BENCHMARKING AND DIAGNOSING CONTINUOUS GEOMETRIC REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) perform well on many spatial benchmarks, but their ability to reason over continuous geometry remains unclear. We introduce two complementary tasks: TangramPose, requiring object-specific prediction of position, rotation, and scale, and CanvasPlace, requiring placements that satisfy geometric constraints. Across 10 VLMs, most models achieve reasonable scene overlap yet fail strict object-level reconstruction, while the strongest model performs substantially better in small number piece task but fail when piece number increase. In-context demonstrations mainly improve scale calibration, whereas fine-tuning yields much larger gains, even with a frozen visual encoder. However, these gains transfer poorly to disjoint rotation ranges, unseen piece counts, and a distinct placement task. Probing shows that angle and scale are decodable from the base models' visual representations far more accurately than they are expressed in final predictions. Counterfactual and activation-patching analyses further show that adaptation improves object-specific responses and strengthens the influence of target-region states on geometric outputs. Together, our results distinguish accurate in-distribution geometric prediction from robust, reusable geometric competence, and provide a framework for analyzing where VLMs succeed and fail in continuous geometric reasoning. Dataset and code will be released upon acceptance.

1 INTRODUCTION

Vision-language models (VLMs) have made progress in spatial understanding (Chen et al., 2024), but the reliability of precise, object-specific predictions of continuous geometric quantities remains an open research question. In applications such as manipulation (Gong et al., 2023) and spatial planning (Xu et al., 2024), predictions must recover not only which objects are present or how they are related, but also precise object-specific quantities such as position, rotation, and scale. Small numerical errors incorrect object–attribute associations, or inconsistent multi-object predictions can make an otherwise plausible configuration unusable.

We study this problem through two complementary tasks. **TangramPose** requires a model to recover the identity and continuous pose of every visible piece in scenes containing one to seven objects. **CanvasPlace** instead asks the model to generate any target pose satisfying a set of geometric constraints, including containment, spatial relations, clearance, orientation, and collision avoidance. The two tasks therefore probe different aspects of continuous geometric reasoning: recovering an observed configuration versus constructing a new geometrically valid one. Our benchmark contains 4,970 Tangram scenes and 1,500 CanvasPlace tasks, together with reproducible generators for controlled pose, composition, and difficulty variation.

Across ten proprietary and open VLMs, we find that continuous geometric capability is highly uneven. Most models produce valid structured outputs and non-trivial scene overlap on Tangram, yet fail strict object-level reconstruction because position, rotation, scale, and identity must all be correct simultaneously. On CanvasPlace, spatial-relation accuracy remains relatively high while strict success decreases substantially as the feasible placement region becomes more constrained. These results reveal a recurring gap between coarse spatial correctness and precise geometric validity.

We next ask whether this gap can be reduced through demonstrations or model adaptation. In-context examples substantially improve some attributes, especially scale, but do not recover exact reconstruction for weaker models. In contrast, parameter adaptation produces much larger gains. A

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.

language-only LoRA variant improves both Qwen and InternVL while keeping the visual encoder and cross-modal interface frozen, and additional visual and cross-modal adaptation further improves reconstruction. However, strong in-distribution performance does not imply robust geometric generalization: accuracy degrades sharply on disjoint rotation ranges and unseen piece counts, and Tangram adaptation does not produce a consistent zero-shot gain on the CanvasPlace task.

Finally, we examine why poor autonomous predictions can coexist with strong adaptation gains. Oracle-assisted probes decode angle and scale from the base models' visual representations far more accurately than the models predict them in their own outputs, indicating that poor performance cannot be explained solely by the absence of geometric information in visual features. Counterfactual experiments show that adaptation improves object-specific responses to geometric changes, while target-region activation patching shows that visual and language states at multiple stages can causally influence geometric outputs, with stronger effects after adaptation. Together, these results distinguish *decodable geometric information* from its reliable use in object-specific prediction.

Our contributions are:

- We introduce a controlled two-task benchmark for continuous geometric reasoning, combining one-to-seven-piece Tangram pose reconstruction with constraint-conditioned CanvasPlace placement.
- We systematically compare prompting, direct geometric heads, language-only adaptation, and joint language–vision adaptation, and evaluate whether their gains extend to unseen rotations, unseen composition lengths, and a distinct geometric task.
- We diagnose the gap between representation and prediction using probing, controlled counterfactual edits, and activation interventions, showing that geometric information can be decodable even when autonomous geometric prediction remains unreliable.

2 Related Work

Geometric benchmarks and Tangram. Spatial evaluations extend from compositional recognition (Johnson et al., 2017; Thrush et al., 2022) to quantitative and robotic spatial understanding (Chen et al., 2024; Song et al., 2025), 3D and synthetic reasoning (Ma et al., 2025; Wang et al., 2025b), visual simulation (Li et al., 2026), and broader or continuous-space evaluation (Jia et al., 2026; Zhu et al., 2025). Tangram research includes specialized solvers and assembly policies (Yamada et al., 2019; 2024; Zhao et al., 2025), abstract visual reference (Ji et al., 2022), and geometric-element recognition/counting rather than assembly (Zhang et al., 2024). Closest to our setting, TangramSR studies continuous geometric prediction and test-time refinement (Zong & Tan, 2026); TangramPuzzle evaluates outlines and seven-piece assemblies with machine-verifiable, multi-solution-aware geometry (Liu et al., 2026). Its TCE-Light variant also predicts rigid transformations, so pose parameterization alone is not our novelty. We emphasize controlled one-to-seven-piece reconstruction of visible configurations, complementary constraint-conditioned placement, and cross-architecture adaptation, diagnosis, and transfer.

Constraint-conditioned placement. Language-grounded manipulation and continuous target states are studied in VLMbench and ARNOLD (Zheng et al., 2022; Gong et al., 2023). Energy-based planning and Set It Up! compose spatial constraints for rearrangement (Gkanatsios et al., 2023; Xu et al., 2024), while PlaceIt3D evaluates language-guided placement in real 3D scenes (Abdelreheem et al., 2025). CanvasPlace isolates predicted geometric feasibility from robot execution in a controlled 2D setting. Paired with Tangram, it tests whether improved reading of an existing pose transfers to constructing a new constraint-satisfying pose.

Adaptation and mechanistic diagnosis. In-context demonstrations (Brown et al., 2020; Alayrac et al., 2022) and LoRA (Hu et al., 2022) motivate complementary interventions without and with parameter updates. Probing distinguishes decodable information from its use (Belinkov, 2022); causal tracing and activation patching test intervention-sensitive computations (Meng et al., 2022; Zhang & Nanda, 2024), including multimodal information flow (Palit et al., 2023; Zhang et al., 2025). We combine these approaches with blank-image, counterfactual, and oracle-isolation controls to relate local geometric decodability to autonomous prediction.

3 METHOD

3.1 TASK CONSTRUCTION AND INFERENCE PIPELINE

3.1.1 TANGRAMPOSE

Task and pose parameterization. Given an image containing $k \in \{1, \ldots, 7\}$ colored pieces, the model predicts every piece's identity and complete pose:

$$I \longmapsto \widehat{\mathcal{P}} = \{(c_i, x_i, y_i, \theta_i, s_i, f_i)\}_{i=1}^{k}. \tag{1}$$

Here $c_i \in \{\mathrm{T1}, \ldots, \mathrm{T5}, \mathrm{S}, \mathrm{P}\}$ identifies two large triangles, one medium triangle, two small triangles, a square, or a parallelogram. The image boundary defines $[0, 10]^2$, with the origin at bottom-left and y increasing upward. Position is the polygon centroid, θ is counter-clockwise rotation, and s scales the supplied canonical template about its centroid. Reflection f is independently meaningful only for P; triangle reflection is absorbed into rotation, and S uses no reflection. This is recovery of the rendered configuration, not assembly of an unseen decomposition from a silhouette.

Construction and rendering workflow. We generate 4,970 Tangram scenes by sampling and augmenting k-piece subsets from 71 seven-piece source configurations, with source pages kept disjoint across training, validation, and test sets. Augmentation applies a continuous scene rotation in $[-180, 180]^\circ$, per-piece angular jitter in $[-30, 30]^\circ$, individual scale perturbations, and a randomized P reflection. A common rescaling and translation fit the complete polygons inside the canvas with a 0.3-unit margin. The resulting configurations need not remain non-overlapping tilings. We render 768×768 RGB images using a fixed color–identity mapping, varying only the background and outline width. No IDs, centroid labels, axes, ticks, or grids are printed. Each image is stored with its exact geometric annotations and source provenance.Full prompts and augmentation details: Appendix A.1; evaluation settings: Section 4.1.

3.1.2 CANVASPLACE

CanvasPlace construction. We construct 1,500 tasks from 500 source families, each with Easy, Medium, and Hard variants. Targets span ten shapes: circle, square, rectangle, triangle, pentagon, hexagon, L, T, U, and arrow. Each scene contains three to eight objects, including a non-solid rectangular tray. Inputs are a 512×512 scene with coordinate-grid cues and a 192×192 preview of the object to place, shown in its canonical orientation at the same pixel scale; a white dot marks its centroid. Polygons, masks, and feasible poses are withheld.

Controlled difficulty levels. The entire target must fit inside the tray. We move its boundaries inward for Medium and Hard, keeping the canvas, solid objects, target, and instructions fixed (Figure 1). For each allowed orientation, feasible target-center sets satisfy Hard $\subseteq$ Medium $\subseteq$ Easy. Summed over the four orientations, their area is reduced to 28% and 5.5% of Easy for Medium and Hard, respectively. Boundaries are selected geometrically, without model outputs, while retaining a feasible solution. Construction and validation: Appendix B.

Placement output. From scene I_{scene}, target preview I_{target}, and textual constraints $\mathcal{C}$, the model returns JSON coordinates and orientation:

$$(I_{\mathrm{scene}}, I_{\mathrm{target}}, \mathcal{C}) \longmapsto (x, y, \theta), \qquad (x, y) \in [0, 1]^2. \tag{2}$$

Here (x, y) is the target centroid, with origin at top-left and y increasing downward. Rotation $\theta \in \{0, 90, 180, 270\}^\circ$ is clockwise from the preview; size is fixed. Instructions define centroid relations, boundary contact, and numerical edge-to-edge clearance (Appendix B.1). Figure 2 shows the prediction and verification pipeline.

3.2 IN-CONTEXT LEARNING PROTOCOL

We compare zero-, five-, and eight-shot prompting on the same 90 Tangram test scenes, 30 each for $k \in \{3, 5, 7\}$. Each demonstration pairs an image with its correct JSON answer as a user–assistant exchange before the query. Demonstrations match the query's piece count and come from training source pages disjoint from test pages, excluding differently augmented copies of test configurations. All models receive identical examples; five-shot uses the first five of the eight-shot sequence. Weights stay fixed and test answers are withheld. Audits, settings, and per-k results: Appendix C.

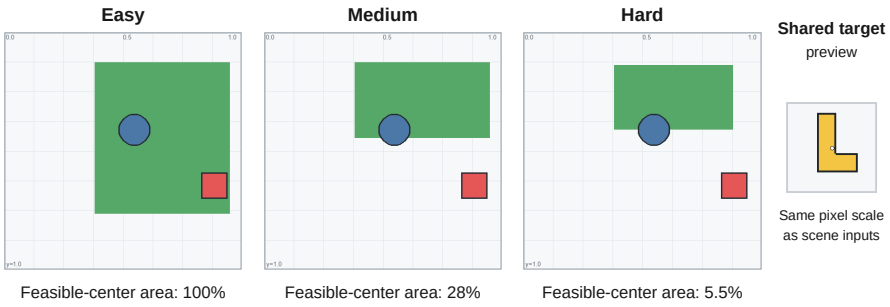


Figure 1: Matched CanvasPlace inputs. **Place the fixed-size yellow L entirely inside the green tray, with its centroid right of the blue circle, its long arm pointing up, no overlap with solid objects, and at least 12-pixel edge-to-edge clearance from the red square.** Only the tray changes across tiers. Percentages report feasible-center area relative to Easy, not tray area.

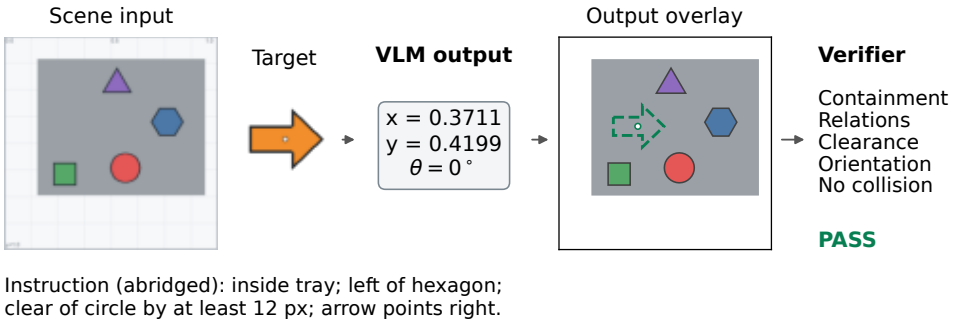


Figure 2: CanvasPlace pipeline Protocol: detailed description are provided in Appendix B.

3.3 Head Training and Language–Vision Adaptation

We use Qwen3-VL-8B-Instruct (Bai et al., 2025) and InternVL3.5-8B-HF (Wang et al., 2025a). M1 trains a randomly initialized head over frozen Base visual features; M2/M3 independently adapt Base. All share a source-page-disjoint Tangram split: 4,200 training, 350 validation, and 420 test scenes. These diagnostic alternatives are not parameter- or compute-matched. Full training settings, count-input and crop ablations, and complete results are provided in Appendix D.

M1: structured head over frozen vision. Seven learned object queries cross-attend to frozen visual features through a three-layer, eight-head, 512-dimensional transformer decoder. Direct heads predict identity (including no-object), normalized position, symmetry-aware angle vectors, log-scale, and reflection. Hungarian matching assigns predictions to targets. The training objective is

$$\mathcal{L}_{\text{head}} = \mathcal{L}_{\text{cls}} + 10\mathcal{L}_{\text{pos}} + 2\mathcal{L}_{\text{angle}} + 2\mathcal{L}_{\log s} + 0.5\mathcal{L}_{\text{flip}}. \tag{3}$$

Position and log-scale use smooth-L_1 losses; angle uses cosine distance; class and reflection use cross-entropy losses. M1 does not use LoRA or autoregressive language decoding. All heads use true k for output selection. In the main comparison, M1 additionally receives k through a learned query embedding, and InternVL uses the same image crops as M2/M3. We train five configurations with initial learning rate 10^{-4}, seed 2027, batch size eight, a validation-driven learning-rate schedule, and early stopping. Training lasts 46–63 epochs, with checkpoints selected solely by validation loss.

M2: language-only LoRA. We apply LoRA (Hu et al., 2022) to language attention query, key, value, and output projections with rank $r = 16$, scaling $\alpha = 32$, and dropout 0.05. The pretrained language weights, visual encoder, and cross-modal interface remain frozen. We compute the autoregressive training loss only over answer tokens; the full objective is provided in Appendix D.

M3: additional visual and cross-modal adaptation. M3 uses the same language LoRA and SFT objective, while directly fine-tuning the last four visual blocks and cross-modal interface: Qwen visual blocks 23–26 and its mergers; InternVL blocks 20–23 and its projector. Both M2/M3 train

for two epochs with batch size one and gradient accumulation eight. Learning rates are 5×10^{-5} for language LoRA, 10^{-5} for interface, and 2×10^{-6} for visual blocks. Checkpoints are selected by validation loss.

4 Experiments and Analysis

4.1 Evaluation Setup and Metrics

Models and evaluation subsets. We evaluate ten proprietary and open VLMs (Table 1); exact model identifiers and API sources are provided in Table 7. For Tangram zero-shot evaluation, all models are tested on the same fixed set of 200 scenes for each piece count $k \in \{1, \ldots, 7\}$, yielding 1,400 scenes per model. Adaptation experiments use the full 420-scene test split. CanvasPlace evaluates the same 200 source families, under all 3 difficulty tiers. This produces 6,000 predictions across 10 models.

API inference protocol. All zero-shot evaluations use a single-pass prediction protocol with no ground-truth poses, geometric feedback, or iterative refinement. Tangram and CanvasPlace use maximum output lengths of 4,000 and 1,200 tokens, respectively. Temperature is set to zero, For GPT-6 Astra, we omit this parameter because its API does not support custom temperature values. Across CanvasPlace tiers, model identifiers, prompts, and inference settings are held fixed. Additional inference and reasoning settings are given in Appendices A.1 and B.1.

Tangram metrics. Predicted pieces are matched to ground truth by their stated identities rather than by geometry-based reassignment. For each matched piece i, we measure

$$e_i^{\text{pos}} = \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2, \qquad e_i^{\text{ang}} = \min_{n \in \mathbb{Z}} |\hat{\theta}_i - \theta_i + nT_i|, \qquad e_i^{\text{scale}} = \frac{|\hat{s}_i - s_i|}{s_i}, \tag{4}$$

where $T_i = 360^\circ, 90^\circ, 180^\circ$ for triangles, the square (S), and the parallelogram (P), respectively; triangle reflections are handled through the effective-angle convention. Position error is measured in canvas units. We additionally report Piece IoU and union-of-pieces Scene IoU using 512×512 polygon masks. Strict *Joint success* requires exact piece set and count, correct parallelogram reflection, and, for every piece, $e^{\text{pos}} \leq 0.5$, $e^{\text{ang}} \leq 15^\circ$, and $e^{\text{scale}} \leq 0.10$.

CanvasPlace metrics. Strict success requires satisfying all applicable geometric constraints, rather than matching a single reference pose. The verifier applies the predicted transformation $(512x, 512y, \theta)$ to the canonical target polygon and evaluates full-polygon containment, centroid-based spatial relations, polygon intersections, shortest edge-to-edge distances, and symmetry-equivalent orientations. Cardinal-orientation constraints use a 1° tolerance, and boundary contact is permitted unless an explicit clearance constraint forbids it. The same strict-success criterion, prompts, parsers, and verifiers are used across Easy, Medium, and Hard; only the tray region changes.

4.2 Benchmarking Reconstruction and Constraint-Conditioned Placement

Task 1: TangramPose. Most evaluated VLMs return parseable Tangram outputs but have large rotation and scale errors (Table 1, left). GPT-6 Astra has best performance, achieveing 87.86% Joint success and 0.973 Scene IoU, with 100% Joint at $k = 1$–3 but significant lower success rate at $k = 4$–7. Other models achieve 0.215–0.416 Scene IoU but zero Joint success except Claude (4.29%). Thus, scene overlap does not ensure correct object-specific poses, and precise reconstruction remains unreliable for most evaluated models.

Task 2: constraint-conditioned placement. Mean strict success falls from 66% (Easy) to 50% (Medium) and 25% (Hard) (Table 1, right). Most models lose at least 40 % points from Easy to Hard.

Correct relations do not guarantee valid placement. From Easy to Hard, containment falls from 92.5% to 49.0% and collision-free placement from 75.0% to 53.0%, while relation accuracy (94.8% to 89.7%) and parsing (96.0% to 94.0%) change less. This pattern is consistent with greater difficulty satisfying all geometric constraints jointly. Full results and examples: Appendices B.2 and B.3.

4.3 From In-Context Learning to Head and Backbone Adaptation

In context learning improve scale estimation but still fail to improve joint success. On the 90 matched scenes, mean scale error across the nine non-GPT-6 models falls from 64.4% (0-shot) to

Table 1: **TangramPose (left) and CanvasPlace placement (right).** TangramPose uses 1,400 scenes per model (200 per $k = 1$–7). IoU and Joint average over scenes. Angle is in degrees; Scale, Joint, and placement success are percentages. Joint requires the exact identity set and all per-piece pose criteria (Section 4.1). CanvasPlace uses 200 source families across 3 difficulty tiers; Δ gives consecutive percentage-point changes. Full results and examples: Appendices A and B, respectively.

	Task 1 TangramPose					Task 2 CanvasPlace Placement				
Model	Position error ↓	Angle error (°) ↓	Relative scale error (%) ↓	Scene IoU ↑	Joint ↑	Easy	Medium	Hard	$\Delta_{E\to M}$	$\Delta_{M\to H}$
GPT-6 Astra	0.011	2.26	0.71	0.973	87.86	100%	100%	90%	+0	-10
GPT-5.6 Luna	0.580	75.23	71.65	0.381	0.00	60%	45%	15%	-15	-30
GPT-4o	1.474	73.06	74.30	0.289	0.00	45%	20%	5%	-25	-15
Gemini 3.1 Pro	1.374	64.39	55.46	0.395	0.00	95%	85%	45%	-10	-40
Gemini Flash Lite	53.359	68.52	63.94	0.347	0.00	80%	45%	20%	-35	-25
Claude Sonnet 4.6	0.835	69.81	50.27	0.416	4.29	85%	75%	40%	-10	-35
Qwen3-VL-8B	2.825	78.74	46.30	0.215	0.00	20%	10%	0%	-10	-10
Qwen3-VL-32B	1.524	74.84	56.13	0.276	0.00	50%	25%	5%	-25	-20
Gemma-4-31B	0.817	70.62	80.16	0.282	0.00	85%	75%	30%	-10	-45
Llama-4 Maverick	1.555	71.54	58.47	0.256	0.00	40%	20%	0%	-20	-20
Mean	6.435	64.90	55.74	0.383	9.21	66%	50%	25%	-16	-25

Table 2: **In-context learning on Tangram.** Each cell reports 0-/5-/8-shot results, each pooled over the same scenes. Scale and Joint are percentages; angle is in degrees. Per-k results: Appendix C.

	0 / 5 / 8 shots			
Model	Relative scale error (%) ↓	Angle error (°) ↓	Scene IoU ↑	Joint ↑
GPT-6 Astra	0.7 / 0.8 / 0.8	2.6 / 2.8 / 2.4	0.972 / 0.969 / 0.969	85.56 / 85.56 / 85.56
GPT-5.6 Luna	72.6 / 22.1 / 20.1	76.2 / 74.6 / 74.3	0.385 / 0.469 / 0.466	0.00 / 0.00 / 0.00
GPT-4o	77.5 / 46.2 / 45.4	72.8 / 74.4 / 74.7	0.310 / 0.333 / 0.327	0.00 / 0.00 / 0.00
Gemini 3.1 Pro	57.0 / 14.6 / 13.7	65.7 / 66.7 / 69.6	0.388 / 0.499 / 0.485	0.00 / 0.00 / 0.00
Gemini Flash Lite	66.4 / 22.2 / 20.6	70.4 / 70.8 / 72.0	0.385 / 0.449 / 0.446	0.00 / 0.00 / 0.00
Claude Sonnet 4.6	54.9 / 18.2 / 15.2	71.1 / 66.4 / 68.3	0.394 / 0.425 / 0.439	0.00 / 0.00 / 0.00
Qwen3-VL-8B	46.4 / 32.0 / 28.9	79.8 / 78.5 / 74.5	0.257 / 0.277 / 0.265	0.00 / 0.00 / 0.00
Qwen3-VL-32B	60.0 / 23.2 / 22.1	75.9 / 71.6 / 72.3	0.312 / 0.283 / 0.308	0.00 / 0.00 / 0.00
Gemma-4-31B	86.6 / 23.7 / 22.2	68.9 / 69.6 / 69.1	0.280 / 0.461 / 0.475	0.00 / 0.00 / 0.00
Llama-4 Maverick	58.6 / 25.3 / 23.0	69.2 / 73.9 / 70.6	0.290 / 0.321 / 0.328	0.00 / 0.00 / 0.00

25.3%/23.5% (5-/8-shot); mean Scene IoU rises from 0.333 to 0.391/0.393, improving in eight of nine models (Table 2). Mean angle error changes little (72.2° to 71.8°/71.7°), and their Joint success remains zero. GPT-6 maintains 85.56% Joint across all three conditions.

Table 3: **Tangram head and LoRA adaptation Evaluation** M1 is a geometry head trained with early stopping and count input; M2 adapts language processing; M3 adapts language, late visual and cross-modal components. Full detail results: Appendix D.

