance. However, improvement is not uniform across axes: gains are strongest for Z-axis rotations, with substantially smaller OOD gains for X and Y. Finally, a comparable gains on same- and mixed-object questions further indicate that the benefit is not limited to rejecting foreign-object distractors.

	Steps in rotation			Rotation axis			Candidates	
Evaluation	1	2	3	X	Y	Z	Same	Mixed
CAD	+40.9	+21.1	+26.3	+26.7	+40.0	+48.5	+30.0	+30.0
OOD CAD	+19.0	+20.0	+31.0	+6.8	+3.7	+36.5	+23.3	+23.3

Table 3: Matched MaRR-SFT gains over Qwen3.5-4B in strict pass@1 accuracy (percentage points). Steps in rotation denotes the minimum number of quarter-turns from the reference pose. Axis results use single-command questions only. "Mixed" pools questions containing one or two foreign-object distractors.

Accuracy and response length. MaRR produces substantially shorter responses than its Qwen3.5-4B base on both CAD evaluations and all four public benchmarks (Table 4). Procedural supervision encourages concise responses. On CAD and OOD CAD, shorter responses accompany higher accuracy, whereas public-benchmark accuracy improves only selectively.

Table 4: Mean response length in generated tokens per question for Qwen3.5-4B and the original MaRR-SFT checkpoint. Means include invalid and length-truncated outputs. Counts are recorded output tokens, excluding input tokens. Both models use the a common generation limit.

Model	CAD objects	OOD CAD objects	SpinBench Fwd.	SRBench Easy	SRBench Hard	OmniSpatial
Qwen3.5-4B	7,491.7	7,472.7	1,387.9	5,499.4	3,687.3	3,650.3
MaRR-SFT	316.3	263.0	210.7	192.4	227.9	247.9

Ablation Study. We test whether the benefit of teacher supervision depends on enforcing our standardized realization of map–rotate–match. We retain the 1,940 training questions for which the direct GPT-5.6-sol response is correct and train two Qwen3.5-4B students on exactly the same question IDs, images, answers, and optimization schedule. The only change is the assistant target. The *explicit MaRR* arm uses our condensed traces, which first map a small set of features, transform their spatial relations, and then verify the predicted pose against the candidates. The *natural direct* arm instead uses the original correct teacher response without rewriting it into this organization.

	Accuracy			Candidate-order diagnostic			Target
Supervision target	CAD	OOD CAD	Blocks	Original	Permuted	Consistency	Tokens
Explicit MaRR	52.3	38.7	29.3	53.1	**57.8**	54.7	275.4
Natural direct	**55.7**	**47.3**	**34.2**	**53.5**	48.4	**66.0**	137.7

Table 5: Matched 1,940-example supervision comparison (in percentages); target length is the mean number of training-target tokens.

The natural targets give higher point estimates on CAD, OOD CAD, and block objects, with the clearest difference on OOD CAD. The paired OOD difference is $+8.7$ percentage points (item-bootstrap 95% interval above 0), whereas the CAD interval includes zero. Natural supervision also yields greater physical-choice consistency. However, the explicit arm is more accurate after candidate permutation. The result is therefore a transfer–robustness tradeoff rather than dominance by one target style.

Importantly, *natural* does not mean structure-free. Inspection shows that the direct traces often perform one or several of the operations as MaRR, but organize them more flexibly. They often do not contain an explicit step of selecting a variable number and type of visual cues, and often exclude candidates immediately from a decisive pose cue, and then transform only the relations needed to distinguish the remaining choices. They may omit an exhaustive candidate-by-candidate comparison or interleave mapping, rotation, and matching. Our explicit traces instead place these operations in a more consistent order and typically verbalize a fuller verification. This organizational difference is visible in the approximately twofold difference in target length.

This comparison does not isolate why the natural traces perform better on transfer. Trace length, wording, feature selection, operation order, and use of early exclusion all change together. It shows that imposing a standardized realization can be weaker than a more flexible realization in the same tasks.

