EGOAUDIT: AUDITING EGOCENTRIC SPATIAL REASONING IN 3D VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D vision-language models (3D VLMs) have made rapid progress in grounding objects and answering questions in 3D indoor scenes. Many queries that people pose are egocentric: "Which lamp is on my left?" has a definite answer only once the model knows where the speaker stands and which way they face. Existing benchmarks and models often provide this viewpoint as an annotated pose, so a high score may reflect how well a model *uses* a given viewpoint rather than whether it can *infer* the viewpoint from the description. We present EGOAUDIT, a framework for auditing egocentric spatial reasoning in 3D VLMs. EGOAUDIT equips a 3D VLM with a viewpoint-conditioned object encoding, through which the viewpoint can be removed, perturbed, or replaced while the scene, the query, and the model stay fixed, and it evaluates the model with three protocols: viewpoint intervention, matched training with and without the viewpoint, and substitution of the viewpoint source. Extensive experiments on public 3D grounding and 3D visual question answering benchmarks show that the audited model, a strong 3D VLM, follows a supplied viewpoint closely but does not infer it: much of its advantage on egocentric queries is lost when the annotated viewpoint is withheld or replaced by the estimate of a learned localizer. These results indicate that scores obtained with annotated viewpoints measure viewpoint use rather than perspective taking, and we propose a reporting protocol for egocentric 3D evaluation.

1 INTRODUCTION

3D vision-language models (VLMs) take a reconstructed indoor scene and a text query and solve tasks such as 3D visual grounding, dense captioning, and 3D visual question answering (VQA) within one model (Hong et al., 2023; Huang et al., 2024b;a; 2026). Many queries that people pose to such a model are egocentric. A user tells a home robot "I am sitting on the bed, facing the window. Which lamp is on my left?" The scene is stored in world coordinates, but "left" is defined by the user: if the user turns around, the answer changes although no object has moved. Answering therefore takes three steps (Figure 1, top): locate the user's *viewpoint*, express the scene relative to this viewpoint and ground or answer. Situated 3D VQA benchmarks are built from such queries (Ma et al., 2023; Linghu et al., 2024; Zhang et al., 2025a).

Benchmarks and models differ in who performs the first two steps. SQA3D describes the user only in text, so the model has to infer the viewpoint (Ma et al., 2023). MSQA provides the location and orientation, and MSR3D rotates the scene into that frame before the language model reads it (Linghu et al., 2024); LEO receives the annotated situation pose (Huang et al., 2024b); SIG3D and Loc3R-VLM estimate the pose with a dedicated module (Man et al., 2024; Qu et al., 2026); and pose-conditioned models, including the one studied here, can be handed the annotated pose. Once the annotated viewpoint is given, locating the user happens outside the model, and the score measures how well the model *uses* a viewpoint, not whether it can *infer* one. The difference is large in practice: on SQA3D, the same architecture reaches 58.7% accuracy when trained and tested without a viewpoint and 64.2% when trained and tested with the annotated one. An aggregate score does not reveal which of the two abilities it measures.

We present EGOAUDIT (egocentric audit), a framework for auditing egocentric spatial reasoning in 3D VLMs that separates the use of a given viewpoint from its inference. EGOAUDIT equips an object-centric 3D VLM with a *viewpoint-conditioned object encoding*: each object's distance

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.

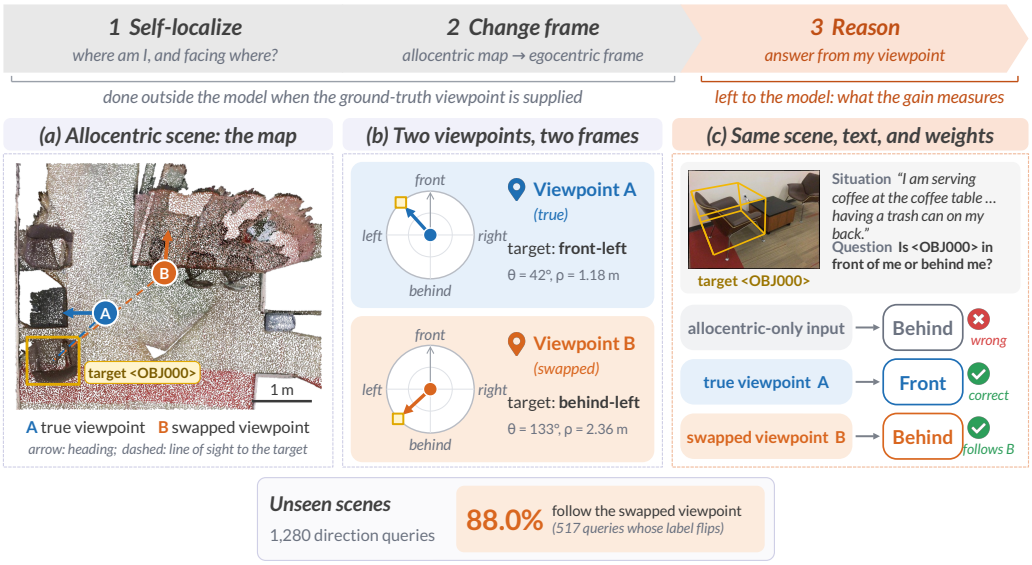


Figure 1: **An egocentric query depends on the viewpoint.** Top: answering takes three steps, and supplying the annotated viewpoint performs steps 1 and 2 outside the model. (a) A 3D scene with the true viewpoint A and a swapped viewpoint B (an illustrative case from a training scene; Appendix E). (b) The same target lies front-left of A and behind-left of B. (c) With the scene, text, and weights fixed, the model answers from whichever viewpoint it receives. Bottom: on 1,280 direction questions from unseen scenes, 88.0% of the answers follow the swapped viewpoint on the 517 questions whose correct answer it changes.