Architecture	Model	Position error ↓	Angle error (°) ↓	Relative scale error (%) ↓	Scene IoU ↑	Joint ↑
Qwen	Base	2.932	76.19	40.62	0.210	0.00
Qwen	M1	0.568	5.15	5.08	0.575	23.10
Qwen	M2	0.149	16.94	4.05	0.829	40.71
Qwen	M3	0.141	8.64	2.78	0.862	52.38
InternVL	Base	3.174	73.64	60.98	0.235	0.00
InternVL	M1	0.620	7.26	6.29	0.532	15.48
InternVL	M2	0.145	38.70	3.73	0.762	23.33
InternVL	M3	0.081	23.69	3.24	0.823	37.86

Direct heads recover rotation accurately but lag in joint reconstruction. M1 has lower angle error but lower Scene IoU and Joint than M2/M3 on both architectures (Table 3). InternVL M1, for

example, achieves 7.26° angle error and 15.48% Joint, versus 23.69° and 37.86% for M3. Positional accuracy is a prominent remaining gap: 0.620 versus 0.081 in this comparison.

Language-model adaptation improves joint reconstruction. M2 trains only language LoRA adapters, keeping the pretrained language weights, visual encoder, and cross-modal interface frozen. Joint rises from 0% to 40.71% on Qwen and 23.33% on InternVL (Table 3), showing that fixed visual features support better downstream prediction. M3 additionally adapts late visual and cross-modal components, reaching 52.38% and 37.86%, respectively. Both outperform the tested M1 heads in Joint, but at $k = 7$, M3 achieves only 13.33% on Qwen and 8.33% on InternVL. Figure 3 shows matched five-piece reconstructions from the checkpoints in Table 3.

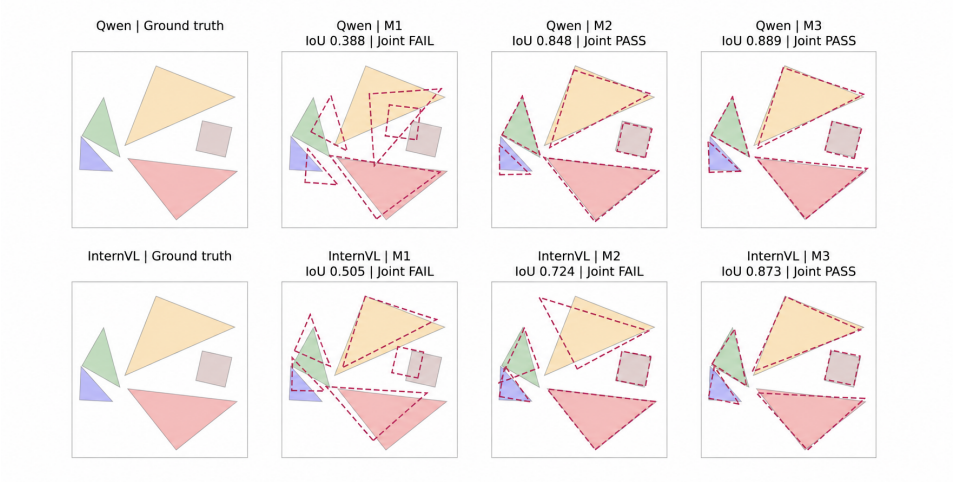


Figure 3: Matched Tangram reconstructions: Filled polygons are ground truth; dashed outlines are predictions for LoRA and Head configurations.

4.4 Generalization Across Angles, Piece Counts, and Tasks

We test whether the gains from Tangram adaptation extend beyond the training distribution along three increasingly broad shifts: unseen rotation ranges, unseen composition lengths, and zero-shot transfer to CanvasPlace.

Table 4: **Generalization and transfer.** Angle-shift rows use separate Qwen checkpoints and exclude squares from angle statistics; piece-count rows use symmetry-aware angle errors. Zero-shot transfer uses 50 CanvasPlace tasks with axes, rotation direction, and coordinate range aligned to Tangram. InternVL Base has a 98% parse-failure rate. Transfer conditions: Appendix E.3.

Evaluation	Model	Scene IoU	Angle (deg.)	Joint / strict success
Angle ID control	Qwen M3	0.882	1.63	70.24% Joint
Angle OOD	Qwen M3	0.588	101.05	3.33% Joint
$k = 1$–$5 \rightarrow 6$–7	Qwen M2	0.745	36.12	0.83% Joint
$k = 1$–$5 \rightarrow 6$–7	Qwen M3	0.802	19.90	6.67% Joint
Aligned transfer	Qwen Base	na	na	36.00% strict
Aligned transfer	Qwen M2	na	na	26.00% strict
Aligned transfer	Qwen M3	na	na	34.00% strict
Aligned transfer	InternVL Base	na	na	2.00% strict
Aligned transfer	InternVL M2	na	na	40.00% strict
Aligned transfer	InternVL M3	na	na	48.00% strict

Rotation accuracy degrades sharply outside the training angle range. We train separate Qwen checkpoints on rotations in approximately $[-179, -91]^\circ$ and evaluate them on both a matched in-support split and the disjoint interval $[91, 179]^\circ$, using source-page-disjoint data. For M3, angle error increases from 1.63° in-distribution to 101.05° under the angle shift, while Scene IoU decreases

from 0.882 to 0.588 and Joint success from 70.24% to 3.33% (Table 4). Thus, accurate in-support reconstruction does not imply reliable generalization to unseen rotation ranges. Full results and symmetry conventions are provided in Appendix E.1.

Approximate scene geometry transfers better than exact reconstruction to unseen composition lengths. Qwen M2 and M3 are trained only on scenes with $k = 1$–5 and evaluated on 120 scenes with unseen $k = 6, 7$. They retain Scene IoU of 0.745 and 0.802, respectively, but achieve only 0.83% and 6.67% Joint success. This gap indicates that the models can preserve substantial scene-level geometry and extend their outputs to longer compositions, while exact object-specific pose recovery remains unreliable. Because Joint success compounds errors across pieces, we additionally report per-piece errors in Appendix E.2.

Tangram adaptation does not yield consistent zero-shot transfer to CanvasPlace. Tangram and CanvasPlace use different coordinate and rotation conventions, so we evaluate Base, M2, and M3 on the same 50 CanvasPlace tasks under progressively aligned input conventions: native inputs, redrawn native labels, aligned axes and rotation direction, and finally an aligned coordinate range of $[0, 10]$. Physical constraints and allowed orientations remain unchanged, and all predictions are mapped to the same evaluator (Appendix E.3). Under full alignment, Qwen achieves 36%, 26%, and 34% strict success for Base, M2, and M3, respectively; the M3–Base difference is -2 percentage points with a paired-task bootstrap 95% interval of $[-18, 14]$. Thus, convention mismatch alone does not explain the absence of a transfer gain. InternVL improves from 2% to 40%/48%, but its Base model has 49/50 parse failures, so this improvement includes better output validity and cannot be attributed solely to geometric transfer. **Overall, strong in-distribution adaptation does not imply robust geometric generalization.** Performance degrades under unseen rotations and longer compositions, while gains on Tangram do not consistently transfer to the distinct CanvasPlace task.

4.5 Decodable Geometry information and its use for object specific prediction

Probing protocol. We test whether angle and scale can be decoded from internal VLM representations, and how this decodability changes after adaptation. Linear and two-layer MLP probes are trained on layer-wise representations using 840/210/420 train/validation/test scenes. Ground-truth polygon masks are used to pool tokens from the target piece, and piece identity is provided to the probe. Pose values are used only as supervision during probe training.

Table 5: Geometry decodability versus autonomous prediction. MLP-probe errors are computed from ground-truth target regions and identities, while final prediction errors are from the model's own JSON outputs. Visual probe layers are selected separately for angle and scale using validation error and evaluated on the held-out test set; the selected visual-layer index is shown in parentheses. Language probes use the fixed final LLM layer (35). Angle is reported in degrees and scale as relative error (%). Full layer-wise results and linear probes are provided in Appendix F.2.

Architecture	Model	Visual probe error		LLM final layer 35 probe error		Final prediction error	
		Angle (°)	Scale (%)	Angle (°)	Scale (%)	Angle (°)	Scale (%)
Qwen	Base	9.77 (v8)	2.92 (v12)	38.51	6.32	76.19	40.62
Qwen	M2	9.77 (v8)	2.92 (v12)	31.16	5.63	16.94	4.05
Qwen	M3	9.77 (v8)	2.92 (v12)	22.18	5.28	8.64	2.78
InternVL	Base	11.26 (v12)	3.89 (v12)	40.89	5.16	73.64	60.98
InternVL	M2	11.26 (v12)	3.89 (v12)	36.80	5.37	38.70	3.73
InternVL	M3	11.26 (v12)	3.89 (v12)	32.31	4.95	23.69	3.24

Geometry in features and final predictions. Table 5 shows that Base visual representations support substantially more accurate angle and scale decoding than autonomous outputs. For Qwen, visual-probe errors are $9.77°$ and 2.92%, compared with $76.19°$ and 40.62% in Base predictions; InternVL shows the same pattern ($11.26°/3.89\%$ versus $73.64°/60.98\%$). M2 substantially improves final predictions with unchanged visual features, indicating that Base failure cannot be explained solely by missing geometric information. At the final language layer, angle-probe error decreases from Base to M2 to M3 in both architectures, suggesting that adaptation makes target geometry more recoverable in

later representations. These assisted results support a gap between geometric information decodable from internal features and its reliable use in accurate object-specific prediction.

We also evaluate **Blank-image controls**, replacing the image with a blank canvas while retaining target masks and identities, sharply worsening angle and scale decoding (Appendix F.3). **Single-object** edits show more selective responses to the edited object after adaptation, especially for position and scale (Appendix F.4). **Target isolation** leaves both Base models at zero target-pose success, showing that removing other objects alone is insufficient under this protocol (Appendix F.5).

4.6 Localizing Layers That Influence Geometric Predictions

We use target-region activation patching to identify which visual and language layers influence geometric predictions. Each of 108 image pairs changes one object's position, angle, or scale (36 pairs per attribute). At one site at a time, we replace target-region activations in the changed-image run with those from the original (clean) image. Visual patches cover the union of the target's original and changed polygons; language patches replace corresponding image-token states during prefill. The same pairs are tested on Base/M2/M3 at 21 Qwen sites and 15 InternVL sites.

Measuring recovery toward the original target. Let e_{d} and e_{p} be target errors before and after patching, both measured against the original image's ground truth:

$$\Delta e = e_{\mathrm{d}} - e_{\mathrm{p}}. \tag{5}$$

Positive Δe indicates that patching moves the prediction closer to the original (clean) target (improving). Table 6 reports mean reductions from source-page bootstrap resampling. Brackets give 95% confidence intervals for the mean effect, not ranges of individual effects. Intervals entirely above or below zero support recovery toward or movement away from the original target, respectively. Intervals crossing zero leave the effect direction uncertain and do not establish absence of an effect.

Table 6: Target-region activation patching at representative visual and language sites. Each cell reports mean error reduction Δe with a 95% source-page bootstrap confidence interval. Units are canvas units, degrees Full layer sweeps are provided in Appendix G.

Checkpoint	Site	Position	Angle	Scale
Qwen Base	`vision.8`	0.122 [-0.037, 0.298]	-5.2 [-12.0, -0.5]	-0.4 [-4.3, 3.4]
Qwen M2	`vision.8`	0.423 [0.394, 0.452]	9.8 [-2.8, 20.7]	15.6 [13.9, 17.3]
Qwen M3	`vision.8`	0.486 [0.457, 0.514]	25.6 [19.2, 29.4]	18.8 [17.4, 20.4]
Qwen Base	`llm.12`	0.039 [-0.071, 0.142]	-0.6 [-2.7, 1.3]	-0.2 [-7.2, 7.9]
Qwen M2	`llm.12`	0.354 [0.328, 0.378]	1.4 [-10.5, 12.3]	14.2 [12.4, 16.3]
Qwen M3	`llm.12`	0.433 [0.400, 0.465]	21.6 [18.4, 24.4]	18.4 [16.7, 20.1]
InternVL Base	`vision.12`	0.052 [-0.137, 0.231]	-0.7 [-7.2, 4.4]	-4.3 [-13.4, 3.0]
InternVL M2	`vision.12`	0.502 [0.487, 0.517]	-7.7 [-30.8, 15.1]	15.0 [13.2, 16.6]
InternVL M3	`vision.12`	0.524 [0.505, 0.539]	5.3 [-9.8, 23.1]	15.9 [13.8, 17.9]
InternVL Base	`llm.18`	-0.020 [-0.164, 0.118]	-1.7 [-4.5, 0.7]	-1.7 [-4.9, 0.0]
InternVL M2	`llm.18`	0.446 [0.416, 0.468]	5.3 [-14.6, 26.4]	14.3 [12.9, 15.7]
InternVL M3	`llm.18`	0.454 [0.422, 0.478]	6.6 [-12.5, 25.7]	14.7 [13.2, 16.2]

Activation patching identifies visual and language stages that influence geometric outputs. In both architectures, M2/M3 show recovery of position and scale across multiple visual and early/middle language sites. Qwen M3 shows particularly large angle recovery at visual layers 4/8 and language layers 6/12/18, from table 6. These results identify stages where target-region interventions affect predictions, rather than a uniquely necessary inference layer.

5 Conclusion

We introduced a controlled benchmark for continuous geometric reasoning with TangramPose and constraint-conditioned CanvasPlace placement. Most evaluated VLMs achieve coarse spatial correctness but remain unreliable in exact object-specific geometry. Fine-tuning substantially improves in-distribution reconstruction, even with frozen visual features, yet these gains degrade under unseen rotations, larger compositions, and cross-task transfer. Probing, counterfactuals, and activation patching further show that geometric information can be recoverable from internal representations without being reliably converted into the correct object-specific output. These results highlight a gap between in-distribution geometric accuracy and robust, reusable geometric generalization.

Ethics Statement

Our benchmark evaluates precise geometric reconstruction and constraint-conditioned placement using synthetic two-dimensional scenes and automated geometric verification. The study does not involve human-subject experiments or collect personally identifiable information. Its purpose is to characterize model capabilities and limitations, rather than establish readiness for deployment in robotics or other safety-critical applications. We plan to release benchmark resources and evaluation code to support reproducibility, while respecting the attribution, licensing, and redistribution requirements of the underlying source materials.

AI Use Statement

Vision-language models are used as experimental subjects to generate predictions for evaluation and analysis, as described in the methodology. Separately, AI assistants supported manuscript drafting, language revision, translation, literature search, reference organization, and LaTeX formatting, as well as discussions of experimental design, hypothesis refinement, and result interpretation. The authors retain full responsibility for the verification of AI-assisted materials and for the accuracy, originality, and integrity of the final manuscript and reported findings.

References

Ahmed Abdelreheem, Filippo Aleotti, Jamie Watson, Zawar Qureshi, Abdelrahman Eldesokey, Peter Wonka, Gabriel Brostow, Sara Vicente, and Guillermo Garcia-Hernando. PlaceIt3D: Language-Guided Object Placement in Real 3D Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6645–6655, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Abdelreheem_PlaceIt3D_Language-Guided_Object_Placement_in_Real_3D_Scenes_ICCV_2025_paper.html.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Anthropic. System Card: Claude Sonnet 4.6. Official system card, February 2026. URL https://anthropic.com/claude-sonnet-4-6-system-card. Published February 17, 2026; revised March 6, 2026. Accessed September 16, 2026.

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*, 2025. URL https://arxiv.org/abs/2511.21631.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

Gemma Team. Gemma 4 Technical Report. arXiv preprint arXiv:2607.02770, 2026. URL https://arxiv.org/abs/2607.02770.

Nikolaos Gkanatsios, Ayush Jain, Zhou Xian, Yunchu Zhang, Christopher G. Atkeson, and Katerina Fragkiadaki. Energy-based Models are Zero-Shot Planners for Compositional Scene Rearrangement. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, 2023. doi: 10.15607/RSS.2023.XIX.030. URL https://www.roboticsproceedings.org/rss19/p030.html.

Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. ARNOLD: A Benchmark for Language-Grounded Task Learning with Continuous States in Realistic 3D Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20483–20495, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Gong_ARNOLD_A_Benchmark_for_Language-Grounded_Task_Learning_with_Continuous_States_ICCV_2023_paper.html.

Google DeepMind. Gemini 3.1 Flash-Lite: Model Card. Official model card, March 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/. Published March 3, 2026. Accessed September 16, 2026.

Google DeepMind. Gemini 3.1 Pro: Model Card. Official model card, February 2026b. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Published February 19, 2026. Accessed September 16, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract Visual Reasoning with Tangram Shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 582–601. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.38. URL https://aclanthology.org/2022.emnlp-main.38/.

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://openreview.net/forum?id=6nZKT2rL0H.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Johnson_CLEVR_A_Diagnostic_CVPR_2017_paper.html.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/96810b6d4752abe7bfb91f234c51e9e6-Abstract-Conference.html.

Daixian Liu, Jiayi Kuang, Yinghui Li, Yangning Li, Di Yin, Haoyu Cao, Xing Sun, Ying Shen, Hai-Tao Zheng, Liang Lin, and Philip S. Yu. TangramPuzzle: Evaluating Multimodal Large Language Models with Compositional Spatial Reasoning. arXiv preprint arXiv:2601.16520, 2026. URL https://arxiv.org/abs/2601.16520.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Meta. Llama 4 Model Card. Official model card, April 2025. URL https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md. Documents Llama 4 Scout and Maverick, released April 5, 2025. Accessed September 16, 2026.

OpenAI. GPT-4o Model. OpenAI API model documentation, n.d.a. URL https://developers.openai.com/api/docs/models/gpt-4o. Includes the gpt-4o-2024-11-20 snapshot. Accessed September 16, 2026.

OpenAI. GPT-5.6 Luna Model. OpenAI API model documentation, n.d.b. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed September 16, 2026.

OpenAI. GPT-6 Astra Model. OpenAI API model documentation, n.d.c. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 16, 2026.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2856–2861, 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/CLVL/html/Palit_Towards_Vision-Language_Mechanistic_Interpretability_A_Causal_Tracing_Tool_for_BLIP_ICCVW_2023_paper.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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_RoboSpatial_Teaching_Spatial_Understanding_to_2D_and_3D_Vision-Language_Models_CVPR_2025_paper.html.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

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*, 2025a. URL https://arxiv.org/abs/2508.18265.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A Diagnostic Benchmark for 6D Spatial Reasoning of Large Multimodal

Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24669–24679, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Spatial457_A_Diagnostic_Benchmark_for_6D_Spatial_Reasoning_of_Large_CVPR_2025_paper.html.

Yiqing Xu, Jiayuan Mao, Yilun Du, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and David Hsu. Set It Up!: Functional Object Arrangement with Compositional Generative Models. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024. doi: 10.15607/RSS.2024.XX.095. URL https://www.roboticsproceedings.org/rss20/p095.html.

Fernanda Miyuki Yamada, João Paulo Gois, and Harlen Costa Batagelo. Solving Tangram Puzzles Using Raster-Based Mathematical Morphology. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 116–123. IEEE, 2019. doi: 10.1109/SIBGRAPI.2019.00024. URL https://doi.org/10.1109/SIBGRAPI.2019.00024.

Fernanda Miyuki Yamada, Harlen Costa Batagelo, João Paulo Gois, and Hiroki Takahashi. Generative approaches for solving tangram puzzles. *Discover Artificial Intelligence*, 4(1), 2024. doi: 10.1007/s44163-024-00107-6. URL https://link.springer.com/article/10.1007/s44163-024-00107-6. Article 12.

Chao Zhang, Jiamin Tang, and Jing Xiao. Tangram: Benchmark for Evaluating Geometric Element Recognition in Large Multimodal Models. arXiv preprint arXiv:2408.13854, 2024. URL https://arxiv.org/abs/2408.13854.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19781–19791, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

Chao Zhao, Chunli Jiang, Lifan Luo, Guanlan Zhang, Hongyu Yu, Michael Yu Wang, and Qifeng Chen. Master Rules from Chaos: Learning to Reason, Plan, and Interact from Chaos for Tangram Assembly. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11503–11509. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127610. URL https://doi.org/10.1109/ICRA55743.2025.11127610.

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. VLMbench: A Compositional Benchmark for Vision-and-Language Manipulation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 665–678. Curran Associates, Inc., 2022. doi: 10.52202/068431-0048. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/04543a88eae2683133c1acbef5a6bf77-Abstract-Datasets_and_Benchmarks.html.

Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. CoSpace: Benchmarking Continuous Space Perception Ability for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29569–29579, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhu_CoSpace_Benchmarking_Continuous_Space_Perception_Ability_for_Vision-Language_Models_CVPR_2025_paper.html.

Yikun Zong and Cheston Tan. TangramSR: Can Vision-Language Models Reason in Continuous Geometric Space? arXiv preprint arXiv:2602.05570, 2026. URL https://arxiv.org/abs/2602.05570.

A Complete Tangram API Results by Piece Count

A.1 Construction, Inference, and Scoring Protocol

End-to-end workflow. The pipeline is (i) partition source pages, (ii) choose a subset and perturb its geometry, (iii) fit and render the configuration while saving exact poses, (iv) request a joint JSON prediction from the image, and (v) parse, identity-match, reconstruct polygons, and score offline. Numerical test annotations are never included in the benchmark prompt. The fixed color key distinguishes the two geometrically identical large triangles and the two small triangles; the task does not require identifying these instances from shape alone. The supplied k also means this is not a blind object-counting benchmark.

Augmentation details. Source-page order and subset selection are seeded. Each piece initially receives a scale multiplier drawn log-uniformly from $[0.72, 1.28]$. If $k > 1$ and the multiplier range is below 0.28, two selected multipliers are replaced by draws from $[0.70, 0.79]$ and $[1.20, 1.30]$, respectively. A shared fitting factor subsequently changes the absolute scales; $[0.70, 1.30]$ therefore describes perturbation multipliers, not a bound on final target scale. Source centroids are rotated about their mean, while each piece orientation receives additional independent jitter. P reflection is sampled independently. The joint configuration is scaled to fit and uniformly translated within the remaining feasible intervals, keeping every vertex between 0.3 and 9.7 on both axes. No collision rejection is imposed, so overlap and partial occlusion can occur. The renderer uses four light backgrounds and outline widths from $\{2, 2, 3, 4\}$ pixels. Stored pose annotations retain more precision than the supervised JSON answers, which round position to two decimals, angle to one, and scale to three. Evaluation uses the stored geometric annotations.

Exact shared prompt. The following is the Tangram-v2 prompt; {k} is replaced by the known piece count. Line wrapping below is typographic only. The single JSON object at the end specifies the schema rather than providing the scene's answer.

```
You are given a square 2D Tangram canvas. The full image boundary is
the coordinate frame: x increases left-to-right from 0 to 10 and y increases
bottom-to-top from 0 to 10. There are no printed IDs, axes, ticks, or grid.
There are exactly {k} pieces. Use this fixed color key:
T1=red, T2=orange, T3=green, T4=blue, T5=purple, S=brown, P=pink.

Recover every piece independently:
- pos=[x,y] is its polygon centroid in canvas coordinates;
- angle is counter-clockwise in continuous degrees;
- scale is relative to the canonical piece template;
- triangles use flip=false and absorb reflection into angle;
- square uses flip=false and its canonical angle is modulo 90 degrees;
- parallelogram P uses angle modulo 180 degrees and has a meaningful flip.

Canonical templates before centering at their vertex mean:
- T1,T2: (0,0),(5.6569,0),(0,5.6569)
- T3: (0,0),(4,0),(0,4)
- T4,T5: (0,0),(2.8284,0),(0,2.8284)
- S: (0,0),(2.8284,0),(2.8284,2.8284),(0,2.8284)
- P: (0,0),(4,0),(6,2),(2,2)

Return strict JSON only, with exactly {k} objects and no prose:
{"pieces":[{"id":"T1","pos":[5.0,5.0],"angle":12.3,
"scale":0.72,"flip":false}]}
```

API requests and retained responses. The benchmark runner sends the image followed by the prompt through an OpenAI-compatible interface, requests a JSON object, and sets `max_tokens=4000`. Temperature is zero except for GPT-6, where it is not sent. Reasoning effort is low for GPT-6 and Gemini Pro, none for GPT-5.6 Luna, and minimal for Gemini Flash-Lite; Claude's reasoning is disabled. These five requests exclude reasoning text from the returned answer. No reasoning override is sent for the remaining models. These are requested settings, not a claim of equal internal compute across providers. The runner allows transport retries and resumption of API/parsing failures; repeated records are reduced to the latest saved response per sample, never the response with the best geometric score. Final predictions, provider information, usage, and per-sample metrics are retained. The exact model identifiers and references appear in Table 7.

