SPAEX: TARGET-ALIGNED SPATIAL EVIDENCE FOR MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

For spatial reasoning, multimodal agents built on vision language models (VLMs) must interpret measurements relative to the objects and reference frames specified by a goal. Aligned states still leave candidate comparisons to the model. We introduce SpaEx, a target-aligned interface coupling object-centered perception with explicit spatial execution and fusing candidate evidence with native VLM scores in the original answer space. With the VLM and perception tools fixed, the reader fits only four scalar parameters. Across SpatialMQA, SPAR-Bench single-view reasoning and CV-Bench 3D, SpaEx improves two backbones by 5.3–18.9 percentage points over matched task-adapted VLMs and attains the highest accuracy and semantic Macro-F1 among compared methods in every benchmark–backbone setting. Under shared observations, readers supplied with executed comparisons outperform those given aligned states even with longer training and compact inputs. Candidate-level fusion also recovers correct answers ranked first by neither branch. Although execution and fusion contribute differently across tasks, these results support explicit computation as an effective interface between spatial perception and native visual judgment.[1]

1 INTRODUCTION

Spatial relations depend on the reference frame: an object on the camera's left can lie on a person's right. A question specifies both the objects to compare and the viewpoint from which their relation should be judged. It can instead posit an imagined observer or ask for distance to an anchor. Benchmarks test the difficulty of converting visual measurements into these distinct relational decisions (Liu et al., 2025; Zhang et al., 2025; Tong et al., 2024). We study explicit spatial computation for multimodal agents whose reasoning relies on task-adapted vision–language models (VLMs).

Object-centered representations identify entities and attributes, while visual programs and Program of Thoughts supply computation over them (Andreas et al., 2016; Yi et al., 2018; Locatello et al., 2020; Gupta & Kembhavi, 2023; Surís et al., 2023; Chen et al., 2023). APC transforms object state into a specified reference viewpoint (Lee et al., 2025). This leaves an interface question: should a VLM derive answer comparisons from **aligned state**, or receive **comparisons already computed for its original candidates**? Compound answers such as "left and behind" expose the distinction. Execution supplies comparisons, but faulty geometry can accompany a correct VLM answer.

SpaEx joins these sources in a **target-aligned state–execution–score interface** (Figure 1). The question specifies the agent's reasoning goal: its object roles bind pretrained visual measurements to the target, comparison reference, optional observer and selection conditions. Typed operations then project displacements or compare distances for the original candidates. A positive-weight reader combines the resulting evidence with native VLM scores, retaining native output whenever execution is unsupported. This construction connects structured perception and spatial computation to the VLM's independent visual judgment.

The evaluation spans observed actor viewpoints in SpatialMQA, imagined frames and compound relations in SPAR-Bench single-view reasoning, and metric distances in CV-Bench 3D. SpaEx improves over matched LoRA for both VLMs across these tasks. Native evidence adds to execution

[1] Code and supplementary materials: `https://anonymous.4open.science/r/spaex-7EC5/`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

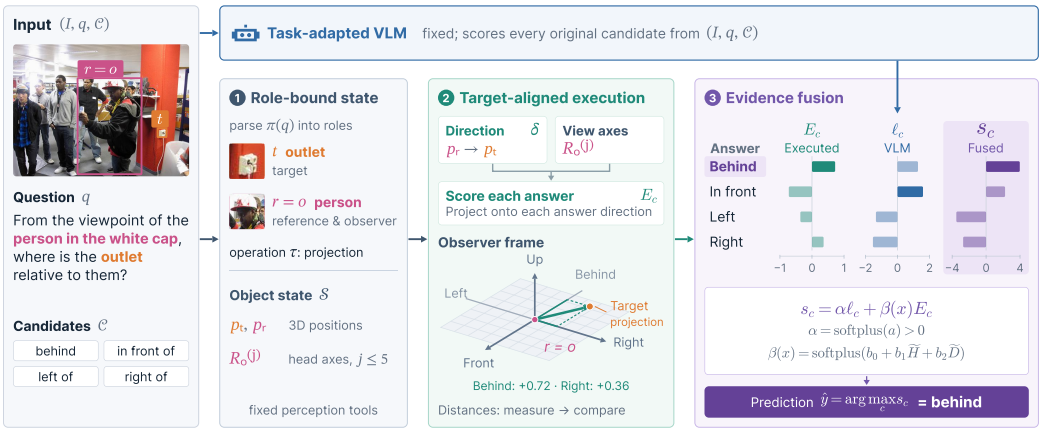

Figure 1: **SpaEx overview.** Question-bound object states provide candidate evidence through frame projection or metric comparison; fusion adds independent VLM scores. The person serves as reference r and observer o, and the outlet as target t. This SpatialMQA example uses predicted regions and LLaVA candidate scores. The abbreviated question and schematic geometry illustrate the computation; mean-centered bars use separate display scales. Unsupported inputs retain VLM scores.

on SpatialMQA and SPAR, whereas execution accounts for the CV-Bench gain. Shared-observation SpatialMQA comparisons **favor execution over aligned vectors** at both training budgets.

Our contributions are:

- A target-aligned interface coupling object-centered perception and spatial execution, followed by candidate-level fusion with native VLM evidence.
- Controlled reader comparisons showing the value of executed candidate comparisons over aligned object states in the evaluated settings.
- Evidence across three benchmarks identifying when native predictions complement execution and when execution accounts for the gain.

2 Related Work

Learning spatial representations. Spatial supervision teaches VLMs to recognize and reason about object relations. SpatialSense and VSR document persistent difficulties in visual relation recognition (Yang et al., 2019; Liu et al., 2023), motivating training data that directly encode spatial concepts. SpatialVLM uses generated spatial annotations, and EmbSpatial-SFT supplies embodied instruction data (Chen et al., 2024; Du et al., 2024). Training strategies organize these abilities: SpatialLadder progresses from localization to reasoning (Li et al., 2026), while SR-REAL jointly optimizes linguistic and detect-then-reason paths (Ji et al., 2026). These approaches develop the backbone's spatial competence through supervision. External measurements raise a complementary issue: how to connect that learned competence to the computation needed for a particular question.

Object-centered perception and geometric grounding. Object-centered interfaces make individual entities available for subsequent reasoning. Slot Attention groups visual features into exchangeable object representations (Locatello et al., 2020), while Set-of-Mark prompting uses region identifiers to connect language references to image regions (Yang et al., 2023). Selecting the intended entity and measuring its spatial attributes are separate requirements. SpatialRGPT learns regional representations and incorporates depth through a visual-encoder module (Cheng et al., 2024). SpatialPIN supplies reconstructed 3D scene information through prompting, drawing on priors from several perception models (Ma et al., 2024). These interfaces expose geometry through features or scene descriptions. To answer a spatial question, the reasoning stage must relate those observations to the requested objects and distinguish measurements of position, orientation and distance.

Executable reasoning and multimodal agents. Executable reasoning assigns explicit operations to intermediate representations. Neural module networks and neural-symbolic VQA compose question-dependent operations over visual representations (Andreas et al., 2016; Yi et al., 2018). CoT elicits intermediate reasoning steps (Wei et al., 2022; Kojima et al., 2022); Program of Thoughts assigns numerical computation to an interpreter (Chen et al., 2023). VISPROG and ViperGPT extend this division of labor to visual tasks through executable programs and calls to perception modules (Gupta & Kembhavi, 2023; Surís et al., 2023). Visual Sketchpad adds drawing and image operations as intermediate reasoning steps (Hu et al., 2024). For 3D reasoning, VADAR synthesizes reusable Python functions (Marsili et al., 2025), and SpaceTools uses reinforcement learning to coordinate perception tools (Chen et al., 2026). These systems give multimodal agents procedures for obtaining and manipulating visual evidence. The resulting computation depends on the semantics of its inputs: spatial operators need consistent object bindings and coordinate conventions, even when the program itself executes successfully.

Viewpoint alignment and prediction evidence. Geometric intermediates differ in how much computation they supply before the VLM chooses an answer. SpatialReasoner learns explicit 3D representations shared across perception, computation and reasoning (Ma et al., 2025). APC aligns object positions and orientations with a requested viewpoint, supplying a numerical or rendered scene to the VLM (Lee et al., 2025). OrientSAM addresses camera-centric shortcuts with orientation tokens and curriculum learning (Fan et al., 2026). A scene aligned to the requested viewpoint still leaves the VLM to derive the relevant answer comparison. A separate line acts directly on predictions: VCD contrasts original and distorted image distributions to mitigate object hallucinations (Leng et al., 2024). SpaEx connects geometric computation to prediction through evidence for the original answer candidates. Independent VLM scores then contribute visual judgment at the same decision level as the executed comparisons.

3 METHOD

For image I, question q and original ordered candidates $\mathcal{C} = (c_1, \ldots, c_K)$, SpaEx fuses executed candidate evidence E_c with the adapted VLM's native score $\ell_c(I, q, \mathcal{C})$:

$$s_c = \begin{cases} \alpha \ell_c + \beta E_c, & \text{input supported,} \\ \ell_c, & \text{otherwise,} \end{cases} \qquad \hat{y} = \arg\max_{c \in \mathcal{C}} s_c. \tag{1}$$

The fitted weights satisfy $\alpha, \beta > 0$, with β either constant or input-dependent. Constructing E_c requires agreement between object roles, coordinates and candidate meanings. Once those semantics are aligned, fusion combines the two sources at the candidate-score level.

3.1 QUESTION-CONDITIONED STATE AND CANDIDATE SEMANTICS

The question determines which measurements are relevant. From its text and any public task category, the parser constructs $\pi(q) = (\mathcal{V}, t, r, o, \tau, \mathcal{K})$, where $\mathcal{V}$ identifies objects, t the target, r the comparison reference, o an optional observer, τ the operation, and $\mathcal{K}$ selection or hypothetical conditions. Relations compare t with r; selection binds candidates to objects. SpatialMQA supports observed actor views with shared observer/reference and no extra facing/position clauses. SPAR constructs look-at frames from observer positions and facing anchors, allowing distinct comparison references. Object sharing preserves role meanings; neither frame construction moves objects.

Perception instantiates the corresponding numerical state:

$$\mathcal{S} = \{(p_v, \rho_v, \sigma_v, \{R_v^{(j)}\}_{j=1}^{J}, m_v) : v \in \mathcal{V}\}, \tag{2}$$

Here p_v is position, ρ_v camera range, σ_v regional spread, $R_v^{(j)}$ reference axes, and m_v field availability. Support requires implemented scope, resolved bindings, valid measurements and passed admission checks. Person-frame projection uses head axes, with body-axis and person-overlap checks controlling admission (Appendix B). Camera-space points, ordinal depth and axes preserve distinct meanings. Failed support returns native scores without assigning a default direction.

Appendix A.4 specifies the operational contracts and inference sequence. Candidate semantics define the comparison, and image measurements determine its value. Candidate permutations reorder constructed evidence, but may change the native scorer's predictions.