and direction relative to the viewpoint are computed outside the model and fused into the object's 3D token (Figure 2). Because the viewpoint enters only through this encoding, it can be removed, perturbed, or replaced while the scene, query, prompt, token layout, and, where required, the weights stay fixed. On top of this encoding, EGOAUDIT evaluates the model with three protocols. **Viewpoint intervention** perturbs or swaps the viewpoint of a fixed model and measures whether the answers follow it. **Matched training** trains identical models with and without the viewpoint and measures whether its benefit depends on training with it. **Viewpoint-source evaluation** replaces the annotated viewpoint with no viewpoint or with the estimate of a learned localizer and measures how much of the benefit remains. We instantiate EGOAUDIT on a 3D VLM built on the Chat-Scene design (Huang et al., 2024a) with a Qwen3-VL-8B backbone (Bai et al., 2025), which achieves strong results on standard 3D grounding and captioning benchmarks, and evaluate it on public 3D grounding and situated 3D VQA benchmarks.

The audited model uses a given viewpoint but does not infer one. Given a viewpoint, it answers from it: on unseen scenes, most of its answers follow a swapped viewpoint, and a reversed heading hurts more than removing the viewpoint altogether. Without the annotated viewpoint, much of the benefit is lost. A matched model trained without the viewpoint falls behind only on the queries that depend on it, so training does not teach it to recover the viewpoint from the description, and a localizer trained with pose supervision recovers only part of the gain.

Our contributions are: (i) EGOAUDIT, a framework that attributes the egocentric behavior of a 3D VLM to the viewpoint it receives by controlling the viewpoint's value, its presence in training, and its source within one model (Section 3); (ii) a comprehensive evaluation on public 3D grounding and 3D VQA benchmarks showing that the audited model uses a given viewpoint but does not infer it, with consistent training effects across datasets (Section 5); and (iii) a reporting protocol for egocentric 3D evaluation: state the viewpoint source (none, predicted, or annotated), any transformation applied to the scene before the model, and whether the model was trained with the viewpoint, and report the predicted-viewpoint score with the annotated-viewpoint score as an upper reference (Section 6).

2 Related Work

3D vision-language models. ScanRefer, Multi3DRefer, Scan2Cap, and ScanQA ground language in reconstructed scenes (Chen et al., 2020; Zhang et al., 2023; Chen et al., 2021; Azuma et al., 2022); 3D-LLM, LEO, Chat-Scene, Grounded 3D-LLM, and GPT4Scene connect such scenes to language models (Hong et al., 2023; Huang et al., 2024b;a; Chen et al., 2024c; Qi et al., 2026). Situated QA makes the questioner's viewpoint explicit: SQA3D asks a model to interpret a verbal situation (Ma et al., 2023); MSR3D studies how to encode the situation, including transforming scene coordinates (Linghu et al., 2024); Spartun3D adds situated alignment objectives (Zhang et al., 2025a); and RoboSpatial distinguishes ego-, world-, and object-centered relations and finds 3D VLMs trailing on ego-centric relations even after fine-tuning (Song et al., 2025). Navigation policies such as Robostral Navigate act in image space, where the viewpoint is the camera (Mistral AI, 2026); reasoning over a scene map instead has to locate the viewpoint in scene coordinates.

Spatial Reasoning in VLMs. Spatial Reasoning benchmarks attribute many errors of multimodal models to perspective taking and egocentric-allocentric transformation (Yang et al., 2025; Li et al., 2026), and VLMs resolve ambiguous frames of reference inconsistently (Zhang et al., 2025b) and default to egocentric interpretations when another viewpoint is required (Lee et al., 2025). In these image- and video-based studies the egocentric frame is the camera's and the hard case is adopting another entity's; in map-based 3D reasoning the scene is allocentric and the target viewpoint belongs to a person described in text, but both require the same change of frame. The benchmarks measure the failure at the output, and Lee et al. (2025) transform the scene into the target viewer's frame before answering; EgoAudit instead intervenes on the viewpoint to test what the model does with it. SIG3D estimates a language-grounded position and orientation before situated QA and finds that a ground-truth situation helps only when re-encoded into scene tokens, not when fed directly (Man et al., 2024); Loc3R-VLM jointly trains layout reconstruction, situation estimation, and language modeling (Qu et al., 2026); direction-aware 3D LMMs recover ego poses for directional queries and align the point cloud to them (Liu et al., 2026); PointAlign regularizes intermediate point-cloud tokens to preserve geometry (Su et al., 2026), and SpatialReasoner reasons through explicit 3D representations (Ma et al., 2025). Real-3DQA filters language shortcuts and rotates viewpoints (Ma et al., 2026), SCOPE tests egocentric-allocentric consistency (Kim et al., 2026), and SURPRISE3D suppresses object-name shortcuts (Huang et al., 2025). The localization works show that viewpoint information helps and propose ways to obtain and encode it; EgoAudit asks what the resulting gains measure, by intervening on the viewpoint at fixed weights, matching training, and varying the viewpoint source within one architecture.

3 EgoAudit

EgoAudit has two components: a 3D VLM whose object tokens are conditioned on an explicit, controllable viewpoint (Section 3.2), and three evaluation protocols that vary this viewpoint while holding everything else fixed (Section 3.3). We first formulate the task.