Parser and identity accounting. Although the prompt requests strict JSON, the scorer can extract a JSON value from a fenced or prose-wrapped response and accepts a `pieces` object or a top-level list. It normalizes supported identity aliases and rejects non-finite numerical fields. A valid identity and two-component position are required; omitted angle, scale, and flip fields default to zero, one, and false, respectively. Malformed piece records are discarded and flagged. If an ID is duplicated, the first accepted record is used for attribute and mask calculations, but the scene fails exact-ID accuracy and Joint. Missing or unexpected IDs likewise fail Joint. Parsing rate indicates successful extraction of a pieces list, not that every entry is valid. Piece recall is the total number of matched IDs divided by the total number of ground-truth pieces.

Scoring conventions. For triangles, the effective angle is $(\theta + 90^\circ f) \bmod 360^\circ$; S and P use angle modulo 90° and 180°. Reflection correctness is scored separately only for P. Numerical positions are not clamped to $[0, 10]^2$, so an out-of-range prediction can incur a large position error. Raster masks cover the fixed canvas and thus clip out-of-frame geometry; non-positive-scale predicted polygons are omitted from masks. Piece IoU uses the complete polygon for each matched ID, not only its visible part; Scene IoU compares binary unions and does not penalize incorrect internal piece boundaries when the union is preserved. The IoU≥ 0.8 column in Appendix A.3 reports the percentage of retained scenes whose Scene IoU is at least 0.8. This threshold is reported separately and is not part of the Joint criterion. All retained scenes, including parsing failures, contribute to scene-level means. Matched-piece means exclude missing objects, which is why identity/recall and Joint must accompany the continuous errors. None of these rules use a geometric matching oracle or post-hoc pose correction.

A.2 Model Identifiers

Table 7: Evaluated API models: short names, exact request identifiers, and primary references. The same ten identifiers are used in both task benchmarks.

Name	API identifier	Reference
GPT-6 Astra	`openai/gpt-6-astra`	(OpenAI, n.d.c)
GPT-5.6 Luna	`openai/gpt-5.6-luna`	(OpenAI, n.d.b)
GPT-4o	`openai/gpt-4o-2024-11-20`	(OpenAI, n.d.a)
Gemini 3.1 Pro	`google/gemini-3.1-pro-preview`	(Google DeepMind, 2026b)
Gemini Flash Lite	`google/gemini-3.1-flash-lite`	(Google DeepMind, 2026a)
Claude Sonnet 4.6	`anthropic/claude-sonnet-4.6`	(Anthropic, 2026)
Qwen3-VL-8B	`qwen/qwen3-vl-8b-instruct`	(Bai et al., 2025)
Qwen3-VL-32B	`qwen/qwen3-vl-32b-instruct`	(Bai et al., 2025)
Gemma-4-31B	`google/gemma-4-31b-it`	(Gemma Team, 2026)
Llama-4 Maverick	`meta-llama/llama-4-maverick`	(Meta, 2025)

For controlled adaptation and mechanism experiments, the checkpoints are `Qwen/Qwen3-VL-8B-Instruct` (Bai et al., 2025) and `OpenGVLab/InternVL3_5-8B-HF` (Wang et al., 2025a). InternVL is an additional diagnostic architecture, not an eleventh entry in the ten-model API comparison. Official model/system cards and API documentation are identified as such in the bibliography rather than presented as research papers. Undated documentation uses "n.d." with an explicit access date.

A.3 Complete Results by Piece Count

The Tangram part of Table 1 in the main paper is an aggregate over all $k = 1, \ldots, 7$ evaluation scenes; it is not a single-complexity result. The consolidated table below reports every continuous-pose metric separately for every evaluated model and piece count. Each cell is computed from 20 scenes at that value of k. Position, angle, and scale are piece-level errors over matched identities. Piece IoU averages matched-piece IoU, whereas Scene IoU measures the union of all predicted and ground-truth pieces. Joint is the strict scene-level criterion used in the main text.

Table 8: Complete Tangram API results by piece count (20 scenes per model and k). Position is in canvas units; angle is in degrees; Scale, IoU$\geq$ .8, and Joint are percentages.

k	Model	Pos.↓	Angle↓	Scale↓	Piece IoU↑	Scene IoU↑	IoU$\geq$ .8	Joint↑
1	GPT-6 Astra	0.005	0.08	0.41	0.988	0.988	100.0%	100.0%
1	GPT-5.6 Luna	0.414	70.81	34.96	0.407	0.400	0.0%	0.0%
1	GPT-4o	0.947	84.13	27.26	0.315	0.315	0.0%	0.0%
1	Gemini 3.1 Pro	1.023	63.11	28.46	0.367	0.367	0.0%	0.0%
1	Gemini Flash Lite	1.032	64.27	34.17	0.280	0.280	0.0%	0.0%
1	Claude Sonnet 4.6	0.408	51.31	14.01	0.549	0.549	15.0%	30.0%
1	Qwen3-VL-8B	2.744	87.98	32.56	0.128	0.122	0.0%	0.0%
1	Qwen3-VL-32B	1.636	87.06	28.77	0.237	0.237	0.0%	0.0%
1	Gemma-4-31B	0.884	71.24	50.63	0.337	0.337	0.0%	0.0%
1	Llama-4 Maverick	1.725	85.11	28.97	0.193	0.186	0.0%	0.0%
2	GPT-6 Astra	0.005	0.13	0.45	0.983	0.986	100.0%	100.0%
2	GPT-5.6 Luna	0.590	70.67	61.01	0.321	0.346	0.0%	0.0%
2	GPT-4o	1.542	74.20	43.30	0.162	0.203	0.0%	0.0%
2	Gemini 3.1 Pro	0.656	53.21	48.99	0.413	0.447	0.0%	0.0%
2	Gemini Flash Lite	21.799	63.07	52.14	0.271	0.299	0.0%	0.0%
2	Claude Sonnet 4.6	0.716	53.97	36.46	0.360	0.400	5.0%	0.0%
2	Qwen3-VL-8B	3.284	65.32	50.28	0.069	0.107	0.0%	0.0%
2	Qwen3-VL-32B	1.931	70.05	51.39	0.124	0.186	0.0%	0.0%
2	Gemma-4-31B	0.829	68.30	76.78	0.280	0.314	0.0%	0.0%
2	Llama-4 Maverick	1.603	62.31	52.36	0.155	0.171	0.0%	0.0%
3	GPT-6 Astra	0.005	0.16	0.41	0.983	0.985	100.0%	100.0%
3	GPT-5.6 Luna	0.509	77.40	53.92	0.336	0.374	0.0%	0.0%
3	GPT-4o	1.423	66.30	49.07	0.191	0.261	0.0%	0.0%
3	Gemini 3.1 Pro	0.994	64.63	55.06	0.316	0.384	0.0%	0.0%
3	Gemini Flash Lite	39.247	76.48	49.37	0.325	0.363	0.0%	0.0%
3	Claude Sonnet 4.6	0.852	63.13	35.47	0.344	0.375	0.0%	0.0%
3	Qwen3-VL-8B	2.732	88.33	40.81	0.102	0.201	0.0%	0.0%
3	Qwen3-VL-32B	1.415	79.00	51.19	0.172	0.258	0.0%	0.0%
3	Gemma-4-31B	0.773	71.65	80.98	0.259	0.317	0.0%	0.0%
3	Llama-4 Maverick	1.529	63.22	46.03	0.183	0.254	0.0%	0.0%
4	GPT-6 Astra	0.009	1.99	0.64	0.961	0.956	90.0%	90.0%
4	GPT-5.6 Luna	0.533	79.14	72.36	0.296	0.377	0.0%	0.0%
4	GPT-4o	1.511	78.78	73.43	0.146	0.265	0.0%	0.0%
4	Gemini 3.1 Pro	1.309	65.38	60.77	0.253	0.340	0.0%	0.0%
4	Gemini Flash Lite	0.648	75.90	59.38	0.295	0.383	0.0%	0.0%
4	Claude Sonnet 4.6	0.757	83.79	45.71	0.292	0.371	0.0%	0.0%
4	Qwen3-VL-8B	3.264	86.72	56.15	0.072	0.220	0.0%	0.0%
4	Qwen3-VL-32B	1.647	84.89	49.55	0.123	0.250	0.0%	0.0%
4	Gemma-4-31B	0.556	78.98	69.20	0.296	0.235	0.0%	0.0%
4	Llama-4 Maverick	1.556	83.28	64.90	0.160	0.244	0.0%	0.0%
5	GPT-6 Astra	0.011	1.32	0.63	0.965	0.972	100.0%	85.0%
5	GPT-5.6 Luna	0.670	87.05	78.54	0.247	0.337	0.0%	0.0%
5	GPT-4o	1.284	66.92	79.63	0.179	0.297	0.0%	0.0%
5	Gemini 3.1 Pro	1.377	67.96	51.81	0.296	0.389	0.0%	0.0%
5	Gemini Flash Lite	150.301	71.36	71.80	0.239	0.330	0.0%	0.0%
5	Claude Sonnet 4.6	0.844	68.64	46.16	0.292	0.377	0.0%	0.0%
5	Qwen3-VL-8B	2.655	81.51	49.24	0.092	0.248	0.0%	0.0%
5	Qwen3-VL-32B	1.485	73.91	60.48	0.141	0.299	0.0%	0.0%
5	Gemma-4-31B	0.702	64.39	80.50	0.278	0.191	0.0%	0.0%
5	Llama-4 Maverick	1.440	68.73	64.87	0.169	0.305	0.0%	0.0%
6	GPT-6 Astra	0.017	5.23	0.96	0.946	0.954	95.0%	60.0%
6	GPT-5.6 Luna	0.641	74.18	79.26	0.277	0.394	0.0%	0.0%
6	GPT-4o	1.535	72.57	91.43	0.151	0.308	0.0%	0.0%
6	Gemini 3.1 Pro	1.314	67.30	57.54	0.252	0.425	0.0%	0.0%
6	Gemini Flash Lite	95.401	66.66	76.32	0.221	0.329	0.0%	0.0%

Continued on next page

Table 8 – continued

k	Model	Pos.↓	Angle↓	Scale↓	Piece IoU↑	Scene IoU↑	IoU≥ .8	Joint↑
6	Claude Sonnet 4.6	0.890	66.41	59.08	0.259	0.405	0.0%	0.0%
6	Qwen3-VL-8B	2.686	77.72	46.59	0.058	0.277	0.0%	0.0%
6	Qwen3-VL-32B	1.561	69.02	68.68	0.137	0.322	0.0%	0.0%
6	Gemma-4-31B	0.840	69.25	82.42	0.237	0.337	0.0%	0.0%
6	Llama-4 Maverick	1.447	73.55	70.65	0.168	0.311	0.0%	0.0%
7	GPT-6 Astra	0.010	2.35	0.84	0.927	0.932	90.0%	50.0%
7	GPT-5.6 Luna	0.558	68.02	76.54	0.295	0.437	0.0%	0.0%
7	GPT-4o	1.618	75.50	83.84	0.149	0.373	0.0%	0.0%
7	Gemini 3.1 Pro	1.840	61.77	59.11	0.231	0.412	0.0%	0.0%
7	Gemini Flash Lite	0.740	62.63	64.18	0.283	0.442	0.0%	0.0%
7	Claude Sonnet 4.6	0.912	75.52	63.44	0.248	0.435	0.0%	0.0%
7	Qwen3-VL-8B	2.726	73.39	41.70	0.074	0.327	0.0%	0.0%
7	Qwen3-VL-32B	1.366	72.34	53.97	0.156	0.378	0.0%	0.0%
7	Gemma-4-31B	1.010	71.11	91.01	0.214	0.241	0.0%	0.0%
7	Llama-4 Maverick	1.701	69.36	50.95	0.131	0.323	0.0%	0.0%

A.4 Matched API reconstruction visualizations

For each k, we use the first evaluation sample (index 0000), fixed independently of model scores. Every panel overlays the saved JSON prediction on the same ground truth. Colored polygons are ground truth; magenta dashed outlines are predictions. Axes are clipped to the input canvas, so an off-canvas prediction may be partly invisible. These diagnostic overlays are not the model inputs.

(a) $k = 1$

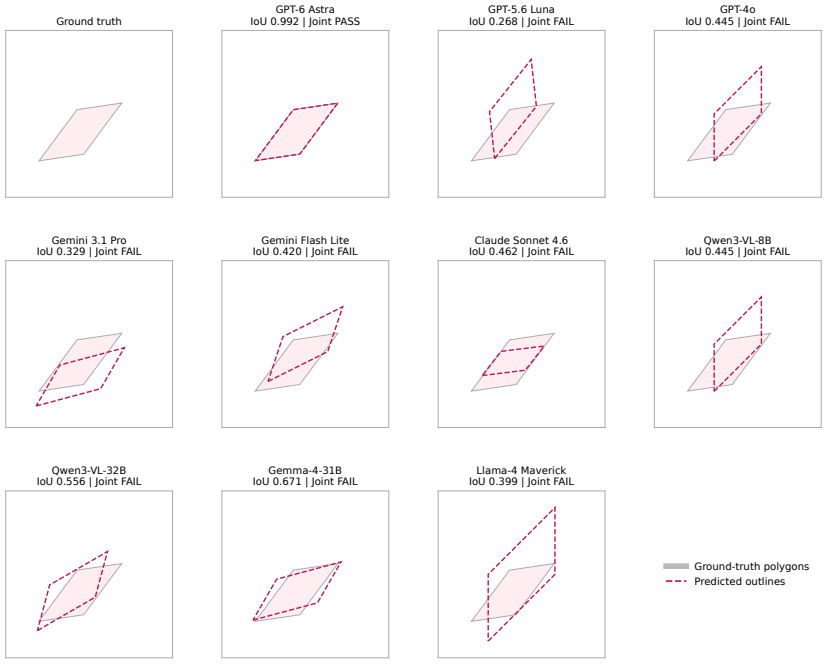


Figure 4: Matched ten-model Tangram reconstructions for $k = 1$. Each uses sample index 0000; IoU and Joint describe the saved predictions. Shared overlay conventions are defined above.

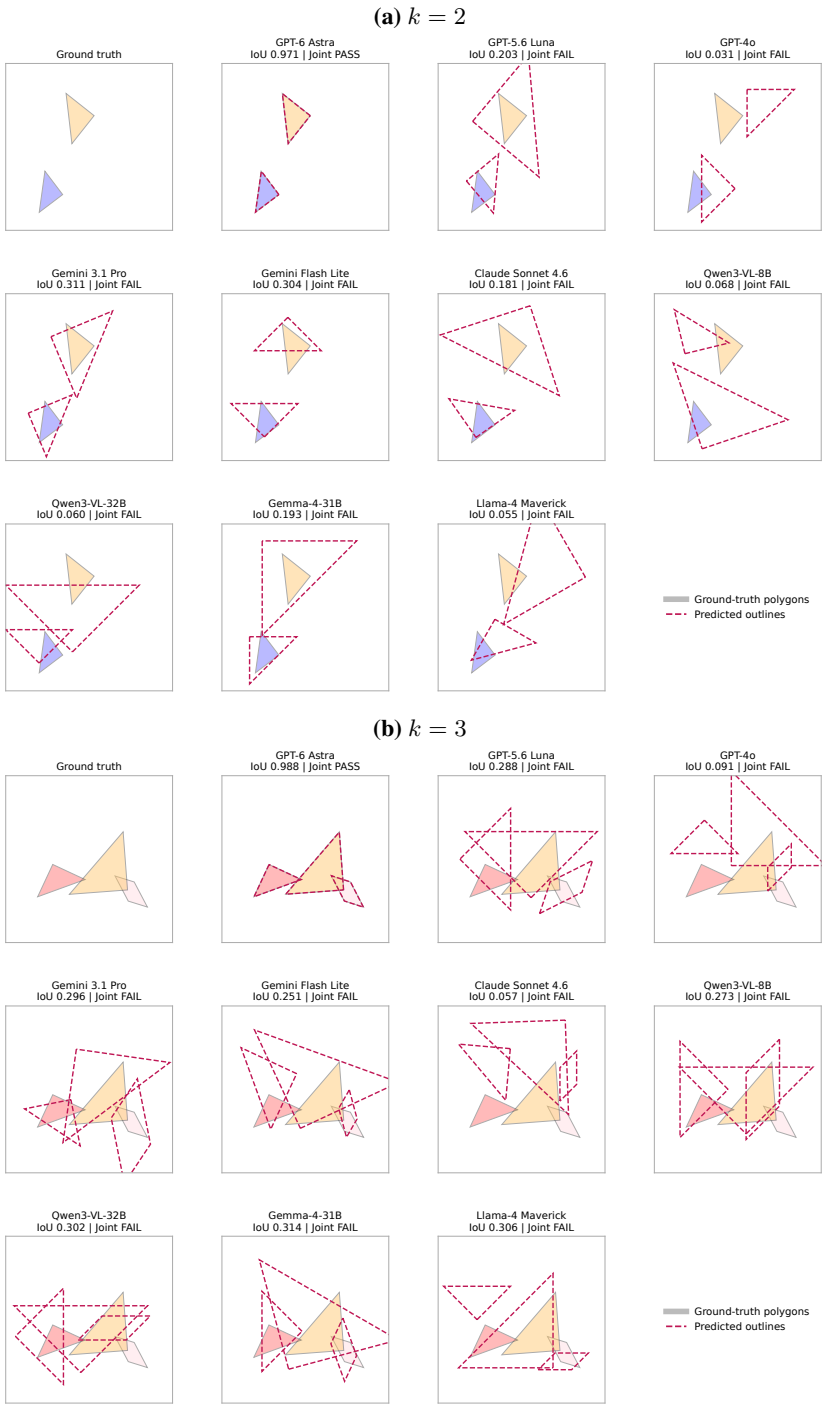


Figure 5: Matched ten-model Tangram reconstructions for $k = 2, 3$. Each uses sample index 0000; IoU and Joint describe the saved predictions. Shared overlay conventions are defined above.

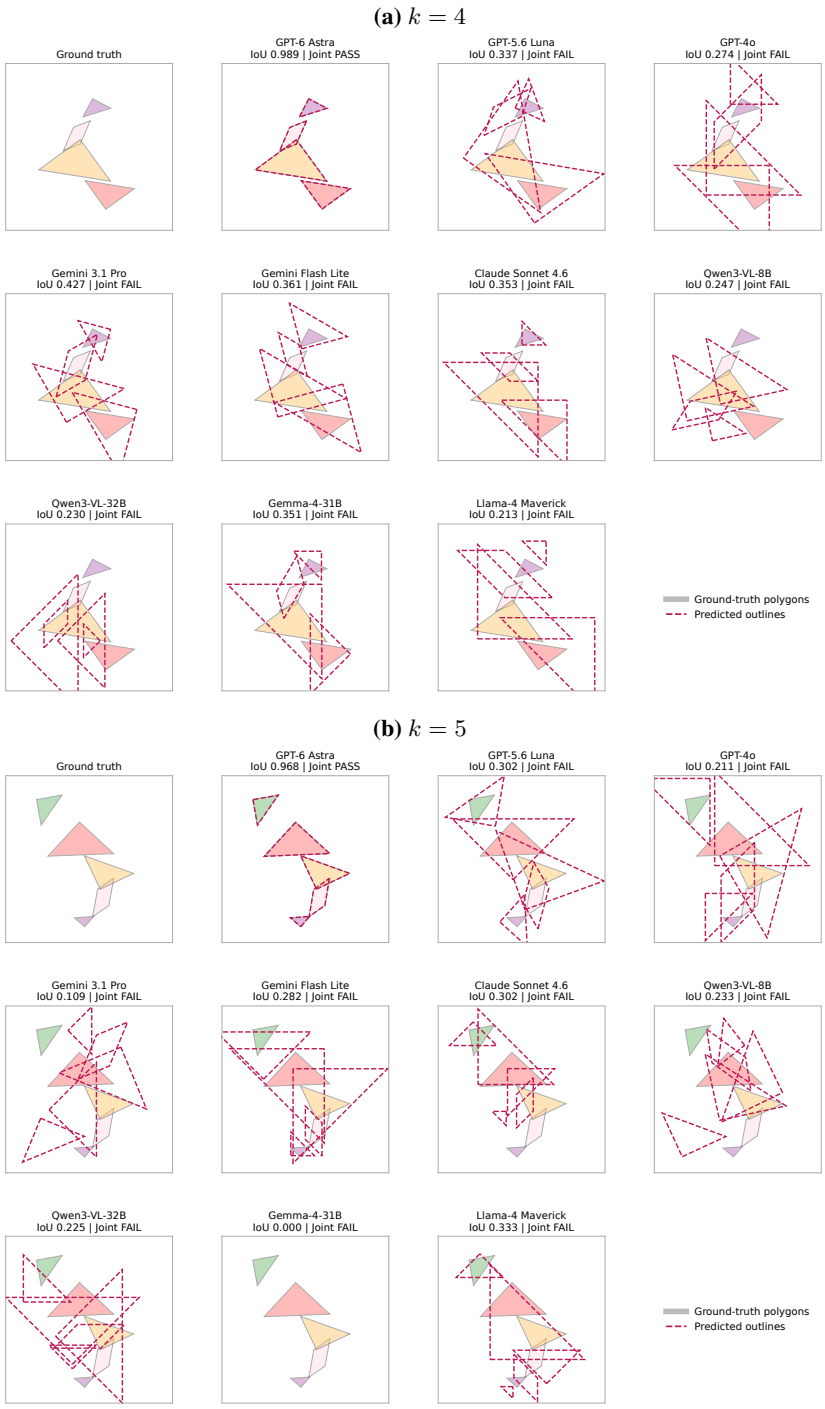


Figure 6: Matched ten-model Tangram reconstructions for $k = 4, 5$. Each uses sample index 0000; IoU and Joint describe the saved predictions. Shared overlay conventions are defined above.

