# QUERY-FRAMED SPATIAL REASONING VIA SPATIAL LATENT TRANSFORMATION IN VIDEO MLLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Answering spatial questions from video requires using observations made at different positions and viewing directions. Remembering what was seen is not enough when a question asks for a relation from another viewpoint. Learning a scene representation is one way to support such reasoning. We investigate a direct alternative: adjust the observed visual evidence for answering, without requiring a video multimodal large language model (Video MLLM) to first generate a separate scene representation. We propose QUASAR (QUery-frAmed SpAtial Reasoning), centered on Spatial Latent Transformation. Given estimated observation geometry and an explicit reference type and numerical pose, it adds bounded spatial corrections to the original visual features. The MLLM reads the full question and answers from this adjusted evidence; geometry remains required at inference. Spatial supervision, including four learned spatial summary tokens, accompanies supervised fine-tuning (SFT), followed by Group Relative Policy Optimization (GRPO). QUASAR exceeds the original Qwen3-VL-8B baseline by 5.54 points on VSI-Bench and 19.88 points on VSTI-Bench; the historical VSI comparison uses different visual inputs and answer options. On both benchmarks, it also outperforms the Visual-only baseline (SFT+GRPO), which shares the backbone and data recipe but differs in pre-GRPO SFT exposure. Fixed-model interventions further support use of the visual corrections and, on parsable target-camera questions, the numerical reference. These results support direct visual-evidence adjustment as a viable route to spatial answering.

## 1 INTRODUCTION

Video multimodal large language models (Video MLLMs) are moving from general video understanding toward spatial reasoning. Such tasks ask an MLLM agent to answer questions about the environment from ego-centric video: whether a sofa is to the left of a bench, how many chairs are present, or how far apart objects seen at different moments are. Video MLLMs address such questions by reasoning over the combination of the visual observations with a language query. But to achieve this, remembering an earlier view is not always enough. As a camera moves, the same sofa can appear to the left of a bench from one position and to its right from another (Klatzky, 1998; Levinson, 2003). Each frame from the ego-centric video records visual evidence at an *observation pose*: the camera's position and orientation when it captured the scene. Answering may require a different *answering reference*, which specifies where to stand and which way to face when judging the relation. Figure 1 illustrates this distinction: a robot relies on the earlier observations to answer the user query from its current position, where both objects are out of view. Counts, object sizes, and distances between object pairs, however, should remain unchanged when the viewing direction changes. Spatial reasoning must therefore connect observations to the reference required for an answer while preserving these invariant quantities.

One of the existing approaches is to learn a representation of the scene: what objects it contains, where they are, and how they relate. This gives the model spatial information to draw on when answering. Such information can be learned within the model's internal states (Qu et al., 2026), generated as a sequence encoding scene geometry for subsequent reasoning (Chen et al., 2026), or organized explicitly as a scene graph of objects and relations (Zemskova & Yudin, 2025; Ma et al., 2025). These approaches need not output a complete layout before answering. They raise a



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

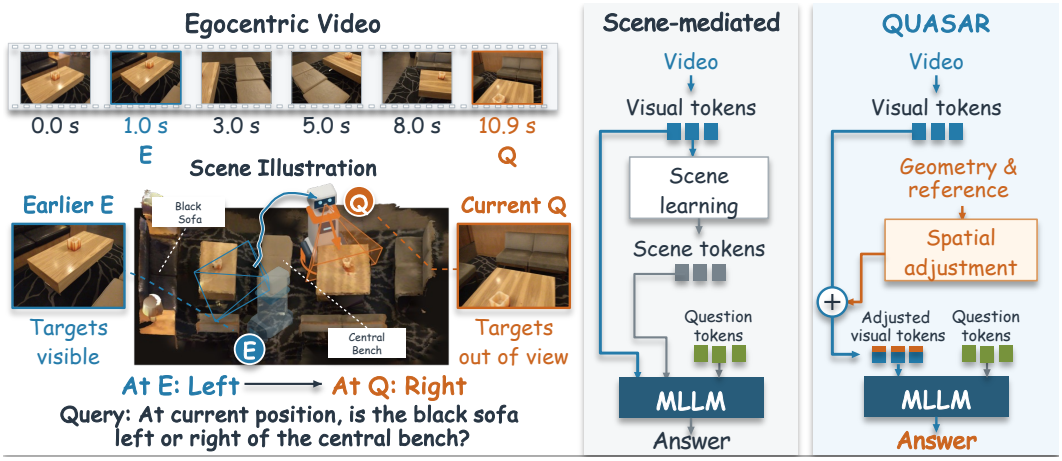



Figure 1: **Past evidence, current reference.** A robot observes the black sofa to the left of the bench at pose $E$. At its current pose $Q$, both targets are out of view, but the question still asks about them: under $Q$, the sofa is to the right of the bench. Earlier observations remain useful, yet their old left–right relation cannot simply be reused. QUASAR retains the original visual features and adds a bounded correction using observation geometry and the supplied reference, then answers the full question from the adjusted evidence without a separate scene-generation stage.

complementary question: *can we instead learn to adjust the observed visual evidence for answering, without requiring the MLLM to first generate a separate scene representation?*

Prior work already re-expresses evidence relative to a reference (Man et al., 2024; Zhu et al., 2025b) and enhances visual features with geometry (Wu et al., 2026; Jiang et al., 2026). We study how to adjust the original visual evidence under a supplied answering reference, without requiring the MLLM to first generate a separate scene representation. We also test whether the model uses these adjustments to answer. We describe the underlying challenge as follows:

**Observation-to-Query (O2Q) Bottleneck.** Observations record evidence under camera poses, while questions ask for particular spatial relations or quantities. The model must express that evidence under the appropriate reference and use it to answer, while preserving quantities that do not depend on viewing direction.

We propose the **QUASAR** (QUery-frAmed SpAtial Reasoning) framework, centered on **Spatial Latent Transformation** (SLT). Given geometry estimated from ego-centric video and an explicit reference type and numerical pose, SLT learns to express the original visual evidence under the supplied reference. It updates a spatial state at each visual position and adds the resulting correction to the original visual features. Limiting the correction's size relative to the original features keeps the native visual signal on the main path. The MLLM separately reads the full question, selects relevant objects and relations, and generates the answer. This changes how existing visual evidence is represented, rather than generating a new image or making a decoded scene the answer interface. Geometry remains required at inference; providing a reference type alone does not supply a valid numerical pose.

Training encourages spatial information to remain available along this answer path. During supervised fine-tuning (SFT), spatial targets supervise both the transformation states and decoder representations. Four learned spatial summary tokens provide global supervision positions, alongside local visual-token readouts; they are summaries, not generated reasoning steps. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) then updates only language adapters while retaining the spatial forward path. Recovering spatial information from a model's states does not by itself show that its answers use that information, so we test this separately by intervening on spatial inputs within fixed models.

With Qwen3-VL-8B, QUASAR achieves 65.36 on VSI-Bench (Yang et al., 2024) and 66.19 on VSTI-Bench (Fan et al., 2026), using eight-task and five-task macro-averages, respectively. It

exceeds the original backbone by $+5.54$/$+19.88$ points and the Visual-only baseline (SFT+GRPO) by $+1.75$/$+3.93$ points. SFT exposure differs between the trained systems; the historical VSI baseline also uses different visual inputs and answer options. We further examine spatial-evidence use through inference-time ablations and qualitative case studies. With model parameters fixed, disabling visual corrections lowers overall scores on both benchmarks, while perturbing numerical references reduces scores on target-camera questions with parsable poses. Case studies show that QUASAR correctly answers left/right questions about objects seen earlier but absent from the reference view, while Visual-only and QUASAR with visual writeback disabled predict the opposite relation. Together, these findings support direct adjustment of visual evidence under a supplied reference as an effective route to spatial reasoning in Video MLLMs. Code and data will be released at `https://anonymous.4open.science/r/QUASAR-Official-D07E`.

Our contributions are:

(1) We study a direct route to spatial question answering: adjusting observed visual evidence under a supplied reference instead of first generating a separate scene representation in the MLLM. We formulate its reference-alignment and evidence-use requirements as the **O2Q Bottleneck**.

(2) We develop the **QUASAR** framework, which retains native visual features and adds bounded corrections computed from observation geometry and a supplied reference. Spatial supervision acts on both the transformation states and decoder representations used for answering.

(3) We conduct a systematic empirical study of direct visual-evidence adjustment through benchmark evaluations, design comparisons, reference interventions, and transfer tests. These experiments demonstrate its effectiveness, provide evidence that learned spatial corrections contribute to answering, and characterize the approach's applicability and computational trade-offs.

## 2 Related Work

**Learning spatial representations.** Spatial MLLMs use geometry both as input and as supervision. Point features, position-enriched visual tokens, and reconstruction objectives provide complementary spatial cues (Hong et al., 2023; Zhu et al., 2025a; Zheng et al., 2025; Wang et al., 2025a); geometry-derived question–answer data provide another learning signal (Chen et al., 2024; Cheng et al., 2024). Loc3R-VLM combines input camera priors with decoder-side spatial learning (Qu et al., 2026). A projected CUT3R camera token precedes each frame's visual sequence. Decoder patch states predict 2D bird's-eye-view (BEV) coordinates, while `<Pos>`/`<Ori>` states predict the described agent's BEV position and orientation. Inserted between the situation description and question, these tokens can support internal viewpoint transformation during answering. By contrast, 3DThinker generates geometric latent tokens conditioned on images, the question, and preceding context; subsequent reasoning reads them (Chen et al., 2026). VGGT (Wang et al., 2025b) supplies feature-alignment targets and reinforcement-learning rewards, but is not an inference-time teacher. Neither approach requires a complete layout to be output and fed back before answering.

**Reference conditions.** Directional relations depend on the answering reference, whereas counting, distance, and size questions need not specify a unique viewing direction. SIG3D predicts an agent's position and orientation from a textual situation and a point cloud, then re-encodes visual tokens relative to that situation before multimodal fusion (Man et al., 2024). Struct2D instead rotates a question-relevant BEV image toward the agent's facing direction for relative-direction reasoning (Zhu et al., 2025b). Reference-dependent re-encoding is thus shared with prior work. QUASAR's Spatial Latent Transformation uses a supplied reference to control bounded corrections of native visual evidence. It retains geometric inputs and spatial supervision; the distinction concerns how representations support answering, not whether geometry is needed.

**Visual updates and temporal memory.** Spatial representation learning can itself modify native visual evidence. SpatioLM adds geometry-aware residuals at visual-token positions in intermediate blocks of a frozen VLM through zero-initialized projections (Wu et al., 2026). Depth Anything V3 (Lin et al., 2025) features and depth/ray targets supervise training; inference needs no external geometry. Spa3-VLM similarly updates visual tokens through residual attention to a pretrained scene latent (Jiang et al., 2026). Temporal methods retain evidence through feature memories or inference-time parameter updates (Liu et al., 2025; Gao et al., 2026; Liu et al., 2026). These choices can coexist

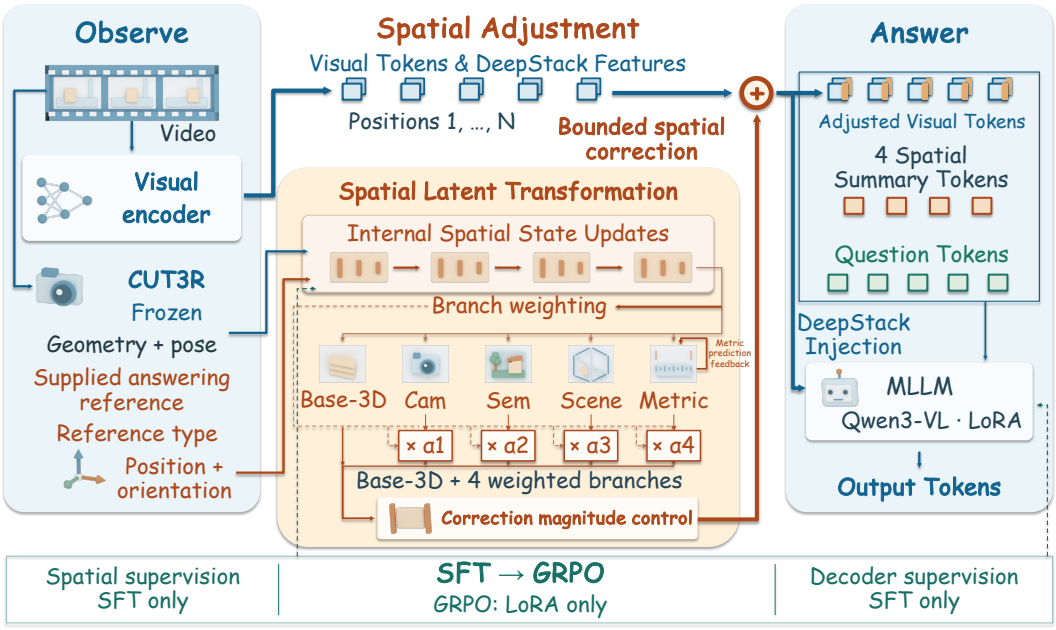



Figure 2: **QUASAR framework.** Frozen visual and geometric frontends process the video. Per-position geometry and an explicit reference type and pose enter Spatial Latent Transformation, which produces norm-bounded visual corrections (Correction magnitude control). Cam, Sem, and Scene denote target-camera, semantic-anchor, and scene-canonical branches. The MLLM processes these features, four spatial summary tokens, and the question. Spatial-module and decoder-side auxiliary losses are applied during SFT; GRPO retains the spatial path but updates only LoRA. Metric prediction feedback denotes the module's metric feedback, retained at inference; decoder-side auxiliary predictions are not needed for answers.

with geometric alignment and reference conditioning. QUASAR focuses on reference-controlled visual correction without adding a question-answering memory bank or test-time parameter updates. Its four spatial summary tokens are summary and supervision positions; their decodability alone does not establish answer dependence.

# 3 Method

## 3.1 Overview

QUASAR turns observation geometry and a supplied answering reference into a correction of native visual features. Fig. 2 shows the QUASAR framework. Spatial Latent Transformation aligns the correction with visual positions; a relative norm bound limits its magnitude. The MLLM reads the corrected features and full question, retaining responsibility for object selection and answering. It does not first decode a separate scene representation to use as an answer input. Spatial targets supervise both the transformation states and decoder representations during SFT. This training encourages spatial information to remain available along the answer path; decoder-side spatial predictions are not required at inference.

## 3.2 Explicit Reference and Geometry Inputs

Write frame $i$'s observation pose as $(R_i^{\text{obs}}, c_i^{\text{obs}})$ and the answering reference as $Q = (R^{\text{qry}}, c^{\text{qry}})$. Rotations map local axes into the shared reconstruction system; translations locate origins. A supplied

type $t$ identifies the reference convention:

$$(R^{\mathrm{qry}}, c^{\mathrm{qry}}) = \begin{cases} (R^{\mathrm{obs}}_{i^\star}, c^{\mathrm{obs}}_{i^\star}) & \text{target-camera,} \\ (R_A, c_A) & \text{semantic-anchor,} \\ (I, \mu) & \text{scene-canonical.} \end{cases} \tag{1}$$

Here $i^\star$ selects an observed frame, and $(R_A, c_A)$ specifies an object-anchored heading and position. For each video, $\mu$ is the mean of valid input 3D points; $I$ retains their shared axes. This reference is constructed from observed geometry without a prebuilt scene model. Figure 3 illustrates these supplied conventions. Canonical coordinates also support counting, distance, and size without a unique viewing direction. A type label does not establish a valid numerical pose: consistent coordinates must be supplied, not automatically inferred from language. Appendix I details the assignments and their validity boundaries.

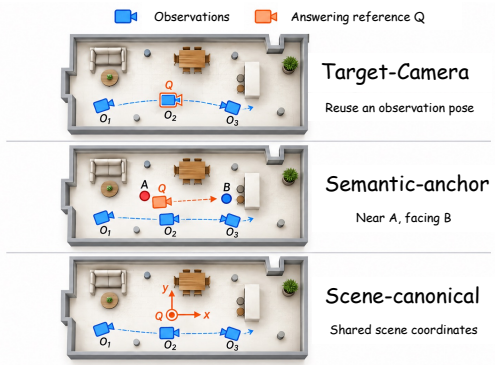



Figure 3: **Supplied reference conventions.** Blue cameras mark observations; orange $Q$ marks the answering reference. Semantic-anchor and scene-canonical references need not be captured views. The room is illustrative, not a predicted scene.