3.2 SHORT TYPED EXECUTION

The executor maps numerical state to $E_c = \mathrm{Exec}_\tau(\mathcal{S}, c)$. SELECT binds roles to observed objects. RELATIVE forms $\delta = \mathrm{unit}(p_t - p_r)$, where $\mathrm{unit}(v) = v/\|v\|_2$; zero or invalid displacements remain unknown. REFRAME selects the requested frame, and PROJECT evaluates displacement against candidate semantics using signed axes a_c in measurement coordinates:

$$n_c = \frac{1}{J}\sum_{j=1}^{J} R^{(j)} a_c, \qquad E_c = n_c^\top \delta. \tag{3}$$

The columns of $R^{(j)}$ express the question's reference frame in measurement coordinates; observer-relative queries use $R_o^{(j)}$. The quantity $(R^{(j)}a_c)^\top\delta$ therefore projects displacement into that frame. Averaging signed axes does not require their mean to be a rotation matrix.

Coordinate agreement makes this projection meaningful across perception tools. Applying the same proper rotation to displacement and candidate axes preserves their dot product; positive rescaling of positions leaves the normalized displacement unchanged. Changing only the observer's orientation can change candidate ordering, as required for viewpoint-dependent questions. Unit normalization removes scale but cannot repair mismatched axis conventions between geometry and pose. We establish a common measurement frame before projection (Appendix B).

For metric selection, MEASURE obtains $d_c = \rho_{v(c)}$ for camera closeness or $d_c = \|p_{v(c)} - p_r\|_2$ for anchor distance. COMPARE converts the candidate distances into evidence:

$$E_c = -\frac{d_c - \bar{d}}{\max(\|\mathbf{d} - \bar{d}\mathbf{1}\|_2, \epsilon)}, \qquad \bar{d} = K^{-1}\sum_c d_c. \tag{4}$$

Higher evidence favors nearer objects, with the sign reversed for farthest queries. This operator supplies the CV-Bench comparisons.

SPAR answers can combine horizontal, vertical and depth relations. For frame-relative measurement u_j and candidate c's interval $\mathcal{I}_{c,j}$ on axis j, COMPARE applies an interval-distance penalty:

$$E_c = \mathrm{clip}\left(-\frac{\sum_j \mathrm{dist}(u_j, \mathcal{I}_{c,j})}{\sigma}, -12, 0\right). \tag{5}$$

Intervals preserve the task's $0.1\,\mathrm{m}$ indistinguishability convention. Regional spread with a fixed floor sets penalty scale σ; SPAR distance selection uses an analogous centered gap. Scores quantify candidate compatibility, not calibrated truth probabilities. Appendix C defines frames, intervals and spread. All operators use estimated geometry, excluding test ground-truth depth and 3D boxes.

3.3 FUSION WITH NATIVE VISUAL EVIDENCE

Native and executed evidence differ in scale and can rank candidates differently. With perception and the VLM fixed, we fit the positive weights in Equation 1. The default SpaEx reader is adaptive:

$$\alpha = \mathrm{softplus}(a), \quad \beta(x) = \mathrm{softplus}(b_0 + b_1\widetilde{H} + b_2\widetilde{D}). \tag{6}$$

The four trainable parameters are a, b_0, b_1, b_2. Native entropy H is normalized by $\log K$. Dispersion D measures repeated-projection variation for observer frames or relative within-region range spread for metric geometry; tildes indicate training-set standardization. These fitting features are not calibrated correctness probabilities, and range spread may reflect object extent or background. Interfaces without repeated or regional observations set $D = 0$.

Constant fusion and free operators are controls; all readers retain native scores outside support.

Why a constant weight can still be selective. For native winner c_0 and challenger c, the challenger exceeds c_0 after fusion precisely when

$$\beta(x)(E_c - E_{c_0}) > \alpha(\ell_{c_0} - \ell_c). \tag{7}$$

A constant positive β still induces example-dependent decisions because both the native margin and evidence gap vary. With $W_E = \max_c E_c - \min_c E_c$, a native margin exceeding $\beta(x)W_E/\alpha$ preserves a unique native winner. Projection bounds the range by $W_E \le 2$, and clipped SPAR compatibility by $W_E \le 12$. This margin condition explains constant-weight selectivity without implying that the preserved prediction is correct.

3.4 Training and Controlled Variants

Supported training examples calibrate both sources through candidate cross-entropy, with quadratic penalties on native scale, evidence weight and adaptive slopes (Appendix A). Answers supervise fitting, never inference. SpatialMQA and CV-Bench select $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ by cross-entropy over three image-grouped training folds, then refit on all supported rows. SPAR uses a separate calibration cohort and $\lambda = 10^{-2}$ throughout development. All benchmarks share the objective and optimizer; each seed retains its exact adapted VLM.

The two-parameter *constant-weight* reader sets $\beta = \mathrm{softplus}(b_0)$. Comparing it with the adaptive form tests the value of conditioning evidence weights on native entropy and geometric dispersion. The *free operator* learns $n_c^\top B\delta$, with identity initialization and regularization of B, while retaining role semantics. It tests whether fitting the geometric comparison itself adds value beyond calibrating the two score sources. For scalar metric evidence, this control reduces to rescaling. Both variants share the default reader's observations, support and native scores, keeping the comparison focused on readout of the same evidence. Appendix A specifies both variants.

4 Experimental Setup

4.1 Benchmark and Baseline

We evaluate three complementary spatial reasoning settings. **SpatialMQA** (Liu et al., 2025) tests object relations from question-specified viewpoints in natural images, using a curated split without training–test image overlap. **SPAR-Bench** (Zhang et al., 2025) covers four single-view tasks on ScanNet (Dai et al., 2017) and ScanNet++ (Yeshwanth et al., 2023): object distance, camera-frame relations, and imagined-view relations with observer or object references. **CV-Bench 3D** (Tong et al., 2024) comprises the complete Depth and Distance tasks, testing camera closeness and distance to an object anchor. All benchmarks retain the original questions and candidates, including every test question regardless of execution support.

Baselines include the released instruction-tuned VLM (**Vanilla**), **task-adapted LoRA**, **CoT** (Kojima et al., 2022), **Set-of-Mark (SoM)** (Yang et al., 2023) and **VCD** (Leng et al., 2024). CoT elicits a rationale, SoM adds automatic SAM-B region marks, and VCD contrasts clean-image and diffusion-noised scores. All adapted methods share the corresponding native checkpoint, source image, question and candidate order. Constant fusion, free operators and execution-only readout test the contributions of computation and fusion under common perception and support (Section 5.2); Appendix D gives the geometric and answer-prior controls.

4.2 Implementation

We use Qwen3-VL-4B (Bai et al., 2025) and LLaVA-OneVision-1.5-4B (An et al., 2025), adapting each with LoRA (Hu et al., 2022). SPAR uses separate SPAR-7M scenes for training, calibration and development; CV-Bench uses external EmbSpatial-SFT data (Du et al., 2024), with no evaluation-image overlap. We fit the reader on supported examples with frozen VLMs and perception; SPAR and CV-Bench use separate calibration cohorts. Appendix A.2 and Table 5 specify training.

Both backbones share the same perceptual state. On SpatialMQA, a text-only Qwen3-VL planner binds question roles; Rex-Omni, SAM and MoGe-2 estimate object regions and camera-space geometry (Jiang et al., 2026; Kirillov et al., 2023; Wang et al., 2025). MeTRAbs supplies observer axes (Sárándi et al., 2021), with Grounding DINO checking person overlap (Liu et al., 2024). SPAR and CV-Bench recover their public RGB markers and use fixed parsers with MoGe-2 geometry; SPAR also uses SAM regions and predicted up axes for imagined viewpoints. Inference uses only the image, question, candidates and public markings. Coordinate construction and validity checks appear in Appendices B and C.

Every method scores the original numeric answer suffix by its mean token log-likelihood (Appendix D.1). Accuracy averages all questions for SpatialMQA and CV-Bench, and the four task accuracies for SPAR; we also compute semantic Macro-F1. Adapted methods use three-seed means and sample SDs. Appendix A.4 defines metrics and grouped uncertainty estimates.

Table 1: **Matched benchmark performance** (%). Acc. follows benchmark aggregation; F1 is semantic Macro-F1 (Appendix A.4). Adapted methods share each LoRA and original candidates, reporting three-seed mean ± sample SD. Vanilla uses one released checkpoint. Bold marks the highest mean within each model, benchmark and metric, including ties.

		SpatialMQA		SPAR-Bench		CV-Bench 3D	
Model	Method	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Qwen	Vanilla	38.20	39.56	44.03	38.64	91.33	91.33
	LoRA	73.61 ± 0.98	75.11 ± 0.88	66.01 ± 0.48	58.46 ± 0.63	87.33 ± 0.46	87.30 ± 0.47
	CoT	71.78 ± 4.70	73.47 ± 4.77	65.53 ± 0.31	58.32 ± 0.67	86.53 ± 1.24	86.52 ± 1.25
	SoM	74.07 ± 0.97	75.52 ± 1.57	63.00 ± 0.85	54.42 ± 1.25	84.31 ± 0.70	84.25 ± 0.70
	VCD	72.68 ± 0.46	74.24 ± 0.85	65.37 ± 0.27	57.55 ± 0.20	86.39 ± 0.29	86.35 ± 0.30
	SpaEx	**79.12** ± 0.94	**79.64** ± 0.96	**82.18** ± 0.29	**77.71** ± 0.67	**94.25** ± 0.08	**94.25** ± 0.08
LLaVA	Vanilla	35.87	36.28	38.39	30.89	75.42	74.70
	LoRA	71.81 ± 0.63	72.14 ± 0.75	62.17 ± 0.30	50.95 ± 0.59	76.97 ± 2.08	76.96 ± 2.09
	CoT	69.33 ± 1.21	69.05 ± 0.98	61.14 ± 1.11	50.75 ± 1.80	73.50 ± 1.83	73.42 ± 1.85
	SoM	71.38 ± 0.19	71.36 ± 0.23	60.43 ± 0.42	48.50 ± 0.62	75.56 ± 1.47	75.54 ± 1.47
	VCD	71.47 ± 0.28	71.83 ± 0.39	61.11 ± 0.28	51.28 ± 1.05	74.50 ± 1.31	74.48 ± 1.34
	SpaEx	**77.08** ± 0.44	**76.78** ± 0.57	**81.10** ± 0.07	**75.98** ± 0.10	**92.53** ± 0.48	**92.53** ± 0.48

Table 2: **Controls separating execution and fusion.** Full-cohort accuracy (%, three-seed means); SPAR uses task-macro accuracy. Within-cell comparisons share native adapters and support. Execution retains native scores outside support. See Appendix D for controls and per-seed results.

	SpatialMQA		SPAR		CV-Bench 3D	
Reader	Qwen	LLaVA	Qwen	LLaVA	Qwen	LLaVA
Native VLM	73.61	71.81	66.01	62.17	87.33	76.97
Execution + fallback	73.82	72.15	78.88	78.60	**94.25**	**92.53**
Constant fusion	**79.28**	**77.08**	81.67	80.93	**94.25**	**92.53**
SpaEx (adaptive)	79.12	**77.08**	**82.18**	**81.10**	**94.25**	**92.53**

5 Results and Analysis

5.1 Matched-Benchmark Performance

SpaEx exceeds matched LoRA in all eighteen model–seed comparisons (Table 1). SpatialMQA mean gains are 5.51/5.27 points for Qwen/LLaVA, with paired image-cluster intervals [4.06, 7.06]/[3.61, 7.01]. SPAR macro gains reach 16.18/18.93 points, and CV-Bench gains 6.92/15.56.

SpaEx also leads the six benchmark–backbone comparisons in both accuracy and semantic Macro-F1. CoT, SoM and VCD use the same adapted VLM and candidate-likelihood protocol (Appendix D.1); subsequent controls examine supplied computation under shared observations. Removing 15 development-shared SpatialMQA rows retains 5.40/5.34-point gains over matched LoRA.