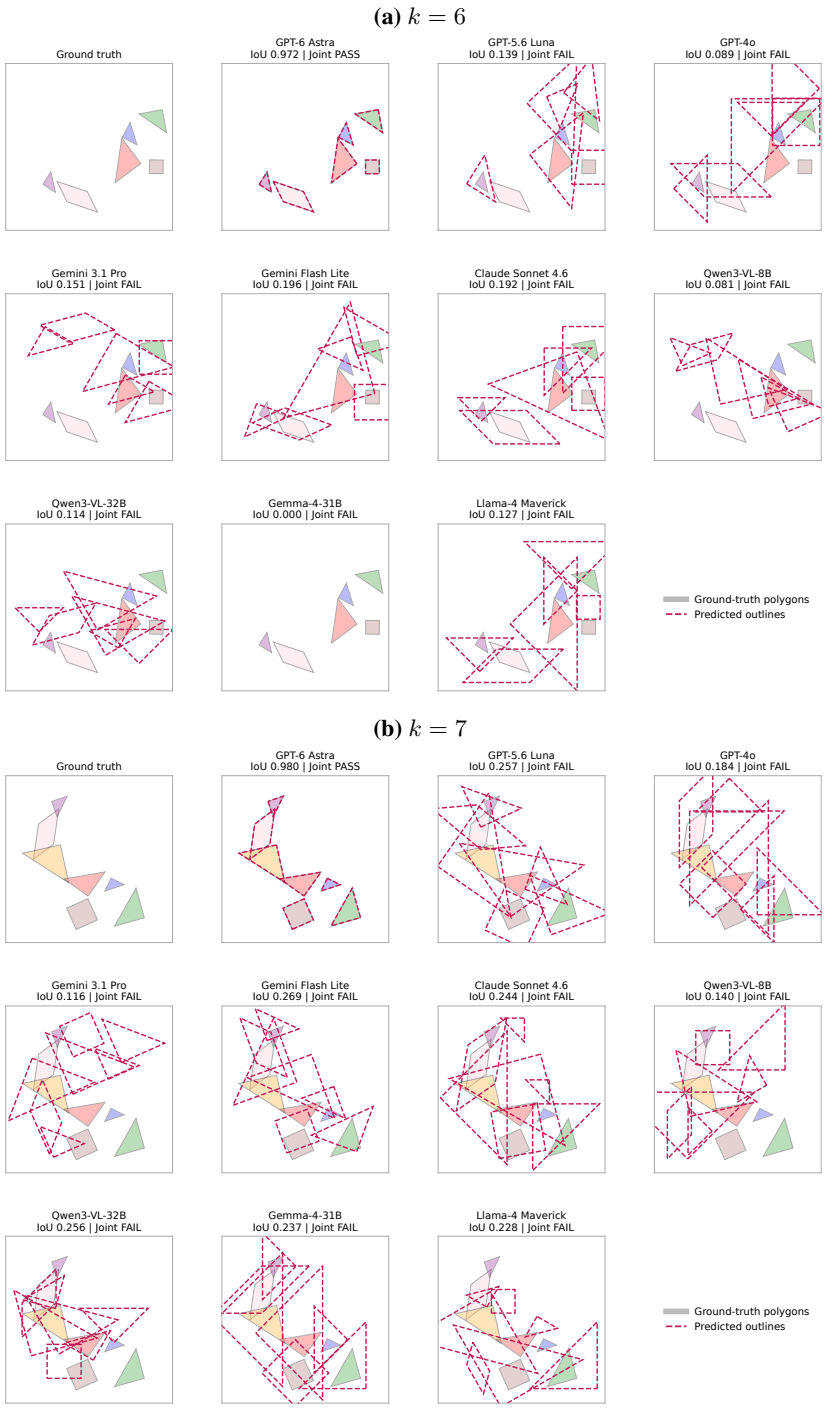


Figure 7: Matched ten-model Tangram reconstructions for $k = 6, 7$. Each uses sample index 0000; IoU and Joint describe the saved predictions. Shared overlay conventions are defined above.

B CanvasPlace Difficulty Extension: Setup, Metrics, and Visual Comparisons

Return to the main CanvasPlace results table.

B.1 Matched Families and Inference Setup

Dataset and evaluated subset. The current CanvasPlace protocol, denoted geometry-v2 (v2) in this appendix and archived artifacts, contains 500 matched source families, each with Easy, Medium, and Hard variants (1500 tasks; 500 per tier). The ten target shapes each occur five times per tier. The API comparison uses exactly source Tasks 01–200, two per shape. Each model has 200 freshly generated outputs per v2 tier, for 6000 model–task predictions in total. These predictions share only 200 independent source families and must not be treated as 6000 independent scenes.

Controlled geometric intervention. Targets, scale, solid-object polygons and positions, and the complete v2 instruction are identical within a family. Only the rectangular tray region is reduced by moving its boundaries inward; reference objects may consequently lie outside or across its boundary, as stated in the prompt. For cardinal angle a, let F_a be the continuous set of target-centroid positions satisfying all constraints and define $V = \sum_{a\in\{0,90,180,270\}} \mathrm{area}(F_a)$. Medium and Hard retain respectively $0.28V_{\mathrm{Easy}}$ and $0.055V_{\mathrm{Easy}}$, with $F_a^{\mathrm{Hard}} \subseteq F_a^{\mathrm{Medium}} \subseteq F_a^{\mathrm{Easy}}$. These are sums of feasible-center areas, not tray-area fractions or estimates of model accuracy. The construction considers exact target extent, nonconvexity, obstacles, and clearance through polygon configuration spaces; conservative distance buffers exceed the current thresholds by less than 0.005 pixels. Area calibration uses cardinal-angle slices, not the angular tolerance bands used in scoring.

Each task retains a witness admitting any fixed-angle translation of up to 6 pixels. The seed is 20260916. Candidate crops are selected geometrically, preferring those for which the tray center is invalid at every cardinal angle and then larger visible trays at the requested feasible-area ratio. No model outputs enter construction or selection. This is a deliberate intervention against a center-placement shortcut, not an unbiased sample of tabletop layouts. In the full 50-family v2 bank, an oracle choosing among four angles at the tray center succeeds on 96% Easy, 40% Medium, and 8% Hard tasks; these are geometric diagnostics, not VLM scores. Validation checks all 150 witnesses, 2,400 six-pixel perturbations, nested feasible sets, and 15,000 random-pose comparisons between direct scoring and feasible-region membership.

API inputs and settings. Requests use the same ten OpenRouter model identifiers listed in Table 7. Only scene PNG, target PNG, and English instruction are sent; polygons, masks, feasible maps, witnesses, and evaluation results are withheld. The scene is 512×512; the 192×192 target preview uses the same pixel scale. Outputs are normalized (x, y) and clockwise θ at fixed target size. All tiers use JSON-object response mode and a 1,200-token cap. Temperature is zero except that GPT-6 omits it. GPT-6 and Gemini Pro request low reasoning, Luna none, Flash Lite minimal, and Claude disables reasoning; the remaining models have no explicit reasoning override. There are no demonstrations, geometric feedback, or answer-based retries. API failures are retried, and rate-limited requests are resumed with one tier per model at a time and 3.2 seconds between attempts. Completed format or geometry failures are not resampled. All 6000 model–task pairs have API responses; 29 fail parsing (8 Easy, 9 Medium, and 12 Hard). Recorded latency includes retries and local pacing and is not a controlled inference-speed comparison.

Protocol comparability. The main three-tier API comparison uses freshly generated responses under the same prompt, inference settings, parser, and verifier. V2 discloses each numerical minimum distance, explicitly defines centroid relations and boundary contact, checks direct polygon intersections including concave targets, accepts symmetry-equivalent orientations, and enforces the four-angle output space on every task. The main comparison and all tables and figures in this appendix use only v2 outputs. Current API Qwen models are not the separately trained M2/M3 checkpoints.

Explicit v2 instruction definitions. The original task sentence is followed by the definitions below; any applicable clearance rule then states its reference object and threshold (12 or 14 pixels in the evaluated bank). The same generic coordinate/output prompt is used across all three v2 tiers.

> Definitions: left/right/above/below compare polygon centroids; between means the target centroid is horizontally between the two reference centroids. The entire target must stay in the tray region; reference objects may be outside that region. No positive-area overlap with any solid object is allowed; boundary-only contact is allowed unless a distance rule says otherwise. The scene is 512 by 512 pixels; the target preview is at the same pixel scale. Do not resize the target.

For example, Task 01 additionally requires edge-to-edge separation from the red square of at least 12 pixels. The same full instruction is used for its Easy, Medium, and Hard variants.

B.2 Evaluation Metrics and Complete Comparison

Metrics and denominators. Strict success is the conjunction of all task constraints and global checks, accepting any feasible pose. It is not distance to a witness or raster IoU. Containment includes tray-boundary contact; relations require strictly positive centroid margins; clearance uses the shortest polygon distance. Global collision avoidance forbids positive-area overlap with every solid object, while boundary-only contact is allowed unless an explicit clearance rule forbids it. Cardinal-angle and task-specific orientation checks use a 1° circular tolerance and polygon checks use numerical tolerance 10^{-7}. Coordinates must be finite and in $[0, 1]$; angles are normalized modulo 360° before checking. Missing, malformed, or duplicate predictions fail all applicable checks and remain in the denominators.

Per model and tier, containment, collision avoidance, parsing, API success, and global discrete rotation each use 20 cases. In the tables, Rotation denotes the global cardinal-angle check. Relations pool 31 instances: 4 left, 6 right, 6 above, 9 below, and 6 horizontal-between. Distance uses 7 and task-specific orientation 14. Cross-model rates below pool passed/applicable counts; because each model sees the same constraints, they also equal model-mean rates. The 200 model–task observations per tier share 20 source scenes and should not be treated as 200 independent scenes.

Table 9: Pooled attribute accuracy (%) across ten models on 20 matched tasks per v2 tier.

Metric	Easy (v2)	Medium (v2)	Hard (v2)
Strict success	66.0	50.0	25.0
Inside tray	92.5	75.0	49.0
Relations	94.8	94.2	89.7
Distance	88.6	87.1	95.7
Orientation	77.1	75.0	72.9
No collision	75.0	69.0	53.0
Parseable output	96.0	95.5	94.0

What changes as geometric difficulty increases? Mean strict success decreases from 66% on Easy to 50% on Medium and 25% on Hard. Containment falls from 92.5% to 75.0% to 49.0%, and collision-free placement from 75.0% to 69.0% to 53.0%, whereas relations remain 94.8%, 94.2%, and 89.7%. Parsing decreases only from 96.0% to 95.5% to 94.0%. This pattern is consistent with loss of precise feasible placement beyond coarse relational compliance, but is not proof of complete semantic understanding. Clearance is non-monotonic (88.6%, 87.1%, 95.7%): cropping may favor positions farther from a reference, and aggregate rates also depend on invalid outputs. Difficulty need not increase for every individual constraint.

Table 10: Complete CanvasPlace v2 results on the same 20 families per tier. All metrics, including strict success, are percentages. Constraint denominators follow the definitions above.

Tier	Model	Parse	Inside	Relation	Distance	Orient.	No coll.	Rotation	Strict
Easy	GPT-6 Astra	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Easy	GPT-5.6 Luna	100.0	85.0	100.0	100.0	85.7	65.0	100.0	60.0
Easy	GPT-4o	100.0	85.0	96.8	85.7	71.4	50.0	100.0	45.0
Easy	Gemini 3.1 Pro	100.0	100.0	100.0	100.0	92.9	100.0	100.0	95.0

Continued on next page

Table 10 – continued

Tier	Model	Parse	Inside	Relation	Distance	Orient.	No coll.	Rotation	Strict
Easy	Gemini Flash Lite	100.0	100.0	100.0	100.0	85.7	90.0	100.0	80.0
Easy	Claude Sonnet 4.6	100.0	100.0	100.0	100.0	100.0	85.0	100.0	85.0
Easy	Qwen3-VL-8B	80.0	80.0	74.2	42.9	42.9	40.0	80.0	20.0
Easy	Qwen3-VL-32B	80.0	80.0	77.4	71.4	42.9	70.0	80.0	50.0
Easy	Gemma-4-31B	100.0	100.0	100.0	100.0	92.9	90.0	100.0	85.0
Easy	Llama-4 Maverick	100.0	95.0	100.0	85.7	57.1	60.0	100.0	40.0
Medium	GPT-6 Astra	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Medium	GPT-5.6 Luna	100.0	80.0	96.8	100.0	78.6	60.0	100.0	45.0
Medium	GPT-4o	100.0	50.0	100.0	100.0	71.4	50.0	100.0	20.0
Medium	Gemini 3.1 Pro	100.0	100.0	100.0	100.0	92.9	90.0	100.0	85.0
Medium	Gemini Flash Lite	100.0	75.0	100.0	100.0	85.7	65.0	100.0	45.0
Medium	Claude Sonnet 4.6	100.0	95.0	100.0	100.0	92.9	80.0	100.0	75.0
Medium	Qwen3-VL-8B	70.0	40.0	58.1	14.3	35.7	35.0	70.0	10.0
Medium	Qwen3-VL-32B	85.0	50.0	87.1	71.4	50.0	65.0	85.0	25.0
Medium	Gemma-4-31B	100.0	90.0	100.0	100.0	92.9	90.0	100.0	75.0
Medium	Llama-4 Maverick	100.0	70.0	100.0	85.7	50.0	55.0	100.0	20.0
Hard	GPT-6 Astra	100.0	95.0	100.0	100.0	100.0	95.0	100.0	90.0
Hard	GPT-5.6 Luna	100.0	50.0	93.5	100.0	85.7	35.0	100.0	15.0
Hard	GPT-4o	100.0	25.0	93.5	100.0	71.4	40.0	100.0	5.0
Hard	Gemini 3.1 Pro	100.0	60.0	100.0	100.0	85.7	75.0	100.0	45.0
Hard	Gemini Flash Lite	100.0	50.0	96.8	100.0	85.7	60.0	100.0	20.0
Hard	Claude Sonnet 4.6	90.0	75.0	90.3	85.7	85.7	50.0	90.0	40.0
Hard	Qwen3-VL-8B	85.0	5.0	61.3	100.0	35.7	45.0	85.0	0.0
Hard	Qwen3-VL-32B	65.0	25.0	67.7	71.4	35.7	25.0	65.0	5.0
Hard	Gemma-4-31B	100.0	75.0	100.0	100.0	92.9	50.0	100.0	30.0
Hard	Llama-4 Maverick	100.0	30.0	93.5	100.0	50.0	55.0	100.0	0.0

B.3 Actual Three-Tier Predictions and Failure Cases

Rows show saved Easy/Medium/Hard predictions at unchanged object scale. Green/red outlines indicate success/failure; dots mark predicted anchors. Invalid outputs have no overlay. All ten models are shown for Task 01, whose L-shaped target must fit inside the tray, right of the blue circle, at least 12 pixels from the red square, with its long arm upward. Tasks 06 and 14 are GPT-6's only Hard failures and illustrate error types rather than their prevalence.

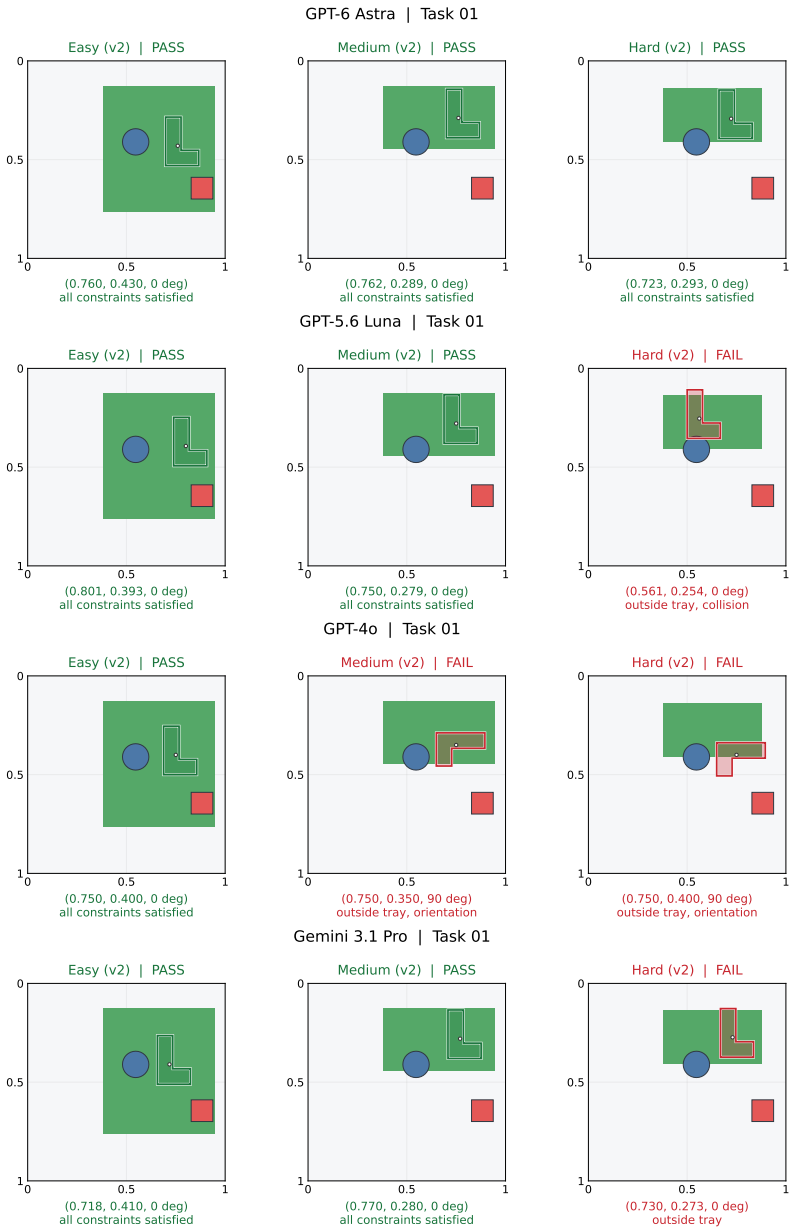


Figure 8: Task 01 predictions for GPT-6 Astra, GPT-5.6 Luna, GPT-4o, and Gemini 3.1 Pro. Columns show Easy, Medium, and Hard.

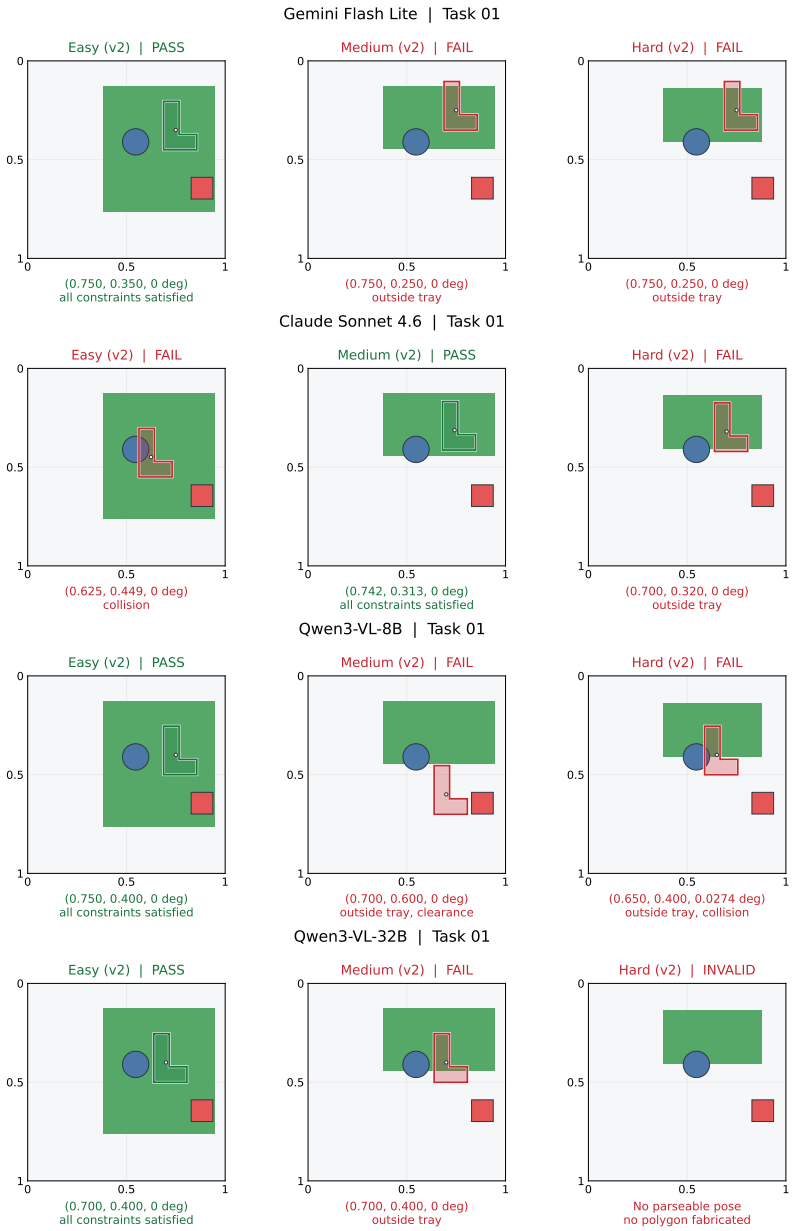


Figure 9: Task 01 predictions for Gemini Flash Lite, Claude Sonnet 4.6, Qwen3-VL-8B, and Qwen3-VL-32B.

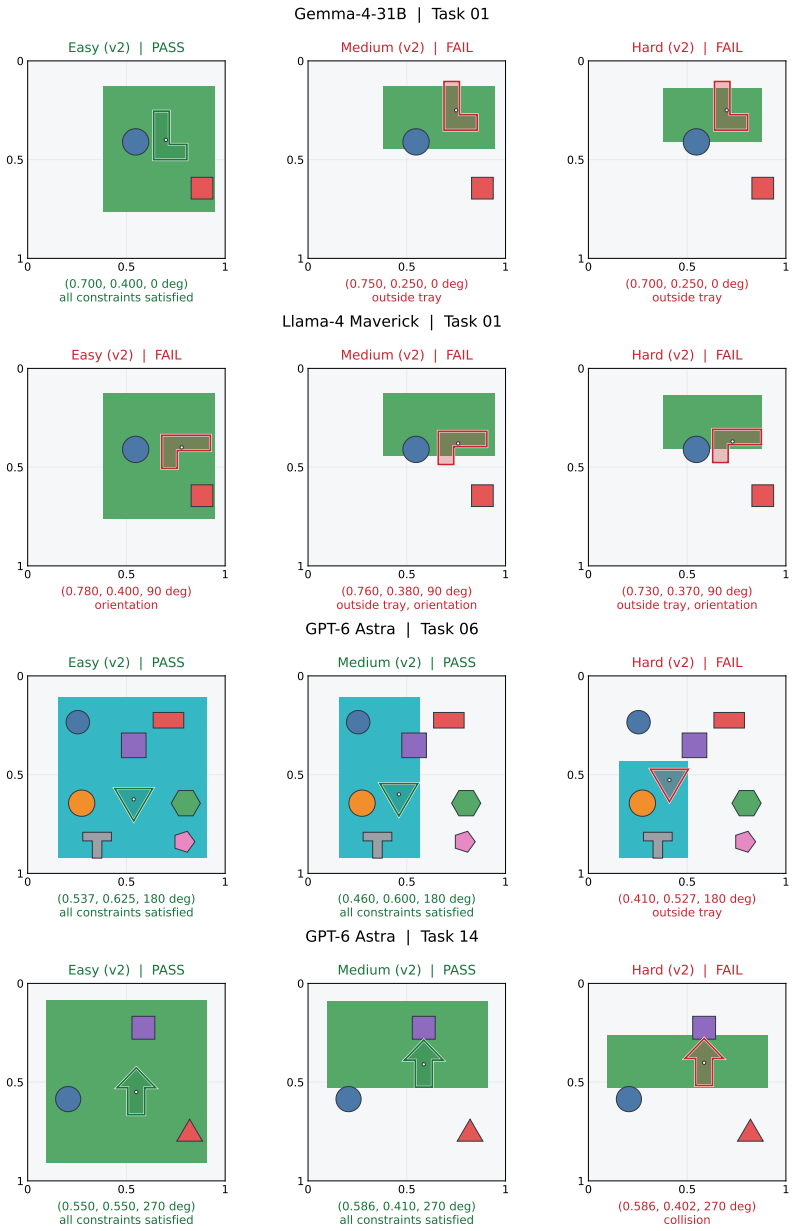


Figure 10: Top two rows: Task 01 predictions for Gemma-4-31B and Llama-4 Maverick. Bottom two rows: the two GPT-6 Hard failures, with Easy/Medium context. Task 06 violates containment. In Task 14, the arrow must be inside the tray, below the purple square, and right of the blue circle; its tip overlaps the square by 3.33 square pixels, failing strict collision scoring.

C Complete Tangram In-Context-Learning Results

Table 2 in the main paper averages over the matched $k \in \{3, 5, 7\}$ evaluation. The consolidated table below reports every metric for each piece count, model, and ICL condition. Every cell uses 30 held-out test scenes. Demonstrations come only from the source-page-disjoint training split, every request uses demonstrations with the same k as its test scene, all models receive identical demonstrations, and the 5-shot set is nested within the 8-shot set. The data audit found zero train/test overlap in sample IDs, source pages, image SHA-256 hashes, and annotation SHA-256 hashes.