6 Related Work

Spatial reasoning and mental rotation. Spatially grounded VLMs learn relations and geometric properties through region supervision, depth signals, or intermediate scene representations (Cheng et al., 2024; Cai et al., 2025; Wang et al., 2025). Rotation benchmarks examine complementary abilities, including recognition across poses, commanded transformations, and cross-object relational transfer (Ramakrishnan et al., 2025; Zhang et al., 2026; Yiu et al., 2025). Forward object rotation differs from recognizing a visible pose. The model must predict a state that has not been shown. It also differs from viewpoint change, because the camera remains fixed, and from transformation inference, because the rotation is supplied rather than inferred from an image pair. We complement benchmark evaluation with controlled diagnostics of whether rotation-relevant information is available, whether a supplied transformation can act on it, and whether transformations can be inferred and reused across objects.

Structured multimodal supervision. Visual chain-of-thought and grounded intermediate representations expose evidence that final-answer supervision leaves implicit (Shao et al., 2024; Ranasinghe et al., 2024). Program distillation and process supervision can similarly make intermediate operations explicit (Hu et al., 2024; Li et al., 2026). Spatial approaches use cognitive maps or orthographic descriptions to connect perception with downstream reasoning (Wang et al., 2025; Zhan et al., 2026). These studies establish the value of intermediate supervision; our question is how tightly that intermediate reasoning should be specified. We examine the degree of constraint imposed on a shared procedure, rather than treating supervision as a binary choice between structured and unstructured reasoning.

7 Conclusion

Our diagnostics show that VLM representations contain usable rotation-relevant information, yet reliably inferring and reusing transformations across objects remains difficult. MaRR teaches a map–rotate–match procedure through object-adaptive teacher traces. This supervision improves rotation accuracy on test CAD objects, transfers to OOD renderings and block geometries, with

shorter responses, although gains on public benchmarks remain selective. The matched supervision comparison further shows that standardizing the trace is not uniformly beneficial.

Future directions. An important next step is to generalize reasoning across various tasks and strengthen robustenss to candidate order. Future studies should also vary teacher quality and trace-selection criteria, and independently control trace length, feature selection, and wording to test whether representational flexibility itself improves transfer.

AI USAGE DISCLOSURE

The author used AI assistance in five aspects of this research.

Retrieval and discovery. AI-assisted literature discovery led the author to a recent spatial-reasoning benchmark that highlighted substantial limitations in VLM rotation performance. To contextualize this observation, the author asked AI to compile source links for ICLR and ICML 2025–2026 oral and poster papers into a spreadsheet and filter for work on 3D reasoning and multimodal learning. The author used the resulting collection to survey recent progress and understand the research landscape.

Research ideation and execution. The author used AI to brainstorm rotation tasks and object families that would extend controlled cube experiments toward more realistic geometry. This discussion identified the ABC CAD dataset as a candidate source. AI also assisted with experimental implementation and execution, including the probing experiments on the cube-based tasks and later SFT and ablation studies.

Synthetic dataset generation. AI assisted in constructing the initial cube-object rotation dataset for probing experiments. Following the discussion of ABC CAD, the author asked AI to help construct an analogous rotation task using individual objects retrieved from that dataset. These activities supported the development of the author's synthetic data-construction pipeline; the generation of training teacher-traces is described in Section 3.

Writing and language revision. The author used AI to refine the wording and organization of manuscript drafts, improve clarity and academic style, and correct grammatical errors.

Draft sections of the paper. The author used AI to draft manuscript sections from existing research materials, experimental results, and author-provided outlines. This assistance included organizing the argument and developing initial descriptions of the methods and findings, beyond polishing existing prose.

REFERENCES

Zhipeng Cai, Ching-Feng Yeh, Xu Hu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025. URL https://arxiv.org/abs/2509.25413.

Zhipeng Cai, Zhuang Liu, Yunyang Xiong, Zechun Liu, Vikas Chandra, and Yangyang Shi. VLM3: Vision language models are native 3d learners. *arXiv preprint arXiv:2605.30561*, 2026. URL https://arxiv.org/abs/2605.30561.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.01584.