3.1 Problem Formulation

A scene is a set of object proposals $S = \{(v_i, x_i, s_i)\}_{i=1}^{N}$ with visual features v_i, center $x_i \in \mathbb{R}^3$, and size $s_i \in \mathbb{R}^3$ in gravity-aligned world coordinates. A situation description d ("I am sitting on the bed, facing the window") specifies a viewpoint $p = (t, \psi)$ with position t and heading ψ, and a query q asks the model to ground an object or answer a question from this viewpoint. The three steps of Figure 1 read

$$\widehat{p} = L(S, V, d), \qquad g_i = g(x_i, s_i; \widehat{p}), \qquad \widehat{y} = F_\theta(S, \{g_i\}, d, q), \tag{1}$$

where L *infers* the viewpoint from the scene, the description, and any available observations V; g expresses each object relative to the viewpoint; and the 3D VLM F_θ *uses* the result to produce the answer $\widehat{y}$. With heading zero along world $+x$, the viewpoint-relative position and bearing of object i are

$$r_i(p) = R_z(-\psi)\,(x_i - t), \qquad b_i(p) = \operatorname{atan2}(r_{i,y}, r_{i,x}), \tag{2}$$

so that the components of r_i point forward, left, and up. These quantities, not the world coordinates, decide answers such as "left" or "behind". Supplying the annotated p replaces L by the annotation,

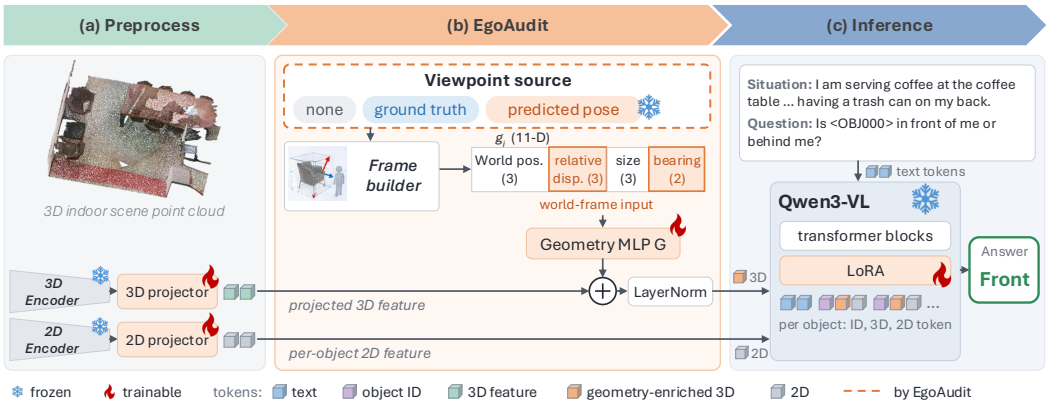


Figure 2: **Overview of EGOAUDIT.** (a) Object proposals of a 3D scene provide per-object 3D and 2D features. (b) Viewpoint-conditioned object encoding: a viewpoint from the chosen source (none, GT, or predicted) is converted outside the model into viewpoint-relative geometry g_i, whose viewpoint-dependent entries (orange) are zero without a viewpoint; a trainable MLP fuses g_i into the object's projected 3D feature. (c) The Qwen3-VL-8B language model answers from the object tokens and the text. Viewpoint intervention alters the viewpoint of a fixed model; matched training and viewpoint-source evaluation compare models trained with different viewpoint inputs.

and the evaluation then measures F_θ alone. The camera poses used to reconstruct the scene do not solve this step: registering images into a map does not locate the person named in the query.

3.2 VIEWPOINT-CONDITIONED 3D VLM

Object-centric scene tokens. Following Chat-Scene (Huang et al., 2024a), each scene is represented by at most 100 object proposals (Figure 2). Each proposal contributes three tokens to the language model: an identifier token, which lets the model refer to the object in its output, a 3D token obtained by projecting a 128-dimensional CLASP point-cloud feature (Chen et al., 2024c), and a 2D token obtained by projecting a 1024-dimensional multi-view DINOv2 feature (Oquab et al., 2024). The language model is Qwen3-VL-8B (Bai et al., 2025); its own vision tower does not process raw images in this pathway, so the 3D and 2D object features are its only visual evidence.

Viewpoint-conditioned geometry. The viewpoint enters only through an 11-dimensional geometry vector per object,

$$g_i(p) = \big[\, x_i,\ \ \rho_i \cos\hat{\theta}_i,\ \ \rho_i \sin\hat{\theta}_i,\ \ \mathrm{clip}(\Delta z_i, -5, 5),\ \ \ s_i,\ \ \sin\hat{\theta}_i,\ \ \cos\hat{\theta}_i \,\big], \tag{3}$$

which concatenates the world position x_i, the viewpoint-relative displacement, the size s_i, and the bearing. Here ρ_i is the horizontal distance from the viewpoint, Δz_i the relative height, and $\hat{\theta}_i$ the center of the 45° bin, out of eight, that contains the bearing $b_i(p)$. The quantization matches the coarse directions used in egocentric language (front, front-left, left, and so on) while keeping the distance. A two-layer MLP G ($11 \to 256 \to 4096$) maps the geometry vector into the embedding space and adds it to the projected 3D feature,

$$h_i = \mathrm{LN}\left(P_{\mathrm{3D}}(v_i^{\mathrm{3D}}) + G(g_i(p))\right), \tag{4}$$

while the identifier and 2D tokens keep their paths. The viewpoint therefore adds no token and leaves the sequence layout unchanged, which is what allows the viewpoint to be manipulated in isolation.