Table 11: Complete ICL results on matched scenes (30 scenes per piece count and shot condition). Position is in canvas units; angle is in degrees; Scale, IoU$\geq$ .8, and Joint are percentages.

k	Shots	Model	Pos.↓	Angle↓	Scale↓	Piece IoU↑	Scene IoU↑	IoU$\geq$ .8	Joint↑
3	0	GPT-6 Astra	0.006	0.66	0.4	0.977	0.978	96.7%	96.7%
3	0	GPT-5.6 Luna	0.562	77.36	55.0	0.331	0.378	0.0%	0.0%
3	0	GPT-4o	1.452	70.34	56.6	0.188	0.265	0.0%	0.0%
3	0	Gemini 3.1 Pro	1.261	60.52	52.3	0.286	0.372	0.0%	0.0%
3	0	Gemini Flash Lite	26.406	75.72	55.1	0.312	0.369	0.0%	0.0%
3	0	Claude Sonnet 4.6	0.848	62.62	36.8	0.320	0.367	0.0%	0.0%
3	0	Qwen3-VL-8B	2.724	90.89	42.2	0.101	0.209	0.0%	0.0%
3	0	Qwen3-VL-32B	1.511	81.44	51.1	0.164	0.260	0.0%	0.0%
3	0	Gemma-4-31B	0.748	71.58	84.0	0.258	0.321	0.0%	0.0%
3	0	Llama-4 Maverick	1.541	69.04	51.1	0.180	0.251	0.0%	0.0%
3	5	GPT-6 Astra	0.007	2.14	0.4	0.971	0.975	96.7%	96.7%
3	5	GPT-5.6 Luna	0.535	72.37	23.0	0.427	0.459	0.0%	0.0%
3	5	GPT-4o	1.307	72.76	27.3	0.206	0.283	0.0%	0.0%
3	5	Gemini 3.1 Pro	0.469	63.07	13.1	0.478	0.502	0.0%	0.0%
3	5	Gemini Flash Lite	0.641	79.22	21.5	0.361	0.426	0.0%	0.0%
3	5	Claude Sonnet 4.6	0.824	56.89	15.7	0.309	0.369	0.0%	0.0%
3	5	Qwen3-VL-8B	2.818	78.99	27.5	0.084	0.211	0.0%	0.0%
3	5	Qwen3-VL-32B	1.955	74.85	24.1	0.151	0.242	0.0%	0.0%
3	5	Gemma-4-31B	0.475	70.51	23.1	0.395	0.436	0.0%	0.0%
3	5	Llama-4 Maverick	1.324	76.13	22.3	0.184	0.256	0.0%	0.0%
3	8	GPT-6 Astra	0.008	1.14	0.4	0.977	0.976	96.7%	96.7%
3	8	GPT-5.6 Luna	0.498	66.89	18.4	0.413	0.454	0.0%	0.0%
3	8	GPT-4o	1.338	74.60	22.9	0.187	0.274	0.0%	0.0%
3	8	Gemini 3.1 Pro	0.847	61.89	11.9	0.413	0.469	0.0%	0.0%
3	8	Gemini Flash Lite	0.673	76.67	17.5	0.358	0.421	0.0%	0.0%
3	8	Claude Sonnet 4.6	0.667	67.64	15.8	0.370	0.412	0.0%	0.0%
3	8	Qwen3-VL-8B	2.659	82.90	25.5	0.071	0.201	0.0%	0.0%
3	8	Qwen3-VL-32B	1.826	78.41	22.8	0.167	0.261	0.0%	0.0%
3	8	Gemma-4-31B	0.473	73.80	19.6	0.403	0.451	0.0%	0.0%
3	8	Llama-4 Maverick	1.222	66.41	20.0	0.201	0.273	0.0%	0.0%
5	0	GPT-6 Astra	0.013	1.55	0.6	0.963	0.969	100.0%	86.7%
5	0	GPT-5.6 Luna	0.678	87.93	76.3	0.254	0.331	0.0%	0.0%
5	0	GPT-4o	1.346	71.99	82.4	0.167	0.290	0.0%	0.0%
5	0	Gemini 3.1 Pro	1.559	73.38	52.4	0.262	0.358	0.0%	0.0%
5	0	Gemini Flash Lite	121.420	73.91	76.5	0.233	0.332	0.0%	0.0%
5	0	Claude Sonnet 4.6	0.779	69.32	56.1	0.279	0.376	0.0%	0.0%
5	0	Qwen3-VL-8B	2.877	82.36	53.2	0.078	0.234	0.0%	0.0%
5	0	Qwen3-VL-32B	1.510	74.59	69.4	0.135	0.291	0.0%	0.0%
5	0	Gemma-4-31B	1.001	63.44	89.7	0.231	0.215	0.0%	0.0%
5	0	Llama-4 Maverick	1.489	72.06	67.6	0.159	0.279	0.0%	0.0%
5	5	GPT-6 Astra	0.008	1.95	0.6	0.965	0.969	100.0%	83.3%
5	5	GPT-5.6 Luna	0.437	78.92	25.6	0.398	0.466	0.0%	0.0%
5	5	GPT-4o	1.314	76.11	45.0	0.173	0.324	0.0%	0.0%
5	5	Gemini 3.1 Pro	0.408	64.05	15.7	0.463	0.526	0.0%	0.0%
5	5	Gemini Flash Lite	0.538	69.52	22.7	0.367	0.454	0.0%	0.0%
5	5	Claude Sonnet 4.6	0.606	74.11	21.9	0.348	0.456	0.0%	0.0%
5	5	Qwen3-VL-8B	2.421	75.25	38.5	0.094	0.280	0.0%	0.0%
5	5	Qwen3-VL-32B	2.308	70.81	24.3	0.102	0.238	0.0%	0.0%

Continued on next page

Table 11 – continued

k	Shots	Model	Pos.↓	Angle↓	Scale↓	Piece IoU↑	Scene IoU↑	IoU≥ .8	Joint↑
5	5	Gemma-4-31B	0.611	71.50	24.1	0.330	0.430	0.0%	0.0%
5	5	Llama-4 Maverick	1.202	73.81	25.7	0.188	0.318	0.0%	0.0%
5	8	GPT-6 Astra	0.010	1.57	0.8	0.962	0.973	100.0%	86.7%
5	8	GPT-5.6 Luna	0.458	77.69	24.4	0.396	0.466	0.0%	0.0%
5	8	GPT-4o	1.263	76.08	43.7	0.177	0.320	0.0%	0.0%
5	8	Gemini 3.1 Pro	0.431	67.45	13.2	0.470	0.524	0.0%	0.0%
5	8	Gemini Flash Lite	0.525	69.21	23.8	0.367	0.452	0.0%	0.0%
5	8	Claude Sonnet 4.6	0.610	70.93	15.0	0.348	0.453	0.0%	0.0%
5	8	Qwen3-VL-8B	2.649	77.41	37.1	0.092	0.250	0.0%	0.0%
5	8	Qwen3-VL-32B	1.861	75.46	24.4	0.132	0.273	0.0%	0.0%
5	8	Gemma-4-31B	0.557	68.93	23.5	0.355	0.447	0.0%	0.0%
5	8	Llama-4 Maverick	1.159	70.60	24.6	0.189	0.322	0.0%	0.0%
7	0	GPT-6 Astra	0.013	4.22	0.9	0.955	0.969	93.3%	73.3%
7	0	GPT-5.6 Luna	0.522	69.23	77.7	0.296	0.445	0.0%	0.0%
7	0	GPT-4o	1.617	74.37	82.9	0.148	0.376	0.0%	0.0%
7	0	Gemini 3.1 Pro	1.528	62.23	62.1	0.251	0.433	0.0%	0.0%
7	0	Gemini Flash Lite	0.709	65.68	64.1	0.292	0.455	0.0%	0.0%
7	0	Claude Sonnet 4.6	0.894	75.93	61.6	0.248	0.437	0.0%	0.0%
7	0	Qwen3-VL-8B	2.711	74.49	43.4	0.072	0.328	0.0%	0.0%
7	0	Qwen3-VL-32B	1.335	74.59	57.1	0.161	0.385	0.0%	0.0%
7	0	Gemma-4-31B	0.919	70.53	86.2	0.231	0.305	0.0%	0.0%
7	0	Llama-4 Maverick	1.625	67.33	55.4	0.136	0.339	0.0%	0.0%
7	5	GPT-6 Astra	0.018	3.73	1.2	0.948	0.963	96.7%	76.7%
7	5	GPT-5.6 Luna	0.629	72.53	19.3	0.350	0.483	0.0%	0.0%
7	5	GPT-4o	1.601	73.79	55.2	0.152	0.392	0.0%	0.0%
7	5	Gemini 3.1 Pro	1.219	70.08	14.6	0.331	0.470	0.0%	0.0%
7	5	Gemini Flash Lite	0.698	68.01	22.1	0.321	0.467	0.0%	0.0%
7	5	Claude Sonnet 4.6	0.917	64.69	16.5	0.267	0.449	0.0%	0.0%
7	5	Qwen3-VL-8B	2.708	80.47	29.5	0.067	0.340	0.0%	0.0%
7	5	Qwen3-VL-32B	1.699	70.81	21.9	0.135	0.370	0.0%	0.0%
7	5	Gemma-4-31B	0.447	67.79	23.7	0.383	0.517	0.0%	0.0%
7	5	Llama-4 Maverick	1.362	72.97	26.2	0.162	0.389	0.0%	0.0%
7	8	GPT-6 Astra	0.015	3.43	0.9	0.949	0.960	96.7%	73.3%
7	8	GPT-5.6 Luna	0.655	75.01	17.8	0.344	0.478	0.0%	0.0%
7	8	GPT-4o	1.609	73.68	56.3	0.135	0.387	0.0%	0.0%
7	8	Gemini 3.1 Pro	1.348	74.37	14.9	0.326	0.462	0.0%	0.0%
7	8	Gemini Flash Lite	0.717	71.94	19.6	0.321	0.467	0.0%	0.0%
7	8	Claude Sonnet 4.6	0.865	66.62	15.0	0.276	0.452	0.0%	0.0%
7	8	Qwen3-VL-8B	2.554	69.12	24.8	0.075	0.342	0.0%	0.0%
7	8	Qwen3-VL-32B	1.666	67.35	20.2	0.162	0.390	0.0%	0.0%
7	8	Gemma-4-31B	0.433	67.28	22.4	0.391	0.528	0.0%	0.0%
7	8	Llama-4 Maverick	1.331	72.43	23.0	0.180	0.389	0.0%	0.0%

D Complete Adaptation Settings and Results

M1 trains a randomly initialized head over frozen visual features, while M2/M3 are independently initialized from their architecture's Base. All use 4,200 training scenes and 350 validation scenes, with 420 held-out test scenes and seed 2027. M1 uses batches of eight and validation-selected extended training, detailed below. M2/M3 use microbatch size one and accumulation eight for two epochs. The differing architectures, initialization, and objectives mean this is not a matched-compute or parameter-efficiency comparison.

Head assignment and losses. For predicted slot j and target i, the implemented Hungarian cost is

$$C_{ji} = -p_j(c_i) + 5\|\widehat{\mathbf{p}}_j - \mathbf{p}_i\|_1 + 2(1 - \widehat{\mathbf{a}}_j^\top \mathbf{a}_i) + 2|\widehat{\log s}_j - \log s_i|. \tag{6}$$

Here positions are normalized to $[0, 1]^2$ and angle vectors use the piece-specific symmetry period (90/180/360 degrees). Matching is not the same as the weighted training loss in Eq. 3. Classification uses a no-object weight of 0.1. The flip binary cross-entropy is applied to parallelograms only. Seven

learned queries, the memory projection, decoder, and output heads are trainable; the visual backbone is frozen.

D.1 Extended M1 Training and Input Controls

We evaluate five configurations with extended training on the same source-page-disjoint 4,200/350/420 train/validation/test split. Here, M1 denotes the head without count input; M1+k adds a learned count embedding to the object queries. All variants use true k for output selection. InternVL single-view heads receive one 448-pixel view (256 visual tokens); the matched-tiles configuration receives four local crops and a thumbnail (1,280 tokens), matching the M2/M3 pixel tensors. Table 3 uses Qwen M1+k and InternVL M1+k+tiles, both abbreviated M1, for input comparability.

All five runs use AdamW, initial learning rate 10^{-4}, seed 2027, and batches of eight cached feature sets. Validation loss selects each checkpoint. ReduceLROnPlateau halves the learning rate after three epochs without improvement, down to 10^{-6}. Training stops after eight epochs without validation improvement, after at least 15 epochs, with an 80-epoch hard cap. All five runs stopped by validation patience after 46–63 epochs. An initially broader learning-rate/seed search was reduced to these five configurations after inspecting validation curves but before production test scoring; this was a budget adjustment, not a preregistered selection. Earlier search artifacts are retained as auxiliary results.

The same optimization recipe and loss reduction are used across the five configurations. The flip loss averages over images containing a parallelogram within each batch. One seed does not estimate training variability, and these controls do not equalize the head's random initialization with language-model pretraining. Reported comparisons concern these tested configurations and do not establish the intrinsic superiority of either modeling paradigm.

Table 12: Training completion and validation-selected epochs for the five M1 configurations.

Architecture	Configuration	Epochs run	Selected epoch	Validation loss
Qwen	M1	55	47	0.1533
Qwen	M1+k	55	47	0.1906
InternVL	M1	63	55	0.1815
InternVL	M1+k	46	38	0.2742
InternVL	M1+k+tiles	48	40	0.1117

Table 13 reports all 420 test scenes (60 per k). Parse, ID, Recall, Scale, IoU80, and Joint are percentages; position is in canvas units and angle in degrees.

Main-table identity accuracy. Exact-ID accuracy is the percentage of scenes with the correct identity set and count. For Base/M1/M2/M3 in Table 3, it is 74.29/99.05/100.00/100.00% on Qwen and 79.29/100.00/100.00/100.00% on InternVL.

Interpretation. The plain heads reach 21.67%/20.48% Joint on Qwen/InternVL. Explicit count input changes Joint to 23.10%/14.29%; matching InternVL input crops gives 15.48%. These are configuration-specific outcomes, not evidence of a consistent benefit from either input change. Across all five configurations, Joint at $k = 5, 6, 7$ is only 0–1.67%. Residual multi-piece failures do not establish an inherent limit of direct regression.

Table 13: Complete M1 controls: aggregate and per-k results for all five configurations (420 test scenes; 60 per k). Units follow Table 14.

k	Parse	ID	Recall	Pos.	Angle	Scale	Piece IoU	Scene IoU	IoU80	Joint
Qwen M1										
overall	100.0	99.8	99.9	0.598	5.53	4.89	0.426	0.553	5.95	21.67
1	100.0	100.0	100.0	0.259	1.62	2.47	0.709	0.709	31.67	96.67
2	100.0	100.0	100.0	0.501	2.48	3.57	0.510	0.578	5.00	33.33
3	100.0	100.0	100.0	0.555	3.01	3.97	0.469	0.524	3.33	15.00
4	100.0	100.0	100.0	0.578	3.44	4.56	0.436	0.520	1.67	5.00

Continued on next page

Table 13 – continued

k	Parse	ID	Recall	Pos.	Angle	Scale	Piece IoU	Scene IoU	IoU80	Joint
5	100.0	98.3	99.7	0.612	4.61	4.79	0.407	0.495	0.00	0.00
6	100.0	100.0	100.0	0.611	6.39	5.10	0.400	0.525	0.00	1.67
7	100.0	100.0	100.0	0.684	9.17	6.09	0.372	0.520	0.00	0.00
Qwen M1+k										
overall	100.0	99.0	99.8	0.568	5.15	5.08	0.446	0.575	9.29	23.10
1	100.0	100.0	100.0	0.203	1.45	2.16	0.752	0.752	45.00	100.00
2	100.0	100.0	100.0	0.439	1.62	3.57	0.548	0.600	11.67	41.67
3	100.0	100.0	100.0	0.548	2.59	4.51	0.483	0.544	5.00	13.33
4	100.0	100.0	100.0	0.540	2.78	4.39	0.449	0.527	1.67	5.00
5	100.0	100.0	100.0	0.552	3.79	4.71	0.440	0.524	0.00	0.00
6	100.0	93.3	98.9	0.570	6.09	5.23	0.425	0.547	1.67	1.67
7	100.0	100.0	100.0	0.691	9.31	6.69	0.376	0.531	0.00	0.00
InternVL M1										
overall	100.0	99.8	99.9	0.573	8.22	5.69	0.441	0.570	4.76	20.48
1	100.0	100.0	100.0	0.290	1.97	3.03	0.685	0.685	28.33	90.00
2	100.0	100.0	100.0	0.435	2.38	3.68	0.551	0.590	3.33	38.33
3	100.0	100.0	100.0	0.489	4.23	4.71	0.514	0.578	0.00	10.00
4	100.0	100.0	100.0	0.533	4.92	5.05	0.464	0.552	0.00	3.33
5	100.0	100.0	100.0	0.548	6.25	5.21	0.445	0.531	0.00	1.67
6	100.0	98.3	99.7	0.624	10.82	6.08	0.396	0.525	0.00	0.00
7	100.0	100.0	100.0	0.688	13.58	7.46	0.367	0.530	1.67	0.00
InternVL M1+k										
overall	100.0	97.6	99.4	0.713	10.81	6.47	0.379	0.516	3.10	14.29
1	100.0	100.0	100.0	0.375	1.70	3.54	0.605	0.605	11.67	70.00
2	100.0	100.0	100.0	0.583	2.04	4.47	0.464	0.527	6.67	23.33
3	100.0	100.0	100.0	0.554	5.79	5.23	0.487	0.539	3.33	6.67
4	100.0	95.0	98.8	0.648	5.39	5.49	0.404	0.491	0.00	0.00
5	100.0	95.0	99.0	0.613	7.62	6.27	0.390	0.482	0.00	0.00
6	100.0	93.3	98.9	0.749	12.86	6.70	0.344	0.492	0.00	0.00
7	100.0	100.0	100.0	0.942	20.34	8.51	0.286	0.473	0.00	0.00
InternVL M1+k+tiles										
overall	100.0	100.0	100.0	0.620	7.26	6.29	0.411	0.532	2.86	15.48
1	100.0	100.0	100.0	0.401	2.57	3.70	0.611	0.611	11.67	70.00
2	100.0	100.0	100.0	0.515	2.82	4.53	0.503	0.562	5.00	31.67
3	100.0	100.0	100.0	0.585	4.22	5.18	0.469	0.517	1.67	3.33
4	100.0	100.0	100.0	0.547	5.69	5.96	0.447	0.539	1.67	3.33
5	100.0	100.0	100.0	0.632	5.88	6.42	0.384	0.477	0.00	0.00
6	100.0	100.0	100.0	0.620	8.91	6.43	0.394	0.525	0.00	0.00
7	100.0	100.0	100.0	0.731	10.95	7.64	0.344	0.493	0.00	0.00

D.2 Complete Base and Language-Adaptation Results

Autoregressive SFT objective. M2 and M3 minimize the negative log-likelihood of the target answer tokens y, conditioned on the image I and prompt q:

$$\mathcal{L}_{\mathrm{SFT}}(\phi) = -\sum_{t=1}^{|y|} \log p_\phi(y_t \mid I, q, y_{<t}). \tag{7}$$

Only answer tokens contribute to the loss. The trainable parameters ϕ comprise the language LoRA adapters in M2, and additionally the selected visual blocks and cross-modal interface in M3.

M3 directly trains the last four visual blocks and the interface, rather than applying visual LoRA. Its language adapters have the same rank and target modules as M2.

Metric conventions. ID is exact identity-set accuracy; recall is matched-piece recall. Errors are over matched pieces, IoU80 is the proportion of scenes with Scene IoU at least .8, and Joint is strict all-piece pose success. Parse, ID, recall, scale, IoU80, and Joint columns are percentages. Each k has 60 scenes.

Table 14: Complete Base/M2/M3 adaptation results: 420 test scenes per checkpoint, 60 per k. Parse, ID, Recall, Scale, IoU80, and Joint are percentages; angle is in degrees and position in canvas units.

Checkpoint	k	Parse	ID	Recall	Pos.	Angle	Scale	Piece IoU	Scene IoU	IoU80	Joint
Qwen Base	overall	100.0	74.3	93.4	2.932	76.19	40.62	0.070	0.210	0.00	0.00
Qwen Base	1	100.0	81.7	81.7	2.522	89.04	28.14	0.127	0.122	0.00	0.00
Qwen Base	2	100.0	83.3	91.7	3.103	73.73	41.37	0.091	0.147	0.00	0.00
Qwen Base	3	100.0	65.0	88.3	3.039	83.92	39.98	0.089	0.193	0.00	0.00
Qwen Base	4	100.0	73.3	92.9	3.142	81.15	38.91	0.062	0.209	0.00	0.00
Qwen Base	5	100.0	68.3	93.0	2.910	73.20	41.91	0.062	0.225	0.00	0.00
Qwen Base	6	100.0	48.3	91.4	2.826	76.23	39.81	0.072	0.279	0.00	0.00
Qwen Base	7	100.0	100.0	100.0	2.879	71.73	42.81	0.059	0.294	0.00	0.00
Qwen M2	overall	100.0	100.0	100.0	0.149	16.94	4.05	0.771	0.829	73.81	40.71
Qwen M2	1	100.0	100.0	100.0	0.061	4.44	2.51	0.866	0.866	88.33	90.00
Qwen M2	2	100.0	100.0	100.0	0.080	6.11	2.87	0.815	0.828	71.67	76.67
Qwen M2	3	100.0	100.0	100.0	0.084	14.11	3.47	0.800	0.825	71.67	45.00
Qwen M2	4	100.0	100.0	100.0	0.447	12.57	3.88	0.791	0.832	75.00	33.33
Qwen M2	5	100.0	100.0	100.0	0.089	14.75	4.18	0.771	0.817	68.33	23.33
Qwen M2	6	100.0	100.0	100.0	0.098	22.56	4.41	0.752	0.821	73.33	11.67
Qwen M2	7	100.0	100.0	100.0	0.125	22.28	4.56	0.739	0.816	68.33	5.00
Qwen M3	overall	100.0	100.0	100.0	0.141	8.64	2.78	0.808	0.862	84.29	52.38
Qwen M3	1	100.0	100.0	100.0	0.042	2.03	1.16	0.896	0.896	90.00	91.67
Qwen M3	2	100.0	100.0	100.0	0.062	3.61	1.92	0.844	0.849	75.00	71.67
Qwen M3	3	100.0	100.0	100.0	0.073	6.02	2.23	0.850	0.874	85.00	76.67
Qwen M3	4	100.0	100.0	100.0	0.446	8.05	2.35	0.821	0.861	86.67	46.67
Qwen M3	5	100.0	100.0	100.0	0.085	5.67	2.72	0.811	0.857	80.00	41.67
Qwen M3	6	100.0	100.0	100.0	0.088	10.11	3.22	0.795	0.858	91.67	25.00
Qwen M3	7	100.0	100.0	100.0	0.119	13.33	3.40	0.768	0.840	81.67	13.33
InternVL Base	overall	100.0	79.3	96.2	3.174	73.64	60.98	0.082	0.235	0.00	0.00
InternVL Base	1	100.0	95.0	95.0	2.199	77.02	52.51	0.162	0.165	0.00	0.00
InternVL Base	2	100.0	91.7	95.8	2.816	68.79	56.57	0.126	0.186	0.00	0.00
InternVL Base	3	100.0	95.0	97.8	3.454	75.47	50.91	0.098	0.207	0.00	0.00
InternVL Base	4	100.0	76.7	93.3	4.373	76.17	60.47	0.083	0.226	0.00	0.00
InternVL Base	5	100.0	55.0	89.3	2.931	76.96	65.93	0.072	0.252	0.00	0.00
InternVL Base	6	100.0	41.7	98.9	2.966	72.57	67.00	0.074	0.279	0.00	0.00
InternVL Base	7	100.0	100.0	100.0	2.979	71.16	59.57	0.066	0.332	0.00	0.00
InternVL M2	overall	100.0	100.0	100.0	0.145	38.70	3.73	0.711	0.762	39.52	23.33
InternVL M2	1	100.0	100.0	100.0	0.065	29.53	2.17	0.785	0.785	61.67	73.33
InternVL M2	2	100.0	100.0	100.0	0.072	33.26	2.71	0.749	0.754	41.67	45.00
InternVL M2	3	100.0	100.0	100.0	0.082	37.87	2.72	0.729	0.744	35.00	23.33
InternVL M2	4	100.0	100.0	100.0	0.458	37.58	3.48	0.728	0.765	33.33	16.67
InternVL M2	5	100.0	100.0	100.0	0.086	41.15	3.61	0.708	0.757	36.67	5.00
InternVL M2	6	100.0	100.0	100.0	0.097	44.65	4.19	0.689	0.755	25.00	0.00
InternVL M2	7	100.0	100.0	100.0	0.107	35.70	4.51	0.691	0.776	43.33	0.00
InternVL M3	overall	100.0	100.0	100.0	0.081	23.69	3.24	0.763	0.823	66.19	37.86
InternVL M3	1	100.0	100.0	100.0	0.056	7.09	1.95	0.868	0.868	85.00	91.67
InternVL M3	2	100.0	100.0	100.0	0.058	14.24	2.11	0.813	0.820	68.33	61.67
InternVL M3	3	100.0	100.0	100.0	0.073	25.21	2.42	0.780	0.812	63.33	38.33
InternVL M3	4	100.0	100.0	100.0	0.077	15.98	2.75	0.789	0.828	73.33	36.67
InternVL M3	5	100.0	100.0	100.0	0.073	22.92	3.12	0.763	0.812	60.00	21.67
InternVL M3	6	100.0	100.0	100.0	0.086	26.07	3.65	0.748	0.812	63.33	6.67
InternVL M3	7	100.0	100.0	100.0	0.097	31.04	4.10	0.726	0.805	50.00	8.33

E Complete Generalization Evaluations

E.1 Angle Support Shift

The angle experiment uses separate Qwen training runs with training/validation angles approximately in $[-179, -91]^\circ$ and test angles in $[91, 179]^\circ$. The in-support control and OOD test each have

420 scenes, balanced over one to seven pieces. The aggregate angle tables exclude squares, while retaining the evaluator's piece-dependent symmetry convention. In particular, geometric symmetry means disjoint numerical angle intervals are not uniformly disjoint appearance classes. These results establish failure on the tested support shift, not a universal impossibility of rotation reasoning.