### 3.3 Visual-Token-Level Geometry Encoding

On the frozen CUT3R (Wang et al., 2025c) path, native visual token $x_n \in \mathbb{R}^H$ receives an aligned point $p_n$, confidence $\gamma_n$, and validity $m_n$. Pointmaps and confidence are resampled to the spatially merged visual grid, preserving frame and patch order. These estimates are not guaranteed metric ground truth. Geometry is precomputed for training, also from depth and camera-pose sources, and remains required at inference.

Using scene scale $\ell > 0$, we form $d_n \in \mathbb{R}^{28}$ from centered position, source/reference-origin offsets, camera translation and axes, confidence, validity, and patch coordinates. In particular,

$$\bar{p}_n = (p_n - \mu)/\ell, \qquad \rho^s_n = (p_n - c^{\mathrm{src}})/\ell, \qquad \rho^q_n = (p_n - c^{\mathrm{qry}})/\ell,$$

where $c^{\mathrm{src}}$ is the observing camera's origin. These offsets retain reconstruction axes; reference orientation enters the learned condition below. Auxiliary targets and gold answers are not inference inputs. Appendix Q gives the full descriptor, initialization, and readout dimensions.

### 3.4 Spatial Latent Transformation

Encode $Q$ in row-major order as $q = \mathrm{vec}_{\mathrm{row}}([R^{\mathrm{qry}} \mid c^{\mathrm{qry}}]) \in \mathbb{R}^{12}$ and form $h_Q = \mathrm{Emb}(t) + \mathrm{MLP}_Q(q) \in \mathbb{R}^F$. Each position starts with

$$z^{(0)}_n = \mathrm{MLP}_{\mathrm{init}}(d_n) \in \mathbb{R}^F. \tag{2}$$

Four separately parameterized residual blocks ($K = 4$) combine the state, condition, and descriptor:

$$z^{(k+1)}_n = z^{(k)}_n + \frac{1}{K}\Delta_{\theta_k}\left([z^{(k)}_n; h_Q; h_Q - z^{(k)}_n; z^{(k)}_n \odot h_Q; d_n]\right). \tag{3}$$

Here $k = 0, \ldots, K-1$, $\odot$ denotes element-wise multiplication, and $\Delta_{\theta_k}$ is a two-layer MLP with input normalization, GELU, and dropout. We obtain $\bar{z}_n = \mathrm{LN}(z^{(K)}_n)$. These are learned representation updates, not analytic rigid transformations or explicit reasoning steps.

Five branches produce $H$-dimensional components: base 3D, target-camera, semantic-anchor, scene-canonical, and metric-scale. Each uses a residual MLP on $[\bar{z}_n; h_Q; d_n]$, followed by code normalization and projection. All branches, including base, receive the reference. The metric branch pools its code to predict 17 scalar/vector statistics covering seven metric families. Projected predictions are added back to its per-position code before normalization: this *metric feedback* remains active at inference, unlike training-only readouts.

A gate maps the sample mean of $\bar{z}_n$ to five softmax weights $\alpha^e$. The residual combines an unweighted base with four weighted branches:

$$b_n = b_n^{\text{base}} + \sum_{e \in \mathcal{E}_{\text{dyn}}} \alpha^e b_n^e, \quad \mathcal{E}_{\text{dyn}} = \{\text{cam}, \text{sem}, \text{bev}, \text{met}\}. \tag{4}$$

Thus dynamic weights sum to $1 - \alpha^{\text{base}}$, but the base itself is not attenuated. Branch-mixture targets supervise the gate during SFT; its inference inputs exclude the full question.

### 3.5 Norm-Bounded Residual Write

To limit geometric perturbations while retaining appearance, use learned scale $s = 0.25\,\sigma(a)$, with sigmoid $\sigma$, scalar $a$, and relative budget $\eta = 0.15$:

$$\begin{aligned} v_n &= \max(\|x_n\|_2, 10^{-6}), \\ \tilde{b}_n &= s b_n \min\left(1, \frac{\eta v_n}{\max(\|s b_n\|_2, 10^{-12})}\right), \quad x'_n = x_n + \tilde{b}_n. \end{aligned} \tag{5}$$

The direction is preserved and $\|\tilde{b}_n\|_2 \le \eta v_n$. This bounds the feature write, not its influence on answers. Zero-initialized branch output projections initially leave native features unchanged.

The same $\tilde{b}_n$ is added to aligned DeepStack features, the intermediate visual features supplied to multiple decoder layers. Their budget is not recomputed separately. Native attention and multimodal rotary positional embeddings (M-RoPE) remain unchanged. Appendix D examines budget sensitivity.

### 3.6 Answer-Path Spatial Supervision and Training

Four learned $H$-dimensional embeddings replace the default embeddings of the spatial summary tokens after visual tokens and before the question. Their causal context contains preceding visual evidence, not the later question. Global decoder readouts use the normalized mean of their final hidden states; local readouts use normalized final visual states. Answers can attend to both. These four summary positions are independent of the four latent updates.

Spatial-module readouts supervise latent states and branch codes; decoder-side spatial readouts supervise the corresponding information after writeback. Both predict reference type/pose, branch mixtures, bird's-eye-view (BEV) occupancy/object centers, local coordinates, and metric quantities. Patch consistency targets $\rho_n^q$, whereas reference-coordinate heads target $(R^{\text{qry}})^\top \rho_n^q$. Reference-type and validity masks restrict applicable losses; annotation availability controls object/metric supervision. Weighted auxiliary terms join answer-span cross-entropy:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ans}} + \mathcal{L}_{\text{flow_aux}} + \mathcal{L}_{\text{decoder_aux}} + \mathcal{L}_{\text{stab}}. \tag{6}$$

Here $\mathcal{L}_{\text{flow_aux}}$ and $\mathcal{L}_{\text{decoder_aux}}$ aggregate spatial-module and decoder readout losses. The stability term combines Sketched Isotropic Gaussian Regularization (SIGReg) (Balestriero & LeCun, 2025) on projected summaries and answer-distribution KL. Its stopped-gradient reference disables language low-rank adaptation (LoRA) (Hu et al., 2022) but retains spatial writes and summary embeddings, so it is not visual-only Base. Targets and coefficients appear in Appendix P.

SFT freezes the backbone/frontends and trains LoRA, the spatial module/scale, summary embeddings, auxiliary readouts, and regularization projection. GRPO (Shao et al., 2024) starts from SFT, updates only LoRA, freezes the spatial module and summaries, and disables auxiliary/stability losses; Appendix R gives training lineage, reward settings, and configurations; Appendix P lists the stage-specific trainable components. Inference retains geometry, explicit reference, metric feedback, bounded writes, and summaries, but needs no auxiliary losses or decoder readout predictions. Decodability is not proof of answer use; §4.3 tests that dependence.

## 4 Experiments

We test the direct evidence-adjustment route in two ways. Trained-system comparisons measure its spatial-answering performance; fixed-checkpoint interventions test whether answers depend on the corrections and supplied reference. Structure and supervision comparisons examine implementation choices, while transfer tests and runtime measurements characterize benefits, limitations, and cost.

### 4.1 SETUP

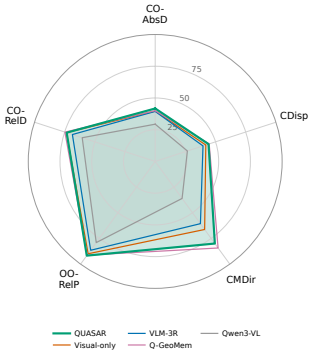
QUASAR combines Qwen3-VL-8B-Instruct (Bai et al., 2025a), geometry inputs, and four spatial summary tokens. The CUT3R (Wang et al., 2025c) inference path and mixed training-geometry sources are detailed in Appendix Q. The Visual-only baseline shares the backbone and data recipe, and reads the video and full question without geometry inputs, the spatial module, or spatial summary tokens; Base is the released model without project-specific training. Both trained systems use 426,916 SFT examples and 2,048 all-task GRPO updates, but differ in SFT exposure. Displayed SFT checkpoints are not the GRPO parents: these comparisons isolate neither equal-budget architectural nor pure GRPO effects (Appendix R).

VSI-Bench (Yang et al., 2024) uses 5,130 questions/eight tasks; historical VSTI-Bench (Fan et al., 2026) uses 6,042/five tasks. Main scores are task macro-averages of numeric Mean Relative Accuracy (MRA) and multiple-choice accuracy. Trained models use at most 32 cached frames, 50,176 pixels per frame, and deterministic decoding. Main confidence intervals (CIs) use 5,000 paired question resamples, not training seeds. Differences are percentage points computed before rounding. Diagnostic aggregates and intervention sampling units remain separate, as specified in their tables.

### 4.2 MAIN RESULTS

Table 1: **VSTI-Bench results.** 6,042 questions/five-task macro. CO-AbsD: camera–object absolute distance; CDisp: camera displacement; CMDir: camera movement direction; OO-RelP: object–object relative position; CO-RelD: camera–object relative distance. CO-AbsD and CDisp use MRA, others accuracy. Local results are rounded to two decimals. Literature rows retain reported precision and differing training/input protocols. Radar: QUASAR and Visual-only are SFT+GRPO endpoints, Qwen3-VL is Base; VLM-3R and Q-GeoMem are reported references. Axes: 0–100.

| Method | Avg | CO-AbsD | CDisp | CMDir | OO-RelP | CO-RelD |
|---|---|---|---|---|---|---|
| *Qwen3-VL-8B Baselines* | | | | | | |
| Base | 46.31 | 29.35 | 26.66 | 36.04 | 79.03 | 60.46 |
| Visual-only (SFT) | 61.86 | 39.16 | 42.60 | 62.98 | 90.58 | 73.97 |
| Visual-only (SFT+GRPO) | 62.26 | 39.65 | 41.92 | 66.27 | 89.91 | 73.56 |
| QUASAR (SFT) | 65.11 | 38.80 | 40.87 | 82.15 | 90.70 | 73.05 |
| QUASAR (SFT+GRPO) | 66.19 | 41.72 | 44.18 | 79.96 | 91.73 | 73.33 |
| *External Spatial Memory Methods* | | | | | | |
| Q-GeoMem (Gao et al., 2026) | 67.0 | 40.0 | 44.7 | 84.3 | 91.2 | 74.8 |
| VLM$^2$-7B (Liu et al., 2025) | 65.3 | 43.1 | 44.1 | 76.8 | 87.7 | 74.9 |
| *Geometry-enhanced, without external spatial memory* | | | | | | |
| VLM-3R-7B (Fan et al., 2026) | 58.8 | 39.4 | 39.6 | 60.6 | 86.5 | 68.6 |
| *General MLLMs* | | | | | | |
| Qwen2.5VL-72B (Bai et al., 2025b) | 40.3 | 18.0 | 10.0 | 41.0 | 74.2 | 58.4 |
| GPT-4o | 38.2 | 29.5 | 23.4 | 37.3 | 58.1 | 42.5 |




QUASAR (SFT+GRPO) reaches **65.36/66.19** on VSI/VSTI (Tables 1–2), exceeding Visual-only (SFT+GRPO) by $+1.75/+3.93$ points (95% CIs $[0.56, 2.90]/[2.83, 4.99]$). Selected SFT differences are $+2.06/+3.26$. SFT+GRPO models share saved VSI inputs; historical Base does not (Table 2). Its $+5.54$ VSI gap is descriptive, while VSTI's $+19.88$ combines training and the complete system.

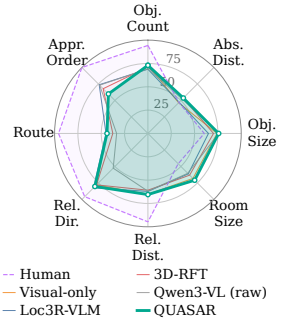
The largest gains over Visual-only (SFT+GRPO) occur in VSTI camera-motion direction and VSI absolute distance, consistent with using geometry to adjust spatial evidence. Camera–object relative distance and counting decline. Appearance order trails Visual-only by 4.05 points at SFT; the SFT+GRPO $+0.16$ difference has CI $[-3.63, 3.75]$, leaving its direction uncertain. Full intervals and group results are in Appendices S and E.

### 4.3 DESIGN AND MECHANISM

**Structure and supervision.** Table 3 examines how spatial structure, readout placement, and auxiliary supervision affect answering. The Basic spatial model applies four reference-conditioned updates with visual writeback, but has no branch mixing, spatial summary tokens, or decoder-side auxiliary readouts. Its counterpart trained without visual writeback uses the same recorded training recipe and scores higher on VSI but lower on VSTI. Complete SFT structure results are in Table 9. Decoder reference-coordinate supervision improves VSTI while slightly reducing VSI. Among the

Table 2: **VSI-Bench results.** 5,130 questions/eight-task macro. Columns: count, absolute distance, size, room size, relative distance/direction, route planning, appearance order. †: historical Base uses raw video, trained models cached frames. Radar: QUASAR and Visual-only are SFT+GRPO endpoints; Human, Loc3R-VLM, and 3D-RFT are reported references. Axes: 0–100.

| Method | Avg | Count | AbsD | Size | Room | RelD | RelDir | Route | Appr. |
|---|---|---|---|---|---|---|---|---|---|
| *Qwen3-VL-8B zero-shot* | | | | | | | | | |
| Base† | 59.82 | 69.24 | 52.03 | 75.87 | 60.38 | 61.83 | 52.61 | 33.51 | 73.14 |
| *QUASAR vs Visual-only (SFT and SFT+GRPO)* | | | | | | | | | |
| Visual-only (SFT) | 62.79 | 72.04 | 45.00 | 74.31 | 66.25 | 63.38 | 75.06 | 44.33 | 61.97 |
| Visual-only (SFT+GRPO) | 63.62 | 73.40 | 46.53 | 74.08 | 68.19 | 65.07 | 78.13 | 43.81 | 59.71 |
| QUASAR (SFT) | 64.86 | 70.83 | 52.69 | 74.38 | 69.03 | 67.46 | 82.73 | 43.81 | 57.93 |
| QUASAR (SFT+GRPO) | 65.36 | 73.01 | 53.69 | 75.81 | 70.87 | 65.49 | 80.36 | 43.81 | 59.87 |
| *External Spatial Memory / Test-time State Update Methods* | | | | | | | | | |
| Q-GeoMem (Gao et al., 2026) | 70.1 | 76.3 | 54.5 | 76.7 | 76.3 | 71.4 | 85.5 | 46.9 | 73.4 |
| VLM$^2$-7B (Liu et al., 2025) | 68.8 | 72.5 | 59.6 | 70.8 | 69.9 | 69.0 | 87.8 | 52.6 | 68.3 |
| SeeRememberExplore (Wei et al., 2026) | 67.5 | 70.2 | 46.8 | 72.2 | 65.7 | 74.9 | 75.1 | 40.2 | 75.1 |
| Spatial-TTT-2B (Liu et al., 2026) | 64.4 | 70.8 | 47.8 | 71.7 | 65.9 | 61.8 | 73.0 | 47.4 | 77.0 |
| *Geometry-enhanced, without external spatial memory* | | | | | | | | | |
| 3DThinker-S1+S2 (Chen et al., 2026) | 63.7 | – | – | – | – | – | – | – | – |
| Loc3R-VLM (Qu et al., 2026) | 63.2 | 68.9 | 47.3 | 64.9 | 61.2 | 62.1 | 82.4 | 44.9 | 73.8 |
| 3D-RFT-4B (Linghu et al., 2026) | 62.8 | 71.2 | 53.5 | 70.3 | 63.2 | 60.8 | 77.9 | 37.6 | 67.8 |
| *General MLLMs & Human* | | | | | | | | | |
| Human | 79.2 | 94.3 | 47.0 | 60.4 | 45.9 | 94.7 | 95.8 | 95.8 | 100.0 |
| GPT-5 | 55.0 | 53.3 | 34.4 | 73.3 | 47.5 | 63.7 | 48.6 | 50.2 | 68.9 |
| Gemini-2.5-Pro | 53.5 | 46.0 | 37.3 | 68.7 | 54.3 | 61.9 | 43.9 | 47.4 | 68.7 |
| GPT-4o | 34.0 | 46.2 | 5.3 | 43.8 | 38.2 | 37.0 | 41.3 | 31.5 | 28.5 |




four readout controls, visual readout without summary tokens leads on VSI, whereas local-only decoder supervision leads on VSTI. The separate configuration jointly removing summary tokens, their prediction heads and objectives, and SIGReg scores 65.67/66.33. These results show task-dependent effects, not consistent benefits from adding summary tokens or supervision. They do not establish that four tokens or updates are necessary; training without writeback is not a matched one-step substitute (Appendix G).