Vanilla provides an additional, stronger reference on CV-Bench. Qwen's Vanilla accuracy is 91.33%, 4.00 points above its initial LoRA, so part of the SpaEx gain recovers an adaptation loss. Relative to Vanilla itself, the gains remain 2.92/17.11 points for Qwen/LLaVA, with capture-cluster intervals [0.50, 5.87]/[13.91, 20.53].

5.2 Ablations: Executed Evidence and Native Complementarity

Additional-training control. Continuing LoRA on the reader's 1,446 SpatialMQA examples gives 74.35%/72.71% for Qwen/LLaVA. On SPAR's 502 calibration examples, continuation reaches 64.56%/60.97% macro, or 64.65%/60.88% with common native fallback (Appendix C). CV-Bench continuation on the same extra labels yields 91.53%/80.94%; SpaEx retains gains of 2.72/11.58 points and capture-cluster intervals [0.17, 5.68]/[8.52, 13.97] (Appendix D.5). These comparisons match additional labels while allowing computational budgets to differ.

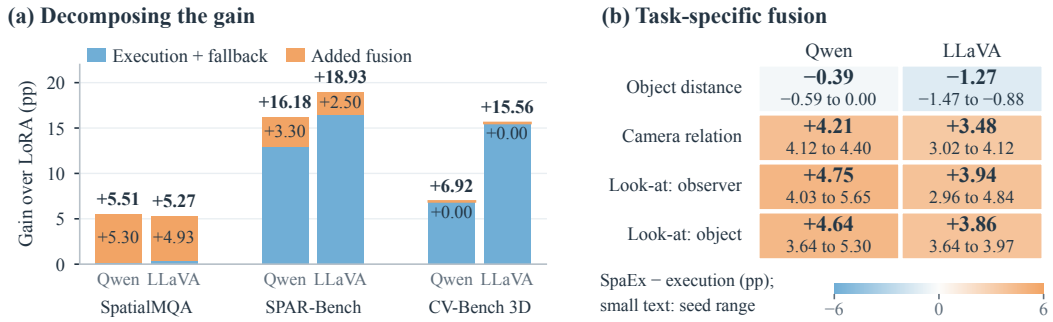

Figure 2: **Gains from execution and fusion.** (a) Gains over matched LoRA separate execution with native fallback and additional fusion; totals appear above bars. (b) SpaEx minus execution by SPAR task. Large values are three-seed means; small values are seed ranges, not confidence intervals. SPAR uses task-macro accuracy in (a).

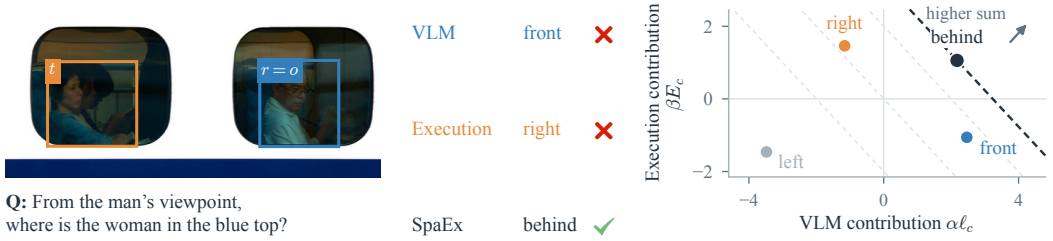

Figure 3: **Recovering a third candidate** (Qwen, seed 13). VLM selects front, execution right, and fusion annotated behind. Mean-centered weighted scores sum equally along diagonals. Across three seeds and 380 supported questions, fusion repairs 29/111 Qwen and 26/113 LLaVA both-wrong outcomes (Appendix D.4). Predicted boxes mark target t and reference/observer $r = o$.

The core ablations separate measurement, execution and fusion within SpaEx (Table 2). SpatialMQA supports contributions from both the measured reference frame and native scores. Substituting camera axes for observer axes lowers accuracy to 73.17%/71.84%, leaving a 5.95/5.24-point gap to SpaEx, despite refitting with identical bindings, displacement, native scores and support. Raw execution with native fallback attains 73.82%/72.15%, compared with fusion's 79.12%/77.08%. The gain reflects branch complementarity and is not reproduced by semantic-bias correction alone.

Image-disjoint native training and reader calibration preserve this complementarity: fusion adds 4.86/5.27 points over the newly trained native models and 5.14/4.96 over execution. The allocation changes native-training budget and calibration size together (Appendix B.2).

SPAR exhibits a similar overall advantage: execution with common fallback reaches 78.88%/78.60% macro, while fusion achieves 82.18%/81.10%. Each of the six model–seed pairs improves over both branches. Gains cover the three relation/imagination tasks, alongside mean losses of 0.39/1.27 points on object distance. Complementarity varies by task (Figure 2).

Reader flexibility. Constant fusion achieves competitive accuracy: 79.28%/77.08% on SpatialMQA and 81.67%/80.93% SPAR macro. A free operator also slightly exceeds the SpatialMQA primary (Table 9). The evidence supports preserving role and candidate semantics without establishing an advantage for dynamic gating or a fixed projection matrix. Constant weights retain example-dependent selectivity through the margin condition (Equation 7).

Positive weights preserve shared branch winners. Complementarity therefore requires recovering alternatives under disagreement while limiting losses of correct predictions from either branch.

Score fusion can recover a correct third candidate even when neither branch ranks it first (Figure 3). Across supported SPAR inputs pooled over three seeds, it repairs 40 such cases per backbone and retains 363/482 Qwen and 363/496 LLaVA native-only successes (Appendix C). A selector restricted to the two branch winners thus does not upper-bound score fusion.

Table 3: **SPAR state readout, execution and fusion.** Task-macro accuracy (%, three-seed mean ± sample SD). Additional examples follow native fitting. Trained readers share 502 examples; execution uses none. Observations, support and native fallback are common.

Readout	Additional examples	Qwen	LLaVA
Native VLM	0	66.01 ± 0.48	62.17 ± 0.30
Aligned state	0	64.18 ± 0.41	60.74 ± 0.64
Native continuation	502	64.65 ± 0.66	60.88 ± 0.34
Aligned state	502	71.47 ± 2.31	68.22 ± 1.01
Execution + fallback	0	78.88 ± 0.08	78.60 ± 0.05
SpaEx	502	**82.18** ± 0.29	**81.10** ± 0.07

Coverage and predictive reliability impose distinct limits. On SpatialMQA's 380 supported questions, Qwen/LLaVA native accuracy is 70.70/70.35%, execution reaches 71.32%, and fusion reaches 86.32/85.26%. Unsupported questions retain native predictions, so coverage constrains full-cohort gains. Most lack an admissible actor-view plan or have unresolved frame conditions, requiring broader role/frame handling and better perception. Even with admission, fusion overturns 43/48 correct native answers across three seeds. Appendix D.4 reports coverage and paired errors.

5.3 Computation before VLM Readout

For 1,446 additional updates, SpatialMQA readers share images, questions, scorer, optimizer and LoRA capacity. Both inputs include native scores and dispersion. Facts supply candidate dot products; aligned vectors supply normals and displacement from which the VLM must compute them. Observations are fixed while supplied computation and serialization differ (Figure 4).



Q: From the blue-shirted person's viewpoint, where is the person without a hat?

Aligned vectors

$\delta = (-0.62, -0.37, +0.70)$

Candidate direction n_c

	x	y	z
front	−0.41	+0.16	−0.90
behind	+0.41	−0.16	+0.90
left	+0.89	−0.16	−0.43
right	−0.89	+0.16	+0.43

Executed facts

$E_c = \langle n_c, \delta \rangle$

Candidate compatibility

front	-0.429
behind	+0.429
left	-0.789
right	+0.789

Qwen	left	✗	right	✓
LLaVA	left	✗	right	✓

Figure 4: **Aligned vectors and executed evidence.** Inputs provide displacement δ and axes n_c, or their dot products, with identical image, question, candidates, VLM scores, entropy and dispersion. After 1,446 updates, both seed-13 backbones choose left from vectors and annotated right from facts. Predicted boxes mark target t and reference/observer $r = o$; measurements are rounded.

Facts outperform vectors in all six pairs at both full-JSON budgets. After 1,446 updates, mean accuracy is 78.31/76.24% with facts versus 74.60/72.24% with vectors (Table 10). Compact serialization cuts mean lengths from 333/567 to 190/290 tokens while retaining gains of 3.41/4.12 points (paired image intervals [2.26, 4.57]/[2.94, 5.43]). External execution of the four-decimal vectors preserves candidate ordering and ties on every supported training and test row: their precision retains the geometric decision, yet readers benefit from receiving the computation. SpaEx further exceeds compact facts by 1.02/0.93 points ([0.09, 1.92]/[0.12, 1.80]), without additional VLM training or readout. Intervals condition on three fitted seeds and characterize these formats and schedules (Appendix D.3).

Training strengthens aligned-state readout (Table 3), yet execution with fallback exceeds it by 7.42/10.38 points from unrounded means, and fusion adds 3.30/2.50. These gaps describe the contributions of candidate computation and fusion without causally isolating them: compatibility conventions and native-score access also differ (Appendix C.6).

Reading evidence inside the VLM. A trained internal reader inserts candidate-bound trace memories into the decoder before answer scoring, sharing observations, executor and candidates. It exceeds matched LoRA in all six benchmark–backbone comparisons but does not consistently exceed output fusion. SpatialMQA differences are +0.12/+0.25 points for Qwen/LLaVA, with paired intervals spanning zero; both decline on SPAR and CV-Bench. Matched development controls favor execution over state-only input and native continuation in all six seed-13 comparisons. Execution also benefits internal readout; representation, capacity and VLM training differ (Appendix E).

5.4 Fusion Boundaries and Native Adaptation

A SPAR development intervention raises native training from 128 to 1,024 examples/updates while fixing perception, execution and the calibration cohort and refitting each reader. Fusion then exceeds execution by 9.24/9.38 points over three seeds. The result identifies native adaptation as a constraint on complementarity, with data volume and optimization steps changed jointly.

Formal SPAR evaluation shows stronger execution but weaker native accuracy and confidence discrimination. ScanNet supplies most net repairs; ScanNet++ gains are small/near-zero (Appendix C). These source differences are descriptive; all questions enter evaluation.

CV-Bench exposes a fusion boundary: execution with native fallback and fusion predict identically in all six cells, at 94.25%/92.53%. Two-candidate normalized metric evidence expresses ordering at constant magnitude, with fitted weights dominating conflicting native margins (Appendix A). Without LoRA, execution with Vanilla fallback attains 94.75%/90.92%; independently calibrated constant/adaptive fusion again predicts identically. Qwen thus obtains the geometric benefit without adaptation under this evidence encoding (Appendix D.5).

Component replacements compare spatial evidence with image-plane cues and answer preferences under the same native adapters, candidates and execution support (Appendix D, Table 9). On CV-Bench, vertical image position for camera closeness and image-center distance for anchor comparison yield 87.50%/78.08% accuracy for Qwen/LLaVA, versus SpaEx's 94.25%/92.53%. The gaps to this image-plane control are 6.75/14.45 percentage points. Two bias controls replace geometric evidence with fitted offsets for candidate indices or answer meanings, retaining the native score in both cases. They achieve 87.81%/74.28% and 87.44%/76.56%, respectively. On SpatialMQA, candidate-meaning biases reach 73.76%/71.84%, versus SpaEx's 79.12%/77.08%. These semantic biases improve native SpatialMQA accuracy by only 0.15/0.03 points. The same pattern across metric and observer-relative tasks supports the contribution of executed geometric comparisons beyond image-plane proxies and learned answer preferences.