Table 15: Angle ID: complete aggregate results. Angle excludes squares; Scale, IoU80 and Joint are percentages.

Model	Scenes	Angle pieces	Pos.	Angle	Scale	Scene IoU	IoU80	Joint
Base	420	1396	3.026	126.75	39.81	0.195	0.00	0.00
M2	420	1431	0.098	2.01	4.26	0.861	87.14	57.62
M3	420	1431	0.084	1.63	2.91	0.882	92.62	70.24

Table 16: Angle OOD: complete aggregate results. Angle excludes squares; Scale, IoU80 and Joint are percentages.

Model	Scenes	Angle pieces	Pos.	Angle	Scale	Scene IoU	IoU80	Joint
Base	420	1405	2.852	104.51	39.41	0.213	0.00	0.00
M2	420	1431	3.534	96.93	4.75	0.577	3.33	3.33
M3	420	1431	0.509	101.05	3.53	0.588	4.29	3.33

The following per-k table reproduces the standard evaluator summaries, whose angle averages include all matched pieces with symmetry-aware error. They therefore should not be averaged to reconstruct the square-excluded angle columns above. Parse, ID, recall, scale, IoU80, and Joint are percentages.

Table 17: Complete Qwen angle-support results by piece count: 60 scenes per k and condition. These standard-evaluator angle errors include squares; the aggregate tables above exclude them. Units follow the adaptation table.

k	Parse	ID	Recall	Pos.	Angle	Scale	Piece IoU	Scene IoU	IoU80	Joint
ID, Base										
1	100.0	81.7	81.7	2.623	117.93	35.71	0.131	0.146	0.00	0.00
2	100.0	85.0	92.5	3.284	118.54	34.15	0.069	0.129	0.00	0.00
3	100.0	83.3	94.4	3.197	111.57	39.59	0.083	0.157	0.00	0.00
4	100.0	71.7	92.9	3.109	113.40	33.24	0.066	0.193	0.00	0.00
5	100.0	60.0	90.0	3.076	118.26	44.86	0.049	0.223	0.00	0.00
6	100.0	58.3	93.1	2.957	114.86	40.63	0.056	0.233	0.00	0.00
7	100.0	100.0	100.0	2.915	112.63	41.47	0.056	0.282	0.00	0.00
ID, M2										
1	100.0	100.0	100.0	0.063	0.67	1.99	0.894	0.894	90.00	95.00
2	100.0	100.0	100.0	0.089	0.90	3.75	0.835	0.859	83.33	78.33
3	100.0	100.0	100.0	0.087	1.81	3.60	0.823	0.858	81.67	60.00
4	100.0	100.0	100.0	0.091	1.80	3.96	0.828	0.864	90.00	66.67
5	100.0	100.0	100.0	0.098	2.09	4.06	0.805	0.855	85.00	48.33
6	100.0	100.0	100.0	0.103	2.18	4.47	0.798	0.855	91.67	38.33
7	100.0	100.0	100.0	0.110	2.42	5.16	0.774	0.840	88.33	16.67
ID, M3										
1	100.0	100.0	100.0	0.050	0.47	1.25	0.905	0.905	91.67	95.00
2	100.0	100.0	100.0	0.067	0.45	2.39	0.865	0.887	90.00	86.67
3	100.0	100.0	100.0	0.072	1.49	2.63	0.858	0.894	95.00	75.00
4	100.0	100.0	100.0	0.080	1.39	2.62	0.840	0.877	86.67	75.00
5	100.0	100.0	100.0	0.082	1.63	2.90	0.835	0.876	96.67	56.67
6	100.0	100.0	100.0	0.087	1.77	3.08	0.824	0.874	98.33	60.00
7	100.0	100.0	100.0	0.100	2.05	3.46	0.801	0.863	90.00	43.33
OOD, Base										
1	100.0	81.7	81.7	2.497	71.58	30.65	0.139	0.153	0.00	0.00
2	100.0	83.3	91.7	2.976	99.80	37.04	0.086	0.143	0.00	0.00
3	100.0	85.0	95.0	2.970	99.45	36.49	0.083	0.176	0.00	0.00

Continued on next page

Table 17 – continued

k	Parse	ID	Recall	Pos.	Angle	Scale	Piece IoU	Scene IoU	IoU80	Joint
4	100.0	80.0	95.0	3.047	94.09	33.51	0.065	0.210	0.00	0.00
5	100.0	58.3	90.3	2.873	100.41	45.31	0.052	0.240	0.00	0.00
6	100.0	50.0	91.7	2.796	96.84	41.39	0.054	0.260	0.00	0.00
7	100.0	100.0	100.0	2.738	91.58	40.09	0.059	0.313	0.00	0.00
OOD, M2										
1	100.0	100.0	100.0	0.098	86.31	2.93	0.566	0.566	20.00	21.67
2	100.0	100.0	100.0	0.753	81.44	3.72	0.539	0.546	3.33	1.67
3	100.0	100.0	100.0	0.705	84.39	4.30	0.536	0.561	0.00	0.00
4	100.0	100.0	100.0	0.131	81.64	4.32	0.553	0.578	0.00	0.00
5	100.0	100.0	100.0	0.434	81.14	4.60	0.544	0.576	0.00	0.00
6	100.0	100.0	100.0	0.562	83.75	5.06	0.544	0.605	0.00	0.00
7	100.0	100.0	100.0	12.739	83.16	5.58	0.527	0.603	0.00	0.00
OOD, M3										
1	100.0	100.0	100.0	0.082	85.48	2.85	0.582	0.582	21.67	21.67
2	100.0	100.0	100.0	1.623	87.62	2.75	0.542	0.550	3.33	1.67
3	100.0	100.0	100.0	0.647	88.45	3.40	0.545	0.571	3.33	0.00
4	100.0	100.0	100.0	0.508	84.70	3.18	0.555	0.588	1.67	0.00
5	100.0	100.0	100.0	0.434	86.26	3.47	0.545	0.579	0.00	0.00
6	100.0	100.0	100.0	0.357	86.13	3.69	0.550	0.614	0.00	0.00
7	100.0	100.0	100.0	0.375	86.01	4.01	0.544	0.631	0.00	0.00

E.2 Unseen Piece Counts

The training set contains $k = 1$–5 only (3,000 training and 250 validation scenes); test scenes have $k = 6, 7$ (60 each). These are the completed complexity-shift results. They do not include seven-to-nine-piece tests.

Table 18: Complete complexity-OOD results for Qwen. Percentages follow the adaptation-table convention.

Model	k	Parse	ID	Recall	Pos.	Angle	Scale	Piece IoU	Scene IoU	IoU80	Joint
Base	overall	100.0	74.2	96.0	2.855	73.54	41.47	0.063	0.286	0.00	0.00
Base	6	100.0	48.3	91.4	2.842	76.65	39.63	0.068	0.275	0.00	0.00
Base	7	100.0	100.0	100.0	2.865	71.12	42.90	0.058	0.298	0.00	0.00
M2	overall	100.0	100.0	100.0	0.141	36.12	5.90	0.665	0.745	22.50	0.83
M2	6	100.0	100.0	100.0	0.125	35.26	5.82	0.672	0.751	25.00	0.00
M2	7	100.0	100.0	100.0	0.155	36.87	5.97	0.659	0.739	20.00	1.67
M3	overall	100.0	100.0	100.0	0.136	19.90	4.51	0.718	0.802	55.83	6.67
M3	6	100.0	100.0	100.0	0.117	18.56	4.35	0.730	0.814	61.67	8.33
M3	7	100.0	100.0	100.0	0.152	21.05	4.63	0.708	0.790	50.00	5.00

E.3 Tangram-to-CanvasPlace Transfer

We evaluate both architectures' original Base/M2/M3 checkpoints on the same 50-task CanvasPlace transfer bank, without placement training, demonstrations, or parameter updates. The four conditions preserve physical objects, constraints, fixed target scale, and the cardinal-angle support: (i) saved native images with top-left origin, downward y, clockwise rotation, and $[0, 1]$ coordinates; (ii) consistently redrawn labels with the same convention; (iii) redrawn labels with bottom-left origin, upward y, counter-clockwise rotation, and $[0, 1]$ coordinates; and (iv) the same aligned convention with range $[0, 10]$. Thus the last two conditions align coordinate axes and rotation direction, and then the coordinate range, without changing the geometric task.

Predictions are mapped to the native evaluator by the fixed transformation $x = x'/a$, $y = 1 - y'/a$, $\theta = (-\theta') \bmod 360$, where $a = 1$ or 10 for the Cartesian conditions. Four hundred round-trip/scoring-invariance checks passed. Decoding is deterministic with a 1,200-token cap, and completed failures are retained. An input audit found different small-target resizing in the saved Qwen Base processor; all 200 Base condition–task evaluations were rerun with the M2/M3 processor and no adapter weights. The formal comparison uses these matched inputs, totaling 1,200 evaluations. The 200 initial unmatched Base records are archived separately.

Table 19: Convention-controlled transfer: successful tasks (parse failures), each out of the same 50 tasks. Native denotes top-left/CW; aligned denotes bottom-left/CCW. All comparisons use matched preprocessing within architecture.

Architecture	Model	Saved native	Redrawn native	Aligned $[0, 1]$	Aligned $[0, 10]$
Qwen	Base	16 (0)	14 (0)	18 (0)	18 (0)
Qwen	M2	13 (0)	16 (0)	16 (0)	13 (0)
Qwen	M3	16 (0)	15 (0)	19 (0)	17 (0)
InternVL	Base	7 (36)	7 (29)	8 (24)	1 (49)
InternVL	M2	22 (0)	21 (0)	24 (0)	20 (0)
InternVL	M3	23 (0)	24 (0)	25 (0)	24 (0)

Qwen shows no consistent aggregate adaptation gain across conventions. Under full alignment, M3 minus Base is -2 percentage points, with a paired-task bootstrap 95% interval of $[-18, 14]$; this does not establish equivalence or absence of transferable information. InternVL improves end-to-end success under every condition, but Base has substantial parse failures, including 49/50 under full alignment, while adapted models have none. Its gains therefore combine output validity and placement performance, rather than identifying pure geometric transfer.

Tables 20–21 use 5,000 paired-task bootstrap resamples for pointwise, descriptive 95% intervals.

Table 20: Qwen paired adapted-minus-Base success differences (percentage points).

Checkpoint	Condition	Difference	95% interval
M2	Saved native	-6.0	[-20.0, 8.0]
M2	Redrawn native	4.0	[-12.0, 18.0]
M2	Aligned $[0, 1]$	-4.0	[-16.0, 8.0]
M2	Aligned $[0, 10]$	-10.0	[-24.0, 4.0]
M3	Saved native	0.0	[-14.0, 14.0]
M3	Redrawn native	2.0	[-12.0, 16.0]
M3	Aligned $[0, 1]$	2.0	[-12.0, 16.0]
M3	Aligned $[0, 10]$	-2.0	[-18.0, 14.0]

Table 21: InternVL paired adapted-minus-Base success differences (percentage points).

Checkpoint	Condition	Difference	95% interval
M2	Saved native	30.0	[16.0, 44.0]
M2	Redrawn native	28.0	[12.0, 44.0]
M2	Aligned $[0, 1]$	32.0	[18.0, 46.0]
M2	Aligned $[0, 10]$	38.0	[24.0, 52.0]
M3	Saved native	32.0	[16.0, 48.0]
M3	Redrawn native	34.0	[18.0, 50.0]
M3	Aligned $[0, 1]$	34.0	[18.0, 50.0]
M3	Aligned $[0, 10]$	46.0	[32.0, 60.0]

The full paired contrasts and adaptation-by-convention interactions accompany the source as `transfer_paired_contrasts.csv`. All main-text transfer claims use these matched comparisons.

F Complete Representation Probes and Behavioral Controls

F.1 Experimental scope and interpretation

Table 22: Complete scope of the mechanistic experiments. "Piece samples" counts target-piece representations; counterfactual and patching pairs differ in exactly one factor of one target piece.

Test	Models	Samples	Question answered
Layer probes	Qwen/InternVL Base, M2, M3	1,680 test pieces	Is local geometry externally decodable at each layer?
Blank-image control	Same	1,680 test pieces	Does decodability depend on pixels rather than the oracle mask/identity alone?
Counterfactual binding	Base, M2, M3	108 pairs/model	Does output selectively follow the changed target attribute?
Oracle isolation	Base, M2, M3	180 targets/model	Does removing other objects suffice for accurate target-pose prediction?
Activation patching	Qwen/InternVL Base/M2/M3	108 pairs/checkpoint	At which stages does target-region state causally influence output?

The representation-probe evaluation uses 840/210/420 train/validation/test scenes, corresponding to 3,360/840/1,680 piece samples, split by source page. Every probe receives a ground-truth polygon-mask pooled representation and ground-truth piece identity but no target pose values, vertices, or explicit token coordinates. The oracle mask supplies localization and can itself carry geometric cues; position decoding is especially sensitive to this assistance. Angle and scale decoding must likewise be interpreted with the matched blank-image controls, which test image dependence beyond the supplied mask and identity. Neither low probe error nor its contrast with autonomous output establishes that the VLM already constructs complete, correctly bound poses. Linear and two-layer MLP probes are trained separately at every site. Lower values are better in all probe tables. M2 freezes the visual encoder and cross-modal interface, so its gains occur without changes to their features. M3 changes the visual tail and interface; unchanged early-layer scores do not establish identical geometry throughout its processing pipeline.

F.2 Complete Layer-Wise Geometry Probes

Table 23: Qwen layer-wise probes, with linear and MLP errors side by side. Each probe receives the ground-truth region and identity. Position is in canvas units, angle in degrees, and scale in percent.

	Linear			MLP		
Layer	Pos.	Angle	Scale	Pos.	Angle	Scale
Base						
`vision.0`	0.526	66.16°	9.48%	0.641	15.56°	7.72%
`vision.4`	0.521	44.65°	4.51%	0.531	12.67°	3.67%
`vision.8`	0.479	21.58°	3.60%	0.540	9.77°	3.05%
`vision.12`	0.692	19.54°	3.75%	0.618	10.24°	2.92%
`vision.16`	0.760	20.07°	3.58%	0.710	10.07°	3.15%
`vision.20`	0.692	20.24°	3.59%	0.689	10.66°	3.36%
`vision.23`	0.697	19.79°	3.62%	0.700	11.16°	3.30%
`vision.24`	0.715	18.33°	3.85%	0.630	10.89°	3.22%
`vision.25`	0.670	19.33°	3.84%	0.630	10.44°	3.18%
`vision.26`	0.893	25.64°	4.25%	0.893	15.85°	3.89%
`vision.final_merger`	1.314	27.11°	7.02%	1.240	20.35°	4.66%
`vision.deepstack_merger.0`	1.093	23.77°	5.77%	1.054	14.80°	4.85%
`vision.deepstack_merger.1`	1.374	18.30°	4.82%	1.056	13.40°	4.17%
`vision.deepstack_merger.2`	1.301	19.74°	5.35%	1.078	16.05°	4.29%
`llm.0`	1.144	23.09°	5.21%	0.852	13.07°	3.94%
`llm.6`	1.319	18.43°	4.68%	0.855	12.75°	3.66%
`llm.12`	1.137	21.73°	5.77%	0.907	16.66°	4.19%
`llm.18`	1.509	26.55°	5.61%	0.828	22.63°	4.51%
`llm.24`	1.246	29.85°	5.64%	0.941	27.86°	5.29%

Continued on next page

Table 23 – continued

	Linear			MLP		
Layer	Pos.	Angle	Scale	Pos.	Angle	Scale
llm.30	1.284	31.72°	5.76%	1.261	35.87°	5.83%
llm.35	1.558	38.47°	6.92%	1.383	38.51°	6.32%
M2						
vision.0	0.526	66.16°	9.48%	0.641	15.56°	7.72%
vision.4	0.521	44.65°	4.51%	0.531	12.67°	3.67%
vision.8	0.479	21.58°	3.60%	0.540	9.77°	3.05%
vision.12	0.692	19.54°	3.75%	0.618	10.24°	2.92%
vision.16	0.760	20.07°	3.58%	0.710	10.07°	3.15%
vision.20	0.692	20.24°	3.59%	0.689	10.66°	3.36%
vision.23	0.697	19.79°	3.62%	0.700	11.16°	3.30%
vision.24	0.715	18.33°	3.85%	0.630	10.89°	3.22%
vision.25	0.670	19.33°	3.84%	0.630	10.44°	3.18%
vision.26	0.893	25.64°	4.25%	0.893	15.85°	3.89%
vision.final_merger	1.314	27.11°	7.02%	1.240	20.35°	4.66%
vision.deepstack_merger.0	1.093	23.77°	5.77%	1.054	14.80°	4.85%
vision.deepstack_merger.1	1.374	18.30°	4.82%	1.056	13.40°	4.17%
vision.deepstack_merger.2	1.301	19.74°	5.35%	1.078	16.05°	4.29%
llm.0	1.130	23.26°	5.25%	0.865	13.49°	3.94%
llm.6	1.282	18.38°	4.82%	0.835	12.06°	3.67%
llm.12	1.191	20.03°	4.94%	0.847	14.91°	4.06%
llm.18	1.163	22.78°	4.88%	0.800	15.91°	4.02%
llm.24	0.944	24.25°	4.82%	0.545	20.74°	4.56%
llm.30	1.121	24.71°	5.98%	0.685	23.06°	4.62%
llm.35	1.331	32.67°	6.18%	0.787	31.16°	5.63%
M3						
vision.0	0.526	66.16°	9.48%	0.641	15.56°	7.72%
vision.4	0.521	44.65°	4.51%	0.531	12.67°	3.67%
vision.8	0.479	21.58°	3.60%	0.540	9.77°	3.05%
vision.12	0.692	19.54°	3.75%	0.618	10.24°	2.92%
vision.16	0.760	20.07°	3.58%	0.710	10.07°	3.15%
vision.20	0.692	20.24°	3.59%	0.689	10.66°	3.36%
vision.23	0.698	19.74°	3.62%	0.701	11.18°	3.32%
vision.24	0.720	18.28°	3.82%	0.641	10.51°	3.23%
vision.25	0.665	19.24°	3.78%	0.635	10.13°	3.21%
vision.26	0.822	24.21°	4.06%	1.066	16.53°	4.35%
vision.final_merger	1.221	28.38°	5.81%	1.140	18.06°	4.66%
vision.deepstack_merger.0	1.186	23.25°	5.39%	1.004	14.18°	4.83%
vision.deepstack_merger.1	1.317	18.73°	4.55%	0.959	12.54°	3.89%
vision.deepstack_merger.2	1.332	19.11°	5.19%	1.110	16.07°	4.28%
llm.0	1.092	20.14°	4.65%	0.791	12.45°	3.66%
llm.6	1.053	17.47°	5.06%	0.868	12.66°	3.45%
llm.12	1.039	18.28°	4.91%	0.772	12.52°	3.49%
llm.18	1.136	18.63°	4.70%	0.874	13.29°	3.42%
llm.24	1.035	20.29°	4.75%	0.587	16.75°	4.10%
llm.30	1.168	21.15°	5.21%	0.706	18.36°	4.15%
llm.35	1.446	26.04°	5.15%	0.777	22.18°	5.28%

Table 24: InternVL layer-wise probes, with linear and MLP errors side by side. Each probe receives the ground-truth region and identity. Position is in canvas units, angle in degrees, and scale in percent.

	Linear			MLP		
Layer	Pos.	Angle	Scale	Pos.	Angle	Scale
Base						
vision.0	2.158	66.63°	9.09%	2.028	25.82°	7.43%
vision.4	1.601	61.06°	5.75%	1.407	14.28°	4.78%
vision.8	1.393	22.98°	4.65%	1.268	11.51°	4.02%

Continued on next page

Table 24 – continued

	Linear			MLP		
Layer	Pos.	Angle	Scale	Pos.	Angle	Scale
vision.12	1.448	23.32°	4.58%	1.270	11.26°	3.89%
vision.16	1.368	27.32°	4.56%	1.290	12.32°	4.33%
vision.20	1.365	27.86°	4.87%	1.300	13.45°	4.27%
vision.23	1.332	29.89°	5.17%	1.238	14.23°	4.44%
projector.final	1.614	29.90°	6.42%	1.568	23.44°	5.73%
llm.0	1.356	28.04°	6.22%	1.428	19.39°	5.43%
llm.6	1.231	24.74°	5.33%	1.192	21.23°	4.64%
llm.12	1.333	26.88°	5.00%	0.964	22.14°	4.43%
llm.18	1.345	35.26°	5.52%	1.043	32.52°	4.73%
llm.24	1.378	36.21°	5.85%	0.870	28.00°	4.41%
llm.30	1.544	39.04°	6.70%	1.182	38.44°	5.07%
llm.35	1.788	43.85°	6.93%	1.433	40.89°	5.16%
M2						
vision.0	2.158	66.63°	9.09%	2.028	25.82°	7.43%
vision.4	1.601	61.06°	5.75%	1.407	14.28°	4.78%
vision.8	1.393	22.98°	4.65%	1.268	11.51°	4.02%
vision.12	1.448	23.32°	4.58%	1.270	11.26°	3.89%
vision.16	1.368	27.32°	4.56%	1.290	12.32°	4.33%
vision.20	1.365	27.86°	4.87%	1.300	13.45°	4.27%
vision.23	1.332	29.89°	5.17%	1.238	14.23°	4.44%
projector.final	1.614	29.90°	6.42%	1.568	23.44°	5.73%
llm.0	1.353	28.03°	6.22%	1.421	19.18°	5.41%
llm.6	1.180	25.51°	5.17%	1.055	18.15°	4.67%
llm.12	1.278	27.54°	5.74%	0.666	20.52°	4.12%
llm.18	1.129	26.83°	5.12%	0.769	21.97°	4.18%
llm.24	1.197	28.07°	5.20%	0.587	20.00°	4.27%
llm.30	1.268	33.91°	5.80%	0.688	28.42°	4.82%
llm.35	1.653	40.28°	6.70%	0.905	36.80°	5.37%
M3						
vision.0	2.158	66.63°	9.09%	2.028	25.82°	7.43%
vision.4	1.601	61.06°	5.75%	1.407	14.28°	4.78%
vision.8	1.393	22.98°	4.65%	1.268	11.51°	4.02%
vision.12	1.448	23.32°	4.58%	1.270	11.26°	3.89%
vision.16	1.368	27.32°	4.56%	1.290	12.32°	4.33%
vision.20	1.364	27.85°	4.86%	1.306	13.33°	4.22%
vision.23	1.295	28.79°	4.92%	1.254	14.13°	4.30%
projector.final	1.601	29.70°	6.17%	1.648	24.76°	5.41%
llm.0	1.473	29.11°	6.65%	1.555	22.49°	6.01%
llm.6	1.099	25.72°	4.84%	1.162	19.43°	4.58%
llm.12	1.202	28.28°	5.02%	0.731	19.31°	4.04%
llm.18	1.196	28.53°	4.44%	0.743	19.19°	3.84%
llm.24	1.158	28.18°	5.11%	0.650	18.36°	3.60%
llm.30	1.287	31.37°	5.32%	0.907	24.16°	4.63%
llm.35	1.696	37.31°	6.38%	0.957	32.31°	4.95%

F.3 BLANK-IMAGE CONTROLS

We replace the input image with a blank canvas while retaining the ground-truth target mask and piece identity. Each table compares MLP probe errors for real and blank images at the same layer, with prompt and split also fixed. Angle and scale decoding sharply worsen, showing that the probes use image content beyond these supplied cues. Position can remain deceptively good because the pooling mask itself locates the target polygon.