An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, Hongxu Yin, Xiaolong Wang, and Sifei Liu. 3d aware region prompted vision language model. *arXiv preprint arXiv:2509.13317*, 2025. URL https://arxiv.org/abs/2509.13317.

Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Visual reasoning: From state to transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. doi: 10.1109/TPAMI.2023.3268093. URL https://arxiv.org/abs/2305.01668.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Hu_Visual_Program_Distillation_Distilling_Tools_and_Programmatic_Reasoning_into_Vision-Language_CVPR_2024_paper.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. *arXiv preprint arXiv:2506.03135*, 2025. URL https://arxiv.org/abs/2506.03135.

Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. ABC: A big CAD model dataset for geometric deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9601–9611, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Koch_ABC_A_Big_CAD_Model_Dataset_for_Geometric_Deep_Learning_CVPR_2019_paper.html.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. STAR-R1: Multi-view spatial transformation reasoning by reinforcing multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12041–12051, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_STAR-R1_Multi-View_Spatial_TrAnsformation_Reasoning_by_Reinforcing_Multimodal_LLMs_CVPR_2026_paper.html.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kr"ahenb"uhl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06468.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12977–12987, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ranasinghe_Learning_to_Localize_Objects_Improves_Spatial_Reasoning_in_Visual-LLMs_CVPR_2024_paper.html.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. MindCube: Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. URL https://arxiv.org/abs/2506.21458.

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://arxiv.org/abs/2407.17773.

Shaoxiong Zhan, Yanlin Lai, Zheng Liu, Hai Lin, Shen Li, Xiaodong Cai, Zijian Lin, Wen Huang, and Hai-Tao Zheng. 3ViewSense: Spatial and mental perspective reasoning from orthographic views in vision-language models. *arXiv preprint arXiv:2603.07751*, 2026. URL https://arxiv.org/abs/2603.07751.

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.

A Diagnostic Protocols and Scope

The four diagnostics use controlled block-object data but intervene at different points in the reasoning pipeline. They are therefore analyzed as separate engineering studies rather than pooled as one benchmark. Experiments A–C keep the vision backbone frozen and vary what a lightweight readout or operator is allowed to observe. Experiment D instead updates the full VLM with an explicit coordinate-based target. Figure 2 summarizes these interfaces.

Experiment A: frozen-feature probes. For Qwen3.5-9B and Gemma 4 31B-IT, feature extraction records the patch embedding, six intermediate vision blocks, and the connector output. Valid visual tokens are mean-pooled and normalized. A train-only PCA projection maps each probe point to at most 1,024 dimensions; points with the same source width share one PCA fit. For a reference–candidate pair, the probe receives $[z_1 + z_2, |z_1 - z_2|, z_1 \odot z_2]$, so its input is invariant to swapping the two images. Separate two-, three-, and four-linear-layer MLPs are trained for binary matching, four-choice retrieval with a shared pair scorer, and rotation-step classification. Model stage and probe depth are selected on the object-disjoint validation split, and the selected configuration is then evaluated on held-out objects over five independent initializations. A shuffled-label connector control and a resized-pixel pipeline test whether the readout is exploiting a learned task signal rather than a class prior or a trivial image-space comparison. These probes establish decodability, not a complete 3D representation or causal use by the answering VLM.

Experiment B: supplied transformation operator. Each action example contains a reference object before and after rotation and a different query object before the same rotation. Candidate features are withheld until the operator has produced a predicted query state. At Qwen3-VL vision layer 12, the group-indexed oracle receives the known rotation identity and routes the query feature through the corresponding member of a 24-expert bank. Each active expert is a $1{,}024 \to 8 \to 1{,}024$ residual network, after which cosine similarity scores the predicted state against the candidates. The matched query-only control uses the same cached features and active network dimensions but always selects expert zero without receiving a rotation. The oracle and control share 6,368 training questions, 796 object-disjoint validation questions, four independent initializations, hyperparameters, and validation-based checkpoint selection. A wrong-rotation-code intervention tests whether the prediction actually depends on the supplied operation. The control matches active-expert dimensions but not the number of experts receiving task gradients. This experiment isolates whether a known transformation can be applied at the selected feature boundary.