Table 3: **Component Ablations and Design Comparisons.** Qwen3-VL-8B SFT comparisons; main-result models are separate references. VSI: 5,130 questions/eight tasks; historical VSTI: 6,042 questions/five tasks. Left: spatial structures. Right: independent readout and coordinate-supervision groups. These are not successive deletions from QUASAR (SFT+GRPO). Complete structure results, configurations, and training sources: Appendix G.

| Spatial component ablations | VSI | VSTI |
|---|---|---|
| Visual-only | 63.22 | 59.95 |
| Basic spatial model | 62.64 | 63.84 |
| Basic spatial model, trained without visual writeback | 63.67 | 60.75 |
| QUASAR (SFT) | 64.86 | 65.11 |
| QUASAR (SFT+GRPO) | 65.36 | 66.19 |

| Supervision ablations | VSI | VSTI |
|---|---|---|
| Summary-token readout | 64.58 | 66.25 |
| Visual readout, tokens retained | 64.73 | 65.14 |
| Visual readout, no summary tokens | 65.18 | 65.28 |
| Local-only decoder supervision | 64.54 | 66.56 |
| Decoder reference-coordinate loss on | 64.54 | 65.87 |
| Decoder reference-coordinate loss off | 64.84 | 64.87 |

**Writeback and reference use.** Table 4 tests answer dependence within fixed checkpoints. On the same 512 questions, SFT+GRPO answers 415 correctly with the supplied reference, 390 without visual writeback, and 390–394 under three pose permutations. These controls corroborate that answering benefits from both visual writeback and reference matching. The separate w/o reference-pose conditioning intervention disables numerical-pose conditioning while retaining writeback. Questions and answer labels stay fixed; correct answer changes under a new viewpoint remain untested. Effects vary by task: disabling writeback improves SFT appearance order by 4.05 points (Appendix C).

Our numerical-pose conclusions are restricted to the parsable target-camera subset (Appendix H).

Table 4: **Fixed-checkpoint writeback and reference interventions.** $\Delta$ is normal minus changed inference, in points. Each row has its own matched normal. VSI uses 5,130 questions/eight tasks; historical VSTI (6,042 questions) uses nine subtypes, its parsable subset (5,598) five tasks, and the fixed subset (512) question-level accuracy. Paired 95% CIs and full settings are in Appendices C and H.

| Intervention / set | Normal | Changed | $\Delta$ | 95% CI |
|---|---|---|---|---|
| *QUASAR (SFT): w/o visual writeback* | | | | |
| VSI, 5,130 questions | 65.08 | 63.47 | $+1.61$ | – |
| VSTI, 6,042 questions | 72.87 | 68.64 | $+4.23$ | – |
| *QUASAR (SFT+GRPO): reference interventions* | | | | |
| Reference type/path permutation, VSI (5,130 questions) | 65.57 | 63.88 | $+1.69$ | $[1.00, 2.36]$ |
| Reference type/path permutation, VSTI (6,042 questions) | 72.81 | 70.73 | $+2.08$ | $[1.25, 2.89]$ |
| w/o reference-pose conditioning, VSTI parsable subset (5,598 questions) | 66.28 | 61.60 | $+4.68$ | $[3.88, 5.51]$ |
| *QUASAR (SFT+GRPO): fixed-subset writeback/pose controls* | | | | |
| w/o visual writeback | 81.05 | 76.17 | $+4.88$ | $[1.87, 8.08]$ |
| Pose permutation (type fixed), mean of three | 81.05 | 76.63 | $+4.43$ | $[2.31, 6.56]$ |

**Selected spatial-answering cases.** Figure 4 in Appendix B compares four variants on selected left/right, camera-movement, and object-distance questions; full records appear in Appendix U. QUASAR answers all four correctly, while the other variants fail. These outcome-selected examples complement, rather than replace, the aggregate evaluations.

### 4.4 Transfer, Boundaries, and Cost

QUASAR (SFT) improves over Visual-only on Qwen3-VL-2B and 4B across both benchmarks, extending the benefit beyond the main 8B model (Table 19, Appendix M). SenseNova gains 2.36 on revised VSTI and loses 0.59 on VSI. These single-run results show benefits across model sizes and task-dependent transfer to another backbone. On external tests, QUASAR (SFT+GRPO) trails Visual-only (SFT) on ViewSpatial (Li et al., 2025a); SPAR/SPBench (Zhang et al., 2025; Li et al., 2025b) difference intervals cross zero. Inference w/o visual writeback hurts SPBench multi-view performance, while the fixed image-count rule offers no reliable overall gain.

On CV-Bench (Tong et al., 2024), inference w/o visual writeback raises QUASAR from 81.80 to 85.67 ($+3.87$, CI $[2.62, 5.16]$), matching Base only overall, not on 3D. This identifies a residual-related cost, not recovery of general visual ability (Appendix O).

At SFT+GRPO, QUASAR adds 2.32%/2.92% time (VSI/VSTI) and 0.67–0.68 GB peak allocated memory over Visual-only (Table 22, Appendix N). Cached-input timings exclude geometry extraction, loading, cache I/O, and scoring; historical prefill timing remains unverified.

## 5 Conclusion

In this work, we investigated a direct route to spatial question answering in video MLLMs, in which observed visual evidence is adjusted to a supplied answering reference. QUASAR implements this idea by using observation geometry and reference pose to produce bounded visual corrections, with spatial supervision along the answer path. Evaluations on VSI-Bench and VSTI-Bench show gains over trained visual-only baselines. Ablation studies further support the contribution of the corrections to answering and the use of numerical references in the tested target-camera setting. These findings support spatial reasoning without requiring the MLLM to first generate a separate scene representation. Improving reference grounding and extending the benefits across tasks are natural directions for future work.

### AI USE STATEMENT

Generative AI tools assisted with experimental-code design, the development of visualization tools for manual data annotation and data-quality inspection, and language editing of the manuscript. All AI-assisted work is subject to manual review and verification by at least two people upon completion. The authors take responsibility for the final content of this work, including AI-assisted code, tools, and text.

### ETHICS STATEMENT

QUASAR studies spatial question answering from recorded visual inputs. Its evaluations include indoor videos and rendered views of ScanNet scenes (Dai et al., 2017); the case-study sources and display annotations are described in Appendix U. Such recordings and reconstructions can reveal room layouts and personal belongings, so benchmark availability does not remove privacy concerns. ScanNet data are subject to the dataset's Terms of Use, and reuse of benchmark media and derived illustrations remains subject to the respective source permissions. Incorrect reference assignment, geometry, or object identification can produce misleading directions and distances. The reported evaluations do not establish safety for autonomous navigation or physical interaction; these uses require independent validation and human oversight. Spatial inference could also support intrusive analysis of private environments, a risk that requires separate consideration before deployment.

### REPRODUCIBILITY STATEMENT

Section 3 and Appendices Q and P specify the architecture, geometric descriptors, and auxiliary targets. Appendix R records model identities, SFT/GRPO checkpoint lineages, training settings, visual inputs, decoding limits, and scoring protocols; Appendix I describes the supplied references. Statistical procedures are reported in Appendix S and with each intervention, including Appendix H, while Appendix N defines the timing scope. The historical provenance gaps documented in Appendix R and the unarchived timing implementation noted in Appendix N limit exact reconstruction of some runs.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling Spatial Intelligence with Multimodal Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7879–7890, June 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views, 2026. URL https://arxiv.org/abs/2510.18632.

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, 2024. URL https://arxiv.org/abs/2406.01584.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction, 2026. URL https://arxiv.org/abs/2505.20279.

Xianqiang Gao, Qizhi Chen, Delin Qu, Haoming Song, Zhigang Wang, Bin Zhao, Dong Wang, and Xuelong Li. Q-geomem: Question-guided geometric memory for video spatial reasoning, 2026. URL https://arxiv.org/abs/2605.27318.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023. URL https://arxiv.org/abs/2307.12981.

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

Haoyi Jiang, Liu Liu, Xinjie Wang, Yonghao He, Wei Sui, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Spa3r: Predictive spatial field modeling for 3d visual reasoning, 2026. URL https://arxiv.org/abs/2602.21186.

Roberta L. Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In Christian Freksa, Christopher Habel, and Karl F. Wender (eds.), *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, volume 1404 of *Lecture Notes in Artificial Intelligence*, pp. 1–17. Springer, Berlin, 1998. doi: 10.1007/3-540-69342-4_1.

Stephen C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, Cambridge, 2003. doi: 10.1017/CBO9780511613609.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models, 2025a. URL https://arxiv.org/abs/2505.21500.

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, 2025b. URL https://arxiv.org/abs/2510.08531.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Xiongkun Linghu, Jiangyong Huang, Baoxiong Jia, and Siyuan Huang. 3D-RFT: Reinforcement Fine-Tuning for Video-based 3D Scene Understanding. In *Proceedings of the International Conference on Machine Learning*, 2026.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-ttt: Streaming visual-based spatial intelligence with test-time training, 2026. URL https://arxiv.org/abs/2603.12255.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-language memory for spatial reasoning, 2025. URL https://arxiv.org/abs/2511.20644.

Haodi Ma, Vyom Pathak, and Daisy Zhe Wang. Bridging Vision Language Models and Symbolic Grounding for Video Question Answering, 2025. URL https://arxiv.org/abs/2509.11862.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3D vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13678–13688, 2024. doi: 10.1109/CVPR52733.2024.01298. URL https://arxiv.org/abs/2406.07544.

Kevin Qu, Haozhe Qi, Mihai Dusmanu, Mahdi Rad, Rui Wang, and Marc Pollefeys. Loc3r-vlm: Language-based localization and 3d reasoning with vision-language models, 2026. URL https://arxiv.org/abs/2603.18002.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, 2024.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness, 2025a. URL https://arxiv.org/abs/2504.01901.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state, 2025c. URL https://arxiv.org/abs/2501.12387.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. MindCube: Spatial Mental Modeling from Limited Views, 2026. URL https://arxiv.org/abs/2506.21458.

Yuxi Wei, Wei Huang, Qirui Chen, Lu Hou, and Xiaojuan Qi. See, remember, explore: A benchmark and baselines for streaming spatial reasoning, 2026. URL https://arxiv.org/abs/2603.23864.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards general physical spatial intelligence in vision-language models, 2026. URL https://arxiv.org/abs/2608.01899.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces, 2024. URL https://arxiv.org/abs/2412.14171.

Tatiana Zemskova and Dmitry Yudin. 3DGraphLLM: Combining Semantic Graphs and Large Language Models for 3D Scene Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8885–8895, October 2025.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yujie 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. *arXiv preprint arXiv:2503.22976*, 2025.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding, 2025. URL https://arxiv.org/abs/2412.00493.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness, 2025a. URL https://arxiv.org/abs/2409.18125.

Fangrui Zhu, Hanhui Wang, Yiming Xie, Jing Gu, Tianye Ding, Jianwei Yang, and Huaizu Jiang. Struct2d: A perception-guided framework for spatial reasoning in mllms, 2025b. URL https://arxiv.org/abs/2506.04220.

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

# APPENDIX

The supplementary experiments distinguish trained-system comparisons from interventions within a fixed model. Checkpoints, question sets, and aggregation rules are specified with each result. In particular, the historical VSTI five-task main score and nine-subtype diagnostic score summarize the same 6,042 questions differently; the revised 5,736-question protocol changes both the set and the aggregation relative to the main table.

## A. LIMITATIONS

QUASAR uses supplied references, not automatically inferred anchors. Historical target-camera inputs include last-frame fallbacks; numerical-pose evidence is restricted to parsable target-camera questions (Appendix I). Geometry quality, scale calibration, object identification, and numerical formatting constrain metric answers.

Benefits are task-dependent. Inference w/o visual writeback improves CV-Bench at GRPO and VSI appearance order at SFT; recovery to Base is limited to CV-Bench overall, not its 3D subset. Transfer to SenseNova-SI-1.5-InternVL3-8B improves revised VSTI but declines on VSI; QUASAR (SFT+GRPO) trails Visual-only (SFT) on ViewSpatial, and SPAR/SPBench differences from Visual-only have intervals crossing zero. Cross-stage comparisons do not isolate residual effects, and fixed single-/multi-image switching provides no reliable overall benefit.

Training exposure differs between QUASAR and Visual-only; displayed SFT checkpoints are not GRPO parents. Historical Base inputs also differ, and MindCube lacks a Visual-only control. Coupled design controls do not establish four tokens or updates as necessary; no same-information one-step substitute is tested. Intervals describe fixed checkpoints, not training-seed variation. Runtime measurements cover generation calls after cached-input preparation, not end-to-end video answering (Appendix N).

## B. Spatial Question-Answering Case Study

Figure 4 compares four model variants on left/right relations, camera movement, and object distance. QUASAR recovers both object relations outside the specified reference views, identifies sustained forward movement relative to the starting orientation, and selects chair as closest to the TV. Full questions, reference conditions, and two-run answers appear in Appendix U.

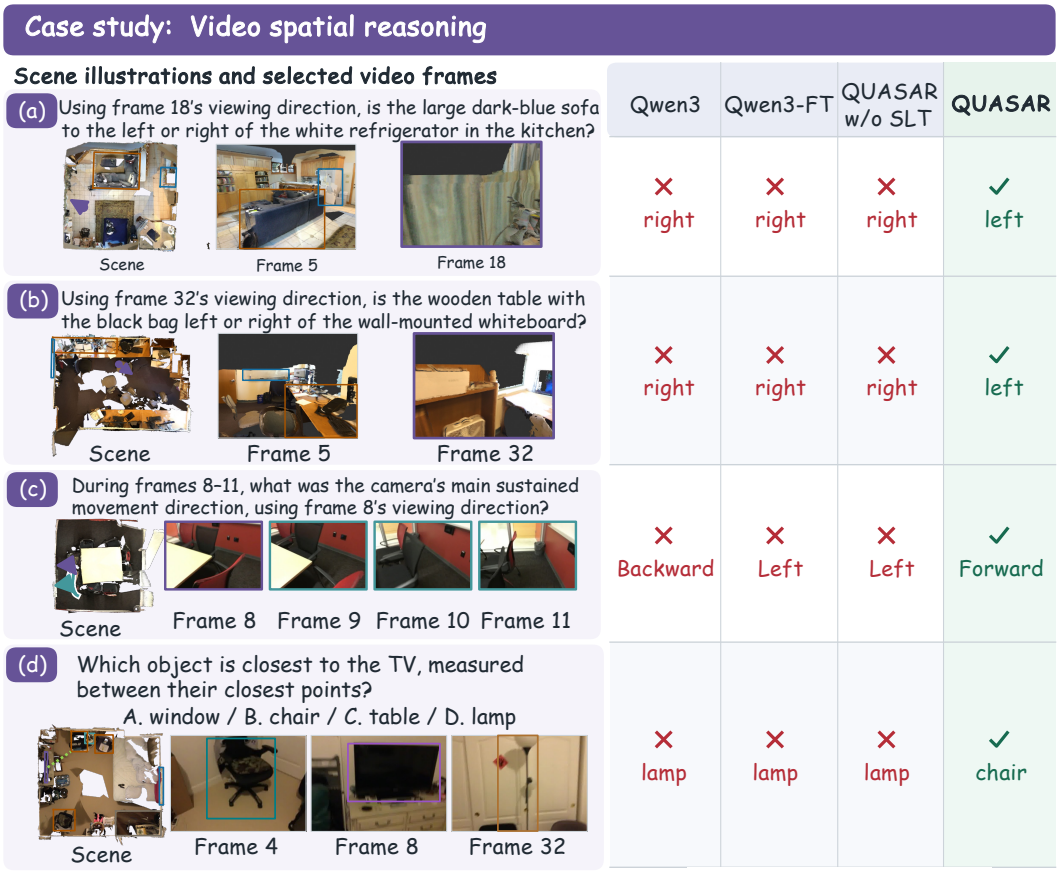



Figure 4: **Comparison between QUASAR and baselines in Spatial Reasoning Tasks.** Rows (a–b) use rendered videos; (c–d) use original ScanNet videos. All models receive 32 frames at $320 \times 224$ pixels. Qwen3 denotes Qwen3-VL-8B; Qwen3-FT denotes Visual-only (SFT+GRPO). Here, *w/o SLT* means disabling visual writeback while retaining QUASAR's parameters and four spatial summary tokens. Purple borders mark reference frames in (a–c); the purple object box marks the TV in (d). The green dotted line is a visual guide, not a distance measurement. Illustrations and annotations are not model inputs. We select these examples after observing their outcomes; each answer matches both recorded runs. Questions are abbreviated for display: (b) uses frame 32's viewing direction, and (c) uses frame 8's. Full questions and settings appear in Appendix U.