The encoding accepts three viewpoint inputs of identical shape. Without a viewpoint, the five viewpoint-dependent entries are zero and only the world position and size remain. The *GT viewpoint* uses the annotated pose. The *predicted viewpoint* uses a frozen Loc3R localizer (Qu et al., 2026), which estimates the planar position and heading from ordered ScanNet RGB frames, CUT3R camera tokens (Wang et al., 2025), and the situation text, but not the answer; its estimate is registered into scene coordinates with the known camera pose of the first frame, and the height is fixed at

1.5 m. Loc3R is trained with pose supervision, including SQA3D training data, so it represents a favorable case of learned localization. In every case Equation 2 is evaluated outside the model, so a supplied viewpoint arrives already converted into object directions; Section 5.2 examines a variant that receives the raw pose instead.

3.3 Evaluation Protocols

Viewpoint intervention. We hold the weights of a model trained with GT viewpoints fixed and change only the viewpoint it receives. *Corruptions* replace the annotated viewpoint with a position shifted by 1 m, a heading rotated by $\pm 90^\circ$ or 180°, the scene centroid, or a random pose inside the scene. *Permutation* gives each situation the viewpoint of another situation in the same scene, a bijection without fixed points; it keeps the set of viewpoints of the scene but breaks their match with the descriptions. All interventions act before the geometry MLP and recompute Equation 3 from the altered pose, so object identities, token layout, and tensor shapes are unchanged, and each is scored by the change in accuracy. An accuracy drop shows that the model depends on the viewpoint, but not that it answers from it. We therefore also measure *viewpoint following* on binary direction questions such as "Is ⟨OBJ000⟩ in front of me or behind me?", which name the target object explicitly and come from 40 validation scenes absent from all training annotations. Their labels are computed from GT geometry under the original viewpoint p and under the viewpoint p' of another situation in the same scene. On the set $\mathcal{D}$ of questions whose label differs between p and p',

$$\text{Follow} = \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \mathbb{I}\big[F_\theta\big(S, \{g(x_i, s_i; p')\}, d, q_j\big) = y_j(p')\big]\,. \tag{5}$$

The description d stays unchanged, so the probe sets the text against the supplied viewpoint and reveals which of the two determines the answer.

Matched training. We train two models with identical data, schedule, checkpoint selection, and decoding, one without a viewpoint and one with GT viewpoints, and evaluate each with both inputs. Let A_{ab} be the accuracy of the model trained with input a and evaluated with input b, where 0 denotes no viewpoint and 1 the GT viewpoint. The interaction

$$I = (A_{11} - A_{10}) - (A_{01} - A_{00}) \tag{6}$$

measures how much more the GT viewpoint helps the model trained with it. The viewpoint-dependent entries are zero throughout training without a viewpoint, so the weights that read them receive no gradient and injecting the viewpoint into this model should have little effect. The protocol therefore verifies that the gain of the GT model comes from learned use of the viewpoint rather than from the extra input, and it shows whether the encoding can be attached to a model trained without it.

Viewpoint-source evaluation. A deployed system has to obtain the viewpoint itself. We compare three matched models that differ only in the viewpoint source used in both training and evaluation: none, predicted, or GT. The model trained without a viewpoint sees only the description and the scene and thus has an incentive to infer the viewpoint; its deficit relative to the GT model measures what the 3D VLM fails to infer by itself. To locate this deficit, we split the questions into viewpoint-dependent ones, whose interpretation requires the user's position or heading, and viewpoint-independent ones, using post-hoc human-reviewed labels (Appendix D.5). The fraction of the GT gain that the predicted viewpoint recovers is

$$\text{Recovery} = \frac{A_{\text{pred}} - A_{\text{none}}}{A_{\text{GT}} - A_{\text{none}}}, \tag{7}$$

and replacing either the predicted position or the predicted heading by its GT value in the fixed predicted-viewpoint model locates the remaining error.

4 Experimental Setup

Datasets and metrics. We report 3D visual grounding on ScanRefer (accuracy at IoU 0.5) and Multi3DRefer (F1 at IoU 0.5) and dense captioning on Scan2Cap (CIDEr at IoU 0.5) (Chen et al., 2020; Zhang et al., 2023; Chen et al., 2021), all on the official validation splits. Egocentric reasoning

is evaluated on the situated 3D VQA benchmark SQA3D (Ma et al., 2023), using all 3,261 validation questions and its standard metric, exact-match accuracy (EM): the percentage of predicted answers that exactly match a reference answer. Viewpoint following uses 1,280 binary direction questions from 40 held-out validation scenes, of which 517 change label under the swapped viewpoint. Post-hoc human-reviewed labels divide the SQA3D questions into 2,892 viewpoint-dependent and 308 viewpoint-independent ones; the remaining 61, 37 unresolved and 24 used to pilot the labeling rubric, are excluded (Appendix D.5). For transfer, the models are evaluated without further training on the ScanNet, 3RScan, and ARKitScenes domains of MSQA (Linghu et al., 2024) and on Spartun3D (Zhang et al., 2025a) (Appendix F).

Models. The comparison with published methods (Table 1), viewpoint following, and the layer probes use our 3D VLM trained with GT viewpoints. The remaining SQA3D experiments, the transfer experiments, and the visual results use three matched models, M_{none}, M_{pred}, and M_{GT}, which share data, schedule, checkpoint selection, and decoding and differ only in the viewpoint source used in training and evaluation (Appendix B).

Implementation details. The Qwen3-VL-8B language model is frozen. LoRA adapters (Hu et al., 2022) (rank 16, scale 16, dropout 0.05), the input and output token embeddings, the projectors, and the geometry MLP are trained jointly on 3D grounding, captioning, and QA data with a global batch size of 32 and learning rates of 1.6×10^{-4} (3.2×10^{-3} for the geometry modules). Decoding uses beam search with five beams, a repetition penalty of 3.0, and at most 64 new tokens. Main-text results come from one training run per model; replications, a Vicuna-7B model line with a different geometry encoding, and statistical analyses are in Appendix D.