Experiment C: inferred cross-object transformation. The operator and candidate-blind scoring rule are retained, but the rotation identity is no longer supplied. A reader T must infer one of the 24 rotation codes from the ordered before–after pair of the reference object, and the operator Φ applies that code to the different query object. The frozen development protocol compares this inferred system with query-only, relation-comparator, shuffled-label, same-rotation-reference-substitution, and wrong-code controls; it also maps predictions back to physical candidates after answer-option permutation. Representation point and checkpoints are selected only on the development validation split. The separate confirmation objects and the legacy validation and test splits remain locked. The observed pattern is sensitivity to the inferred code without reliable operation selection, so the experiment does not establish cross-object reuse.

Experiment D: coordinate-tracking SFT. This study uses uniform six-cube questions with one reference and four candidate images. The engineered target first reconstructs the complete set of six normalized voxel coordinates, then applies the commanded rotation, simulates the transformed state, compares it with every candidate, and ends with a boxed answer. Qwen3.5-4B is updated with full-parameter SFT rather than held fixed behind a probe. Evaluation uses a frozen 2,116-question validation set and a deterministic candidate permutation whose predictions are mapped back to physical candidates. Transfer checks reuse the selected checkpoint without CAD-specific fine-tuning: the CAD pilot changes only the terminal boxed-answer instruction, while the public SRBench diagnostic uses a secondary terminal-answer reparse. Because these checks differ in domain, prompt, and parser, they are kept separate and do not isolate representation choice as the cause of transfer behavior.

B Dataset and Teacher-Trace Construction

B.1 Training-Corpus Composition

Table 6: Training-corpus composition.

Item	Full corpus	Matched direct-correct subset
Object-disjoint questions	2,114	1,940
Images	10,570	–
Appearance composition (full-corpus questions)		
Palette-rendered	1,480	–
Stronger appearance variation	423	–
White-object controls	211	–
Candidate composition (questions)		
Same questions	1,918	1,758
Mixed: one foreign-object distractor	129	116
Mixed: two foreign-object distractors	67	66

Both matched arms use identical question text and images; the assistant target is replaced by either the retained MaRR trace or the original correct direct teacher response. The direct-correct selection limits the population to which the style comparison applies.

B.2 Teacher-Trace Construction

Table 7 reports the full construction flow. The two candidate paths overlap and therefore do not form a single decreasing funnel; the branch-outcome rows instead partition the complete question pool.

Table 7: Teacher-trace construction flow. "Parent" identifies the set from which each conditional rate is computed. Checking is conditioned on a completed map and rotation and can recover an initially unparsed comparison.

Checkpoint	Parent	Passed	Cond. (%)
Parallel candidate paths			
Questions processed	—	2,350	—
Direct response completed	Question pool	2,350	100.0
Direct answer correct	Question pool	2,006	85.4
Staged map and rotation completed	Question pool	2,157	91.8
Initial comparison parsed	Map and rotation	1,419	65.8
Checked comparison completed	Map and rotation	2,157	100.0
Staged answer correct	Checked comparison	1,661	77.0
Branch outcomes			
Both paths correct	Question pool	1,482	63.1
Direct path only correct	Question pool	524	22.3
Staged path only correct	Question pool	179	7.6
Neither path correct	Question pool	165	7.0
At least one path correct	Question pool	2,185	93.0
Post-merge funnel			
Condensed trace returns the known answer	At least one correct path	2,149	98.4
Final SFT trace retained	Gold-consistent condensation	2,114	98.4

These are answer and construction checks, not exhaustive geometric certification of every sentence.