Table 25: Qwen blank-image controls. Each cell gives real / blank MLP error with mask, identity, prompt, and split fixed. Position is in canvas units, angle in degrees, and scale in percent.

Layer	Position real / blank	Angle real / blank	Scale real / blank
Base			
vision.0	0.641 / 0.563	15.56 / 68.46	7.72% / 6.77%
vision.4	0.531 / 0.630	12.67 / 68.13	3.67% / 5.98%
vision.8	0.540 / 0.578	9.77 / 69.71	3.05% / 6.37%
vision.12	0.618 / 0.537	10.24 / 69.24	2.92% / 7.31%
vision.16	0.710 / 0.435	10.07 / 68.56	3.15% / 8.52%
vision.20	0.689 / 0.470	10.66 / 67.33	3.36% / 8.27%
vision.23	0.700 / 0.605	11.16 / 69.11	3.30% / 8.35%
vision.24	0.630 / 0.486	10.89 / 70.06	3.22% / 9.39%
vision.25	0.630 / 0.548	10.44 / 71.10	3.18% / 9.72%
vision.26	0.893 / 0.631	15.85 / 71.72	3.89% / 17.86%
vision.final_merger	1.240 / 1.018	20.35 / 69.91	4.66% / 14.97%
vision.deepstack_merger.0	1.054 / 1.148	14.80 / 69.15	4.85% / 15.99%
vision.deepstack_merger.1	1.056 / 1.030	13.40 / 71.92	4.17% / 12.68%
vision.deepstack_merger.2	1.078 / 0.844	16.05 / 71.22	4.29% / 12.33%
llm.0	0.852 / 0.919	13.07 / 71.70	3.94% / 12.79%
llm.6	0.855 / 0.772	12.75 / 68.19	3.66% / 11.39%
llm.12	0.907 / 0.833	16.66 / 68.49	4.19% / 10.94%
llm.18	0.828 / 0.921	22.63 / 69.62	4.51% / 11.28%
llm.24	0.941 / 0.950	27.86 / 70.01	5.29% / 11.57%
llm.30	1.261 / 0.933	35.87 / 69.06	5.83% / 12.11%
llm.35	1.383 / 0.993	38.51 / 70.71	6.32% / 13.74%
M2			
vision.0	0.641 / 0.563	15.56 / 68.46	7.72% / 6.77%
vision.4	0.531 / 0.630	12.67 / 68.13	3.67% / 5.98%
vision.8	0.540 / 0.578	9.77 / 69.71	3.05% / 6.37%
vision.12	0.618 / 0.537	10.24 / 69.24	2.92% / 7.31%
vision.16	0.710 / 0.435	10.07 / 68.56	3.15% / 8.52%
vision.20	0.689 / 0.470	10.66 / 67.33	3.36% / 8.27%
vision.23	0.700 / 0.605	11.16 / 69.11	3.30% / 8.35%
vision.24	0.630 / 0.486	10.89 / 70.06	3.22% / 9.39%
vision.25	0.630 / 0.548	10.44 / 71.10	3.18% / 9.72%
vision.26	0.893 / 0.631	15.85 / 71.72	3.89% / 17.86%
vision.final_merger	1.240 / 1.018	20.35 / 69.91	4.66% / 14.97%
vision.deepstack_merger.0	1.054 / 1.148	14.80 / 69.15	4.85% / 15.99%
vision.deepstack_merger.1	1.056 / 1.030	13.40 / 71.92	4.17% / 12.68%
vision.deepstack_merger.2	1.078 / 0.844	16.05 / 71.22	4.29% / 12.33%
llm.0	0.865 / 0.939	13.49 / 70.53	3.94% / 13.53%
llm.6	0.835 / 0.853	12.06 / 68.84	3.67% / 11.86%
llm.12	0.847 / 0.921	14.91 / 70.43	4.06% / 12.91%
llm.18	0.800 / 0.878	15.91 / 69.39	4.02% / 11.19%
llm.24	0.545 / 1.160	20.74 / 70.13	4.56% / 14.28%
llm.30	0.685 / 1.357	23.06 / 72.81	4.62% / 15.18%
llm.35	0.787 / 0.880	31.16 / 68.16	5.63% / 11.17%
M3			
vision.0	0.641 / 0.563	15.56 / 68.46	7.72% / 6.77%
vision.4	0.531 / 0.630	12.67 / 68.13	3.67% / 5.98%
vision.8	0.540 / 0.578	9.77 / 69.71	3.05% / 6.37%
vision.12	0.618 / 0.537	10.24 / 69.24	2.92% / 7.31%
vision.16	0.710 / 0.435	10.07 / 68.56	3.15% / 8.52%
vision.20	0.689 / 0.470	10.66 / 67.33	3.36% / 8.27%
vision.23	0.701 / 0.464	11.18 / 68.18	3.32% / 8.39%
vision.24	0.641 / 0.601	10.51 / 67.09	3.23% / 9.74%
vision.25	0.635 / 0.516	10.13 / 69.04	3.21% / 9.03%
vision.26	1.066 / 0.691	16.53 / 71.24	4.35% / 17.04%
vision.final_merger	1.140 / 1.569	18.06 / 71.52	4.66% / 15.76%
vision.deepstack_merger.0	1.004 / 0.721	14.18 / 71.24	4.83% / 12.15%

Continued on next page

Table 25 – continued

Layer	Position real / blank	Angle real / blank	Scale real / blank
vision.deepstack_merger.1	0.959 / 0.922	12.54 / 69.73	3.89% / 11.42%
vision.deepstack_merger.2	1.110 / 0.855	16.07 / 69.41	4.28% / 13.17%
llm.0	0.791 / 1.046	12.45 / 69.00	3.66% / 14.24%
llm.6	0.868 / 1.067	12.66 / 69.18	3.45% / 11.85%
llm.12	0.772 / 0.816	12.52 / 69.51	3.49% / 10.86%
llm.18	0.874 / 0.867	13.29 / 70.51	3.42% / 13.97%
llm.24	0.587 / 1.252	16.75 / 69.54	4.10% / 12.76%
llm.30	0.706 / 1.080	18.36 / 69.30	4.15% / 13.76%
llm.35	0.777 / 1.037	22.18 / 69.63	5.28% / 12.64%

Table 26: InternVL blank-image controls. Each cell gives real / blank MLP error with mask, identity, prompt, and split fixed. Position is in canvas units, angle in degrees, and scale in percent.

Layer	Position real / blank	Angle real / blank	Scale real / blank
Base			
vision.0	2.028 / 1.335	25.82 / 69.16	7.43% / 8.48%
vision.4	1.407 / 1.153	14.28 / 69.34	4.78% / 6.67%
vision.8	1.268 / 1.071	11.51 / 68.26	4.02% / 6.48%
vision.12	1.270 / 0.952	11.26 / 69.08	3.89% / 6.12%
vision.16	1.290 / 1.043	12.32 / 69.55	4.33% / 6.90%
vision.20	1.300 / 1.068	13.45 / 69.55	4.27% / 7.16%
vision.23	1.238 / 1.044	14.23 / 71.10	4.44% / 7.37%
projector.final	1.568 / 1.551	23.44 / 69.08	5.73% / 10.66%
llm.0	1.428 / 1.524	19.39 / 72.50	5.43% / 18.88%
llm.6	1.192 / 1.582	21.23 / 71.93	4.64% / 16.82%
llm.12	0.964 / 0.973	22.14 / 68.90	4.43% / 9.16%
llm.18	1.043 / 1.011	32.52 / 71.79	4.73% / 13.60%
llm.24	0.870 / 1.325	28.00 / 69.33	4.41% / 11.33%
llm.30	1.182 / 1.031	38.44 / 69.79	5.07% / 10.37%
llm.35	1.433 / 0.918	40.89 / 69.91	5.16% / 10.67%
M2			
vision.0	2.028 / 1.335	25.82 / 69.16	7.43% / 8.48%
vision.4	1.407 / 1.153	14.28 / 69.34	4.78% / 6.67%
vision.8	1.268 / 1.071	11.51 / 68.26	4.02% / 6.48%
vision.12	1.270 / 0.952	11.26 / 69.08	3.89% / 6.12%
vision.16	1.290 / 1.043	12.32 / 69.55	4.33% / 6.90%
vision.20	1.300 / 1.068	13.45 / 69.55	4.27% / 7.16%
vision.23	1.238 / 1.044	14.23 / 71.10	4.44% / 7.37%
projector.final	1.568 / 1.551	23.44 / 69.08	5.73% / 10.66%
llm.0	1.421 / 1.584	19.18 / 72.44	5.41% / 19.08%
llm.6	1.055 / 1.666	18.15 / 70.02	4.67% / 17.03%
llm.12	0.666 / 0.900	20.52 / 67.98	4.12% / 9.90%
llm.18	0.769 / 0.717	21.97 / 68.47	4.18% / 10.36%
llm.24	0.587 / 0.856	20.00 / 68.01	4.27% / 8.77%
llm.30	0.688 / 0.905	28.42 / 68.68	4.82% / 9.70%
llm.35	0.905 / 0.925	36.80 / 70.27	5.37% / 9.39%
M3			
vision.0	2.028 / 1.335	25.82 / 69.16	7.43% / 8.48%
vision.4	1.407 / 1.153	14.28 / 69.34	4.78% / 6.67%
vision.8	1.268 / 1.071	11.51 / 68.26	4.02% / 6.48%
vision.12	1.270 / 0.952	11.26 / 69.08	3.89% / 6.12%
vision.16	1.290 / 1.043	12.32 / 69.55	4.33% / 6.90%
vision.20	1.306 / 1.096	13.33 / 69.95	4.22% / 7.40%
vision.23	1.254 / 1.143	14.13 / 69.62	4.30% / 7.78%
projector.final	1.648 / 2.276	24.76 / 72.63	5.41% / 18.84%
llm.0	1.555 / 1.119	22.49 / 69.77	6.01% / 11.39%
llm.6	1.162 / 1.159	19.43 / 68.35	4.58% / 10.92%
llm.12	0.731 / 0.970	19.31 / 67.97	4.04% / 9.59%

Continued on next page

Table 26 – continued

Layer	Position real / blank	Angle real / blank	Scale real / blank
`llm.18`	0.743 / 0.886	19.19 / 68.41	3.84% / 10.35%
`llm.24`	0.650 / 0.827	18.36 / 68.18	3.60% / 9.40%
`llm.30`	0.907 / 0.951	24.16 / 69.16	4.63% / 10.83%
`llm.35`	0.957 / 1.203	32.31 / 71.38	4.95% / 14.85%

F.4 Object-Level Counterfactual Binding

In 108 image pairs, we change only one piece's position, angle, or scale and compare the model's JSON predictions before and after the edit.

Signed projection is one when the predicted target change has exactly the desired magnitude and direction. Change-vector error is zero at the ideal response; alignment is one when the direction is correct; collateral change is zero when non-target pieces remain unchanged.

M2/M3 track position and scale changes more accurately than Base, with smaller unintended changes in other pieces. Qwen also tracks rotation more accurately; InternVL's rotation responses remain unreliable. This indicates more selective responses to the edited object, consistent with improved object-specific use of visual information. However, these responses do not isolate an internal binding operation or distinguish it from improvements in numerical output and instruction following.

Table 27: Qwen: all counterfactual-binding summary metrics.

Model	Factor	Pairs	Parse	Signed proj.	Change err. ↓	Alignment	Collateral ↓
Base	Angle	36	1.000	-0.064	1.064	-0.111	0.261
Base	Position	36	0.972	0.048	1.438	-0.018	0.785
Base	Scale	36	1.000	0.081	0.972	0.083	0.250
M2	Angle	36	1.000	0.912	0.304	0.944	0.195
M2	Position	36	1.000	0.990	0.130	0.992	0.066
M2	Scale	36	1.000	0.975	0.119	1.000	0.061
M3	Angle	36	1.000	0.864	0.243	0.889	0.107
M3	Position	36	1.000	1.000	0.090	0.996	0.060
M3	Scale	36	1.000	0.961	0.112	1.000	0.044

Table 28: InternVL: all counterfactual-binding summary metrics.

Model	Factor	Pairs	Parse	Signed proj.	Change err. ↓	Alignment	Collateral ↓
Base	Angle	36	1.000	-0.108	1.108	-0.056	0.299
Base	Position	36	0.944	0.031	2.328	-0.055	1.089
Base	Scale	36	0.972	0.088	1.079	0.057	0.219
M2	Angle	36	1.000	0.398	1.713	0.444	0.473
M2	Position	36	1.000	0.966	0.118	0.996	0.054
M2	Scale	36	1.000	0.982	0.095	1.000	0.068
M3	Angle	36	1.000	0.272	1.355	0.556	0.311
M3	Position	36	1.000	0.984	0.117	0.994	0.053
M3	Scale	36	1.000	0.963	0.096	1.000	0.047

F.5 Oracle Target Isolation

For 180 targets, we remove every other piece while preserving the target's absolute pose. This is a target-level test, not a new $k = 1$ benchmark. Both Base models still have zero target-pose success, so removing clutter alone is insufficient under this protocol.

Table 29: Qwen: complete full-scene / oracle-isolated target results.

Model	n	ID match	Position	Angle	Scale	Piece IoU	Pose success
Base	180	87.2% / 75.0%	3.055 / 2.610	67.28 / 80.41	40.68% / 51.45%	0.066 / 0.089	0.0% / 0.0%
M2	180	100.0% / 100.0%	0.138 / 0.563	12.31 / 9.43	4.12% / 3.11%	0.773 / 0.807	73.9% / 77.8%
M3	180	100.0% / 100.0%	0.130 / 0.582	9.71 / 5.68	3.02% / 2.12%	0.791 / 0.840	80.6% / 85.0%

Table 30: InternVL: complete full-scene / oracle-isolated target results.

Model	n	ID match	Position	Angle	Scale	Piece IoU	Pose success
Base	180	96.7% / 96.7%	3.343 / 2.380	63.41 / 66.35	59.91% / 91.69%	0.069 / 0.113	0.0% / 0.0%
M2	180	100.0% / 100.0%	0.093 / 0.064	28.27 / 23.47	3.50% / 2.83%	0.719 / 0.767	57.8% / 70.6%
M3	180	100.0% / 100.0%	0.084 / 0.051	26.69 / 12.46	3.12% / 2.56%	0.748 / 0.813	66.1% / 78.3%

G Complete Target-Region Activation Interventions

Matched expanded protocol. The expanded analysis replaces the earlier 18-pair pilot with the full 108-pair pool: 36 pairs per manipulated attribute, spanning $k = 3, 5, 7$. All six Base/M2/M3 checkpoints reuse the same pairs. Qwen is scanned at 21 sites and InternVL at 15, yielding 11,664 separate single-site interventions; neither these interventions nor the 648 checkpoint–pair combinations are independent scenes. Each attribute spans only 13–14 source pages.

Clean target-region activations replace changed-run activations within the union of the clean and changed target polygons; language sites patch corresponding image-token positions during prefill. The geometric errors use the clean target as reference. Up to seven independent single-site conditions are batched with unchanged clean/changed baseline rows, using one common batching protocol and a 700-token cap. All six self-replacement checks passed; baseline consistency across chunks is asserted. A serial-versus-batched smoke audit was not bitwise identical (maximum observed InternVL position-effect difference approximately 0.023 canvas units), so all reported effects use freshly computed baselines under the common batch protocol.

Effects, uncertainty, and missing outputs. The primary outcome is $\Delta e = e_{\mathrm{d}} - e_{\mathrm{p}}$ (Eq. 5), without a positive-denominator filter. Tables 32–37 report mean Δe [pointwise 95% source-page bootstrap interval] (numeric n). Tables 38–41 report paired adapted-minus-Base effect differences in the same format, using shared numeric cases. All intervals use 5,000 bootstrap resamples of source pages. Position, angle, and scale effects are in canvas units, degrees, and percentage points of relative scale error, respectively. Positive effects reduce error toward the clean target and need not recover the exact ground-truth pose. Per-site samples range from 32 to 36 per attribute; parse or target-output failures remain excluded from numeric effects. Repeated edits from one page and multiple sites are not resampled as independent observations. These intervals are descriptive, without simultaneous correction across sites, and cannot identify unique critical layers.

For comparison with the pilot, let e_{c} denote the clean run's error relative to the clean target. Normalized recovery remains

$$R = \frac{e_{\mathrm{d}} - e_{\mathrm{p}}}{e_{\mathrm{d}} - e_{\mathrm{c}}}, \qquad e_{\mathrm{d}} - e_{\mathrm{c}} > 10^{-4}. \tag{8}$$

Here $R = 1$ restores clean-run error and $R > 1$ indicates lower error than the clean run. Small positive denominators amplify the ratio. The table below shows that expanding the candidate pool does not remove Base denominator attrition. Ratios for Base, especially angle and scale, should not be treated as reliable layer profiles.

Table 31: Coverage from 36 candidates per attribute: numeric clean/changed baselines / positive denominators above 10^{-4}. Per-site numeric patch counts are reported in the following tables.

Architecture	Checkpoint	Position	Angle	Scale
Qwen	Base	35 / 10	36 / 5	36 / 2
Qwen	M2	36 / 36	36 / 32	36 / 35
Qwen	M3	36 / 36	36 / 36	36 / 36
InternVL	Base	34 / 9	36 / 4	35 / 2
InternVL	M2	36 / 36	36 / 24	36 / 35
InternVL	M3	36 / 36	36 / 30	36 / 36

Reading the complete profiles. Both architectures' adapted checkpoints show positive position and scale effects across visual and early/middle language sites. Qwen M3 has clear descriptive angle effects at visual layers 4/8 and language layers 6/12/18; InternVL angle intervals remain wide. Late image-token effects approach zero, which does not exclude geometry at other token positions or in the generation state. Paired adapted-minus-Base effects support intervention-dependent changes, but neither these contrasts nor the ratio curves prove that adaptation repaired a particular layer or a necessary-and-sufficient circuit.

Table 32: Qwen Base: error reduction across all tested sites (shared conventions).

Site	Position	Angle	Scale
`vision.0`	0.015 [-0.121, 0.144] (33)	-0.7 [-8.2, 8.5] (36)	-2.6 [-8.1, 3.0] (35)
`vision.4`	0.119 [-0.040, 0.289] (32)	-6.7 [-17.8, 4.1] (36)	-1.0 [-6.0, 3.1] (35)
`vision.8`	0.122 [-0.037, 0.298] (35)	-5.2 [-12.0, -0.5] (36)	-0.4 [-4.3, 3.4] (35)
`vision.12`	0.100 [-0.007, 0.190] (35)	0.9 [-2.6, 5.7] (36)	2.1 [-0.1, 5.9] (36)
`vision.16`	0.069 [-0.085, 0.219] (35)	-3.3 [-7.9, -0.6] (36)	-1.7 [-6.5, 3.0] (36)
`vision.20`	-0.013 [-0.146, 0.112] (35)	-7.5 [-15.6, -0.3] (36)	-1.5 [-5.4, 1.4] (36)
`vision.23`	-0.042 [-0.188, 0.109] (35)	-0.1 [-7.2, 7.4] (36)	-3.5 [-10.3, 1.5] (36)
`vision.24`	0.009 [-0.161, 0.156] (35)	-3.7 [-10.6, 1.0] (36)	1.3 [-2.1, 5.5] (36)
`vision.25`	-0.004 [-0.171, 0.152] (35)	-0.2 [-7.2, 7.3] (36)	-1.3 [-6.7, 4.3] (36)
`vision.26`	-0.029 [-0.152, 0.067] (35)	-2.9 [-7.9, 0.3] (36)	0.2 [-6.6, 8.3] (36)
`final_merger`	-0.021 [-0.147, 0.116] (35)	-2.1 [-3.9, -0.3] (36)	0.9 [-5.7, 8.7] (36)
`deepstack.0`	0.000 [-0.071, 0.063] (35)	-1.5 [-3.4, -0.2] (36)	3.0 [0.0, 8.5] (36)
`deepstack.1`	-0.020 [-0.148, 0.133] (34)	1.9 [-2.3, 8.5] (36)	1.6 [-3.1, 7.6] (36)
`deepstack.2`	0.024 [-0.072, 0.100] (35)	-1.5 [-10.4, 7.3] (36)	2.3 [-3.2, 9.6] (36)
`llm.0`	0.004 [-0.110, 0.105] (35)	-7.7 [-14.7, -1.5] (36)	-1.7 [-4.3, 0.3] (36)
`llm.6`	0.016 [-0.092, 0.130] (35)	1.5 [-3.5, 8.5] (36)	-0.7 [-7.9, 7.7] (36)
`llm.12`	0.039 [-0.071, 0.142] (35)	-0.6 [-2.7, 1.3] (36)	-0.2 [-7.2, 7.9] (36)
`llm.18`	-0.026 [-0.125, 0.068] (34)	2.2 [-2.2, 9.1] (36)	-0.6 [-6.5, 6.3] (36)
`llm.24`	-0.044 [-0.146, 0.060] (35)	-0.8 [-2.5, 0.0] (36)	1.0 [-5.4, 9.2] (36)
`llm.30`	0.007 [-0.068, 0.094] (35)	-0.9 [-2.2, 0.0] (36)	0.7 [-5.3, 7.1] (36)
`llm.35`	0.000 [0.000, 0.000] (35)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 33: Qwen M2: error reduction across all tested sites (shared conventions).