## C. Inference without Visual Writeback

Does the spatial residual contribute to answers after the model has been trained? We compare normal inference with inference w/o visual writeback at the full QUASAR SFT endpoint used for writeback diagnostics (Table 27). Both conditions retain the same weights, questions, geometry, prompt, and scorer. Setting the inference residual multiplier to zero removes the written signal without removing its training history or the decoder-side spatial summary tokens.

Table 5 shows that inference w/o visual writeback lowers both overall point estimates. Four spatial groups have negative intervals excluding zero; the largest observed group drop is 4.99 points on VSTI motion-and-relations questions. VSI counting-and-routes questions have an inconclusive interval, while appearance order improves by 4.05 points. Thus the residual affects answers and benefits

Table 5: **Inference without visual writeback (SFT).** Differences are w/o visual writeback minus Normal. VSI uses the eight-task macro; historical VSTI uses the nine-subtype macro, not the five-task main score. Group intervals use 5,000 paired resamples within raw question-type strata and recompute the same nested task macro as the displayed point estimate. Counts are sample counts, not macro weights.

| Set | Group / protocol | $n$ | Normal | w/o visual writeback | $\Delta$ | 95% CI |
|---|---|---|---|---|---|---|
| VSI | Eight-task macro | 5130 | 65.08 | 63.47 | $-1.61$ | – |
| VSTI | Nine-subtype macro | 6042 | 72.87 | 68.64 | $-4.23$ | – |
| VSI | Object relations | 1678 | 74.16 | 70.41 | $-3.75$ | $[-5.77, -1.75]$ |
| VSI | Distance and size | 2075 | 66.40 | 64.33 | $-2.07$ | $[-3.05, -1.10]$ |
| VSI | Counting and routes | 759 | 58.81 | 57.19 | $-1.62$ | $[-4.49, +1.21]$ |
| VSI | Appearance order | 618 | 55.50 | 59.55 | $+4.05$ | $[+0.65, +7.44]$ |
| VSTI | Distance and displacement | 1744 | 42.19 | 40.60 | $-1.59$ | $[-3.07, -0.14]$ |
| VSTI | Motion and relations | 4298 | 81.64 | 76.65 | $-4.99$ | $[-6.40, -3.61]$ |

several spatial groups, but its contribution is not uniformly positive. The appearance-order result concerns this SFT checkpoint, not an intrinsic inability of the architecture to represent time. The condition w/o visual writeback is not the Visual-only training baseline, and this intervention does not assign the full QUASAR–Visual-only difference to the residual.

## D. Norm-Bound Sensitivity of Residual Injection Strength

The residual budget limits how strongly geometry can modify native visual evidence. As defined in Eq. (5), the norm limit is relative to the original visual feature, not an absolute vector norm. The same bounded correction is added to the main visual features and their aligned DeepStack features. We vary the inference multiplier and relative norm budget at the same SFT checkpoint used for the writeback diagnostic, without changing the learned parameters.

Table 6: **Sensitivity to visual-writeback strength (SFT).** VSI: 5,130 questions/eight tasks. VSTI: historical 6,042 questions/nine subtypes. The default relative budget is 0.15. Differences are setting minus default, computed from unrounded scores. These are point estimates, not a training-seed study or an equivalence test.

| Setting | Multiplier | Norm budget | VSI | VSTI | $\Delta$ VSI / VSTI |
|---|---|---|---|---|---|
| w/o visual writeback | 0.0 | 0.15 | 63.47 | 68.64 | $-1.61$ / $-4.23$ |
| Near-zero norm | 1.0 | $10^{-6}$ | 63.35 | 68.73 | $-1.73$ / $-4.15$ |
| Budget 0.05 | 1.0 | 0.05 | 64.68 | 70.47 | $-0.40$ / $-2.40$ |
| Default | 1.0 | 0.15 | 65.08 | 72.87 | +0.00 / +0.00 |
| Budget 0.30 | 1.0 | 0.30 | 64.11 | 72.90 | $-0.97$ / $+0.03$ |
| Budget 0.50 | 1.0 | 0.50 | 63.94 | 72.79 | $-1.14$ / $-0.08$ |

In Table 6, zero and near-zero settings score below the default on both aggregates. Among the tested budgets, 0.15 has the highest VSI score; 0.30 has the highest VSTI score, only 0.03 points above the default. Increasing the budget therefore does not consistently help. At group level, the default has the highest observed score for VSI object relations, distance and size, and counting and routes, and for VSTI motion and relations. VSTI distance-and-displacement questions peak at 0.30 (42.45, 0.26 points above default), while VSI appearance order peaks under w/o visual writeback (59.55). These observations describe sensitivity around one trained model, not a universally optimal budget or a guarantee of safe behavior.

## E. Appearance-Order Control

The 618 VSI appearance-order questions require recognition of object categories and their first occurrence along the video. They provide an in-domain temporal-order control, but not a general test of visual-semantic ability. Table 7 retains Visual-only (SFT) as a common historical comparator so that each saved paired interval has the correct baseline.

Table 7: **VSI appearance order: accuracy on 618 questions.** Displayed SFT models are the main-result checkpoints; the two QUASAR GRPO variants share a different SFT parent (Table 27). In the numeric-only variant, only GRPO is restricted to numeric questions. The first panel uses Visual-only (SFT) as baseline; its SFT interval comes from 5,000 paired question resamples of the main evaluation. The second panel uses Visual-only (SFT+GRPO) and the corresponding 5,000-resample SFT+GRPO-pair interval.

| Model | Accuracy | Δ vs panel baseline | 95% CI |
|---|---|---|---|
| Visual-only (SFT) | 61.97 | – | – |
| QUASAR (SFT) | 57.93 | $-4.05$ | $[-7.64, -0.46]$ |
| QUASAR (SFT+GRPO) | 59.87 | $-2.10$ | – |
| QUASAR (SFT+GRPO, numeric-only) | 58.74 | $-3.24$ | – |
| *SFT+GRPO endpoints: Visual-only (SFT+GRPO) baseline* | | | |
| Visual-only (SFT+GRPO) | 59.71 | – | – |
| QUASAR (SFT+GRPO) | 59.87 | $+0.16$ | $[-3.63, +3.75]$ |

The selected QUASAR (SFT) model is lower than Visual-only (SFT) by 4.05 points, with an interval excluding zero. The all-task GRPO endpoint has a smaller negative point difference from that Visual-only (SFT) baseline. This does not establish recovery or non-inferiority. In the separate SFT+GRPO system comparison, Visual-only (SFT+GRPO) scores 59.71 and QUASAR (SFT+GRPO) scores 59.87, a difference of 0.16 points with CI $[-3.63, +3.75]$. This crossing interval is inconclusive, not evidence of equal performance. Neither comparison supports a claim about general visual understanding outside this task.

## F. Geometry-Source Diagnostics

We compare CUT3R (Wang et al., 2025c) and VGGT (Wang et al., 2025b) on camera–object absolute-distance diagnostics (Table 8). These tests use ground-truth object masks and a common distance definition; they do not replace the geometry frontend in a full QUASAR answer-generation experiment. Per-scene scale fitting further uses ground truth and measures an oracle calibration ceiling, not a deployable score.

Table 8: **CO-AbsD geometry oracles with ground-truth masks.** Scores are MRA after per-scene best-fit scale calibration. The confidence-filtered comparison uses the 344 rows retained by both sources (216 videos); no filtering retains 549 rows (267 videos). Differences are VGGT minus CUT3R. Neither oracle interval establishes an advantage or equivalence.

| Setting | Rows | CUT3R | VGGT | Δ | 95% CI |
|---|---|---|---|---|---|
| 70th-percentile confidence filter | 344 | 70.20 | 72.09 | $+1.89$ | $[-0.30, +4.01]$ |
| No confidence filter | 549 | 61.07 | 60.80 | $-0.27$ | $[-1.94, +1.28]$ |

Using a global fitted scalar rather than a separate scale for each scene, the no-filter held-out evaluation contains 274 questions and scores 32.52 for CUT3R and 26.39 for VGGT: a difference of $-6.13$ points with CI $[-10.79, -1.37]$. This is a more restrictive calibration than the per-scene oracle, but ground-truth masks and fitted calibration still distinguish it from end-to-end deployment. The median VGGT/CUT3R ratios are 0.274 for point-cloud extent and 0.322 for trajectory diameter, showing why uncalibrated distance scales require care.

Confidence filtering removes 205 rows with empty VGGT masks. Inspecting 17 rows from eight videos yields no flagged overlays, and changing from 16 to 32 crops gives a median relative distance difference of 0.94%. This small inspection supports those checked cases, not the absence of all extraction or binding errors. The diagnostics provide no demonstrated CO-AbsD replacement gain from VGGT; they do not rank all geometry sources or predict the effect of retraining QUASAR with a different frontend.

Absolute metric reasoning remains difficult: QUASAR (SFT+GRPO) scores 53.69 on VSI absolute distance and 41.72 on VSTI camera–object absolute distance. Scale calibration, object-instance

binding, and numeric formatting remain possible error sources, but these diagnostics do not isolate their contributions to answer errors.

## G. Architecture and Supervision Comparisons

Table 9 gives the complete eight-configuration SFT structure comparison; Table 3 shows three of these configurations alongside the main-result models. The Visual-only endpoint here and QUASAR (SFT, variant B) differ from the selected SFT checkpoints in the main results. Neither the selected QUASAR SFT checkpoint nor variant B is the GRPO parent (Table 27). The Metric-slot model uses seven dedicated tokens for spatial quantities, distinct from QUASAR's four spatial summary tokens. The Branch-gated model (no metric slots) uses the multi-branch metric model without that seven-slot interface. QUASAR (SFT) uses four decoder-side spatial summary tokens and auxiliary readouts. The paired letters identify training records within a model family, rather than component deletions or random seeds; their complete training differences remain unresolved.

**Basic architecture and training-time writeback.** The Basic spatial model pairs Qwen3-VL-8B with a single-path spatial module conditioned on the supplied reference type and pose. Four spatial updates produce a correction for the visual features. It has no branch mixing, spatial summary tokens, metric-prediction heads, or decoder-side auxiliary readouts. Auxiliary supervision acts on spatial-module states: reference type, reference pose, bird's-eye occupancy, and patch consistency use weights 0.20, 0.10, 0.05, and 0.05, respectively. Neither the answer-distribution KL term nor SIGReg is used.

The separately trained counterpart without visual writeback retains the spatial module and its auxiliary objectives, but sets the correction entering the language model to zero during training. The recorded backbone initialization, data paths, seed, optimization settings, and auxiliary weights match. Both runs complete one epoch (53,365 updates) over 426,916 training examples. They form a training-time writeback ablation within the Basic architecture, whereas comparison with full QUASAR also changes the architecture and auxiliary supervision.

**Basic-model evaluation scope.** For the counterpart trained without visual writeback, zero correction is directly recorded in the training audit. The saved evaluation arguments describe the command-line flag, not its effective value after checkpoint loading. The available evaluator restores the checkpoint flag, but the effective setting was not logged in the inspected historical evaluation records.

Table 9: **Complete SFT spatial-structure comparison.** Qwen3-VL-8B; VSI uses 5,130 questions/eight tasks and historical VSTI uses 6,042 questions/five tasks. All eight trained configurations are retained, including both Metric-slot and QUASAR A/B records. Model definitions and training records are given in this section, with QUASAR and Visual-only checkpoint roles in Table 27.

| SFT spatial structure | VSI | VSTI |
|---|---|---|
| Visual-only | 63.22 | 59.95 |
| Basic spatial model | 62.64 | 63.84 |
| Basic spatial model, trained without visual writeback | 63.67 | 60.75 |
| Metric-slot model (variant A) | 64.25 | 66.29 |
| Metric-slot model (variant B) | 64.32 | 66.17 |
| Branch-gated model (no metric slots) | 65.16 | 65.52 |
| QUASAR (SFT, variant A) | 65.10 | 65.28 |
| QUASAR (SFT, variant B) | 65.08 | 66.11 |

The configuration trained without visual writeback scores 1.03 points higher on VSI and 3.09 lower on VSTI than the Basic spatial model. Relative to the Branch-gated model (no metric slots), QUASAR (SFT, variant B) changes VSI by $-0.08$ and VSTI by $+0.59$. This mixed pattern does not support a monotonic progression in which every added component improves both benchmarks. It does support evaluating the full system on both metric and relational tasks. The configuration trained without visual writeback is not a same-information, one-step learned substitute for the original four-step module, so these results do not establish that four updates are necessary.

Configuration comparisons test a trained system, whereas a matched-checkpoint intervention tests its dependence on an input path. Table 10 separates these questions.

Table 10: **What the architecture and mechanism comparisons control.** The main Qwen comparisons use VSI (5,130 questions/eight tasks) and VSTI (6,042 questions/five tasks). Inference w/o visual writeback instead uses historical VSTI (6,042 questions/nine subtypes). Each contrast retains its own training or inference identity.

| Comparison | Controlled contrast | Supported conclusion |
|---|---|---|
| Main QUASAR vs Visual-only | Same backbone and data recipe; selected SFT checkpoints or SFT+GRPO systems, with different SFT exposure. | Positive checkpoint differences, not strict equal-budget or single-module effects. |
| SFT configuration family | Separate runs of the Basic spatial model (including the variant trained without visual writeback), Metric-slot model, Branch-gated model (no metric slots), and QUASAR (SFT). | Mixed configuration results; no monotonic benefit from adding components. |
| Summary-token vs visual readout | Shared initialization/order; global readout and SIGReg representation both change. | A readout-placement package, not a token-only contrast. |
| Token insertion / global decoder supervision | Under visual readout, retaining vs omitting summary tokens changes insertion; local-only decoder supervision disables 12 decoder global/condition objectives. Spatial-module and local decoder objectives remain. | Task-dependent placement and supervision effects. |
| Decoder reference-coordinate loss on/off | Only two decoder reference-coordinate weights change from 0.1 to 0; spatial-module supervision remains. | VSTI benefits from this supervision, while VSI changes in the opposite direction. |
| Without summary tokens, their readouts, and SIGReg | Spatial summary tokens, associated prediction heads and training objectives, and SIGReg are removed together in a separate training configuration. | Does not establish that four spatial summary tokens are necessary or isolate a token-only effect. |
| Inference w/o visual writeback | Same QUASAR SFT checkpoint and inputs as the writeback diagnostic; w/o visual writeback. | Answer dependence on the residual, with positive and negative task effects. |
| Norm-budget sweep | Same QUASAR SFT checkpoint as the writeback diagnostic; only inference strength settings change. | Local sensitivity, not a universally optimal budget. |

**Auxiliary-target revision.** The training-data revision associated with the later four-token run changes availability of object annotations and metric targets, not the number of examples. Both training variants contain 426,916 rows and the same question-type counts; their auxiliary records differ. For absolute-distance, counting, and object-size questions, the number with valid object targets increases from zero to 6,172/46,213, 2,281/8,390, and 3,722/11,607, respectively. Camera-displacement coverage increases from zero to 27,296/28,855 and camera–object absolute-distance coverage to 30,919/30,919. This changes effective supervision even when the configured loss weights are unchanged. It does not establish valid semantic-reference poses at evaluation.

The A/B labels identify separate dated runs: Metric-slot model (variant A) and QUASAR (SFT, variant A) begin on June 13, QUASAR (SFT, variant B) on June 15, and Metric-slot model (variant B) on June 16. QUASAR (SFT, variant B) is the original four-token QUASAR SFT endpoint; endpoint steps are recorded in Appendix R. The A/B comparisons are not treated as single-factor experiments because their complete training differences remain unresolved. Differences between these runs therefore cannot be assigned solely to repaired targets, spatial-summary-token insertion, or the number of latent updates.

**Supervision placement and coordinate controls.** The readout-placement experiment trains four controls from a shared initialization and sample order with one controlled seed. All retain spatial-module supervision and four local decoder objectives. *Summary-token readout* inserts four summary tokens and reads global/condition targets from their final states; SIGReg acts on summary states. *Visual readout, tokens retained* keeps the tokens but reads these targets from the masked mean of final visual states; SIGReg acts on the individual valid visual states. *Visual readout, no summary tokens* omits summary-token insertion while retaining the same visual readout and SIGReg source. *Local-only decoder supervision* also inserts no summary tokens and disables all 12 decoder global/condition objective weights, retaining their heads for diagnostics and SIGReg on visual states. The summary-token versus visual-readout comparison therefore changes both global readout and SIGReg location. The two visual-readout controls isolate token insertion; comparing visual readout without summary tokens with local-only decoder supervision tests additional global/condition decoder supervision. The local-only control retains global supervision in the spatial module.