5 Results

5.1 Main Results

A strong 3D VLM. Table 1 places our model among published 3D models: it obtains the best grounding and captioning scores in the table, so the audit concerns a competitive 3D VLM rather than a weak baseline.

The GT viewpoint helps where the query depends on it. Supplying the GT viewpoint in training and evaluation raises SQA3D EM from 58.69 to 64.21 (Table 2). The gain lies on the queries that need the viewpoint: M_{none} scores 6.1 points below M_{GT} on the 2,892 viewpoint-dependent questions (58.78 vs. 64.87), whereas the two models are on par on the 308 viewpoint-independent ones (61.36 vs. 61.04). M_{none} sees only the description and the scene, so training without the viewpoint does not teach the model to infer it.

Table 2: **Viewpoint source.** SQA3D EM (%) of matched models trained and evaluated with each viewpoint source; Δ is the gain over no viewpoint.

Viewpoint source	EM	Δ
None (M_{none})	58.69	–
Predicted (M_{pred})	60.44	+1.75
GT (M_{GT})	**64.21**	+5.52

A learned localizer recovers only part of the gain. With viewpoints predicted by Loc3R in training and evaluation, EM rises to 60.44, which recovers 31.7% of the gain of the GT viewpoint (Equation 7); an independent replication of the three models recovers about half (Appendix D.3). The localizer is trained with pose supervision on SQA3D and registered with known camera poses, so even this favorable setting recovers only a third to a half of the gain.

5.2 Quantitative Results

The model follows a supplied viewpoint. On the 517 held-out direction questions whose label changes under the swapped viewpoint, 88.0% of the answers of our GT-viewpoint model agree with the swapped viewpoint, against 14.1% when the original viewpoint is kept (Figure 1). Reviewed viewpoint pairs with natural-language references, in which the description changes together with the viewpoint, show the same behavior on a small scale: both answers of a pair are correct for 12 of 14 pairs, against 2 of 14 for the Vicuna-7B model line, although the answer also changes on 5 of 14 control pairs whose label does not change (Appendix H.10).

Table 1: **Comparison with published 3D models.** Published rows are the values reported in the cited papers under their own protocols. Our rows are official validation scores: the first two from our GT-viewpoint model evaluated with and without the viewpoint, the third from M_{none}. The SQA3D column is not a like-for-like comparison: published scores are test-split EM with the situation given as text (LEO also receives the annotated situation pose), whereas ours are validation EM. Best grounding and captioning scores in **bold**.

Method	Venue	Type	LLM	ScanRefer Acc@0.5	Multi3DRefer F1@0.5	Scan2Cap CIDEr@0.5	SQA3D EM
Expert models							
ScanRefer (Chen et al., 2020)	[ECCV'20]	expert	–	24.3	–	–	–
ViL3DRel (Chen et al., 2022)	[NeurIPS'22]	expert	–	37.7	–	–	–
3DJCG (Cai et al., 2022)	[CVPR'22]	expert	–	37.3	–	49.5	–
3D-VLP (Jin et al., 2023)	[CVPR'23]	expert	–	39.5	–	54.9	–
M3DRef-CLIP (Zhang et al., 2023)	[ICCV'23]	expert	–	44.7	38.4	–	–
3D-VisTA (Zhu et al., 2023)	[ICCV'23]	expert	–	45.8	–	66.9	48.5
Vote2Cap-DETR++ (Chen et al., 2024b)	[TPAMI'24]	expert	–	–	–	67.6	–
ConcreteNet (Unal et al., 2024)	[ECCV'24]	expert	–	46.5	–	–	–
LLM-based models							
LL3DA (Chen et al., 2024a)	[CVPR'24]	LLM	OPT-1.3B	–	–	65.2	–
LEO (Huang et al., 2024b)	[ICML'24]	LLM	Vicuna-7B	–	–	72.4	50.0
Scene-LLM (Fu et al., 2024)	[arXiv'24]	LLM	Llama2-7B	–	–	–	54.2
Grounded 3D-LLM (Chen et al., 2024c)	[arXiv'24]	LLM	Tiny-Vicuna-1B	44.0	40.8	70.2	–
Chat-Scene (Huang et al., 2024a)	[NeurIPS'24]	LLM	Vicuna-7B	50.2	52.4	77.1	54.6
Chat-Scene++ w/o 2D (Huang et al., 2026)	[TPAMI'26]	LLM	Vicuna-7B	47.3	51.6	72.8	54.2
Chat-Scene++ (Huang et al., 2026)	[TPAMI'26]	LLM	Vicuna-7B	53.4	58.7	83.0	55.8
Ours							
GT viewpoint	–	LLM	Qwen3-VL-8B	**58.19**	**62.95**	**87.40**	64.21
Viewpoint removed at inference	–	LLM	Qwen3-VL-8B	–	–	–	55.23
No viewpoint in training or inference	–	LLM	Qwen3-VL-8B	–	–	–	58.69

A badly wrong viewpoint is worse than none. Table 3 alters only the values of the viewpoint and keeps the input format. Rotating the heading by 180° costs 17.39 points, far more than removing the viewpoint (6.53), and leaves the model 11.8 points below M_{none}, which was trained without a viewpoint (46.83 vs. 58.66). This is the most damaging perturbation tested, whereas a 1 m shift of the position costs 0.86 points. Permuting viewpoints within a scene costs 10.49 points, so the model relies on the match between a description and its viewpoint rather than on scene-level statistics of viewpoints; the same permutation costs M_{none}, when viewpoints are injected into it, only 0.77 points (Appendix D.3). Consistently, the GT viewpoint raises the EM of M_{GT} by 7.30 points on viewpoint-dependent questions and changes it by -1.62 on viewpoint-independent ones, a difference of 8.92 points that persists when the 2D feature branch is switched off (Appendices D.5 and H.12).