6 Discussion

Cost and iteration. SpatialMQA planning/perception takes 41.12 minutes per cohort, shared across backbones/seeds; all numerical heads take under 0.5 seconds on A100 80 GB (Appendix D). These batch costs do not measure uncached query latency. Frozen perception and native scores allow reader changes to reuse visual computation. CPU refitting reproduces all eighteen main cells from supplied calibration caches (Appendix A).

7 Conclusion

SpaEx connects object-centered perception to spatial decisions through target-aligned execution and fusion with native VLM scores in the original answer space. Across three benchmarks and two backbones, it improves matched task-adapted VLMs by 5.3–18.9 percentage points while fitting four scalar reader parameters with fixed VLMs and perception tools. Under shared observations, executed comparisons outperform aligned states even with longer reader training and compact inputs.

Candidate-level fusion can recover correct answers ranked first by neither source. Native evidence complements execution on viewpoint tasks, whereas execution accounts for the metric-benchmark gain. The current interface covers short frame and distance comparisons, with gains bounded by execution coverage and geometry quality. Future work should extend it to multi-step relations and study how uncertainty in object bindings and frame estimates affects candidate comparisons.

AI ASSISTANCE

In this work, we used generative AI tools for language and grammar checking of the manuscript and for checking code for bugs. We did not use generative AI tools to generate datasets, formulate hypotheses or mathematical claims, design experiments, implement methods, or interpret results. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the method and evaluation protocol; Appendices A–E specify implementation, data and controls. Anonymous supplementary materials provide code, prompts, numerical-reader refitting inputs and result records. Appendix A explains reproduction from these caches and identifies the assets needed for full-pipeline reproduction.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48, 2016.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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 (CVPR)*, pp. 14455–14465, 2024.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. SpaceTools: Tool-augmented spatial reasoning via double interactive RL. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37109–37120, 2026.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of Thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093, 2024.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017.

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

Wenxiao Fan, Hang Yin, and Kan Li. OrientSAM: Mitigating camera-centric shortcut in multimodal spatial reasoning via orientation-aware spatial alignment. *arXiv preprint arXiv:2607.17657*, 2026.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023.

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 (ICLR)*, 2022.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yatai Ji, An-Chieh Cheng, Yang Fu, Yukang Chen, Han Zhang, Zhaojing Yang, Wei Huang, Ka Chun Cheung, Song Han, Vidya Nariyambut Murali, Pavlo Molchanov, Jan Kautz, Simon See, Hongxu Yin, Ping Luo, and Sifei Liu. Reinforcing dual-path reasoning in spatial vision language models. *arXiv preprint arXiv:2606.17539*, 2026.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9241–9251, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 620–632, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. SpatialPIN: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3D priors. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic AI for spatial reasoning with a dynamic API. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19446–19455, 2025.

István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30, 2021.

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, and Xiangyu Yue. SpaceVista: All-scale visual spatial reasoning from mm to km. In *International Conference on Machine Learning (ICML)*, 2026.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

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

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From Flatland to Space: Teaching vision-language models to perceive and reason in 3D. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

A Implementation and Reproducibility

A.1 Reproduction Materials

The anonymous `spaex_supplement.zip` supports numerical reproduction through worked examples and a CPU refitting script. Calibration features and labels recover constant and adaptive readers for all eighteen model–seed cells while retaining candidate order, support masks and image folds. Predictions are saved before reporting targets are opened, after which the script reconstructs seed summaries and grouped intervals. A fresh extraction reproduced the selected regularizers and predictions. SpatialMQA refits condition on the supplied native adapters; rerunning the full pipeline additionally requires original images, pretrained perception tools and native checkpoints.

The supplement records sources, recipes and results for image-disjoint native fits, compact VLM readers and trained SPAR state readers in `experiment_inventory.md`. Internal-reader implementation modules and records are provided under `evidence/spaex_internal_readout_20260917/`. Reproducing these additional VLM fits requires checkpoints, complete prediction caches and upstream environments that are not bundled; the eighteen numerical refits do not reproduce them. Underlying per-seed tables are supplied as CSV and original table sources in `reporting_records/`, with repeated aggregate results referenced from the main text.

The walkthrough uses the first supported SpatialMQA test row in source order, selected independently of correctness. The person wearing a white hat is both reference and observer, and the outlet is the target (Figure 1). Supplied assets contain unrounded measurements, coefficients and exact facts/vector payloads. The reported displacement is $\delta = (0.662, 0.188, 0.726)$, yielding behind/front/left/right evidence $(0.723, -0.723, -0.363, 0.363)$. The bound regions pass admission, but have not undergone independent semantic verification.

A.2 Native VLM Adaptation

Both backbones use seeds 13/21/42 and rank-16 LoRA, with scale 32, dropout 0.05 and attention $q/k/v/o$ projections. Each adapter receives one epoch on its native-fitting cohort (Table 5). SpatialMQA adaptation uses all 3,640 training questions, and reader fitting uses the 1,446 supported examples from those labels; grouped reader validation therefore conditions on a VLM adapted to the complete training split. SPAR and CV-Bench separate native fitting from reader calibration as specified in Appendices C and B. Adapted comparisons retain their native checkpoint.

A.3 Numerical Reader Fitting

Numerical readers are fitted in float64 with full-batch L-BFGS, at most 200 iterations, history size 20 and strong-Wolfe line search. Gradient and objective-change tolerances are 10^{-7} and 10^{-10}. Accepted fits require finite parameters and scores and a maximum final gradient of at most 10^{-4}. Native and evidence scales initialize to one, adaptive slopes to zero, and the free matrix to identity. For supported training rows $\mathcal{T}_s$, the objective is

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_s} \log \frac{\exp s_{i,y_i}}{\sum_{c \in \mathcal{C}_i} \exp s_{i,c}} + \lambda \Omega(\theta), \tag{8}$$

$$\Omega(\theta) = (\alpha - 1)^2 + \frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_s} \beta(x_i)^2 + b_1^2 + b_2^2. \tag{9}$$

Constant readers remove the adaptive-slope terms; free-operator controls add $\|B - I\|_F^2$. SpatialMQA and CV-Bench select regularization by pooled cross-entropy across three image-grouped training folds (seed 3407), preferring larger regularization in a tie, then refit on all supported rows. SPAR retains the development-fixed $\lambda = 10^{-2}$ on its separate 502-row calibration cohort.

Entropy is normalized by $\log K$. SpatialMQA defines dispersion D as the average population SD of candidate projections across five poses; CV-Bench averages relative range SD over the required regions, including the anchor. Training-fold means and SDs standardize these features, with SD floor 0.05 and clipping to $[-5, 5]$; final refits use all supported training rows. Metric normalization sets $\epsilon = 10^{-8}$. Padded candidates are excluded from softmax, loss and prediction. Semantic-bias

controls use zero-sum, quadratically penalized offsets for candidate meanings; SpatialMQA fits five free offsets together with a positive scale.

For two unequal distances and an inactive normalization floor, Equation 4 produces $\pm 1/\sqrt{2}$, retaining ordering rather than a distance ratio. Geometry reverses a disagreeing native winner when its margin is below $\sqrt{2}\beta/\alpha$. Execution and fusion can consequently agree despite a nonzero native contribution. With four candidates, normalized evidence also preserves relative gaps.

Axis magnitude. Averaging unit pose axes bounds their mean norm by one. We assess the resulting attenuation by normalizing each nonzero mean axis or removing D from the gate, holding observations, support, folds and regularizer selection fixed. The axis-normalized variant retains the original D. Training-axis norms have median 0.999, fifth percentile 0.959 and minimum 0.188;

Table 4: **Axis norm and dispersion.** SpatialMQA accuracy (%, three-seed mean ± sample SD).

	Qwen		LLaVA	
Reader	Mean axes	Unit axes	Mean axes	Unit axes
Execution + fallback	73.82 ± 0.79	73.73 ± 0.79	72.15 ± 0.60	72.06 ± 0.60
Constant fusion	**79.28** ± 1.14	**79.18** ± 1.16	**77.08** ± 0.44	**77.14** ± 0.43
Entropy gate	79.18 ± 0.85	79.12 ± 0.94	77.01 ± 0.54	77.08 ± 0.47
Entropy + dispersion gate	79.12 ± 0.94	**79.18** ± 0.98	**77.08** ± 0.44	77.08 ± 0.44

question-mean norm and D correlate at $r = -0.90$. Normalization shifts accuracy by less than 0.10 point, and removing D by less than 0.07. These small changes do not identify attenuation as the explanation for constant fusion's gains.

A.4 Inference Contract and Statistical Analysis

Native scores are mean log-likelihoods of the original numeric answer suffix, with shared-prefix scoring verified against full per-candidate forwards. All comparisons retain their matched adapter, original candidate order and complete cohort. Inference proceeds as follows:

1. Score image, question and candidates; bind roles and the operation from public inputs.
2. Acquire the required measurements; verify scope, binding, numerical validity and the benchmark-specific admission conditions in Appendices B and C.
3. Evaluate every original candidate in the requested frame. If inputs are missing, the frame is degenerate or a condition is unsupported, return the native score vector unchanged.
4. For supported inputs, compute H, D, combine scores with the fitted positive weights and choose the highest-scoring candidate.

Admission verifies availability and validity, not perceptual correctness. Candidate-energy tests establish permutation equivariance, invariance to common proper rotations of displacement and axes, the distinction between camera range and reference distance, and exact fallback. These properties belong to the numerical operator; they do not imply image-rotation invariance of perception or prompt-order invariance of the VLM. Positive weights preserve even incorrect shared winners.

A repair corrects a native error; a damage overturns a correct native answer. Their count difference, divided by cohort size, is the accuracy gain. Paired intervals use 2,000 bootstrap draws (seed 3407), retaining all questions within sampled SpatialMQA images or CV-Bench capture groups. Fitted-seed correctness differences are averaged within each resampled group, so intervals condition on the three fitted seeds. CV-Bench image-group intervals provide sensitivity checks. Official SPAR test data lack physical-scene identities, precluding scene-level uncertainty estimates; independent-question intervals are not substituted. Appendix E specifies the internal-reader bootstrap.

Accuracy and semantic Macro-F1. Table 1 augments benchmark accuracy with Macro-F1 over answer meanings rather than choice positions. Each class contributes $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, averaged over classes present in the fixed reference labels and shared by all methods and seeds. SpatialMQA has six directional meanings. SPAR first computes task-level F1 and then equally

averages its four tasks: distance classes identify green/blue points, and relation classes are complete three-axis tuples, including empty components, with no partial credit by axis. The tasks contain 2/24/23/24 reference classes; one public observer–camera tuple is absent from the references. CV-Bench equally averages Depth and Distance F1, whose red/blue and blue/green classes follow the marker bindings in public questions. All original questions and fallbacks remain included. Scores are computed per seed before taking the mean and sample SD; Vanilla has one endpoint. This metric is supplementary and never selects models. Confusion counts and the standalone table-regeneration script appear under `evidence/spaex_main_f1_20260917/`.