Table 3 (right) reports these controls. Readout-placement and decoder reference-coordinate loss are separate fresh-SFT families. The decoder reference-coordinate loss pair changes only two decoder reference-coordinate loss weights (0.1 versus 0); spatial-module coordinate supervision remains active. Without summary tokens, their readouts, and SIGReg is a separate jointly modified training configuration, not one of these four controls, a replacement main model, or a token-only ablation.

No placement dominates both benchmarks: visual readout without summary tokens has the highest VSI score, while local-only decoder supervision has the highest VSTI score in this four-control family. Turning off decoder coordinate losses changes VSI by $+0.30$ and VSTI by $-1.01$. The coordinate control therefore supports a task-dependent benefit of that supervision rather than a universal gain. The separate configuration without summary tokens, their readouts, and SIGReg reaches 65.67/66.33 after jointly removing the tokens, associated prediction heads and training objectives, and SIGReg. Its performance does not support the necessity of four tokens, and its bundled changes do not isolate their causal effect. The original four-token model remains the model evaluated in the main results. Supervision makes selected quantities decodable; it does not, by itself, show that the generated answer depends on them.

## H. Composite Evidence for Reference Conditioning

Spatial supervision can make information readable without ensuring that answers use it. Table 11 identifies the experimental settings; Table 12 summarizes their results by question group. These groups describe question content; they do not certify that an object-anchored numerical pose was available.

**Permuting the reference-conditioned path.** At the SFT+GRPO checkpoint, we permute reference types and reconstruct the associated pose path. Table 13 uses the normal run from the same intervention batch, not a different main-table evaluation. A separate uniform-branch-mixing arm tests sensitivity to branch mixing while retaining the other inputs.

Table 11: **Identities of the composite mechanism experiments.** SFT system differences use the main-result checkpoints; the writeback diagnostic uses a separate SFT endpoint (Table 27). Historical VSTI group results average two numeric or seven relation subtypes; they are not the five-task main aggregate. Numerical-pose and fixed-subset results retain their own sets and estimators in Tables 14 and 15.

| Evidence | Model / comparator | Set and interpretation |
|---|---|---|
| SFT system difference | QUASAR (SFT) minus Visual-only (SFT); main-result checkpoints | VSI (5,130 questions); historical VSTI (6,042 questions) diagnostic groups. Between trained systems. |
| w/o visual writeback | QUASAR (SFT), writeback-diagnostic checkpoint; Normal vs w/o visual writeback | Same full sets and group macros; within-checkpoint intervention. |
| Reference type/path permutation | QUASAR (SFT+GRPO); matched normal vs type/path permutation | Same full sets; changes type and associated pose-construction path. |
| Uniform branch mixing | QUASAR (SFT+GRPO); matched normal vs uniform branch mixing | Same full sets; changes mixing weights, not the reference type. |
| Text-only type router | QUASAR (SFT+GRPO); predicted vs rule-assigned reference types | 11,172 questions total; does not test construction of numerical poses. |
| Numerical pose | QUASAR (SFT+GRPO); fixed reference type, perturbed numerical pose | VSTI parser-valid subset (5,598 questions/five tasks). |
| Fixed-subset writeback/pose controls | QUASAR (SFT+GRPO); original reference vs w/o visual writeback and pose permutations | Same 512 historical VSTI multiple-choice questions, 127 scenes; question-level accuracy. |
| Gate distribution | QUASAR (SFT+GRPO); supplied reference | 19 question groups; descriptive mixture statistics. |

Table 12: **Question-group mechanism summary.** The SFT system difference is QUASAR minus Visual-only. Each drop is its matched normal minus the named intervention; a negative drop denotes an increase. Predicted-type change is text-predicted minus rule-assigned type performance. Each column retains its comparator in Table 11; columns must not be added. Branch counts identify the highest-weight branch per question; mean entropy describes the branch-weight distribution. Groups describe question content, not the supplied reference.

| Set | Group | $n$ | QUASAR − Visual-only (SFT) | Writeback-off drop | Type/path-permutation drop | Uniform-mixing drop | Predicted-type change | Dominant-branch counts; mean entropy |
|---|---|---|---|---|---|---|---|---|
| VSI | Object relations | 1678 | +5.87 | +3.75 | +1.79 | −0.12 | +0.30 | Metric: 710; Semantic: 968; 1.17 |
| VSI | Distance and size | 2075 | +3.51 | +2.07 | +0.08 | +0.57 | +0.21 | Metric: 2075; 1.20 |
| VSI | Counting and routes | 759 | −0.86 | +1.62 | +2.36 | +0.43 | −0.16 | Metric: 565; Semantic: 194; 1.18 |
| VSI | Appearance order | 618 | −4.05 | −4.05 | +2.10 | −0.16 | +0.16 | Metric: 618; 1.20 |
| VSTI | Motion and relations | 4298 | +2.35 | +4.99 | +3.19 | +0.26 | −0.14 | Target-camera: 4298; 1.26 |
| VSTI | Distance and displacement | 1744 | −1.05 | +1.59 | +1.82 | +0.60 | +0.05 | Target-camera: 1744; 1.26 |

Table 13: **Reference type/path permutation (SFT+GRPO).** VSI: 5,130 questions/eight tasks. VSTI: historical 6,042 questions/nine subtypes. Normal is matched within this experiment. Permutation changes the reference type and its associated pose-construction path, not the gate alone. CIs use 2,000 paired question-ID resamples.

| Set | Condition | Normal | Intervention | $\Delta$ | 95% CI |
|---|---|---|---|---|---|
| VSI | Reference type/path permutation | 65.57 | 63.88 | −1.69 | [−2.36, −1.00] |
| VSTI | Reference type/path permutation | 72.81 | 70.73 | −2.08 | [−2.89, −1.25] |

Both permutation intervals exclude zero, supporting the use of the supplied reference-conditioned pathway. Figure 5 shows the matched-normal differences and intervals by question group. On VSI, type/path permutation reduces scores on object relations, counting and routes, and appearance order, while distance and size shows little change. The uniform-branch-mixing point differences are smaller, $-0.36$ on VSI and $-0.37$ on VSTI, with intervals crossing zero for most question groups. Normal itself differs from the original main evaluation by $+0.20$ and approximately $-0.004$ points under these diagnostic aggregates; its drift intervals cross zero. All intervention differences use the matched normal run. The drift check alone cannot establish equivalence between runs.

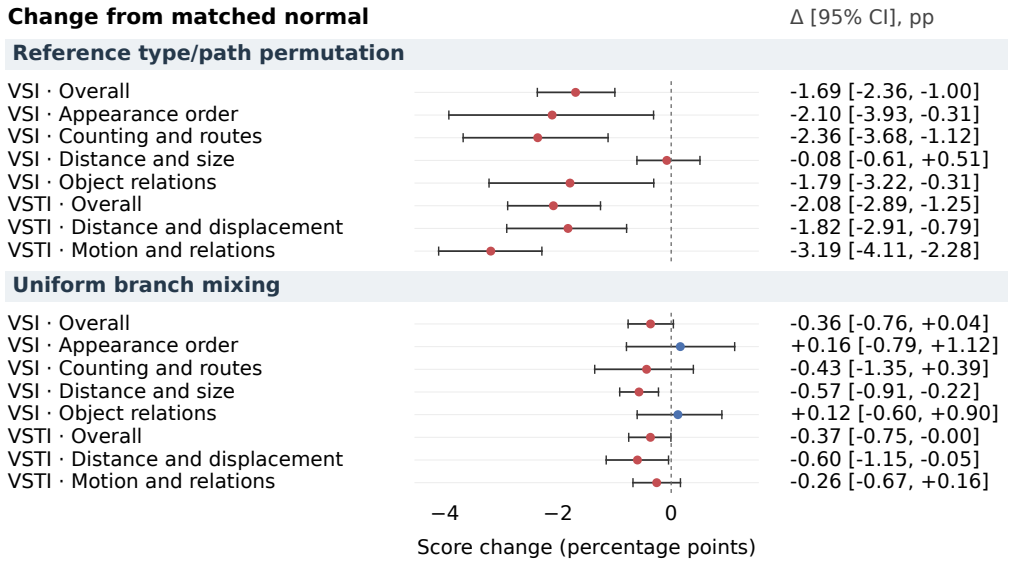



Figure 5: **Reference and branch-mixing interventions (SFT+GRPO).** Points show intervention minus matched-normal scores in percentage points; bars show paired 95% CIs from 2,000 question resamples. Negative values indicate lower scores. VSI uses 5,130 questions/eight tasks; historical VSTI uses 6,042 questions/nine subtypes. Groups describe question content. Reference type/path permutation changes both the type and its associated pose-construction path, measuring their combined effect.

**Numerical pose, with the reference type fixed.** A complementary intervention holds the target-camera type fixed and changes the numerical pose input. It covers 5,598 questions whose specified frame can be parsed in the historical VSTI set, excluding 444 unmatched questions. The condition w/o reference-pose conditioning disables the numerical-pose conditioning path; three additional arms permute poses with different permutation seeds. These are perturbation seeds on one trained model, not independent training seeds.

Table 14: **Numerical-pose dependence at the SFT+GRPO checkpoint.** The parsable target-camera subset has 5,598 historical VSTI questions and a five-task macro. Pose permutations hold the reference type fixed; seeds index perturbations, not training runs. Every difference uses the matching unmodified condition. Intervals use 5,000 paired scene-cluster resamples.

| Subset | Intervention | Normal | Changed | Δ | 95% CI |
|---|---|---|---|---|---|
| Parsable target-camera | w/o reference-pose conditioning | 66.28 | 61.60 | −4.68 | [−5.51, −3.88] |
| Parsable target-camera | Pose permutation (type fixed), seed 20260811 | 66.28 | 61.95 | −4.33 | [−5.14, −3.52] |
| Parsable target-camera | Pose permutation (type fixed), seed 20260812 | 66.28 | 61.81 | −4.47 | [−5.34, −3.66] |
| Parsable target-camera | Pose permutation (type fixed), seed 20260813 | 66.28 | 61.63 | −4.65 | [−5.46, −3.87] |

The target-camera results support numerical-pose dependence on the parser-valid subset. The revised VSTI intersection contains 5,292 questions and uses a different, nine-subtype aggregation; it is not pooled with the 5,598-question result here.

### FIXED-SUBSET WRITEBACK AND REFERENCE-POSE CONTROLS

The preceding residual and numerical-pose experiments differ in checkpoint and question set. This supplementary analysis checks both on the same SFT+GRPO model and 512 fixed questions (Table 15). It connects the existing findings under a common input and generation path, rather than introducing a new method or an independent training replication. The SFT residual results and the 5,598-question numerical-pose results retain their original checkpoints, predictions, and aggregates.

**Model, subset, and scoring.** We retain the original four-token model and all spatial and language weights, using Qwen3-VL-8B-Instruct snapshot `0c351dd01ed87e9c1b53cbc748cba10e6187ff3b`. Of the 5,598 parser-valid historical VSTI questions, 3,854 are multiple-choice candidates; 1,744 numerical questions are excluded. The 512 questions span 127 base scenes and seven raw question types. Selection cycles through types and then base scenes within each type, following original question order within each type–scene pair. The subset was fixed before its original predictions, without selecting for correctness or residual benefit. Each condition retains the original 32 frames, visual grid, CUT3R cache, question, and options. Frozen-weight inference uses BF16, batch size one, scaled dot-product attention, greedy decoding, at most 16 new tokens, and seed 0. Scoring uses historical 6,042-set labels and the original multiple-choice parser. Accuracy weights questions equally; it is neither the full five-task macro nor the revised 5,736-question protocol. Missing, failed, empty, or unparseable outputs count as incorrect within the fixed denominator of 512; none occurred.

**Interventions.** The originally supplied reference condition retains normal visual writeback with the recorded reference type and pose (`provided_full` in the implementation). The w/o visual writeback condition sets the actual bounded residual at the native visual positions and all three DeepStack levels to zero. This is not the w/o reference-pose conditioning intervention in Table 14. The pose-permuted conditions retain the reference type but substitute another question's numerical pose from the same video and coordinate system. They jointly change the 12-dimensional pose and the six reference-dependent channels of the 28-dimensional descriptor, retaining its other 22 channels and all other direct inputs. Writeback is recomputed on the recipient's visual and geometric inputs, not copied as a residual from the donor. The write sites use the residual after scaling, norm bounding, and conversion to inference precision, without applying these operations twice. Each condition builds its own generation cache. The four spatial-summary-token input embeddings remain the same learned parameters; no additional global 3D positional embedding is enabled. Downstream hidden states respond normally to the changed writeback.

Permutation seeds 20260811, 20260812, and 20260813 retain the original 5,598-question mappings rather than re-pairing within the subset. Respectively, 452, 439, and 444 donors lie outside the 512 questions; none of the subset's donor–recipient poses are numerically identical. All conditions are scored against the recipient's original answer, without relabeling for the donor reference. The three permutations perturb one model on the same questions, not three training runs or 1,536 independent questions.

**Statistics and task dependence.** For the permutation comparator, we average the three correctness indicators within each question before aggregating. Paired bootstrap sampling draws 127 base scenes with replacement 10,000 times using Python's `random.Random(0)`. Each resample divides the pooled correct count by the pooled question count, retaining scene multiplicities; it does not average scene-level accuracies. Intervals are the linearly interpolated 2.5th and 97.5th percentiles of the paired differences. The supplementary analysis plan was fixed before reading the new permutation predictions, but historical aggregate and original subset results were already known. We therefore report exploratory intervals, not a fresh confirmatory test or training-seed variability. Both primary difference intervals have positive lower bounds, satisfying the analysis plan's joint rule for subset-level writeback benefit and reference-match advantage.

Normal writeback repairs 44 errors under w/o visual writeback but introduces 19, yielding 25 additional correct answers. It exceeds the three permutations by 22, 25, and 21 correct answers,

Table 15: **Writeback and reference-pose controls on the same fixed subset.** QUASAR (SFT+GRPO); 512 historical VSTI multiple-choice questions from 127 base scenes. Accuracies are percentages, and differences are percentage points. The permutation mean averages within each question; its correct count is an average, not an additional sample. Paired 95% intervals are exploratory and use 10,000 base-scene bootstrap resamples.

| Condition | Correct / 512 | Accuracy (%) |
|---|---|---|
| Originally supplied reference | 415 | 81.0547 |
| w/o visual writeback | 390 | 76.1719 |
| Pose permutation (type fixed), seed 20260811 | 393 | 76.7578 |
| Pose permutation (type fixed), seed 20260812 | 390 | 76.1719 |
| Pose permutation (type fixed), seed 20260813 | 394 | 76.9531 |
| Within-question mean of three pose permutations | 392.3333 | 76.6276 |
| **Paired contrast** | **Difference (pp)** | **95% interval (pp)** |
| Originally supplied − w/o visual writeback | +4.8828 | [+1.8692, +8.0844] |
| Originally supplied − permutation mean | +4.4271 | [+2.3137, +6.5606] |

respectively (mean 22.6667). Table 16 describes the task dependence: camera-motion direction is sensitive to pose permutation, whereas left–right performance is close to the permutation mean and relative-distance subtype v2 scores lower with normal than w/o visual writeback. Overall benefit does not imply that every question or type improves.

Table 16: **Descriptive breakdown of the same 512 questions.** Question-level accuracy (%); originally supplied reference, w/o visual writeback, and the within-question mean of three permutations. Subtype v1/v2/v3 denotes dataset labels, not QUASAR versions. No subgroup significance claim is made.

| Raw question type | $n$ | Original reference | w/o visual writeback | Permutation mean |
|---|---|---|---|---|
| Camera-motion direction | 74 | 82.43 | 75.68 | 62.61 |
| Camera–object relative distance, v1 | 73 | 65.75 | 52.05 | 63.47 |
| Camera–object relative distance, v2 | 73 | 58.90 | 63.01 | 58.45 |
| Camera–object relative distance, v3 | 73 | 80.82 | 73.97 | 75.80 |
| Object–object left–right | 73 | 91.78 | 86.30 | 91.32 |
| Object–object near–far | 73 | 90.41 | 87.67 | 87.67 |
| Object–object up–down | 73 | 97.26 | 94.52 | 97.26 |

**Prediction provenance.** The original-reference and w/o visual writeback conditions reuse the first recorded predictions from the fixed-subset residual evaluation. Only the three permutation conditions were newly generated along that evaluation's input and generation path. All 512 original inputs and all 1,536 donor mappings and non-reference inputs were checked. Replaying Normal and w/o visual writeback inference on the first 16 questions reproduced all 32 parsed answers; write-site, zeroing, self-replay, restoration, and cache-isolation checks passed. All 1,536 new records were accepted without selecting favorable repeats, and independent scoring reproduced the results. This replay does not mean that all 512 Normal/w/o visual writeback predictions were regenerated. An earlier numerical-pose run disagrees with the original subset run on four unperturbed parsed answers, and incomplete historical package-version records leave the cause unresolved. Its permutation predictions are not pooled into Table 15. This check is not an additional independent replication of the historical results.