C Training and Model Selection

C.1 Supervised Fine-Tuning Prompt

Each SFT and GRPO example uses the following user-prompt template. The five image placeholders are populated, in order, by the reference image and candidates A–D. Only the rotation command changes across examples.

```
You will see five orthographic images of the same rigid CAD object.
Image 1: <image> is the reference.
Image 2: <image> is candidate A.
Image 3: <image> is candidate B.
Image 4: <image> is candidate C.
Image 5: <image> is candidate D.

The object rotates while the camera, lighting, image scale, and world axes
remain fixed. The object's material color is uniform; it does not label
individual parts.

The coordinate frame is fixed to the room and camera. It does not turn with
the object:
- the positive z direction points vertically upward in the image;
- the positive x direction projects diagonally downward and rightward;
- the positive y direction projects diagonally upward and rightward;
- each negative direction points opposite its corresponding positive
  direction.

Interpret every rotation command in this fixed-world coordinate frame.

Rotation command:
{ROTATION_COMMAND}

Which candidate matches the rotated reference? Explain your reasoning in
your own words, then end with exactly one final line in the form \boxed{X},
where X is A, B, C, or D.
```

C.2 Optimization Configuration

All Qwen ablation runs use the same Qwen3.5-4B revision, three epochs, learning rate 10^{-5}, effective batch size 16, warmup ratio 0.1, and full parameter updates. Full-data runs complete 399 optimizer steps and matched-subset runs complete 366.

Table 8: Full-data MaRR-SFT optimization configuration.

Item	Value
Trainable parameters	Full model
Epochs	3
Effective batch size	16
Learning rate	10^{-5}
`warmup_ratio`	0.1
`lr_scheduler_type`	`cosine`
`bf16`	`true`
Optimizer steps	399

C.3 Checkpoint Selection

The original full-data model uses checkpoint 360 from an evaluation of checkpoints every 30 steps and the final step 399 on the frozen 256-question validation cohort. The retained comparison records 86/256 correct under both orders for checkpoints 360 and 390; checkpoint 360 has higher physical consistency (57.0% vs. 55.5%). Its original and permuted accuracies are 48.8% and 49.6%.

D Evaluation Protocol

D.1 Rendered Evaluation Cohorts and Integrity

Table 9: Validation and evaluation-set composition. Object counts are reference objects followed by foreign objects used only as distractors. Dashes mark candidate-composition fields that do not apply.

Set	Questions	Objects	Same	Mixed		Balance
				One foreign object	Two foreign objects	
Primary ABC CAD test	300	300/0	210	60	30	75 per answer
Colored block objects n06	300	43/0	–	–	–	75 per answer; 100 per turn
Generic block objects n08	300	300/0	–	–	–	75 per answer; 100 per turn
OOD CAD	300	300/90	240	30	30	75 per answer

Repeated colored-block objects preclude treating all 300 questions as independent object samples. Training and evaluation object separation and image availability are checked mechanically. Full human visual certification remains incomplete; OOD results should in particular be treated as preliminary distribution-shift evidence.

D.2 Public-Benchmark Selection

The public evaluation retains 218 SpinBench forward-rotation questions and 189 SpinBench rotation-direction questions, 200 questions from each of the SRBench Easy and Hard splits, and 28 reviewed object-rotation questions from OmniSpatial. We exclude components that rotate the camera or observer and tasks based on folding, projection, or cross-sections. The retained units remain separate during inference, scoring, and reporting.

D.3 Models, Inference, and Parsing

Within each evaluation set, every model receives the same ordered images and question text; prompts do not disclose whether foreign-object distractors are present. We use deterministic decoding when supported and retain the model revision, prompt and image hashes, decoding parameters, raw response, and parser version for every question. Specialized checkpoints are additionally checked for disclosed overlap between their post-training data and the selected evaluation questions.

The rendered-task prompt supplies the reference and candidates A–D, the fixed-world axis convention, and the ordered rotation command. Positive Z projects upward, positive X down-right, and positive Y up-right; commands specify viewing from the positive axis endpoint toward the origin. The prompt requests an explanation followed by exactly one final line `\boxed{X}`, where X is A–D. The matched Qwen evaluations use SGLang, deterministic generation with thinking disabled, and a 16,384-token output cap. GPT-5.1 uses the pinned `gpt-5.1-2025-11-13` snapshot through the Responses API, original-detail images, medium reasoning effort, the same 16,384-token output cap, and at most 32 concurrent requests.