B Data Protocols and Perceptual State

Table 5: **Cohorts for training and evaluation.** Native fitting, reader calibration, development and test counts. CV-Bench uses external EmbSpatial-SFT examples for training and calibration.

Dataset	Role in this paper	Fit and evaluated scope
SpatialMQA	Matched main test	3,640 native / 1,446 reader fit; 536 dev; 1,076 test
SPAR-Bench	Matched main test, single-view reasoning	1,024 native / 502 reader fit; 512 dev; 1,378 test in four tasks
CV-Bench 3D	Main test and fusion boundary	1,200 test; external training only
EmbSpatial-SFT	External CV-Bench training source	566 native / 286 supported calibration; 429 dev proxy

B.1 SpatialMQA Observer-Aligned State

Our SpatialMQA split contains 3,640 overlap-filtered training, 536 development and 1,076 test questions, preserving the original two/four/six candidates. Only images, questions and candidates enter inference; gold programs, traces, answers and role boxes remain excluded. Filtered training contains no test image. Section 5 tests exclusion of 15 questions sharing development images.

A question-only plan distinguishes the target, comparison reference, observer, selection constraints and explicit facing or position conditions. The vanilla Qwen3-VL-4B planner runs in text-only mode with greedy decoding, a 320-token output limit and batches of four. It receives only the question; downstream candidate interpretation handles the visible options. Every claimed mention must occur in the public question, and plans may contain neither answer labels nor candidate rankings nor a gold relation. The plan requests perceptual observations without passing the answer to localization, segmentation, geometry or pose estimation.

Common coordinates and repeated observations. MoGe-2 positions and MeTRAbs poses share the right-handed OpenCV camera convention: x points right, y down and z away from the camera. MeTRAbs receives MoGe intrinsics scaled to original pixel coordinates, together with identity extrinsics. Its internal crop/flip transforms are reversed before poses are returned, with no additional axis swap, reflection or fitted rotation. Poses use millimeters, whereas geometry retains its estimated metric scale. Unit normalization removes scale from direction vectors; the camera conventions, rather than normalization, establish coordinate agreement.

The five pose estimates come from deterministic test-time augmentations. The pinned implementation pairs gamma values $(.6, .7, .8, .9, 1)$, in-plane rotations $(-25, -12.5, 0, 12.5, 25)$ degrees and crop scales $(.8, .9, 1, 1.05, 1.1)$, applying horizontal flips to the second and fourth views. Settings are `num_aug=5`, `average_aug=False`, and `antialias_factor=1`. No sampling seed is involved, and the estimates are not independent random samples. We retain all five poses in original camera coordinates, average their candidate projections and summarize dispersion through D.

For a shared observer/reference, let e_L, e_R be the estimated left/right ear locations and n the nose in pose coordinates. The head-right axis is defined by

$$u = \mathrm{unit}(e_R - e_L). \tag{10}$$

Defining $f_0 = n - (e_L + e_R)/2$ gives the forward axis

$$f = \mathrm{unit}(f_0 - (f_0^\top u)u), \qquad d = \mathrm{unit}(f \times u). \tag{11}$$

The visible candidates select from directions $-u, u, -d, d, f, -f$, corresponding to left, right, above, below, front and behind. Option-aligned projections are averaged across the five repeated poses. The head-down axis supplies a pose-based gravity approximation; it neither measures true gaze nor defines arbitrary intrinsic object frames.

Component-wise medians of MoGe-2 points within selected SAM regions define target and reference positions. Projection uses their normalized displacement and requires resolved bindings, valid points, nondegenerate displacement and head axes. Person admission additionally requires valid head/body frames in all five poses and at least 0.5 IoU between a Grounding DINO person detection and the observer region. Body axes and detection overlap gate admission without adding arguments to Equation 3. The current observer-frame implementation thus needs an accepted person observation, although the state interface represents broader object roles. Extra facing/position conditions, unresolved distinct observer/reference roles and unavailable visual dependencies remain unsupported. Admission is answer-independent, covering 1,446/223/380 train/dev/test examples.

Support failure stages. The first recorded stopping condition partitions the 1,076 test rows. Required actor-view plans are absent for 489 rows (395 unspecified, 58 camera, 36 hypothetical); 41 add unsupported frame conditions, 66 have unresolved/colliding bindings, 13 bind distinct observer/reference objects, five lack valid displacement, and 82 fail person/frame validity. The remaining 380 execute. All 82 person/frame failures have numerically valid head/body axes but fail person overlap. Binding failures comprise 63 missing-role cases and three collisions, potentially arising in planning/coreference or localization. First-stop records identify where processing ends, not the underlying error cause or the amount of missed support.

The frame-removed ablation replaces measured observer axes with signed camera-coordinate candidate axes and refits the reader. It keeps the primary support set, isolating frame removal without expanding coverage to previously unsupported examples. Raw execution retains the same supported state and candidate projections while omitting native scores on supported inputs.

B.2 SpatialMQA with Separate Calibration Images

To separate reader calibration images from native adaptation, a fixed, label- and support-independent hash ordering allocates one fifth of the original training image identities to calibration. Fresh adapters train for one epoch on the remaining 2,907 questions from 2,784 images. Calibration comprises 733 questions from 696 disjoint images, with 310 satisfying the unchanged support rules. The split is shared by both backbones and seeds 13/21/42. Frozen perception is reused, while new native scores form the calibration features. Constant/adaptive readers retain image-grouped cross-entropy regularizer selection; no additional labeled data are introduced.

Evaluation retains all 1,076 questions, the same 380 supported observations and exact fallback to the new native model on the other 696. Per-seed records give development counts, heads and costs.

Table 6: **Calibration on images disjoint from native training.** SpatialMQA full-test accuracy (%, three-seed mean ± sample SD). Readers share observations and each seed's new adapter.

Reader	Qwen	LLaVA
New native	73.95 ± 0.56	70.76 ± 0.62
Execution + fallback	73.67 ± 0.62	71.07 ± 0.62
Constant fusion	**79.00** ± 0.65	**76.08** ± 0.33
Adaptive fusion	78.81 ± 0.52	76.02 ± 0.43

Adaptive fusion improves over its matched native branch by 4.86/5.27 points, with paired image-bootstrap intervals [3.34, 6.38]/[3.64, 6.91], and over execution by 5.14/4.96 points, with intervals [3.70, 6.69]/[3.51, 6.46]. All six pairs outperform both branches; intervals condition on the three fitted seeds. Constant fusion reaches 79.00%/76.08%, compared with adaptive fusion's 78.81%/76.02%.

Disjoint calibration images therefore retain branch complementarity on the main SpatialMQA evaluation cohort. Because the allocation also changes native-update count and calibration size, differences from the primary configuration cannot be assigned to image separation alone.

B.3 CV-BENCH METRIC STATE AND EXTERNAL TRAINING

CV-Bench's public RGB markings identify red/blue candidate regions for Depth and a red anchor with blue/green candidates for Distance. A fixed outline detector recovers those markings directly from RGB without consulting hidden box metadata. MoGe-2 processes the same marked image. Valid regional points must be finite, have positive camera z and satisfy the model's validity mask. Each region requires at least 16 pixels and valid fraction 0.5. Duplicate, unresolved or invalid regions leave the entire operation unknown.

For valid regional points $\{P_j\}$, the stored measurements are $p = \mathrm{mean}_j\, P_j$, $\rho = \mathrm{mean}_j\, \|P_j\|_2$, and relative spread $\mathrm{std}_j(\|P_j\|_2)/\rho$. Camera closeness compares ρ, whereas anchor distance compares $\|p_c - p_r\|_2$. These regional means approximate benchmark comparisons constructed from separated ground-truth 3D box vertices (Tong et al., 2024) and may contain background. Camera range is Euclidean distance rather than camera z alone. With two unequal candidates, Equation 4 yields equal-magnitude opposite signs, encoding ordering rather than an absolute distance margin. The free operator can only rescale this scalar evidence.

External adaptation uses a fixed prefix of released EmbSpatial-SFT (Du et al., 2024), restricted to four-object nearest/farthest questions. Both branches see the same four marked, annotated objects in a fixed candidate order, with no answer-dependent highlighting. These annotations are training inputs and never enter CV-Bench inference. Disjoint native-fitting, calibration and development partitions contain 128/128/96 image groups and 566/536/429 questions. Exact and connected near-duplicate checks exclude cross-partition image matches. The 536 calibration rows contain 286 supported reader examples, and external development fixes the interface before CV-Bench reporting.

Native adaptation uses rank-16 attention LoRA with scale 32 and dropout 0.05, trained by AdamW at learning rate 5×10^{-5} and weight decay 0.01 for one source-ordered epoch with batch size one. Calibration images are unseen by that adapter. The additional-label control continues the exact native weights at seeds 13/21/42 for one epoch on the same 286 supported calibration examples, restarting the optimizer but retaining the native recipe. Evaluation includes both unrestricted continuation and original-native fallback on the common unsupported set. This control matches additional examples without matching computation or establishing an optimal continuation schedule.

CV-Bench's 674 underlying images form 189 capture groups, some containing multiple views or questions. Evaluation retains all 1,200 original 3D questions, including 238 unknown-state fallbacks. Exact and near-image comparisons against the 352 fine-tuning/calibration/development groups find no overlap, without establishing absence from VLM or perception-tool pretraining. The evaluation does not include the 2D configuration.

C SPAR-BENCH SINGLE-VIEW REASONING

C.1 DECLARED SCOPE, TRAINING DATA AND SEPARATION

Evaluation covers the complete original `distance_infer_center_oo`, `obj_spatial_relation_oo`, `spatial_imagination_oc` and `spatial_imagination_oo` tasks, containing 340/364/372/302 questions. We retain all 1,378 original IDs, candidate lists and row ordering across ScanNet (699) and ScanNet++ (679). The scope excludes other single-view perception and multi-view tasks. Equal-task macro accuracy follows official task aggregation, with ordinary accuracy and task counts reported alongside it. Numeric-candidate likelihood scoring differs from the official generated-response parser. Comparisons are matched and SPAR-adapted, rather than full-benchmark or zero-shot results.

Official SPAR-7M training questions provide 1,024/512/512 task-balanced rows for native fitting, calibration and development. Both physical scenes (14/15/15) and images (531/372/390) are disjoint across these partitions. Fixed ID ordering selects at most one question per image/task without consulting answers, model scores or support. All 44 scenes occur on the official ScanNet training list and none on its validation/test lists. RGB comes from the public SpaceVista (Sun et al., 2026) RGB repack, matched to official SPAR training filenames. SpaceVista answers, depth, poses, scene graphs and reasoning traces are excluded. Original image bytes are unpublished, so matching uses released filenames.

The official test release includes marked RGB but omits per-row physical-scene IDs and source-image paths. Test separation consequently also relies on the SPAR authors' statement that benchmark images come from validation data; an independent per-test scene-overlap audit cannot be reconstructed. Adapters and fusion heads were frozen using training/calibration data, and complete predictions for all three seeds were saved and accepted before official answer reporting. CoT, SoM, VCD and additional-label controls apply fixed recipes after the primary results are known, without changing original adapters, heads or primary predictions.

C.2 Visible Markers and Estimated Object State

A public task category selects the parser/executor template, while the question binds objects, comparator and viewpoint. Marked RGB supplies measurements, and scoring preserves the original options. The native branch sees the image, question and options. Both training branches receive identically rendered public markings; evaluation uses the original marked JPEGs. Test role coordinates, 3D annotations, pose, depth, programs, traces and answers never enter perception or inference.