**Interpretation and limits.** The results support a useful contribution of visual writeback that depends on reference matching on this fixed subset. "Originally supplied" refers to the historical convention, not a newly audited ground-truth reference for each question. Because permutations retain the original answers, a performance drop does not establish that changing to another valid viewpoint produces its independently determined correct answer, nor does it exclude all generic perturbation explanations. No same-information, same-budget alternative architecture is tested, so the result does not establish the superiority or necessity of four updates, four tokens, or the present

structure over simpler alternatives. It is not a full-benchmark score or evidence of generalization to new scenes.

## I. Reference Type Mapping and External Assignment

The reference interface supplies a discrete type and a numerical pose separately, as described in §3.2. Table 17 lists the assignments used for evaluation. A type label describes the intended spatial convention; it does not guarantee that the corresponding numerical pose has been grounded correctly. The full question is read by the MLLM, whereas Spatial Latent Transformation receives geometry and the explicit reference condition.

Table 17: **Reference assignments in the evaluation inputs.** Counts describe the supplied reference types. Target-camera records marked valid can still use a parsing fallback.

| Type / count | Question groups | Numerical-reference meaning and observed coverage |
|---|---|---|
| Scene-canonical / 3968 | VSI count, absolute/relative distance, object/room size, appearance order | Origin at the mean of valid video-estimated 3D points, retaining reconstruction axes. This is not an inferred viewpoint for each question. |
| Semantic-anchor / 1162 | VSI relative direction (all difficulties) and route planning | Reference position and orientation anchored to the designated object or situation. |
| Target-camera / 6042 | All historical VSTI tasks | Uses the estimated pose of the specified observation frame. The parser-valid subset has 5,598 questions. |

The target-camera, semantic-anchor, and scene-canonical references correspond to the implementation identifiers `target_camera_pose`, `semantic_query_pose`, and `canonical_bev_scene_pose`, respectively. Automatically predicting a reference type is different from identifying the referenced object instance or a valid numerical pose. Likewise, a convention assigned to appearance-order questions does not explain the sign of a performance difference on that task.

**External inputs.** External image and multi-image evaluations use a separate assignment interface because their questions lack VSI/VSTI task annotations. Its heuristic assigns metric/count questions to the scene-canonical reference and camera/viewpoint/direction questions to the last input camera; other questions default to the scene-canonical reference. The interface also supports fixed scene-canonical or last-camera assignments. These are supplied conventions, not learned semantic anchoring. The external results evaluate the chosen assignment together with the trained model, without a ScanNet (Dai et al., 2017) object-annotation bank or benchmark-specific parameter updates. Inference w/o visual writeback changes the written signal but retains those other input and training choices.

## J. Spatial Gate Distribution

Across 19 question groups (5,130 VSI and 6,042 historical VSTI questions), mean gate probabilities at the SFT+GRPO checkpoint largely follow the supplied reference type, with mean entropy of approximately 1.15–1.26. These distributions describe supplied-reference conditioning, not autonomous question understanding or valid numerical-pose construction.

Uniform branch mixing gives changes of $-0.36$ on VSI (95% CI $[-0.76, +0.04]$) and $-0.37$ on historical nine-subtype VSTI (CI $[-0.75, -0.00]$, rounded). Both use the matched normal run at the SFT+GRPO checkpoint and 2,000 paired question resamples. Their point differences are smaller than the corresponding reference type/path permutation differences, $-1.69$ and $-2.08$. This contrast does not rank the intrinsic importance of the gate and reference: the perturbations change different parts of the computation.

## K. Spatial Latent Transformation Diagnostics

Figures 6 and 7 show diagnostic examples from a 72-example validation subset. They visualize relationships between spatial cues, annotated object-center neighborhoods, and attention for a shared

next-token probe. They are not a quantitative localization benchmark or examples of correcting previously wrong answers.

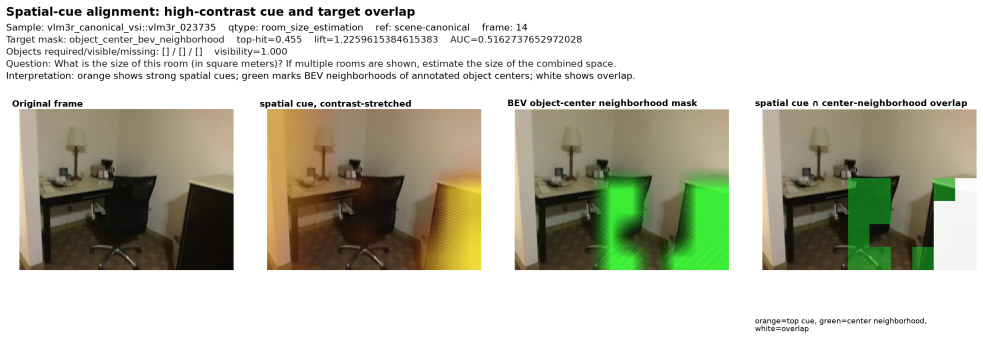



Figure 6: **Spatial-cue alignment on a room-size question.** Orange shows strong spatial cues, green marks patches within BEV neighborhoods of annotated object centers, and white shows their overlap. The token-level AUC over this example's input sequence is 0.5163. This scene-level example does not establish accurate localization of named objects or a strong alignment result.




Figure 7: **Residual sensitivity on an SQA-style question.** Red/blue show attention increases/decreases for the same "yes" probe token after enabling visual writeback; white marks overlap between strong spatial cues and attention gains. The highest-probability next token is "yes" in both conditions. Its log-probability changes by 0.0005; the JS divergence between frame-level attention distributions is 0.0008. The example illustrates a small representational change, not an answer correction or an official VSI/VSTI success case.

## L. Text-Only Reference-Type Router

A text-only Qwen3-VL-8B router predicts the three reference-type labels from question text without images, answers, or geometry. Across 11,172 VSI/VSTI questions it agrees with the rule-assigned type labels 99.27% of the time and produces no invalid labels. This is label agreement, not the accuracy of estimated camera poses or semantic anchors.

The overall point changes in Table 18 are small in this setup, but their uncertainty remains compatible with changes in either direction. The router tests replacement of a discrete input interface; it does not show that QUASAR can obtain arbitrary numerical reference conditions from language.

Table 18: **Replacing rule-assigned reference types with text-only predictions (SFT+GRPO).** Differences are text-predicted-type minus rule-assigned-type inference. VSI uses its eight-task aggregate; historical VSTI uses the nine-subtype aggregate. Group counts are those in Table 12. All intervals cross zero; they do not establish equivalence.

| Set | Scope | $\Delta$ | 95% CI |
|---|---|---|---|
| VSI | Eight-task macro | $+0.16$ | $[-0.11, +0.44]$ |
| VSI | Object relations | $+0.30$ | $[-0.18, +0.78]$ |
| VSI | Distance and size | $+0.21$ | $[-0.04, +0.47]$ |
| VSI | Counting and routes | $-0.16$ | $[-0.60, +0.27]$ |
| VSI | Appearance order | $+0.16$ | $[-0.65, +0.99]$ |
| VSTI | Nine-subtype macro | $-0.17$ | $[-0.47, +0.05]$ |
| VSTI | Motion and relations | $-0.14$ | $[-0.35, +0.05]$ |
| VSTI | Distance and displacement | $+0.05$ | $[-0.38, +0.49]$ |

## M. Backbone and External Spatial Transfer

**Backbone and training-recipe transfer.** Table 19 reports completed SFT endpoints. The Qwen3-VL 2B/4B comparisons are against Visual-only; the Qwen2.5-VL (Bai et al., 2025b) comparisons instead use the Metric-slot model, which already has a spatial module and must not be called visual-only. InternVL (Zhu et al., 2025c) results use the revised VSTI set and are displayed in a separate panel. Each pair compares configurations within its own backbone and recipe, not a common cross-backbone training budget. The SenseNova-SI-1.5-InternVL3-8B and plain InternVL3-8B rows load different pretrained backbones, not merely different training data.

Table 19: **Full-schedule SFT transfer.** Qwen rows use VSI (5,130 questions/eight tasks) and VSTI (6,042 questions/five tasks); InternVL rows use VSI (5,130 questions/eight tasks) and revised VSTI (5,736 questions/nine subtypes). QUASAR (SFT) denotes the transferred spatial architecture. Each difference is QUASAR minus the stated comparator. No endpoint paired difference is available for plain InternVL3-8B.

| Backbone | Configuration | VSI | VSTI | $\Delta$ VSI | $\Delta$ VSTI |
|---|---|---|---|---|---|
| Qwen3-VL-2B | Visual-only | 58.22 | 57.11 | – | – |
| | QUASAR (SFT) | 61.47 | 61.73 | +3.25 | +4.62 |
| Qwen3-VL-4B | Visual-only | 62.46 | 57.54 | – | – |
| | QUASAR (SFT) | 63.75 | 60.63 | +1.29 | +3.09 |
| Qwen2.5-VL-3B | Metric-slot model | 35.09 | 53.79 | – | – |
| | QUASAR (SFT) | 52.74 | 57.86 | +17.65 | +4.08 |
| Qwen2.5-VL-7B | Metric-slot model | 37.33 | 55.21 | – | – |
| | QUASAR (SFT) | 58.09 | 61.35 | +20.76 | +6.14 |
| *InternVL3-8B: different VSTI set and aggregation* | | | | | |
| SenseNova-SI-1.5-InternVL3-8B (Cai et al., 2026) | Visual-only | 67.26 | 72.37 | – | – |
| | QUASAR (SFT) | 66.67 | 74.74 | $-0.59$ | +2.36 |
| Plain InternVL3-8B | QUASAR (SFT) | 58.31 | 75.36 | – | – |

The architecture can be trained across these backbones, but the transfer effect is not uniformly positive: on the SenseNova-SI-1.5-InternVL3-8B backbone, QUASAR (SFT) lowers VSI by 0.59 while raising revised VSTI by 2.36 points. These are single-run endpoint differences without training-seed intervals. The plain InternVL3-8B result is an absolute score, not evidence of an endpoint paired improvement. The high InternVL VSTI scores also cannot be read as improvements over Qwen's historical five-task scores.

**External spatial benchmarks.** We evaluate SPAR-Bench (Zhang et al., 2025), ViewSpatial-Bench (Li et al., 2025a), and SPBench (Li et al., 2025b) without benchmark-specific SFT, GRPO, or prompt tuning (Table 20). All use a shared evaluator (numeric answers: MRA; counts: exact match; choices: accuracy), not each benchmark's official scorer. QUASAR Normal and w/o visual writeback use the same all-task GRPO checkpoint; Visual-only is its full SFT endpoint (Table 27), not

Visual-only (SFT+GRPO) from the main tables. Thus Normal versus Visual-only mixes model design and training stage, while w/o visual writeback versus Normal is a within-checkpoint intervention.

Table 20: **External spatial evaluation with the shared evaluator.** Visual-only is SFT; QUASAR is SFT+GRPO. Normal activates the residual, w/o visual writeback zeros it, and the fixed image-count rule combines w/o visual writeback for single-image samples with Normal for multi-image samples. The policy is fixed, not learned or selected from ground-truth answers.

| | | | | QUASAR (SFT+GRPO) | | |
|---|---|---|---|---|---|---|
| Benchmark | $n$ | Base | Visual-only (SFT) | Normal | w/o visual writeback | fixed image-count rule |
| SPAR-Bench | 7211 | 38.23 | 39.79 | 40.80 | 39.57 | 40.95 |
| ViewSpatial-Bench | 5712 | 41.04 | 46.69 | 43.79 | 43.40 | 44.08 |
| SPBench | 1328 | 57.99 | 60.63 | 59.13 | 58.00 | 58.77 |

On SPAR, Normal improves over Base by $+2.57$ (95% paired CI $[+1.56, +3.57]$). The difference from Visual-only is $+1.02$, with CI $[-0.0014, +2.04]$ crossing zero. Turning the residual off lowers the score by 1.23 points (CI $[-2.11, -0.33]$), supporting a beneficial contribution from the residual on this benchmark. The fixed image-count rule increase over Normal is only $+0.14$ ($[-0.35, +0.63]$), not a reliable overall improvement.

ViewSpatial exposes a negative transfer result. Although Normal exceeds Base by $+2.75$ ($[+1.63, +3.90]$), it trails Visual-only (SFT) by $-2.91$ ($[-4.03, -1.75]$). Most examples are single-image (4,607/5,712), but that composition alone does not identify why Visual-only performs better: removing the residual changes Normal by $-0.39$, with an interval crossing zero. The fixed image-count rule raises the point estimate by 0.30 ($[-0.56, +1.12]$), also inconclusively. These results do not justify attributing the entire Visual-only gap to geometric interference.

On SPBench, Normal minus Visual-only is $-1.50$ ($[-3.33, +0.29]$). A crossing interval does not establish parity. Table 21 separates 1,009 single-image (SI) from 319 multi-view (MV) examples. Normal is lower than Visual-only on SI by $-2.51$ ($[-4.62, -0.43]$); on MV its $+1.69$ difference has CI $[-1.82, +5.11]$. Within the QUASAR checkpoint, inference w/o visual writeback decreases MV performance by 3.20 points ($[-5.77, -0.72]$). This supports residual use on MV examples without establishing an MV advantage over Visual-only. The fixed image-count rule is lower than Normal overall by 0.36 points ($[-1.62, +0.87]$), so the fixed rule does not consistently improve external performance.

Table 21: **SPBench SI/MV decomposition.** Scores use the shared evaluator. The fixed image-count rule exactly reuses w/o visual writeback on SI and Normal on MV; it does not close the SI gap to Visual-only. The MV advantage over Visual-only is a point estimate with a CI crossing zero.

| Model / inference condition | SI ($n = 1009$) | MV ($n = 319$) | All ($n = 1328$) |
|---|---|---|---|
| Base | 54.80 | 68.09 | 57.99 |
| Visual-only (SFT) | 56.66 | 73.17 | 60.63 |
| *QUASAR (SFT+GRPO)* | | | |
| Normal | 54.15 | 74.86 | 59.13 |
| w/o visual writeback | 53.68 | 71.66 | 58.00 |
| fixed image-count rule | 53.68 | 74.86 | 58.77 |

All external CIs above use 5,000 paired question-level resamples of fixed predictions. They describe checkpoint comparisons, not a distribution over training runs. Appendix T gives the separate MindCube-Tiny absolute evaluation; it is not a paired architectural transfer test.

## N. Cached-Input Generation Time and Memory

The recorded benchmark times the answer-generation call (`generate_one`) with CUDA synchronization after batch preparation and GPU transfer, with batch size 1, 10 warm-up examples, 100 timed examples, and at most 16 generated tokens. Model loading, data loading and batch assembly,

cache I/O, scoring, prediction merging, and evaluator replay are outside that boundary. The cached visual and geometric inputs avoid measuring their extraction from a new video.

Tokenization and batch preparation precede the timed call; model prefill is a different operation. In the available implementation, the call constructs prompt embeddings, performs the first full-prompt language-model forward pass (prefill), and then generates tokens autoregressively. QUASAR also computes its spatial transformation and residual writeback inside the call. Thus this interface is not decode-only. The historical evaluator implementation is not archived, so its exact internal timing decomposition remains unverified; the current implementation does not retrospectively certify that run. The reported figures are cached-input generation costs, not raw-video-to-answer latency.