For CAD questions, the strict parser requires exactly one boxed A–D choice on the final nonempty line. A wrong, missing, ambiguous, or malformed answer is incorrect. The separately reported relaxed parser accepts an otherwise unambiguous choice in a different syntax and is used only as a format diagnostic. Public-benchmark parsers accept one unambiguous terminal answer in the task's native label space. We also retain format validity and predicted answer distributions.

The SpatialStack complementarity ablation uses the same retained public units except SpinBench direction, giving 646 questions per arm. Its primary loose parser accepts exact answer lines, answer tags, boxed or bare terminal labels, and the systematic terminal form `Answer: X}`. Exact `Answer: X` compliance is stored separately. Ambiguous and length-limited responses remain invalid and count as wrong. The two arms cover all 1,292 expected responses with zero request errors; independent rescoring, image hashes, checkpoint hashes, and geometry and fusion hook counts all pass.

On CAD, 3DThinker and SpatialStack have no strict-format-valid CAD responses: the former returns answer tags and the latter bare labels, so scores obtained from those formats are relaxed format diagnostics. Length truncation occurs on 270/300 DeepSeek, 238/300 MultiHopSpatial, 10/300 Qwen3.5-4B, and 1/300 MaRR-SFT CAD responses; on block objects it occurs on all 600 DeepSeek responses and 559/600 MultiHopSpatial responses. GPT-5.1 has 296/300 strict-format-valid CAD responses and 585/600 strict-format-valid block responses; 2/300 CAD and 3/600 block responses are length-truncated. Its strict block counts are 93/300 on colored n06 and 70/300 on generic n08. MaRR-SFT has 298/300 strict-format-valid CAD answers; in addition to the truncated response, one response ends with `\boxed{None}`. Coverage and independent rescoring agree for all 7,200 responses from the 12 models evaluated on block objects. Both matched trace models have 300/300 valid CAD and OOD answers. Relaxed parsing is not substituted for any main-table score.

D.4 Candidate Permutations and Consistency Metrics

One frozen nonidentity permutation is applied per question. Predictions in the permuted label space are mapped back to the original physical candidate. Physical consistency requires two valid predictions selecting that same candidate, whether correct or wrong; invalid pairs count as inconsistent. Correct-under-both requires correctness in both orders. The 256-question validation cohort was filtered by teacher answerability and is not a population sample. Its use for original-model checkpoint selection also means those order statistics are not an independent test of that selection. The original full-data model has a separate 300-question held-out order test: 149 original correct, 150 permuted

correct, 105 correct under both, and 174 consistent pairs. Those counts are not combined with the 256-question ablation diagnostic.

D.5 Response-Length Accounting

For response-length analyses, we preserve each provider's raw usage record and normalize only reported input, cached-input, image-input, reasoning, output, and total-token fields; unavailable fields remain null. For open-weight models, generated tokens and GPU inference time are recorded separately. Controlled comparisons between Qwen3.5-4B and MaRR-SFT use their shared tokenizer and backbone. Cross-family token totals are treated descriptively rather than as directly comparable computation.

E Supplementary Results

E.1 Candidate-Order Robustness

We apply one deterministic nonidentity permutation to every held-out CAD question and map predictions back to physical candidates. Table 10 reports the paired results.

Table 10: Candidate-order robustness on the held-out CAD test set. "Both correct" requires correctness under both orders; consistency requires both responses to identify the same physical candidate.

Model	Original	Permuted	Both correct	Consistency	Valid (orig.)	Valid (perm.)
MaRR-SFT	49.7	50.0	35.0	58.0	99.3	100.0

The permutation leaves aggregate accuracy nearly unchanged, but only 58.0% of paired responses select the same physical candidate. MaRR-SFT therefore retains substantial question-level instability despite its stable aggregate score.