Visible point/box markers are recovered by a color-component detector that joins compatible fragments of an outline, without access to hidden boxes. A marked box prompts SAM, whose highest-quality mask is retained. A marked point uses surrounding geometry from an annulus of radii 24 and 48 pixels, excluding the painted point. MoGe-2 estimates camera-space points and normals from the same RGB. Each object requires at least 16 finite points satisfying the validity mask and lying outside colored markers. The state retains coordinate-wise median position, camera-z interquartile range and point count. Surface medians approximate the object locations.

World-up estimation uses finite, valid normals outside the markers. Approximately 20,000 normals are subsampled; normalized normals satisfying $|n_y| > 0.7$ are retained and oriented toward negative camera y. At least 128 such normals and median-vector norm above 0.6 are required, after which the normalized median defines estimated world-up. This construction assumes roughly upright cameras and suitable scene surfaces rather than an independently measured gravity direction. Missing required geometry or frames triggers native fallback.

The shared support admits 1,356 of 1,378 official questions, leaving 22 fallbacks: 11 invalid point markers, seven invalid box markers, one missing marker, one unavailable object geometry and two unavailable up axes. Admission is independent of correctness and common to all backbones and seeds. Four supported cases have exact top-score ties and retain their original decision rule. Even changing all four could alter ordinary accuracy by only approximately 0.29 point or task-macro accuracy by 0.30, below the observed 2.50–3.30-point fusion increment.

C.3 Frames, Compound Relations and Distance Evidence

Red-target and blue-reference positions a, b define the camera-frame tuple

$$u = (a_x - b_x,\ -(a_y - b_y),\ \|a\|_2 - \|b\|_2). \tag{12}$$

Its components represent horizontal/upward displacement and range difference. Imagined-view queries bind the green object to observer position o and the blue object to facing anchor f. Estimated up axis g completes the frame:

$$z' = \mathrm{unit}(f - o), \qquad x' = \mathrm{unit}(z' \times g), \qquad y' = \mathrm{unit}(x' \times z'). \tag{13}$$

Horizontal and vertical comparisons use $(a-b)^\top x'$ and $(a-b)^\top y'$; explicit world-coordinate queries instead use g vertically. Reference b denotes the observer for ○○ and the yellow object for ○○. Their respective depth measurements are $(a-o)^\top z'$ and $\|a-o\|_2 - \|b-o\|_2$. Degenerate directions and unresolved facing clauses remain unsupported.

Each original answer specifies signed or indistinguishable relations. At tolerance $\eta = 0.1\,\mathrm{m}$, positive relations occupy $[\eta, \infty)$, negative relations $(-\infty, -\eta]$, and omitted/indistinguishable axes $[-\eta, \eta]$. Equation 5 adds the distances from measurements to these intervals. Candidate words fix the signs: right/above/farther/front are positive on the corresponding measurements, and left/below/closer/behind negative. Regional variation sets the scale:

$$\sigma = \max\left(0.1,\ \frac{1}{2|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathrm{IQR}(z_v)\right). \tag{14}$$

Object-distance queries use the red point as anchor and green/blue points as candidates. For $d_c = \|p_c - p_{\text{red}}\|_2$, set $\kappa = -1$ for nearer and $+1$ for farther, giving $\text{clip}((\kappa d_c - \max_k \kappa d_k)/\sigma, -12, 0)$. This compatibility retains distance gaps, unlike the constant-magnitude binary ordering in Equation 4. Tolerance, spread and clipping specify compatibility rather than calibrated uncertainty.

Adaptive-reader features comprise normalized native entropy and $D = \overline{\text{IQR}(z)}/\max(0.1, \overline{\|p_v\|_2})$. Standardization, clipping and optimization follow Appendix A, with regularization fixed at 10^{-2} before official evaluation. All 502 supported rows in the separate 512-question calibration cohort fit the four adaptive or two constant-reader parameters for each exact adapter. Test results neither select nor refit the primary model.

C.4 Task and Source Differences

Table 7: **SPAR performance by task and source.** Within-task/source ordinary accuracy (%, mean across three seeds). Equally averaging the four task rows recovers the primary macro metric.

Task	N	Qwen exec.	Qwen SpaEx	LLaVA exec.	LLaVA SpaEx
Object distance	340	91.18	90.78	90.10	88.82
Camera-frame relation	364	77.29	81.50	77.20	80.68
Imagined view / observer	372	71.24	75.99	71.51	75.45
Imagined view / object	302	75.83	80.46	75.61	79.47

Source	N	Q LoRA	Q exec.	Q SpaEx	L LoRA	L exec.	L SpaEx
ScanNet	699	70.00	79.26	85.07	67.67	78.87	83.74
ScanNet++	679	62.44	78.25	78.99	56.85	78.11	78.20

Every model–seed pair benefits from fusion over execution, with Qwen/LLaVA mean macro gains of 3.30/2.50 points. Constant fusion is only 0.51/0.18 points below the primary, leaving the general advantage of input-dependent weights unestablished. Net repairs concentrate on ScanNet; ScanNet++ shows a small Qwen gain and near-zero LLaVA gain, including one lost answer for LLaVA seed 13. Training RGB comes only from ScanNet, but does not explain these source differences.

Pooling supported inputs over three fitted seeds, fusion retains 363/482 Qwen and 363/496 LLaVA native-only successes and repairs 40 both-wrong outcomes for each backbone. Pooling repeats seed–question observations and does not enlarge the independent sample. Positive weights preserve a shared branch winner, including its correctness when both branches are correct.

C.5 Native Adaptation and Category Controls

The development intervention extends a 128-example native prefix to 1,024 examples/updates while preserving optimizer/RNG state, perception, execution and the calibration cohort; heads are refitted for each adapter. At seed 13, Qwen/LLaVA fusion rises from 77.54/77.73% to 84.57/83.98% macro. Including the same 1,024-update endpoints at seeds 21/42 yields native/execution/fusion means of 75.20/74.67/83.92% and 70.25/74.48/83.85%. Native-confidence AUC on branch-only successes rises from 0.651/0.675 to 0.810/0.803 at seed 13, while official-test three-seed AUC is lower at 0.716/0.682. AUC measures discrimination; data volume and update count change together.

Additional-label controls train each final native adapter for one epoch with fresh AdamW on the same 502 supported calibration examples. We score both unrestricted predictions and common-support predictions with original-native fallback, without refitting fusion heads to continued scores. Unrestricted macro accuracy is $64.56 \pm 0.61/60.97 \pm 0.42\%$; Table 3 gives common-support results. This schedule neither explains the fusion gain nor establishes an optimal native-training budget.

Public category in the native prompt. The native prompt originally contains only image, question and candidates. The control adds the literal prefix `Task category: {task_type}`, drawn from the same public field used to select execution templates. Images, candidate order, numeric scoring and all six fitted adapters remain unchanged. No further training or generated rationale is introduced, and every task and seed remains included.

Table 8: **Adding the SPAR category cue to native prompts.** Task-macro accuracy (%, three-seed mean ± sample SD). Common fallback retains original native vectors for 22 unsupported questions; unrestricted scoring applies the prefix to all 1,378 questions.

Native input	Qwen	LLaVA
Original question	**66.01** ± 0.48	62.17 ± 0.30
Question + public category	65.80 ± 0.55	**62.29** ± 0.31
Category with common fallback	65.80 ± 0.55	62.27 ± 0.27

Adding the prefix changes Qwen/LLaVA macro accuracy by $-0.21/+0.12$ points, or $-0.21/+0.09$ with common fallback. Under this fixed encoding, category access does not explain the main improvement. The supplement retains task counts for all six cells.

C.6 Shared Aligned-State Readout

APC transforms scenes into a requested viewpoint (Lee et al., 2025); its inspected official release implements visual readout but leaves the numerical branch unfinished. We construct a scoped numerical APC-style control from SpaEx's accepted SAM/MoGe observations. Shared perception permits a comparison of transformed-scene readout, without claiming to reproduce APC's full perception, rendering or prompting system.

The comparison retains all 1,378 SPAR questions, both backbones and three seeds. Camera tasks receive camera-frame centers; imagined-view tasks translate the origin to the bound observer and adopt the same look-at rotation as SpaEx. Object positions, the original camera origin and estimated world-up are expressed in that frame, alongside measured visible-surface depth spreads. A fixed four-decimal table supplements the original image, question, candidates and public category. The state block states "Coordinates are meters: x right, y up, z forward" and identifies either the camera frame or the imagined observer at green facing blue. It includes transformed world-up and camera position. Its "observed depth spread" column preserves regional camera-depth interquartile ranges after reframing without explaining that statistic. It omits the executor's 0.1 m relation tolerance, compatibility penalties and spread normalization, as well as native scores and entropy. Candidate compatibility, fused scores, gold geometry, programs and answers are likewise absent. The interfaces therefore differ in supplied information and conventions as well as numerical computation.

Both state readouts preserve original numeric-suffix scoring. One uses the exact native adapter without further fitting. The other trains from that adapter on the same 502 supported calibration examples as numerical fusion and native continuation. A single source-ordered epoch comprises 502 updates with rank-16 LoRA (scale 32, dropout 0.05), AdamW at 5×10^{-5}, weight decay 0.01 and batch size one. The prompt and endpoint are fixed without search or checkpoint selection, and serialization is identical across fitting, development and reporting. All controls retain the entire original native vector for the 22 unsupported rows.

Table 3 shows that training raises aligned-state macro accuracy from 64.18%/60.74% to 71.47%/68.22%; evaluating only the untrained reader would understate this interface's capacity. Native continuation on the same examples reaches 64.65%/60.88%, execution with fallback reaches 78.88%/78.60%, and numerical fusion reaches 82.18%/81.10%. Fusion exceeds the trained state reader in every model–seed pair, with mean gains of 10.72/12.88 points. These comparisons support explicit candidate computation under the stated recipe while leaving arithmetic, representation, trainable capacity and computation unequal. They do not establish superiority over full APC.

Development accuracy is 76.95%/73.96%, against native accuracy of 75.20%/70.25%; the supplement gives all task/seed counts. Additional fitting takes 3.82–6.19 minutes on A100 80 GB. Worker time totals 14.37–27.06 minutes for loading, fitting, development and final scoring, excluding shared perception and original adaptation. Missing test scene IDs preclude scene-cluster intervals.

D CONTROLLED COMPARISONS AND FUSION BOUNDARIES

D.1 EXTERNAL BASELINE RECIPES

Vanilla evaluates the released checkpoint without LoRA. Every other external baseline retains the exact primary-cell adapter, original question, ordered candidates and complete cohort, with no additional fitting. Baseline outputs condition a common numeric-option likelihood scorer. Table 1 gives main accuracies, and `reporting_records/` supplies the corresponding per-seed counts.

CoT. CoT appends "Let's think step by step." to the image-conditioned question and candidates, greedily generates at most 512 tokens, then adds "Therefore, the answer option is:" before scoring original numeric suffixes (Kojima et al., 2022). The generated response conditions likelihood scoring rather than being parsed as an answer. Neither SpaEx state nor execution is supplied. After answer-supervised adaptation, SPAR median rationale lengths are 9/8/9 Qwen tokens and 3/3/3 LLaVA tokens across seeds; 72/1/46 and 3/1/0 responses reach the cap. Capped responses remain included. The fixed instruction therefore need not produce extended reasoning.