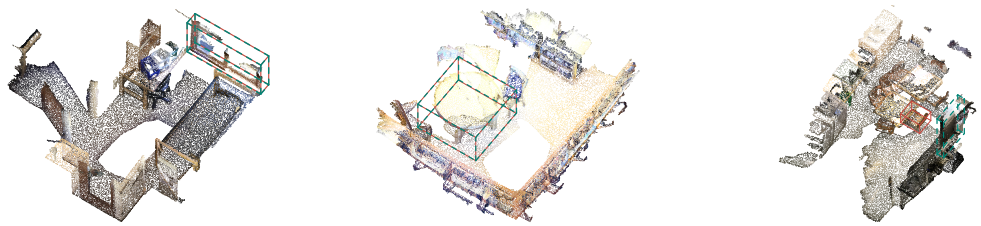
Table 3: **Viewpoint intervention.** SQA3D EM (%) of M_{GT} when only its inference viewpoint is altered; Δ is relative to the GT viewpoint.

Inference viewpoint	EM	Δ
GT	**64.21**	–
None	57.68	-6.53
Position shifted 1 m	63.35	-0.86
Scene centroid	55.04	-9.17
Random pose in scene	54.55	-9.66
Heading rotated $\pm 90°$	54.16	-10.06
Heading rotated 180°	46.83	-17.39
Permuted within scene	53.73	-10.49

The benefit is learned in training. The GT viewpoint raises the EM of M_{GT} by 6.53 points but changes that of M_{none} by -0.03 (Table 4), an interaction of $+6.56$ points (Equation 6). The null effect on M_{none} is the expected outcome for inputs that were zero throughout its training, and it confirms that the gain of M_{GT} comes from learned use of the viewpoint rather than from the extra input, so the viewpoint encoding has to be trained in rather than attached to a model trained without it. The interaction replicates in independent training runs and in the Vicuna-7B model line (Appendix D.2). This use depends on the language adapters learned with the viewpoint: switching off

Table 4: **Matched training.** SQA3D EM (%) by training viewpoint (rows) and inference viewpoint (columns). Gain is the effect of the GT viewpoint at inference. M_{none} differs from Table 2 by one question owing to the evaluation pipeline (Appendix B).

Training	None	GT	Gain
None (M_{none})	58.66	58.63	-0.03
GT (M_{GT})	57.68	**64.21**	$+6.53$



(a) window (b) table (c) refrigerator

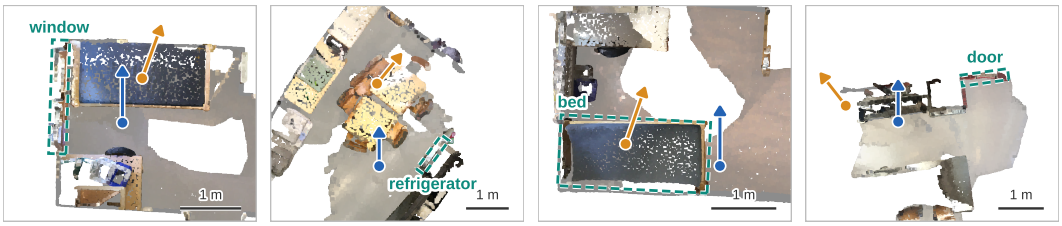
Figure 3: **What the predicted-viewpoint model grounds.** Annotated target box (teal, dashed) and the proposal box of the object identifier the model generates when asked for the referent (red, solid). (a) The model trained without a viewpoint picks a wrong proposal while the predicted-viewpoint model recovers the window; (b) all three models recover the table; (c) the predicted-viewpoint model fails where the GT-viewpoint model is correct. Panels (a) and (c) are the scenes of Figure 4(a) and (b); cases are selected for illustration.

the LoRA of the GT-trained model while keeping all trained input modules lowers joint correctness on the reviewed viewpoint pairs from 12/14 to 4/14, close to the 3/14 of the model trained without viewpoints (Appendix H.13).

A raw pose helps much less than encoded directions. Given the annotated pose as raw values, so that it has to relate each object to the viewpoint itself, a model trained with the same data and schedule reaches at most 60.38 EM (Appendix B.1), close to M_{pred} and well below the 64.21 obtained when the directions are computed outside the model. This agrees with SIG3D, where a ground-truth situation helps only after it is re-encoded into scene tokens (Man et al., 2024). Linear probes over all 36 decoder blocks of our GT-viewpoint model find no block from which viewpoint-relative coordinates can be read reliably when it receives no viewpoint (Appendix G.1).

Correcting the heading helps more than correcting the position. To locate the error of the predicted viewpoint, we keep M_{pred} fixed and replace either component of its inference viewpoint by the annotated value, with the observer height fixed at 1.5 m in every condition. Replacing the predicted heading raises EM from 60.44 to 62.93, whereas replacing the predicted position raises it only to 61.02, and replacing both gives 63.35, within 0.9 points of M_{GT}. The two corrections are nearly additive (interaction -0.15 points), so heading and position errors harm the answer largely independently, and correcting the heading alone closes most of the gap. This ordering agrees with the viewpoint interventions in Table 3, where a heading rotation costs far more than a 1 m shift of the position, and it holds in an independent replication; on the 1,216 questions about horizontal relations, the difference between the two corrections is not reliable (Appendix D.1). Heading estimation is therefore the most direct target for improving learned localizers in egocentric 3D reasoning.