Table 22: **Cached-input generation time and peak allocated memory.** Seconds are rounded to three decimals. The timing comparison uses the main-result SFT checkpoints and the stated SFT+GRPO models (Table 27). Memory is peak allocated memory in the source report's GB units, not total device memory. Paired memory entries are VSI / VSTI; the numeric-only GRPO variant retains its reported range.

| Model | VSI mean | VSI p95 | VSTI mean | VSTI p95 | Peak GB |
|---|---|---|---|---|---|
| Visual-only (SFT) | 0.604 | 1.008 | 0.555 | 0.582 | 17.68 / 17.64 |
| QUASAR (SFT) | 0.631 | 1.059 | 0.565 | 0.578 | 18.36 / 18.32 |
| QUASAR (SFT+GRPO) | 0.620 | 1.032 | 0.559 | 0.567 | 18.36 / 18.32 |
| QUASAR (SFT+GRPO, numeric-only) | 0.634 | 1.069 | 0.563 | 0.576 | 18.32–18.36 |
| Visual-only (SFT+GRPO) | 0.606 | 1.030 | 0.543 | 0.550 | 17.68 / 17.64 |

From unrounded means, the SFT comparisons add 4.39% on VSI and 1.90% on VSTI; SFT+GRPO system overhead is 2.32% and 2.92%. The corresponding peak allocated-memory increase is approximately 0.67–0.68 GB. These results quantify an incremental generation cost for the local setup. They do not include the geometry frontend, establish a cross-hardware speed advantage, or combine a published model score with a different model's timing run.

## O. CV-Bench Transfer without Visual Writeback

CV-Bench (Tong et al., 2024) evaluates 2,638 single-image questions: 1,438 in 2D and 1,200 in 3D. Base is the released backbone, Visual-only is its full SFT endpoint, and QUASAR is the all-task SFT+GRPO model (Table 27). The cross-model comparison therefore mixes architecture and training stage; Normal versus w/o visual writeback is the fixed-checkpoint control.

Table 23: **CV-Bench scores and the inference w/o visual writeback control.** Normal minus Visual-only (SFT) is $-2.92$ points overall, with 95% paired CI $[-4.25, -1.63]$. The difference w/o visual writeback minus Normal is $+3.87$ ($[+2.62, +5.16]$). The fixed image-count rule equals w/o visual writeback because all inputs are single-image. Base and w/o visual writeback have equal overall scores but different subset scores.

| Model / inference condition | 2D ($n = 1438$) | 3D ($n = 1200$) | All ($n = 2638$) |
|---|---|---|---|
| Base | 80.32 | 92.08 | 85.67 |
| Visual-only (SFT) | 80.67 | 89.58 | 84.72 |
| *QUASAR (SFT+GRPO)* | | | |
| Normal | 78.79 | 85.42 | 81.80 |
| w/o visual writeback (also fixed image-count rule) | 81.99 | 90.08 | 85.67 |
| Normal − Visual-only (SFT) | $-1.88$ | $-4.17$ | $-2.92$ |
| 95% CI | $[-3.69, +0.00]$ | $[-6.08, -2.25]$ | $[-4.25, -1.63]$ |

Visual-only minus Base is $-0.95$ points with CI $[-1.97, +0.11]$, an inconclusive interval rather than evidence of capability preservation. Normal minus Base is $-3.87$ ($[-5.16, -2.58]$). The full residual contributes to a cost on this benchmark: disabling it improves both subsets (Table 23). Recovery to the Base overall value is specific to CV-Bench. QUASAR w/o visual writeback remains 2.00 points below Base on 3D and 1.67 above it on 2D, so neither the predictions nor subset behavior have been

restored to Base. This intervention does not establish recovery of general visual ability on other datasets.

## P. Auxiliary Objectives and Inference Path

During SFT, auxiliary spatial objectives act at two locations: the spatial module and decoder hidden states. Global decoder readouts use the four spatial summary positions; local readouts use visual-token positions. The summary positions follow the visual sequence and precede the question, so they summarize earlier visual evidence rather than representing four generated reasoning steps. Table 24 gives the configured coefficients for the original four-token SFT recipe.

Table 24: **Configured SFT auxiliary coefficients.** Each shared weight is applied separately to spatial-module and decoder-side readouts. The seven metric families are scene scale, room extent, camera displacement, object size, object-pair distance, camera–object distance, and object count. Nonzero coefficients do not imply valid targets for every training example. GRPO disables these auxiliary objectives and updates LoRA only.

| Location | Target / term | Loss | Weight |
|---|---|---|---|
| Spatial / decoder | Reference type | CE | 0.2 |
| Spatial / decoder | Numerical reference pose | L1 | 0.1 |
| Spatial / decoder | BEV occupancy | BCE | 0.05 |
| Spatial / decoder | Object-center heatmap | BCE | 0.05 |
| Spatial / decoder | Branch-mixture target | CE | 0.1 |
| Spatial / decoder | Patch consistency | L1 | 0.05 |
| Spatial / decoder | Target-camera coordinates | L1 | 0.1 |
| Spatial / decoder | Semantic-reference coordinates | L1 | 0.1 |
| Spatial / decoder | Canonical BEV coordinates | L1 | 0.1 |
| Spatial / decoder | Seven metric terms | L1 | 0.05 each |
| Stability | SIGReg (128D, 512 projections) | Slice statistics | 0.03 |
| Stability | Answer-distribution KL to the LoRA-off reference | KL | 0.05 |

The answer loss has coefficient 1; Eq. (6) combines it with the spatial, decoder, and stability terms. The auxiliary objective for discretized numerical targets and additional metric-consistency objectives have zero coefficients, and the global 3D positional-encoding option is disabled in this recipe. Target availability and the auxiliary-target revision affect which configured terms supply training signal.

SFT applies LoRA to attention and feed-forward projections; its rank, scaling, and dropout are given in Appendix R. It also trains the spatial module, residual scale, summary embeddings, readouts, and regularization projection while freezing pretrained backbone and geometric-frontend weights. SIGReg (Balestriero & LeCun, 2025) acts on projected summary states. The SFT answer-distribution KL uses a stopped-gradient reference with LoRA disabled but spatial writes and summary embeddings retained; this is not the released visual-only model.

GRPO retains the spatial forward path while freezing the added spatial and readout modules; it optimizes LoRA with the answer reward and KL term. The reference policy disables LoRA but retains the spatial path, and is therefore not the visual-only Base model. At inference, decoder auxiliary predictions are not fed back as answer inputs. The metric predictions inside Spatial Latent Transformation are different: they feed the metric branch and remain part of residual construction. Estimated observation geometry, supplied reference conditions, bounded writeback, and the shared DeepStack correction also remain active.

## Q. Geometry Descriptors and Spatial Readouts

**Geometry and target sources.** Training uses precomputed geometry and auxiliary-target records. The main QUASAR SFT inputs include both CUT3R caches and records supplying ScanNet depth maps, camera intrinsics, and poses. Object statistics come from a hybrid object bank, not a uniformly human-annotated source. These training sources are distinct from the CUT3R inference path; object annotations and auxiliary targets are not additional test-time inputs. Coordinate units depend on the source and cannot uniformly be treated as metres.

**Descriptor components.** For visual position $n$, let $c^{\mathrm{src}}$ be the observing camera's origin and let $f^{\mathrm{src}}, r^{\mathrm{src}}, u^{\mathrm{src}}$ be the third, first, and second columns of its rotation, respectively. With per-video observed-point mean $\mu$ and scale $\ell > 0$, define

$$\bar{p}_n = \frac{p_n - \mu}{\ell}, \quad \rho_n^s = \frac{p_n - c^{\mathrm{src}}}{\ell}, \quad \rho_n^q = \frac{p_n - c^{\mathrm{qry}}}{\ell}, \quad \delta c_{q\to s} = \frac{c^{\mathrm{src}} - c^{\mathrm{qry}}}{\ell}.$$

The 28-dimensional input descriptor is

$$\begin{aligned} d_n = [\bar{p}_n;\, \rho_n^s;\, \rho_n^q;\, \delta c_{q\to s};\, c^{\mathrm{src}}/\ell;\, f^{\mathrm{src}};\, r^{\mathrm{src}};\, u^{\mathrm{src}};\\ \gamma_n;\, m_n;\, u_n^{\mathrm{patch}};\, v_n^{\mathrm{patch}}] \in \mathbb{R}^{28}. \end{aligned} \tag{7}$$

The first eight entries are three-vectors; the remaining four are scalars. Here $\gamma_n$ is geometric confidence, $m_n$ indicates geometric validity, and $(u_n^{\mathrm{patch}}, v_n^{\mathrm{patch}}) \in [-1, 1]^2$ locates the patch within its frame. Only $\bar{p}_n$ is centered at $\mu$: the source-camera term is $c^{\mathrm{src}}/\ell$, not $(c^{\mathrm{src}} - \mu)/\ell$. Source and reference offsets retain the reconstruction axes. Full reference orientation conditions the latent updates, while rotated coordinates provide training targets rather than additional inference inputs.

**Dimensions and initialization.** The Qwen3-VL-8B configuration uses native feature dimension $H = 4096$ and spatial-state dimension $F = 512$. Four residual blocks have separate parameters and a $1/K$ update factor, as in Eq. (3). Each branch forms a residual code from $[\tilde{z}_n; h_Q; d_n]$, then normalizes and projects it into the native visual dimension. The metric branch predicts the seven families listed in Table 24. Their scalar/vector statistics occupy 17 dimensions. A zero-initialized projection maps these predictions back to the branch-code dimension and adds them at every visual position before normalization and output projection. The branch output projections are also zero-initialized. The learned scale $s = 0.25\sigma(a)$ starts at $0.05$; the norm floors in Eq. (5) handle vanishing feature or residual norms.

**Cached metric targets.** The main QUASAR SFT run uses precomputed auxiliary-target records containing scene, camera, and object statistics with availability flags. Its training-set audit reports scene and camera statistics for all 426,916 examples, and object statistics for 296,054; the remaining 130,862 are recorded as lacking an object bank. Target availability is distinct from both answer annotation and reference-pose validity. Table 25 gives the seven families and their 17-component order.

Table 25: **Cached metric-target layout.** Positions are one-based. The listed components are stored after a $\log(1 + x)$ transform (`log1p` in the implementation); the table describes saved training fields.

| Family | Positions | Dims. | Ordered components |
|---|---|---|---|
| Scene scale | 1 | 1 | Scene scale. |
| Room extent | 2–4 | 3 | Extents along the three source-coordinate axes. |
| Camera displacement | 5–7 | 3 | Primary, first-to-reference, maximum-to-reference displacement fields. |
| Object size | 8–10 | 3 | Mean size along each source-coordinate axis. |
| Object-pair distance | 11–13 | 3 | Minimum, mean, maximum center distance. |
| Camera–object distance | 14–16 | 3 | Minimum, mean, maximum reference-to-center distance. |
| Object count | 17 | 1 | Number of retained object-bank entries. |

Spatial quantities retain their source-coordinate units before the transform and are not uniformly divided by the scene scale; object count has no physical unit. Object statistics summarize retained bank entries, not necessarily the object or pair named in the question or the full scene. The field layout does not specify every fallback used when a reference is invalid. Training-target definitions follow the saved records rather than an interchangeable on-the-fly constructor; the historical generation rules are only partly recoverable (Appendix R).

**Readout locations and targets.** The four summary positions have learned input embeddings. Their implementation identifiers range from `<|spatial_query_0|>` to `<|spatial_query_3|>`. For final decoder states $h_j^{\mathrm{sum}}$, global heads read $\mathrm{LN}(\frac{1}{4}\sum_{j=1}^{4} h_j^{\mathrm{sum}})$; local heads read normalized final visual states. Spatial-module heads instead read pooled latent states or branch codes. The two sets of readouts receive corresponding target families and weights (Table 24).

Reference regression uses 12-dimensional poses. Occupancy and object-center maps use $32 \times 32$ bird's-eye-view grids. Patch consistency reconstructs $\rho_n^q$, whereas target-camera and semantic-coordinate heads predict

$$y_n^Q = (R^{\mathrm{qry}})^\top (p_n - c^{\mathrm{qry}})/\ell \in \mathbb{R}^3.$$

The canonical BEV head predicts a two-component projection of normalized shared coordinates. The checked historical target records use the first and third components; this does not establish physical ground-plane alignment across sources. The axes in Figure 3 label an illustrative plane, not an additional ground-alignment step. Reference type and numerical validity determine applicable pose/coordinate losses; object and metric targets are used where available. These supervision masks are distinct from the descriptor's geometric validity entry. Classification and mixture objectives use cross-entropy, map objectives use binary cross-entropy, and numerical objectives use masked L1. Neither occupancy maps, object annotations, metric targets, nor gold answers enter the answer-generation prompt.

## R. Scoring Protocols and Model Identity

Table 26 defines the model names and inference conditions.

Table 26: **Model and diagnostic identities.** "SFT+GRPO" denotes the all-task GRPO endpoint unless explicitly marked numeric-only, not the selected SFT row. Decoder-side spatial readouts are auxiliary heads, not a separate model. The conditions w/o visual writeback and fixed image-count rule retain the training history of the checkpoint being diagnosed.

| Name | Definition |
|---|---|
| Base | Released Qwen3-VL-8B-Instruct, without project-specific SFT or GRPO. |
| Visual-only (SFT) / Visual-only (SFT+GRPO) | Visual-only LoRA SFT / subsequent all-task GRPO. |
| QUASAR (SFT) / QUASAR (SFT+GRPO) | Original four-token model after SFT / subsequent all-task GRPO. |
| QUASAR (SFT+GRPO, numeric-only) | SFT followed by GRPO restricted to numeric questions; this restriction applies only to GRPO. |
| w/o visual writeback | The stated QUASAR checkpoint with its spatial residual zeroed at inference; not a model trained without geometry. External tests use QUASAR (SFT+GRPO). |
| fixed image-count rule | Within the stated QUASAR (SFT+GRPO) checkpoint, reuse w/o visual writeback predictions for one image and normal predictions for multiple images. |

SFT uses 426,916 spatial instruction examples, LoRA rank 16, scaling 32, and dropout 0.05. The backbone is frozen while LoRA and QUASAR's added modules are trained. The full SFT schedule is one epoch with learning rate $3 \times 10^{-5}$, microbatch 1, and gradient accumulation 8. Training geometry is precomputed from CUT3R (Wang et al., 2025c) and depth/camera-pose sources; auxiliary-target records and source-unit boundaries are described in Appendix Q. The full question enters the MLLM, while the spatial module receives the geometric descriptors and explicit reference condition.

Table 27 records both the displayed SFT checkpoints and the actual parents of the GRPO models. QUASAR and Visual-only each complete 2,048 GRPO updates over 8,192 question groups, with four sampled answers per group. They use the same all-task reward recipe: Mean Relative Accuracy (MRA) for numeric answers and accuracy for multiple-choice answers. GRPO updates LoRA only, with auxiliary losses disabled; the spatial forward path remains active. The learning rate is $10^{-6}$, gradient accumulation is 4, and the KL coefficient is 0.02. The reference policy disables LoRA in the same model.

The QUASAR residual-diagnostic endpoint also supplies the norm-budget sweep, architecture variant B, and MindCube-Tiny evaluation. The full Visual-only SFT endpoint supplies the architecture comparison and external spatial/CV-Bench tests. QUASAR SFT+GRPO reference, gate, external, and qualitative conditions use the stated all-task endpoint; numeric-only GRPO shares its QUASAR SFT parent. The plain InternVL3-8B transfer evaluation reports a single SFT endpoint with no paired comparator.

Table 27: **Training lineage and endpoint roles.** The selected main-result SFT rows are not the GRPO parents. Main-result checkpoints also supply the SFT appearance-order, system-difference, and timing comparisons. Full SFT endpoints and the numeric-only GRPO variant are listed separately.

| Model / role | Displayed SFT step | GRPO parent SFT step | GRPO updates |
|---|---|---|---|
| Visual-only | 46,592 | 53,365 | 2,048 |
| QUASAR | 44,544 | 35,840 | 2,048 |
| QUASAR residual diagnostic | 53,365 | – | – |
| Visual-only, full SFT endpoint | 53,365 | – | – |
| QUASAR (SFT, variant A) | 53,248 | – | – |
| QUASAR (SFT+GRPO, numeric-only) | – | 35,840 | 2,048 |

These comparisons share the backbone, training-data recipe, and GRPO update scale, but not the full SFT exposure before GRPO. Accordingly, the SFT+GRPO system difference measures the difference between the two trained systems, not a strictly equal-budget architectural effect. Nor does SFT+GRPO minus displayed SFT isolate a GRPO effect from that displayed checkpoint. Table 29 also reports endpoint differences, with each comparator kept explicit.

**Scoring and evaluation inputs.** For a positive numerical target, MRA averages correctness over ten relative-error tolerances from 5% to 50% in steps of 5%; tables report the result on a 0–100 scale. Table 28 summarizes the question sets and scoring rules.