Set-of-Mark. SoM's automatic SAM route (Yang et al., 2023) uses cached SAM-B in place of the demonstration's SAM-H. Images are resized to a 640-pixel short side; a 32×32 grid generates masks at predicted-IoU/stability thresholds 0.88/0.95 and NMS threshold 0.7. Region numbering follows decreasing area, with labels placed at maximum-interior-distance locations and explicitly distinguished from answer indices in the prompt. RGB alone determines the marks, which are shared across backbones and seeds. Scoring uses marked images without CoT or SpaEx evidence.

VCD. We apply VCD's clean/noisy contrast (Leng et al., 2024) within the original candidate vocabulary, retaining defaults of strength $\gamma = 1$, plausibility ratio $\eta = 0.1$ and diffusion step 500 in a 1,000-step sigmoid schedule. Final-digit logits $z_c, \widetilde{z}_c$ determine admissibility and score:

$$z_c \geq \max_{k \in \mathcal{C}} z_k + \log \eta, \qquad s_c^{\mathrm{VCD}} = (1+\gamma) z_c - \gamma \widetilde{z}_c. \tag{15}$$

For suffix length L with a shared prefix, the mean-score threshold includes $\log \eta / L$; common prefix terms leave final-digit rankings unchanged. An image-keyed generator adds noise to normalized pixels from the unchanged processor, preserving crop/grid metadata. Each branch makes one forward pass, verified against full candidate scoring. Coefficients remain fixed across evaluations. The control assesses spatial selection rather than VCD's original open-generation hallucination task.

D.2 COMPONENT REPLACEMENTS

Table 2 contains constant fusion, execution and the primary. Table 9 extends the comparison to a free bilinear operator, answer biases and geometric replacements. Camera axes remove the measured observer frame without changing primary support. CV-Bench's image-plane control substitutes vertical image position for camera closeness and image-center distance for anchor comparison. Meaning biases follow candidate semantics, whereas position biases follow answer indices.

Table 9: **Component-replacement controls.** Full-cohort accuracy (%, mean over three seeds), holding native adapters and primary support fixed. Replacements are evaluated where applicable.

Benchmark	Replacement	Qwen	LLaVA
SpatialMQA	Free operator	**79.28**	**77.32**
	Candidate-meaning bias	73.76	71.84
	Camera axes	73.17	71.84
CV-Bench 3D	Free operator	**94.25**	**92.53**
	Candidate-meaning bias	87.44	76.56
	Image-plane geometry	87.50	78.08
	Candidate-position bias	87.81	74.28

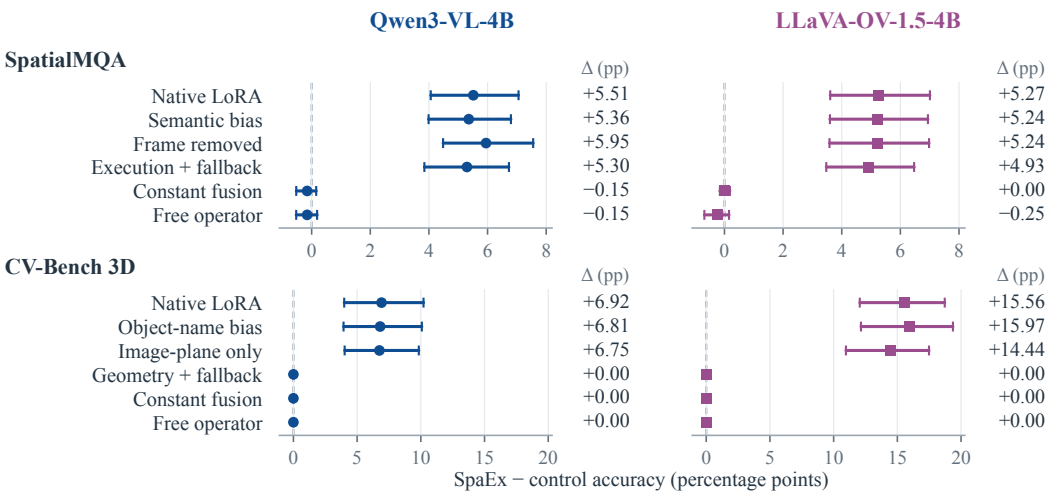

Figure 5: **Paired gains over component controls.** Primary minus control, with 95% image/capture-cluster intervals conditional on the three fitted seeds.

D.3 State, Facts and Training Budget

SpatialMQA readers start from matched native adapters and train for one source-ordered epoch on the same 1,446 supported examples, using AdamW at 5×10^{-5}, weight decay 0.01 and batch size one. They share images, original candidates, numeric scoring and LoRA capacity. Native continuation introduces no measurements. The score-only input contains candidate semantics, original native scores and normalized entropy; facts additionally supply candidate dot products and pose dispersion. Vectors instead supply candidate normals, displacement, dispersion and camera-axis conventions, requiring the VLM to compute dot products. Augmented readers share 380 supported test inputs and 696 original-native fallbacks, while unrestricted continuation scores every question.

Full JSON preserves formatter float precision. The longer-budget facts/vector readers resume for a second identical epoch with AdamW moments, step count and RNG intact. Compact inputs keep field/candidate order and native information in a fixed four-decimal representation: facts round computed dot products, whereas vectors round their components. Mean facts/vector text lengths fall from 333/567 to 190/290 tokens, with compact maxima of 240/368 and no requested truncation. Observations and training data are shared, while computation and input length differ.

Table 10: **SpatialMQA readout by format and budget.** Full-test accuracy (%, three-seed mean ± sample SD; $N = 1,076$). Updates follow native adaptation; the longer budget uses two passes.

Reader	Extra updates	Qwen	LLaVA
Original native	0	73.61 ± 0.98	71.81 ± 0.63
SpaEx	0	**79.12** ± 0.94	**77.08** ± 0.44
Native continuation, support	1,446	74.72 ± 0.90	72.18 ± 0.86
Native continuation, all	1,446	74.35 ± 0.25	72.71 ± 0.99
Native scores only	1,446	74.13 ± 1.08	71.81 ± 0.48
Vectors, full JSON	1,446	74.60 ± 1.30	72.24 ± 0.38
Facts, full JSON	1,446	78.31 ± 0.51	76.24 ± 0.99
Vectors, full JSON	2,892	75.40 ± 1.69	71.62 ± 0.39
Facts, full JSON	2,892	77.88 ± 0.41	76.05 ± 1.40
Vectors, compact	1,446	74.69 ± 0.78	72.03 ± 1.30
Facts, compact	1,446	78.10 ± 0.30	76.15 ± 1.64

Every model–seed pair favors facts over vectors at both full-JSON budgets. Removing 15 questions with development-shared images leaves gains of 3.61/4.02 points after one epoch and 2.45/4.49 after two, with paired image intervals [2.44, 4.93]/[2.77, 5.40] and [1.35, 3.59]/[3.22, 5.83]. The Qwen gap narrows by 1.24 points [−2.26, −0.25], whereas LLaVA's widens by 0.43 [−0.49, 1.30]. These endpoints establish neither convergence nor an inability to learn the computations.

Compact facts exceed compact vectors by 3.41/4.12 points, with intervals [2.26, 4.57]/[2.94, 5.43]. Relative to full-JSON facts, their differences are $-0.22/-0.09$ $[-0.80, 0.31]/[-0.71, 0.49]$, insufficient to establish equivalence. SpaEx improves over compact facts by 1.02/0.93 [0.09, 1.92]/[0.12, 1.80] and over longer-budget facts by 1.24/1.02 [0.31, 2.12]/[0.15, 1.90]. Executing four-decimal vectors in a label-free check preserves every ranking and tie across 1,446 training and 380 supported test rows; maximum deviation from rounded facts is below 1.43×10^{-4}. Rounding preserves these geometric decisions; reader arithmetic and representation remain coupled.

D.4 Coverage, Complementarity and Errors

On 380 supported SpatialMQA questions, native/execution/fusion accuracy is 70.70/71.32/86.32% for Qwen and 70.35/71.32/85.26% for LLaVA. Exact fallback implies $\Delta_{\text{full}} = (380/1076)\Delta_{\text{supported}}$. This separates coverage from net repair: an admitted geometric observation may still support an incorrect answer. Figure 6 reports the group sizes and fusion

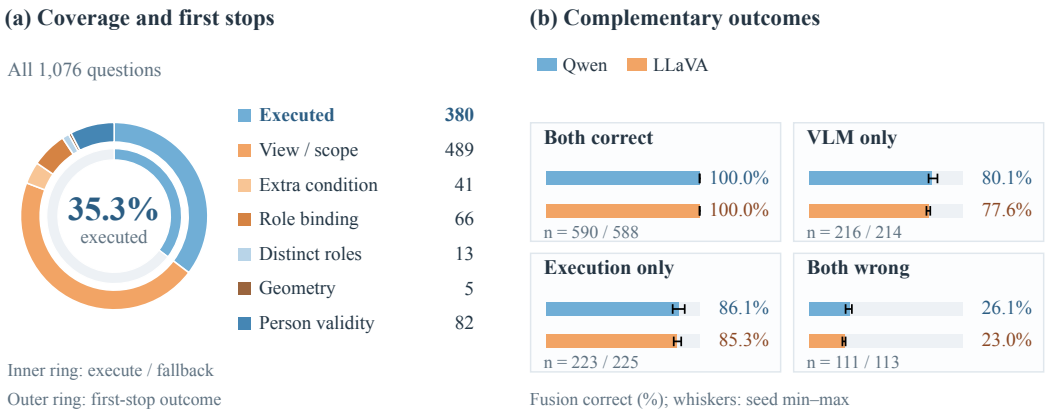

Figure 6: **Coverage and conditional fusion outcomes on SpatialMQA.** (a) Execution support and first-stop conditions. (b) Fusion accuracy by branch correctness, with black whiskers denoting seed ranges. Counts pool three seeds over the same 380 supported questions for each backbone.

success rates. Positive weights guarantee the 100% both-correct outcome, so disagreements and third-choice repairs expose complementarity (Figure 3). Fusion also overturns 43/48 correct native Qwen/LLaVA predictions across the three SpatialMQA seeds. In test row 47, Qwen seed 13 correctly answers "front" for the tall glass relative to the woman in red, whereas execution and fusion answer "behind". Aggregates include all errors; records give per-seed repairs and damages.

D.5 CV-Bench: Execution and Native Adaptation

The complete CV-Bench evaluation retains 1,200 questions and 238 native fallbacks. Execution, constant fusion and the primary predict identically (Table 2). Because Qwen Vanilla exceeds LoRA by 4.00 points, part of the matched-LoRA improvement recovers adaptation-induced loss. SpaEx also surpasses Vanilla Qwen/LLaVA by 2.92/17.11 points, with capture-cluster intervals [0.50, 5.87]/[13.91, 20.53].

Replacing LoRA with the released VLM, we recalibrate numerical readers on the same 286 external examples while fixing observations, folds and support. Execution and both fusion variants again predict identically. Accuracy shifts by $+0.50/-1.61$ points relative to the adapted configuration, entirely on fallback questions (Table 11). Geometry thus helps without native adaptation.