The training effect transfers across datasets. We further evaluate the matched models, without additional training, on the ScanNet, 3RScan, and ARKitScenes domains of MSQA and on Spartun3D, whose scenes do not appear in QA training. MSQA supplies the viewpoint with each question; for Spartun3D we reconstruct it from the situation construction, and the offline 2D features, unavailable for 3RScan, ARKitScenes, and Spartun3D, are masked identically in all compared conditions (Appendix F). In every set, the GT viewpoint helps the model trained with it and does not help the model trained without it: on MSQA-ScanNet, for example, it raises the EM of M_{GT} from 35.49 to 39.93 but changes that of M_{none} from 39.25 to 38.57. The resulting interactions are positive in all four sets ($+5.12$, $+5.35$, $+5.84$, and $+1.29$ points on MSQA-ScanNet, MSQA-3RScan, MSQA-ARKitScenes, and Spartun3D), most reliably on MSQA-ScanNet and Spartun3D. The effect concentrates on directional queries: on MSQA-ScanNet, the gain of the GT viewpoint for M_{GT} grows from 4.29 to 5.62 points when only direction questions are kept. It also persists when unparseable MSQA answers are counted as wrong, and two LLM graders preserve the sign of every pose and injection effect on MSQA. Because MSQA provides the viewpoint, these results show that learned viewpoint use carries over to new scenes and scanners; they do not test viewpoint inference.



(a) *What is on the left side of the bed in front of me?*

None	Desk ✗
Predicted	Window ✓
GT	Window ✓

(b) *What object is at my three o'clock?*

None	Copier ✗
Predicted	Chair ✗
GT	Refrigerator ✓

(c) *What is to my left side?*

None	Bed ✓
Predicted	Recycling bin ✗
GT	Bed ✓

(d) *What is the object furthest away from me on the right?*

None	Chair ✗
Predicted	Refrigerator ✗
GT	Table ✗

Figure 4: **Visual evaluation results.** Selected answers of M_{none}, M_{pred}, and M_{GT} on four validation scenes, each shown top-down and rotated so that the GT viewpoint (blue arrow) faces up; the ochre arrow is the Loc3R estimate and the teal box the annotated instance of the reference answer. (a) The predicted viewpoint repairs an error, (b) only the GT viewpoint succeeds, (c) the predicted viewpoint causes an error, and (d) all three models fail. Situations are listed in Appendix I.

5.3 Visual Evaluation Results

Figure 4 compares the answers of the three matched models on four validation scenes, and Figure 3 shows the objects that the models ground for three of these queries. In Figure 4(a), the Loc3R estimate deviates from the annotated viewpoint, yet the predicted-viewpoint model both answers and grounds the window correctly (Figure 3(a)), repairing the error of M_{none}. In Figure 4(b), the estimate is displaced and rotated, and the predicted-viewpoint model names a chair at three o'clock and grounds a wrong proposal (Figure 3(c)); only the GT viewpoint yields the refrigerator. In Figure 4(c), the estimate falls on the bed itself, so the bed no longer lies to its left: the predicted-viewpoint model answers a recycling bin, whereas M_{none}, which receives no viewpoint, answers correctly, a concrete case of a wrong viewpoint being worse than none. Figure 4(d) combines a direction with a distance comparison, and all three models fail, including the one given the GT viewpoint. For the table in Figure 3(b), all three models ground the referent correctly.

6 Conclusion

We presented EgoAudit, a framework that audits egocentric spatial reasoning in 3D VLMs by making the viewpoint an explicit, controllable input and evaluating the model through viewpoint intervention, matched training, and viewpoint-source evaluation. Applied to a strong 3D VLM on public 3D grounding and 3D VQA benchmarks, EgoAudit shows that the model follows a given viewpoint closely but does not infer it, so scores obtained with annotated viewpoints measure viewpoint use rather than perspective taking. We therefore recommend that egocentric 3D results report the viewpoint source (none, predicted, or annotated), any transformation applied to the scene before the model, and whether the model was trained with the viewpoint, and that progress be measured with predicted viewpoints, with the annotated-viewpoint score as an upper reference. Improving viewpoint inference, and heading estimation in particular, is the most direct path toward 3D VLMs that reason from the user's perspective.

AI Use Statement

Generative AI tools assisted with refining and polishing the text. They were not used to synthesize benchmark observations, ground truth, or reported experimental results. Numerical claims are checked against saved experiment records. The authors remain responsible for verifying the code, scientific claims, citations, and manuscript.

REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *CVPR*, 2022.

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

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. *arXiv preprint arXiv:2111.08897*, 2021.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3DJCG: A unified framework for joint dense captioning and visual grounding on 3D point clouds. In *CVPR*, 2022.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *ECCV*, 2020.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3D object grounding. In *NeurIPS*, volume 35, 2022.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: Visual interactive instruction tuning for omni-3D understanding, reasoning, and planning. In *CVPR*, 2024a.

Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, and Tao Chen. Vote2Cap-DETR++: Decoupling localization and describing for end-to-end 3D dense captioning. *TPAMI*, 2024b.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3D-LLM with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024c.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *CVPR*, pp. 3193–3203, 2021.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-LLM: Extending language model for 3D visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

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. In *NeurIPS*, volume 36, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-Scene: Bridging 3D scene and large language models with object identifiers. In *NeurIPS*, volume 37, 2024a.

Haifeng Huang, Yilun Chen, Zehan Wang, Jiangmiao Pang, and Zhou Zhao. Chat-Scene++: Exploiting context-rich object identification for 3D LLM. *TPAMI*, 48(8):9813–9825, 2026.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *ICML*, volume 235 of *PMLR*, pp. 20413–20451, 2024b.