Site	Position	Angle	Scale
vision.0	0.470 [0.426, 0.513] (36)	-5.9 [-28.2, 15.9] (36)	16.5 [14.5, 18.3] (36)
vision.4	0.464 [0.431, 0.495] (36)	18.6 [6.0, 31.5] (36)	15.6 [13.8, 17.3] (36)
vision.8	0.423 [0.394, 0.452] (36)	9.8 [-2.8, 20.7] (36)	15.6 [13.9, 17.3] (36)
vision.12	0.348 [0.324, 0.371] (36)	-5.2 [-17.3, 7.3] (36)	13.0 [11.3, 14.8] (36)
vision.16	0.333 [0.306, 0.358] (36)	-2.7 [-15.1, 8.4] (36)	12.8 [11.2, 14.7] (36)
vision.20	0.250 [0.222, 0.274] (36)	-6.5 [-19.5, 5.5] (36)	10.3 [9.2, 11.5] (36)
vision.23	0.240 [0.212, 0.264] (36)	-2.4 [-11.6, 6.5] (36)	10.4 [9.3, 11.6] (36)
vision.24	0.230 [0.199, 0.257] (36)	-3.2 [-13.4, 6.9] (36)	10.6 [9.5, 11.7] (36)
vision.25	0.216 [0.185, 0.244] (36)	1.1 [-8.5, 10.2] (36)	9.8 [8.9, 10.8] (36)
vision.26	0.208 [0.179, 0.235] (36)	1.6 [-5.1, 8.8] (36)	9.4 [8.2, 10.7] (36)
final_merger	0.219 [0.190, 0.242] (36)	2.3 [-6.2, 10.5] (36)	9.0 [7.9, 10.2] (36)
deepstack.0	0.044 [0.032, 0.054] (36)	-2.7 [-15.0, 9.3] (36)	1.9 [1.4, 2.5] (36)
deepstack.1	0.049 [0.038, 0.058] (36)	1.4 [-3.4, 7.7] (36)	2.9 [2.5, 3.3] (36)
deepstack.2	0.010 [0.003, 0.017] (36)	-0.1 [-0.9, 0.7] (36)	1.1 [0.5, 1.6] (36)
llm.0	0.218 [0.190, 0.243] (36)	-2.8 [-15.0, 7.5] (36)	9.1 [8.0, 10.3] (36)
llm.6	0.362 [0.330, 0.389] (36)	2.1 [-9.0, 12.3] (36)	14.4 [12.5, 16.5] (36)
llm.12	0.354 [0.328, 0.378] (36)	1.4 [-10.5, 12.3] (36)	14.2 [12.4, 16.3] (36)
llm.18	0.301 [0.269, 0.327] (36)	-2.9 [-14.1, 7.5] (36)	9.6 [8.5, 10.8] (36)
llm.24	0.029 [0.014, 0.047] (36)	-3.4 [-8.9, 1.6] (36)	0.0 [-0.4, 0.3] (36)
llm.30	0.001 [-0.006, 0.007] (36)	-1.1 [-3.9, 1.1] (36)	-0.1 [-0.4, 0.2] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 34: Qwen M3: error reduction across all tested sites (shared conventions).

Site	Position	Angle	Scale
vision.0	0.460 [0.433, 0.487] (36)	6.0 [-8.8, 22.7] (36)	18.1 [16.5, 19.7] (36)
vision.4	0.488 [0.460, 0.517] (36)	23.6 [13.1, 30.3] (36)	19.0 [17.5, 20.5] (36)
vision.8	0.486 [0.457, 0.514] (36)	25.6 [19.2, 29.4] (36)	18.8 [17.4, 20.4] (36)
vision.12	0.346 [0.319, 0.372] (36)	-0.6 [-5.8, 4.4] (36)	13.5 [12.4, 14.6] (36)
vision.16	0.337 [0.315, 0.359] (36)	-1.5 [-7.7, 4.7] (36)	13.1 [12.1, 14.2] (36)
vision.20	0.158 [0.137, 0.179] (36)	1.9 [0.8, 3.0] (36)	7.0 [6.4, 7.6] (36)
vision.23	0.155 [0.134, 0.176] (36)	1.3 [0.7, 2.0] (36)	7.4 [6.8, 8.0] (36)
vision.24	0.155 [0.134, 0.176] (36)	1.1 [0.2, 1.9] (36)	7.2 [6.6, 7.8] (36)
vision.25	0.121 [0.104, 0.139] (36)	2.4 [0.9, 4.5] (36)	5.9 [5.3, 6.4] (36)
vision.26	0.118 [0.101, 0.134] (36)	2.4 [0.9, 4.6] (36)	5.7 [5.2, 6.2] (36)
final_merger	0.122 [0.103, 0.140] (36)	2.3 [0.8, 4.5] (36)	5.6 [4.9, 6.3] (36)
deepstack.0	0.092 [0.077, 0.105] (36)	3.6 [-3.0, 7.6] (36)	5.3 [4.7, 6.0] (36)
deepstack.1	0.097 [0.076, 0.118] (36)	-0.1 [-1.7, 1.1] (36)	5.6 [5.0, 6.2] (36)
deepstack.2	0.020 [0.008, 0.031] (36)	0.2 [0.0, 0.5] (36)	1.4 [1.1, 1.8] (36)
llm.0	0.121 [0.101, 0.139] (36)	2.3 [0.8, 4.5] (36)	5.4 [4.7, 6.1] (36)
llm.6	0.445 [0.409, 0.478] (36)	18.8 [15.3, 22.1] (36)	18.7 [17.1, 20.3] (36)
llm.12	0.433 [0.400, 0.465] (36)	21.6 [18.4, 24.4] (36)	18.4 [16.7, 20.1] (36)
llm.18	0.371 [0.330, 0.408] (36)	21.4 [17.0, 25.4] (36)	14.1 [12.9, 15.3] (36)
llm.24	0.036 [0.028, 0.045] (36)	0.7 [-1.1, 2.3] (36)	0.4 [0.2, 0.7] (36)
llm.30	0.002 [-0.002, 0.007] (36)	0.0 [-0.2, 0.3] (36)	-0.1 [-0.4, 0.2] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 35: InternVL Base: error reduction across all tested sites (shared conventions).

Site	Position	Angle	Scale
vision.0	0.128 [-0.320, 0.517] (35)	-2.1 [-8.1, 5.2] (35)	-2.6 [-12.5, 6.5] (35)
vision.4	0.158 [-0.109, 0.384] (34)	-0.2 [-4.5, 5.0] (36)	-5.8 [-14.3, 0.0] (35)
vision.8	0.100 [-0.108, 0.270] (36)	-0.7 [-3.9, 2.8] (36)	-5.5 [-13.9, 0.0] (35)
vision.12	0.052 [-0.137, 0.231] (36)	-0.7 [-7.2, 4.4] (36)	-4.3 [-13.4, 3.0] (35)
vision.16	0.033 [-0.119, 0.181] (36)	0.4 [-2.1, 3.5] (36)	-5.4 [-13.9, 0.2] (35)
vision.20	0.013 [-0.381, 0.334] (36)	-0.9 [-2.5, 0.0] (36)	-5.5 [-13.9, 0.0] (35)
vision.23	0.079 [-0.090, 0.252] (35)	0.4 [-2.1, 3.5] (36)	-5.5 [-13.9, 0.0] (35)
projector.final	0.009 [-0.328, 0.254] (36)	-1.1 [-5.0, 2.0] (36)	-5.5 [-13.9, 0.0] (35)
llm.0	-0.089 [-0.448, 0.190] (36)	-4.1 [-9.2, 0.0] (36)	-5.5 [-13.9, 0.0] (35)
llm.6	0.228 [-0.117, 0.559] (36)	-2.0 [-4.1, 0.0] (36)	-5.4 [-13.9, 0.2] (35)
llm.12	0.106 [-0.126, 0.337] (35)	-0.7 [-2.4, 0.3] (36)	-5.5 [-13.9, 0.0] (35)
llm.18	-0.020 [-0.164, 0.118] (36)	-1.7 [-4.5, 0.7] (36)	-1.7 [-4.9, 0.0] (36)
llm.24	-0.101 [-0.362, 0.000] (36)	0.1 [0.0, 0.4] (36)	-0.2 [-0.8, 0.3] (36)
llm.30	0.058 [0.000, 0.152] (36)	0.0 [0.0, 0.0] (36)	0.3 [0.0, 0.7] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 36: InternVL M2: error reduction across all tested sites (shared conventions).

Site	Position	Angle	Scale
vision.0	0.526 [0.502, 0.549] (36)	10.0 [-11.0, 24.9] (36)	14.6 [12.6, 16.3] (36)
vision.4	0.538 [0.521, 0.555] (36)	15.7 [-6.8, 35.1] (36)	15.6 [13.7, 17.3] (36)
vision.8	0.517 [0.498, 0.535] (36)	1.6 [-18.3, 20.5] (36)	15.4 [13.5, 17.2] (36)
vision.12	0.502 [0.487, 0.517] (36)	-7.7 [-30.8, 15.1] (36)	15.0 [13.2, 16.6] (36)
vision.16	0.499 [0.483, 0.514] (36)	-9.1 [-27.5, 8.4] (36)	15.2 [13.6, 16.8] (36)
vision.20	0.500 [0.484, 0.515] (36)	4.0 [-15.4, 23.4] (36)	15.5 [13.7, 17.1] (36)
vision.23	0.499 [0.482, 0.515] (36)	-6.4 [-27.7, 15.2] (36)	15.3 [13.3, 17.0] (36)
projector.final	0.473 [0.448, 0.498] (36)	1.1 [-19.7, 20.9] (36)	14.8 [13.1, 16.4] (36)
llm.0	0.473 [0.448, 0.500] (36)	6.3 [-16.2, 29.0] (36)	15.1 [13.2, 16.8] (36)
llm.6	0.476 [0.450, 0.501] (36)	12.4 [-11.2, 34.6] (36)	15.2 [13.4, 16.9] (36)
llm.12	0.474 [0.450, 0.499] (36)	1.2 [-24.2, 26.1] (36)	16.1 [14.5, 17.6] (36)
llm.18	0.446 [0.416, 0.468] (36)	5.3 [-14.6, 26.4] (36)	14.3 [12.9, 15.7] (36)
llm.24	0.124 [0.086, 0.166] (36)	7.0 [-1.9, 16.9] (36)	-0.1 [-0.4, 0.3] (36)
llm.30	0.002 [-0.002, 0.006] (36)	-0.2 [-10.2, 10.0] (36)	0.0 [-0.3, 0.3] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 37: InternVL M3: error reduction across all tested sites (shared conventions).

Site	Position	Angle	Scale
vision.0	0.522 [0.490, 0.552] (36)	2.5 [-14.7, 19.5] (36)	15.3 [13.4, 17.4] (36)
vision.4	0.541 [0.517, 0.564] (36)	11.2 [-9.3, 31.1] (36)	16.3 [14.2, 18.3] (36)
vision.8	0.528 [0.505, 0.547] (36)	-9.8 [-28.8, 15.2] (36)	16.2 [13.8, 18.4] (36)
vision.12	0.524 [0.505, 0.539] (36)	5.3 [-9.8, 23.1] (36)	15.9 [13.8, 17.9] (36)
vision.16	0.520 [0.500, 0.535] (36)	8.6 [-9.7, 29.7] (36)	16.2 [14.3, 18.1] (36)
vision.20	0.522 [0.499, 0.540] (36)	13.7 [-2.6, 33.5] (36)	16.1 [14.0, 18.1] (36)
vision.23	0.520 [0.500, 0.535] (36)	13.6 [-5.9, 33.2] (36)	16.8 [14.7, 18.8] (36)
projector.final	0.491 [0.465, 0.514] (36)	-2.8 [-19.4, 12.3] (36)	16.0 [13.9, 18.1] (36)
llm.0	0.491 [0.465, 0.514] (36)	-3.0 [-19.6, 12.3] (36)	16.1 [14.0, 18.1] (36)
llm.6	0.489 [0.461, 0.514] (36)	1.7 [-10.0, 13.9] (36)	16.4 [14.2, 18.5] (36)
llm.12	0.484 [0.454, 0.508] (36)	0.5 [-19.1, 20.9] (36)	16.7 [14.7, 18.7] (36)
llm.18	0.454 [0.422, 0.478] (36)	6.6 [-12.5, 25.7] (36)	14.7 [13.2, 16.2] (36)
llm.24	0.139 [0.100, 0.180] (36)	4.2 [-7.2, 17.6] (36)	0.1 [-0.1, 0.5] (36)
llm.30	-0.003 [-0.007, 0.000] (36)	-1.7 [-3.9, 0.1] (36)	-0.1 [-0.4, 0.2] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Paired checkpoint differences. Each contrast subtracts the Base patch effect from the adapted effect on the same numeric pairs at the same site. An increased intervention effect is not equivalent to a repaired internal computation.

Table 38: Qwen M2 minus Base: paired effect differences (shared conventions).

Site	Position	Angle	Scale
vision.0	0.454 [0.331, 0.575] (33)	-5.3 [-26.7, 16.7] (36)	19.0 [13.8, 24.5] (35)
vision.4	0.348 [0.182, 0.499] (32)	25.3 [11.7, 40.4] (36)	16.6 [12.6, 21.7] (35)
vision.8	0.303 [0.121, 0.466] (35)	15.0 [2.2, 25.0] (36)	15.9 [12.6, 18.9] (35)
vision.12	0.250 [0.162, 0.358] (35)	-6.1 [-16.9, 6.4] (36)	10.9 [6.9, 14.1] (36)
vision.16	0.265 [0.117, 0.423] (35)	0.6 [-12.6, 13.0] (36)	14.5 [10.0, 18.8] (36)
vision.20	0.263 [0.142, 0.388] (35)	1.1 [-14.3, 15.1] (36)	11.8 [9.0, 15.2] (36)
vision.23	0.282 [0.123, 0.429] (35)	-2.3 [-11.8, 8.9] (36)	13.8 [8.8, 20.4] (36)
vision.24	0.221 [0.078, 0.386] (35)	0.5 [-10.0, 10.7] (36)	9.3 [5.1, 12.4] (36)
vision.25	0.220 [0.062, 0.383] (35)	1.3 [-8.1, 12.9] (36)	11.1 [5.4, 16.4] (36)
vision.26	0.237 [0.126, 0.354] (35)	4.6 [-2.7, 12.9] (36)	9.2 [0.8, 16.1] (36)
final_merger	0.239 [0.095, 0.369] (35)	4.4 [-4.6, 12.8] (36)	8.1 [0.2, 14.2] (36)
deepstack.0	0.044 [-0.026, 0.118] (35)	-1.2 [-13.9, 11.0] (36)	-1.1 [-6.4, 2.1] (36)
deepstack.1	0.067 [-0.086, 0.195] (34)	-0.6 [-6.9, 6.7] (36)	1.3 [-4.7, 5.9] (36)
deepstack.2	-0.015 [-0.088, 0.080] (35)	1.4 [-7.3, 10.4] (36)	-1.2 [-8.3, 4.2] (36)
llm.0	0.215 [0.124, 0.317] (35)	4.9 [-10.0, 17.1] (36)	10.8 [8.8, 13.1] (36)
llm.6	0.347 [0.225, 0.462] (35)	0.6 [-9.6, 10.7] (36)	15.0 [6.5, 23.0] (36)
llm.12	0.316 [0.211, 0.432] (35)	2.0 [-8.9, 12.0] (36)	14.4 [6.2, 22.4] (36)
llm.18	0.330 [0.217, 0.436] (34)	-5.1 [-15.4, 4.9] (36)	10.1 [3.4, 15.6] (36)
llm.24	0.073 [-0.026, 0.176] (35)	-2.6 [-8.2, 2.2] (36)	-1.1 [-9.3, 5.2] (36)
llm.30	-0.005 [-0.092, 0.071] (35)	-0.2 [-3.3, 2.3] (36)	-0.8 [-7.3, 5.1] (36)
llm.35	0.000 [0.000, 0.000] (35)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 39: Qwen M3 minus Base: paired effect differences (shared conventions).

Site	Position	Angle	Scale
vision.0	0.446 [0.329, 0.567] (33)	6.6 [-13.5, 27.6] (36)	20.5 [15.2, 25.6] (35)
vision.4	0.368 [0.206, 0.519] (32)	30.3 [14.3, 44.9] (36)	19.8 [15.7, 24.6] (35)
vision.8	0.364 [0.181, 0.526] (35)	30.8 [22.0, 38.9] (36)	19.0 [15.6, 22.1] (35)
vision.12	0.247 [0.170, 0.346] (35)	-1.5 [-6.1, 3.4] (36)	11.4 [7.4, 14.2] (36)
vision.16	0.267 [0.126, 0.419] (35)	1.8 [-5.0, 9.6] (36)	14.8 [10.4, 19.3] (36)
vision.20	0.169 [0.048, 0.292] (35)	9.4 [1.6, 17.5] (36)	8.5 [5.6, 12.3] (36)
vision.23	0.196 [0.047, 0.332] (35)	1.4 [-5.6, 8.3] (36)	10.9 [5.9, 17.5] (36)
vision.24	0.144 [0.004, 0.306] (35)	4.7 [0.5, 11.6] (36)	5.8 [1.5, 9.2] (36)
vision.25	0.125 [-0.029, 0.291] (35)	2.6 [-4.8, 9.3] (36)	7.2 [1.5, 12.5] (36)
vision.26	0.146 [0.043, 0.269] (35)	5.3 [1.9, 10.1] (36)	5.5 [-2.6, 12.2] (36)
final_merger	0.141 [0.009, 0.262] (35)	4.4 [2.1, 6.6] (36)	4.7 [-3.2, 11.2] (36)
deepstack.0	0.091 [0.021, 0.164] (35)	5.1 [-1.9, 9.3] (36)	2.2 [-3.0, 5.4] (36)
deepstack.1	0.118 [-0.031, 0.241] (34)	-2.1 [-8.5, 2.3] (36)	4.0 [-1.8, 8.2] (36)
deepstack.2	-0.004 [-0.078, 0.091] (35)	1.7 [-7.1, 10.8] (36)	-0.8 [-8.0, 4.4] (36)
llm.0	0.117 [0.022, 0.228] (35)	10.0 [2.7, 18.7] (36)	7.1 [5.3, 9.3] (36)
llm.6	0.429 [0.308, 0.538] (35)	17.4 [8.2, 23.8] (36)	19.4 [10.9, 27.1] (36)
llm.12	0.393 [0.283, 0.511] (35)	22.2 [18.5, 25.4] (36)	18.6 [10.7, 26.1] (36)
llm.18	0.397 [0.288, 0.498] (34)	19.2 [10.3, 26.0] (36)	14.6 [7.8, 20.6] (36)
llm.24	0.079 [-0.021, 0.177] (35)	1.6 [-0.6, 3.4] (36)	-0.6 [-8.9, 5.8] (36)
llm.30	-0.005 [-0.091, 0.069] (35)	0.9 [0.1, 2.1] (36)	-0.8 [-7.3, 5.2] (36)
llm.35	0.000 [0.000, 0.000] (35)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 40: InternVL M2 minus Base: paired effect differences (shared conventions).

Site	Position	Angle	Scale
vision.0	0.396 [0.003, 0.855] (35)	14.4 [-6.3, 29.7] (35)	17.2 [8.0, 27.3] (35)
vision.4	0.381 [0.151, 0.654] (34)	16.0 [-6.3, 35.0] (36)	21.3 [14.6, 30.0] (35)
vision.8	0.417 [0.240, 0.628] (36)	2.3 [-18.5, 22.1] (36)	20.9 [14.3, 29.5] (35)
vision.12	0.450 [0.271, 0.639] (36)	-7.0 [-28.5, 13.9] (36)	19.2 [11.6, 28.5] (35)
vision.16	0.467 [0.320, 0.616] (36)	-9.5 [-27.2, 7.8] (36)	20.6 [14.0, 29.5] (35)
vision.20	0.487 [0.168, 0.881] (36)	4.8 [-14.3, 24.0] (36)	21.0 [14.3, 29.9] (35)
vision.23	0.423 [0.252, 0.594] (35)	-6.8 [-28.1, 14.8] (36)	20.8 [14.0, 29.7] (35)
projector.final	0.464 [0.217, 0.810] (36)	2.3 [-19.5, 23.5] (36)	20.3 [13.8, 28.9] (35)
llm.0	0.562 [0.281, 0.925] (36)	10.4 [-13.0, 32.6] (36)	20.6 [14.0, 29.4] (35)
llm.6	0.248 [-0.082, 0.591] (36)	14.4 [-9.3, 36.6] (36)	20.6 [14.0, 29.5] (35)
llm.12	0.369 [0.140, 0.596] (35)	1.9 [-23.5, 26.4] (36)	21.6 [15.2, 30.4] (35)
llm.18	0.466 [0.323, 0.619] (36)	7.1 [-14.5, 29.7] (36)	16.0 [13.0, 19.7] (36)
llm.24	0.225 [0.099, 0.489] (36)	6.9 [-2.1, 16.8] (36)	0.1 [-0.5, 0.7] (36)
llm.30	-0.056 [-0.150, 0.005] (36)	-0.2 [-10.2, 10.0] (36)	-0.3 [-0.8, 0.2] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Table 41: InternVL M3 minus Base: paired effect differences (shared conventions).

Site	Position	Angle	Scale
vision.0	0.393 [-0.003, 0.852] (35)	3.5 [-16.6, 21.5] (35)	18.0 [9.0, 28.1] (35)
vision.4	0.387 [0.155, 0.664] (34)	11.5 [-11.1, 32.5] (36)	22.0 [14.7, 31.1] (35)
vision.8	0.427 [0.251, 0.640] (36)	-9.0 [-30.2, 18.2] (36)	21.7 [14.4, 31.1] (35)
vision.12	0.473 [0.289, 0.664] (36)	6.0 [-12.2, 29.3] (36)	20.2 [12.7, 29.8] (35)
vision.16	0.487 [0.341, 0.636] (36)	8.2 [-11.4, 31.3] (36)	21.6 [14.6, 30.9] (35)
vision.20	0.509 [0.190, 0.900] (36)	14.5 [-2.4, 35.6] (36)	21.6 [14.5, 30.9] (35)
vision.23	0.446 [0.271, 0.619] (35)	13.2 [-7.9, 34.7] (36)	22.3 [15.2, 31.6] (35)
projector.final	0.482 [0.242, 0.817] (36)	-1.7 [-18.8, 14.5] (36)	21.5 [14.7, 30.4] (35)
llm.0	0.580 [0.304, 0.932] (36)	1.1 [-17.6, 19.6] (36)	21.6 [14.7, 30.8] (35)
llm.6	0.261 [-0.068, 0.600] (36)	3.7 [-9.1, 17.7] (36)	21.8 [14.9, 30.7] (35)
llm.12	0.381 [0.154, 0.606] (35)	1.3 [-19.1, 23.1] (36)	22.2 [15.5, 31.0] (35)
llm.18	0.474 [0.326, 0.628] (36)	8.3 [-11.0, 27.1] (36)	16.4 [13.5, 19.8] (36)
llm.24	0.239 [0.113, 0.497] (36)	4.1 [-7.2, 17.4] (36)	0.3 [-0.1, 0.8] (36)
llm.30	-0.061 [-0.156, 0.000] (36)	-1.7 [-3.9, 0.1] (36)	-0.4 [-1.0, 0.1] (36)
llm.35	0.000 [0.000, 0.000] (36)	0.0 [0.0, 0.0] (36)	0.0 [0.0, 0.0] (36)

Denominator sensitivity. We retain the original threshold 10^{-4} and additionally use 0.01, 0.05, and 0.1 for position; 0.1, 1, and 5 degrees for angle; and 0.001, 0.01, and 0.05 for relative scale error. The following table reports the minimum–maximum numeric valid count across sites under each threshold, exposing attrition without selecting favorable sites. Full per-site median recovery and bootstrap intervals at every threshold, as well as mean/median absolute effects and paired contrasts,

accompany the source as `patch_denominator_sensitivity.csv`, `patch_absolute_effects.csv`, and `patch_paired_checkpoint_contrasts.csv`.

Table 42: Valid recovery counts across all tested sites at the original and three factor-specific thresholds defined above. Every row starts with 36 candidates; ranges describe variation across sites, not uncertainty intervals.

Architecture	Model	Attribute	10^{-4}	Threshold 2	Threshold 3	Threshold 4
Qwen	Base	position	10	10	10	9
Qwen	Base	angle	5	5	5	3
Qwen	Base	scale	2	2	2	2
Qwen	M2	position	36	36	36	36
Qwen	M2	angle	32	32	32	32
Qwen	M2	scale	35	35	35	35
Qwen	M3	position	36	36	36	36
Qwen	M3	angle	36	36	36	36
Qwen	M3	scale	36	36	35	35
InternVL	Base	position	9	9	9	9
InternVL	Base	angle	4	4	4	3
InternVL	Base	scale	2	2	2	2
InternVL	M2	position	36	36	36	36
InternVL	M2	angle	24	24	24	23
InternVL	M2	scale	35	35	35	33
InternVL	M3	position	36	36	36	36
InternVL	M3	angle	30	30	30	29
InternVL	M3	scale	36	36	36	34