Unrestricted continuation on the same 286 extra labels reaches 91.53/80.94%. The primary retains gains of 2.72/11.58 points with capture-cluster intervals [0.17, 5.68]/[8.52, 13.97]; common-support continuation yields gaps of 3.44/11.50 [0.72, 6.50]/[8.42, 13.80]. Qwen continuation achieves 94.39% on Depth and 88.67% on Distance, against primary accuracy of 94.50/94.00%. Its task-level gain intervals $[-1.32, 1.45]/[-0.08, 11.70]$ span zero. LLaVA continuation attains 86.22/75.67%, with positive primary-gain intervals [4.92, 9.95]/[10.62, 19.54]. The stronger native control there-

Table 11: **CV-Bench fusion with released native weights.** Full-cohort accuracy (%), using one released checkpoint and one fitted reader per backbone.

Reader	Qwen	LLaVA
Vanilla VLM	91.33	75.42
Execution + Vanilla fallback	**94.75**	**90.92**
Vanilla + constant fusion	**94.75**	**90.92**
Vanilla + adaptive fusion	**94.75**	**90.92**

Table 12: **CV-Bench performance by task and source.** Mean accuracy over three seeds (%); G denotes capture-group count. Task and source rows form separate partitions of the benchmark.

			Qwen		LLaVA	
Subset	N	G	Native	SpaEx	Native	SpaEx
Depth	600	173	89.67	94.50	81.06	93.56
Distance	600	83	85.00	94.00	72.89	91.50
Hypersim	400	23	89.08	95.17	76.17	93.75
SUNRGBD	400	149	95.25	98.58	82.08	97.67
nuScenes	400	17	77.67	89.00	72.67	86.17

fore largely closes Qwen's camera-closeness gap. Every task outcome remains included, with image-group sensitivity intervals in the reporting records.

D.6 Measured Costs

Table 13: **Cost of additional SpatialMQA reader training and scoring.** Ranges across three seeds on A100 80 GB. Fitting includes 1,446 updates; scoring includes 1,076 questions for unrestricted continuation or 380 supported questions for augmented readers.

	Qwen		LLaVA	
Reader	Fit (min)	Score (s)	Fit (min)	Score (s)
Native continuation (all questions)	8.2–9.0	95–126	9.0–9.6	104–128
Native scores only	8.3–11.7	42–75	9.0–9.2	44–76
Executed facts + native scores	8.2–11.8	39–65	8.2–9.4	43–46
Aligned vectors + native scores	8.7–9.3	46–58	8.2–12.4	48–53

Reader-stage timings include model loading and cache writes, and scoring includes answer-format verification. Shared perception and original-native scoring are excluded. A second epoch requires 8.30–12.34 minutes; compact fitting requires 8.02–8.91 minutes and supported-row scoring 38.48–51.62 seconds. These costs concern fixed implementations, not equally tuned deployment systems.

Shared planning/perception requires 41.12 minutes for SpatialMQA and 2.59 for CV-Bench, increasing to 42.65/4.42 minutes with setup and orchestration. SpatialMQA comprises 975.01 seconds of question planning, 927.25 of grounding and 564.98 of geometry/pose. Original native scoring adds 94–179/185–361 seconds. Applying all cached non-native heads requires 0.19–0.29/0.22–0.42 seconds per complete SpatialMQA/CV-Bench cohort. Reuse is valid only for unchanged observations and scores; these batch measurements do not estimate uncached single-query latency.

E Execution Evidence Inside the VLM

Internal readout shares output-fusion observations, executor, candidates and complete test cohorts across both backbones and three seeds. Representation, capacity and VLM updates differ.

E.1 Internal Reader and Training

The internal reader applies a residual update before zero-based block 18 of the 36-layer language decoder (Figure 7). A width-256 encoder turns numerical operands, operation identifiers, interme-

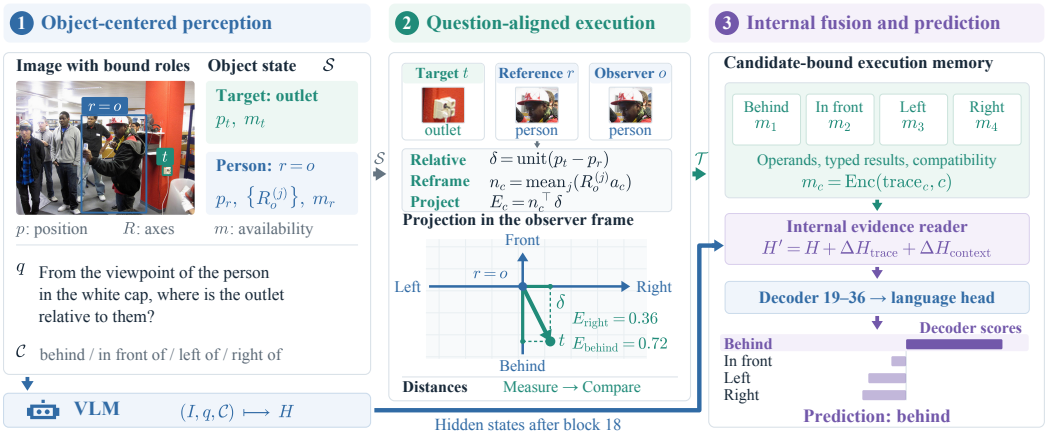

Figure 7: **Reading execution evidence inside the VLM.** The main SpaEx perception and execution pipeline is retained (Figure 1). After decoder block 18, candidate-bound trace memories and an execution-guided contrast of candidate states update hidden representations. Blocks 19–36 and the original language head then score answers. Bars show mean-centered LLaVA internal-reader scores at seed 13; projection geometry is schematic. Unsupported inputs preserve original VLM scores.

diate values and candidate compatibility into memory tokens. SpatialMQA supplies 24 fields and six tokens per candidate: five pose observations plus their aggregate. SPAR and metric inputs use 40 fields with operation-specific token counts. Candidate embeddings attach memories to original answer options, and candidate-routed attention updates option endpoints and the shared decision prefix while leaving other token positions unchanged. A second gated path weights uninjected candidate states by centered execution compatibility, adding their contrast at the decision prefix. Remaining decoder layers and the unchanged language head produce likelihoods without score mixing.

Training updates the reader and existing language LoRA while freezing the vision encoder, base weights and language head. Reader size is 2.177M parameters for SpatialMQA and 2.181M for SPAR/CV-Bench, alongside 11.796M updated LoRA parameters. Residual outputs initialize to zero, preserving the starting VLM. Candidate cross-entropy fits full-memory predictions, and a separate memory-disabled forward uses unit-weight KL to retain the native distribution:

$$\mathcal{L} = -\log p_{\mathrm{mem}}(y \mid I, q, \mathcal{C}) + \mathrm{KL}(p_{\mathrm{original}} \,\|\, p_{\mathrm{off}})\,. \tag{16}$$

The reference distribution is used only during training. Each fit processes the same 1,446/502/286 supported examples as output-reader fitting in one pass, with metric examples external to CV-Bench. AdamW uses 5×10^{-5} for LoRA and 3×10^{-4} for the reader, weight decay 0.01, gradient clipping at 1 and batch size one. This endpoint is fixed across seeds.

E.2 Complete-Test Results

Table 14: **Comparing output fusion and internal readout.** Accuracy ↑ (%, mean ± sample SD at seeds 13/21/42) over all 1,076 SpatialMQA, 1,378 SPAR and 1,200 CV-Bench questions. SPAR reports equal-task macro accuracy. Δ denotes internal minus output fusion. Bold indicates row maxima without a significance claim.

Benchmark	VLM	LoRA	Output (SpaEx)	Internal	Δ (pp)
SpatialMQA	Qwen	73.61 ± 0.98	79.12 ± 0.94	**79.24** ± 0.77	+0.12
	LLaVA	71.81 ± 0.63	77.08 ± 0.44	**77.32** ± 0.25	+0.25
SPAR-Bench	Qwen	66.01 ± 0.48	**82.18** ± 0.29	79.38 ± 2.14	−2.80
	LLaVA	62.17 ± 0.30	**81.10** ± 0.07	79.45 ± 1.58	−1.66
CV-Bench 3D	Qwen	87.33 ± 0.46	**94.25** ± 0.08	93.86 ± 0.35	−0.39
	LLaVA	76.97 ± 2.08	**92.53** ± 0.48	91.14 ± 1.27	−1.39

Internal readout exceeds original LoRA in every model–benchmark combination, but its advantage over output fusion is inconsistent (Table 14). SpatialMQA changes are +0.12/+0.25 points

for Qwen/LLaVA, with conditional image-cluster 95% intervals $[-0.46, 0.68]$/$[-0.37, 0.90]$. SPAR means decline by 2.80/1.66 points with greater seed variation; CV-Bench declines by 0.39/1.39, with capture-cluster intervals $[-1.03, 0.08]$/$[-2.19, -0.67]$. Original native score vectors are retained for all 696/22/238 unsupported inputs.

SPAR losses are concentrated in observer-reference and object-reference imagination: Qwen declines by 6.18/5.41 points and LLaVA by 4.75/3.20, whereas camera-frame object relations improve by 1.37/0.73. CV-Bench Distance declines by 0.83/2.44 points, while Depth changes by +0.06/-0.33. This breakdown locates weaknesses without establishing their causes. Test intervals use 5,000 grouped bootstrap draws (RNG seed 39107), conditional on three fitted seeds.

E.3 Controls on Evidence and Readout

Matched development controls test the contribution of executed input beyond additional VLM training. Native continuation, state-only injection and full execution share extra examples, update counts and loss. State-only injection preserves operands and candidate binding but masks execution outputs. Full execution wins all six seed-13 comparisons (Table 15). Metric development uses an external training-derived cohort, without CV-Bench test examples.

Table 15: **Internal-reader controls at a common training budget.** Correct-answer counts for seed 13; N gives the full development cohort. Continued uses native continuation, State masks execution fields, and Execution supplies the complete trace. Unsupported rows retain original VLM scores.

Development cohort	VLM	N	Continued	State	Execution
SpatialMQA	Qwen	536	396	408	**435**
SpatialMQA	LLaVA	536	387	397	**430**
SPAR	Qwen	512	386	387	**428**
SPAR	LLaVA	512	364	364	**425**
External metric	Qwen	429	290	282	**302**
External metric	LLaVA	429	230	218	**281**

A three-seed SPAR Qwen control restricts memory to executed compatibility while preserving candidate binding, slots, capacity and training schedule. Full traces improve development accuracy by 1.30 points, with a 15-scene interval $[-0.30, 3.24]$, leaving their advantage unestablished. Removing operands, types and intermediate values together leaves their individual effects unresolved.

A frozen-VLM SPAR control maps full traces through a shared $248 \to 256 \to 256 \to 1$ GELU MLP to a candidate-score residual. Inputs comprise six 40-field slots, presence masks, native log probability and entropy. Its 129,793 parameters train for one pass over the same 502 examples using AdamW (learning rate 3×10^{-4}, weight decay 0.01, clipping 1, batch size one). Training means and SDs standardize features without clipping. Macro accuracy is 82.78 ± 0.12/82.27 ± 0.28%, exceeding output fusion in all six seeds. Because internal readout also changes VLM weights and computation, this comparison does not isolate where fusion occurs.

Output fusion is retained as the primary configuration. Across six seed-13 internal fits, training takes 2.5–13.2 minutes on A100 80 GB with peak allocation 12.8–26.7 GB. Shared perception, model loading and scoring are excluded from these training costs.