Table 28: **Question sets and aggregation rules.** Macro-averages weight task components equally, not individual examples. A dataset size alone does not identify its score. Restricted numerical-pose sets are subsets, not additional full benchmarks.

| Evaluation | Questions | Statistic |
|---|---|---|
| VSI main | 5130 | Eight tasks: four numeric MRA, four multiple-choice accuracy. Direction first averages easy/medium/hard subtypes. |
| VSTI historical main | 6042 | Five tasks: two numeric MRA and three multiple-choice tasks; relational subtypes pool within their main task. |
| VSTI historical diagnostics | 6042 | Nine fine question types, averaged equally; used for residual, routing, and corresponding groups. |
| VSTI revised transfer | 5736 | Revised set with nine-subtype aggregation; used for the reported InternVL results. |
| VSTI numerical-pose test | 5598 | Parsable target-frame subset of the historical set; five-task macro. |
| Revised numerical-pose intersection | 5292 | Intersection with the revised set; nine-subtype macro, not pooled with 5,598. |
| External spatial tests | By benchmark | Shared evaluator: numeric MRA, count exact match, choices accuracy; not cross-paper official-score comparisons. |
| CV-Bench / MindCube-Tiny | 2638 / 1050 | Question-level accuracy, with the stated subset counts. |

Every local main-table total and task column is obtained from the same prediction set and scorer. The historical VSI Base score is 59.8247049665833%, displayed as 59.82, with eight task scores of 69.24, 52.03, 75.87, 60.38, 61.83, 52.61, 33.51, and 73.14 in main-table order. Its 5,130 IDs, answer semantics, and scoring rules agree with the SFT+GRPO model evaluation, but Base reads raw videos whereas the SFT+GRPO models read cached frames. Option lists differ for 671 questions: 217 easy-direction, 378 medium-direction, and 76 route-planning questions. QUASAR (SFT+GRPO) exceeds this historical Base by $+5.54$ points. This is a comparison across input settings, not an input-matched method or training gain.

QUASAR (SFT+GRPO, numeric-only) shares the all-task GRPO model's SFT parent in Table 27. Historical VSI Base specifies raw-video input and a 64-token generation limit, but its resolved

model revision, actual sampled frames/resolution, and complete decoding configuration remain undocumented. Saved predictions cannot isolate the effects of its input and option-list differences.

Our SFT and SFT+GRPO model evaluations use at most 32 frames and 50,176 pixels per frame, with deterministic decoding. QUASAR allows 16 generated tokens; Visual-only (SFT) allows 24, and Visual-only (SFT+GRPO) allows 24 on VSI and 20 on VSTI. The longest SFT+GRPO model outputs are 13 and 4 tokens respectively, so the SFT+GRPO comparison shows no output truncation at these limits. Reported confidence intervals (CIs) concern fixed checkpoints: main-result intervals use 5,000 paired question-ID bootstrap resamples, while intervention captions specify their own sampling units. They do not estimate training-seed variability.

**Reproducibility scope.** The results retain their historical model and input identities. The mapping to a separate server reproduction, original run-specific source snapshots, selection criteria for the displayed SFT checkpoints, and equivalence of visual caches across training stages remain unverified. Identical saved VSI inputs for the two SFT+GRPO models do not establish these other conditions. The saved target layout and coverage do not fully recover the historical target-generation and loss code. Exact scale construction, loss reductions, KL calculations, and branch-mixture target rules remain partly unverified, as do source-specific geometry masks, object truncation, and coordinate-unit conversions. QUASAR uses estimated geometry and supplied references without external spatial memory or test-time parameter updates; cross-paper scores do not isolate this interface without matched resources.

## S. Paired Bootstrap Comparisons

Paired intervals resample fixed predictions after matching question IDs. They quantify evaluation-sample uncertainty for the stated checkpoints, not training-seed variability. For a macro score, each resample recomputes the relevant aggregate rather than averaging rounded task columns. Table 29 uses the VSI eight-task and historical VSTI five-task protocols throughout.

Table 29: **Main paired differences: 5,000 question-ID resamples, 95% CIs.** Positive differences favor the first named model. The numeric-only control restricts GRPO, not SFT, to numeric questions. SFT+GRPO minus displayed SFT compares different endpoints, not continued training from that displayed SFT checkpoint.

| Comparison | VSI Δ [95% CI] | VSTI five-task Δ [95% CI] |
|---|---|---|
| QUASAR (SFT) − Visual-only (SFT) | +2.06 [+0.77, +3.34] | +3.26 [+2.14, +4.37] |
| QUASAR (SFT+GRPO) − Visual-only (SFT+GRPO) | +1.75 [+0.56, +2.90] | +3.93 [+2.83, +4.99] |
| QUASAR (SFT+GRPO) − displayed QUASAR (SFT) | +0.51 [−0.50, +1.51] | +1.07 [+0.07, +2.03] |
| QUASAR (SFT+GRPO, numeric-only) − displayed QUASAR (SFT) | +0.44 [−0.56, +1.43] | +1.04 [+0.05, +2.05] |
| Visual-only (SFT+GRPO) − displayed Visual-only (SFT) | +0.82 [−0.22, +1.86] | +0.40 [−0.60, +1.39] |

The selected SFT and SFT+GRPO system comparisons both have positive intervals excluding zero. These support checkpoint differences under the shared backbone and data recipe, not strict equal-budget effects. The SFT+GRPO-versus-displayed-SFT VSTI intervals also exclude zero, but lineage prevents interpreting them as pure GRPO increments from the displayed checkpoint. The VSI and Visual-only endpoint intervals cross zero and are inconclusive; they neither establish equivalence nor show that GRPO cannot help.

Residual group intervals use a different design: 5,000 paired resamples within fixed raw task strata, followed by the nested task macro. Reference type/path permutation and uniform-branch-mixing intervals use 2,000 paired question resamples against their matched normal. Numerical-pose intervals use 5,000 paired scene-cluster resamples; the fixed-512 analysis in Table 15 instead uses 10,000 paired base-scene resamples and question-weighted accuracy. These sampling units are not interchangeable, and perturbation seeds on one checkpoint are not independent training seeds.

**Question-group analysis.** We group tasks by the spatial relations they test to examine where the SFT+GRPO model differences occur. These are analysis groups, not labels of the actual reference-

type input: the "Object relations" group is not evidence that a valid object-anchored pose was supplied. Table 30 compares QUASAR (SFT+GRPO) with Visual-only (SFT+GRPO), keeping the comparator stage explicit.

Table 30: **Question-group results, SFT+GRPO systems.** VSI groups average their constituent task scores. VSTI Distance and displacement averages two numeric subtypes; Motion and relations averages seven multiple-choice subtypes, not three pooled main tasks. These VSTI groups belong to the historical nine-subtype diagnostic convention. Counts describe the underlying questions, not aggregation weights. The table reports descriptive group differences; the appearance-order comparison is examined further in Appendix E.

| Benchmark | Group | $n$ | Visual-only | QUASAR | $\Delta$ |
|---|---|---|---|---|---|
| VSI | Object relations: relative direction/distance | 1678 | 71.60 | 72.92 | +1.33 |
| VSI | Distance and size: absolute distance, object/room size | 2075 | 62.94 | 66.79 | +3.85 |
| VSI | Counting and routes: counting and route planning | 759 | 58.61 | 58.41 | −0.19 |
| VSI | Appearance order | 618 | 59.71 | 59.87 | +0.16 |
| VSTI | Distance and displacement: CO-AbsD and CDisp | 1744 | 40.78 | 42.95 | +2.17 |
| VSTI | Motion and relations: seven relation subtypes | 4298 | 78.28 | 81.35 | +3.07 |

The VSI distance-and-size and object-relations groups have positive differences, while counting and routes is slightly negative. VSTI improves in both displayed groups. This pattern is consistent with the use of reference-conditioned geometric evidence, but task grouping alone cannot identify the mechanism. The near-zero appearance-order difference in the SFT+GRPO comparison also differs from the $-4.05$ SFT difference. Temporal-order behavior therefore depends on the evaluated checkpoint; a deficit is neither proof of a designed capability boundary nor evidence that a particular memory-based alternative would perform better under matched conditions.

## T. MindCube-Tiny Cross-Domain Evaluation

MindCube-Tiny (Wang et al., 2026) contains egocentric around-relations, rotated viewpoints, and object–object relations. The evaluated original four-token QUASAR SFT endpoint uses CUT3R geometric descriptors and receives no MindCube-specific training, external memory, or test-time parameter updates. The result measures performance on these question formats, not a matched estimate of the contribution of geometry.

Table 31: **MindCube-Tiny absolute accuracy (SFT).** The model answers 413 of 1,050 questions correctly. Overall accuracy weights questions, not the three groups equally. No Base or Visual-only difference is inferred from this evaluation.

| Model | All ($n = 1050$) | Around ($n = 250$) | Rotation ($n = 200$) | Among ($n = 600$) |
|---|---|---|---|---|
| QUASAR (SFT) | 39.33 | 54.40 | 30.50 | 36.00 |
| Correct answers | 413 | 136 | 61 | 216 |

In Table 31, around-relations have the highest observed accuracy, while rotation remains harder. Without an input-matched control in this evaluation, the absolute scores do not isolate transfer from SFT, geometric input, or decoder-side supervision. They complement the mixed external-benchmark results rather than establishing a general zero-shot advantage.

## U. Qualitative Case Studies

The selected cases in Appendix B illustrate differences in spatial answers under the references recorded below.

Each model run uses all 32 video frames at $320 \times 224$ pixels. The apartment, kitchen, and office cases use rendered videos with $640 \times 480$ source frames; the camera-movement and distance cases use original ScanNet videos. Frame numbers in Figure 4 index the 32-frame input, not the original recording. For the apartment, kitchen, and office, scene illustrations use copies of the original scanned-room meshes, preserving their geometry and scan holes. Object markers use recorded annotation bounds; reference markers project recorded camera positions onto the floor, with purple arrows indicating horizontal viewing directions. Camera axes determine left/right independently of the illustration's page orientation. Illustrations, boxes, and object numbers never enter the models. The office whiteboard appears before its frame-32 reference view.

In Figure 4, Qwen3 denotes the original Qwen3-VL-8B-Instruct model, and Qwen3-FT denotes Visual-only, using its separately trained SFT+GRPO checkpoint (Table 27). The label *w/o SLT* denotes the condition w/o visual writeback: QUASAR retains its SFT+GRPO parameters and four spatial summary tokens, but zeros the main visual writeback and all three DeepStack writebacks. Qwen3-VL-8B and Visual-only receive the video and question without geometry, poses, or writeback features.

For the rendered-video cases, the Qwen3-VL-8B and Visual-only prompts omit the spatial-summary-token line but retain the remaining original text and image order. That text includes "Use the video frames and the injected 3D spatial evidence to answer the question." despite their video-only inputs. These runs use greedy decoding with a 16-token output limit.

**Selection scope.** We select the apartment, office, camera-movement, and distance examples in Figure 4 after observing their outcomes. Only QUASAR answers these four questions correctly among the four tested variants. The rendered cases combine original QUASAR runs with supplementary Qwen3-VL-8B and Visual-only runs. The two ScanNet cases use separate four-method runs on fixed benchmark questions and are not the main-table prediction records. These outcome-selected examples illustrate individual answer differences and complement the aggregate evaluations.

**Rendered-video questions and references.** All three questions offer A. left and B. right, with A. left as the recorded correct answer.

**Apartment: reference frame 18.** Using the viewing direction in frame 18 of this video as your reference, is the large dark-blue sofa to the left or to the right of the white refrigerator in the kitchen? Frames 5, 9, 13, and 18 connect the sofa and kitchen to the curtain-facing reference at frame 18.

**Office: reference frame 32.** Using the viewing direction in frame 32 of this video as your reference, is the wooden table with the black bag to the left or to the right of the wall-mounted whiteboard? Frames 5, 11, 20, and 32 connect the table and wall-mounted whiteboard to the window-side reference at frame 32.

**Kitchen supplement: reference frame 32.** Using the viewing direction in frame 32 of this video as your reference, is the microwave to the left or to the right of the black water dispenser? Frames 5, 12, 22, and 32 connect the microwave and black water dispenser to the wall-facing reference at frame 32. Qwen3-VL-8B and QUASAR both answer correctly.

**Camera movement: original question and reference.** VSTI question 4268 uses the original ScanNet video `scene0354_00` (Fan et al., 2026). Input slots 8–11 correspond to zero-based video frames 254, 290, 327, and 363 at 10.583, 12.083, 13.625, and 15.125 seconds. The controlled reruns use the orientation at slot 8 as the reference, replacing the historical endpoint reference at slot 11. QUASAR estimates its reference pose and reference-dependent descriptors from RGB. Recorded camera poses support verification and the route illustration only; the models receive no ground-truth poses. The question concerns sustained movement relative to the starting orientation, independently of subsequent turning.

**Original question.** During the sequence between frame 8 and frame 11 of 32, what was the primary consistent direction of the camera's movement relative to its orientation at the start? The options are Left, Right, Backward, and Forward.

**Options.** A. Left; B. Right; C. Backward; D. Forward. The official answer is D. Forward. Figure 4(c) shows input slots 8–11 in chronological order.

**Object distance: original question and input frames.** VSI question 2983 uses the original ScanNet video `scene0426_00` (Yang et al., 2024). The source video contains 2,301 frames at 24 fps and $640 \times 480$ pixels; each model receives the same 32 sampled $320 \times 224$ inputs. QUASAR uses a video-derived scene-canonical reference for this question, without a reference camera frame. The TV stands on a drawer cabinet; the desk monitors are separate objects. Figure 8 shows all candidate categories and three chair views in chronological order. The purple TV box marks an object, whereas purple borders in Figure 4(a–c) mark reference views.

**Original question.** Measuring from the closest point of each object, which of these objects (window, chair, table, lamp) is the closest to the tv?

**Options.** A. window; B. chair; C. table; D. lamp. The official answer is B. chair.

**Closest-point measurement and instance scope.** The distance measure uses the closest points between the official oriented three-dimensional annotation boxes. For each candidate category, it takes the minimum over all instances, including all three chair instances. This measures box-to-box distance, rather than object-center distance or a gap between image pixels. The display boxes locate visible objects only; they do not establish a correspondence to the three-dimensional chair instance that attains the minimum. Scene-canonical reference construction is independent of the official annotation boxes used to verify the answer. Display annotations never enter the model inputs.

**Recorded answers.** Each cell gives the identical answer from both recorded runs of that model.

| Case | Qwen3-VL-8B | Visual-only | QUASAR w/o visual writeback | QUASAR |
|---|---|---|---|---|
| Apartment | B. right | B. right | B. right | A. left |
| Office | B. right | B. right | B. right | A. left |
| Kitchen | A. left | B. right | B. right | A. left |
| Camera movement | C. Backward | A. Left | A. Left | D. Forward |
| Object distance | D. lamp | D. lamp | D. lamp | B. chair |

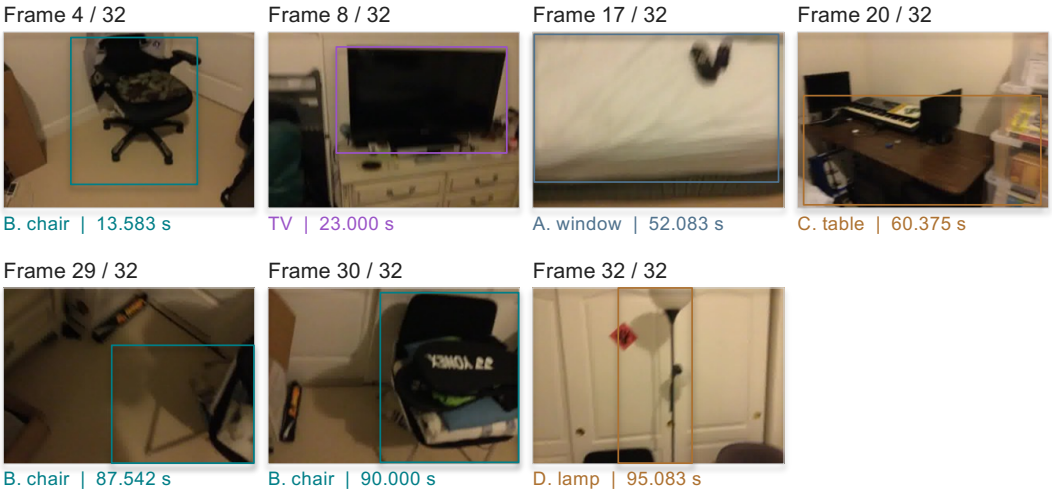



Figure 8: **Object distance: seven selected inputs preserve all candidates and three chair views.** Frame numbers index the 32-frame model input. Colored object boxes and labels aid reading only. Each panel preserves the input aspect ratio; the models receive unannotated frames.