Jiaxin Huang, Ziwen Li, Hanlue Zhang, Runnan Chen, Zhengqing Gao, Xiao He, Yandong Guo, Wenping Wang, Tongliang Liu, and Mingming Gong. Surprise3D: A dataset for spatial understanding and reasoning in complex 3D scenes. In *NeurIPS*, volume 38, 2025.

Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3D-language pre-training. In *CVPR*, 2023.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In *ACL*, pp. 32803–32827, 2026.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *ICCV*, pp. 9241–9251, 2025.

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. In *ECCV*, 2026.

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3D scenes. In *NeurIPS*, volume 37, pp. 140903–140936, 2024.

Quan Liu, Weihao Xuan, Junjue Wang, Naoto Yokoya, Ling Shao, and Shijian Lu. Direction-aware 3D large multimodal models. In *CVPR*, pp. 9668–9678, 2026.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan L. Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *NeurIPS*, volume 38, 2025.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *ICLR*, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *ICLR*, 2023.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, et al. OpenEQA: Embodied question answering in the era of foundation models. In *CVPR*, pp. 16488–16498, 2024.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3D vision language reasoning. In *CVPR*, pp. 13678–13688, 2024.

Mistral AI. Robostral navigate. *arXiv preprint arXiv:2607.20785*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. GPT4Scene: Understand 3D scenes from videos with vision-language models. In *ICLR*, 2026.

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. *arXiv preprint arXiv:2603.18002*, 2026.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *CVPR*, pp. 15768–15780, 2025.

Yuanhao Su, Shaofeng Zhang, Xiaosong Jia, and Qi Fan. PointAlign: Feature-level alignment regularization for 3D vision-language models. In *CVPR*, 2026.

Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. Four ways to improve verbo-visual fusion for dense 3D visual grounding. In *ECCV*, 2024.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *ICCV*, pp. 7658–7667, 2019.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, pp. 10510–10522, 2025.

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. In *CVPR*, 2025. arXiv:2412.14171.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding text description to multiple 3D objects. In *ICCV*, pp. 15225–15236, 2023.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. SPARTUN3D: Situated spatial understanding of 3D world in large language model. In *ICLR*, 2025a.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *ICLR*, 2025b.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In *ICCV*, 2023.

A Representation and Optimization

Object-centric visual inputs. Each scene contains at most 100 proposals. Each proposal provides an identifier, a 128-dimensional CLASP 3D feature, a 1024-dimensional offline multi-view 2D feature, and geometry. The offline visual features enter the language backbone through learned projectors. The Qwen3-VL visual tower does not process raw RGB in this QA path; the 3D and offline 2D features provide its visual evidence. For a valid proposal i, geometry is fused into the 3D token as

$$\mathbf{h}_i^{\mathrm{3D}} = \mathrm{LN}(P_{\mathrm{3D}}(\mathbf{f}_i) + G(\mathbf{g}_i)), \tag{8}$$

where G is an $11 \to 256 \to 4096$ MLP. The 2D feature and identifier retain their corresponding object-token paths. Geometry adds no token and uses the same dimensions in every main intervention.

Observer-relative encoding. The 11 channels concatenate world position (three), observer-relative displacement (three), object size (three), and sine/cosine bearing (two). Let ρ_i be horizontal distance from the observer, Δz_i relative height, and $\hat{\theta}_i$ the center of the object's bearing bin. The encoded displacement is

$$\Big(\rho_i \cos\hat{\theta}_i,\ \rho_i \sin\hat{\theta}_i,\ \mathrm{clip}(\Delta z_i, -5, 5)\Big). \tag{9}$$

The horizontal plane is divided into eight 45° bearing bins. Quantization therefore affects both displacement and the explicit bearing channels. The world-frame fallback preserves world position and size and sets the other five channels to zero. The deterministic geometry builder subtracts observer position and rotates object coordinates before applying this encoding. Consequently, the supplied-frame condition exposes the downstream model to the transformed relations themselves.

Training and inference. The original diagnostic checkpoint is trained on nine task configurations for three epochs, with global batch size 32. Logged optimizer learning rates are 1.6×10^{-4} for the base group and 3.2×10^{-3} for the geometry group. LoRA uses rank 16, scale 16, and dropout 0.05. Transformer and vision weights are frozen; LoRA, input/output token embeddings, projectors, and geometry/fusion modules train. The original manifest records 1,326,820,352 trainable parameters, including 43,646,976 LoRA parameters; a listed 528,384-parameter positional branch is disabled in this configuration. The matched siblings share training recipe, data, and schedule and are evaluated at epoch 2, step 26,346. All main cross-inference cells use beam size 5, repetition penalty 3.0, and 64 maximum new tokens.

Learned viewpoint provider. Loc3R consumes RGB frames and the situation description; the QA answer is not a provider input. Its RGB and camera representations produce planar position and orientation estimates. These estimates are registered from the first-frame local coordinates to scene coordinates using the first frame's camera-to-world transform and the scene axis alignment. Because Loc3R predicts only planar position and heading, observer height is fixed to 1.5 m in the cached predictions. The provider uses supervised training and known scene registration, while supplying predicted observer viewpoints to the downstream QA model at inference. This distinction specifies exactly what information replaces the annotated SQA3D viewpoint.

B Checkpoint Roles and Benchmark Context

Main-text names. M_{GT}, M_{none}, and M_{pred} in the main text are the matched pose-trained (GT-relative), no-pose (world-frame), and predicted-pose siblings of Table 5. The GT-viewpoint model of Table 1, also used for the direction probe, is the original diagnostic checkpoint. The world-frame input, labeled "Global" in some tables below, is the no-viewpoint input of the main text.

The original diagnostic checkpoint and the matched GT-trained sibling happen to have the same rounded native SQA3D EM, but they are separate training runs. Fixed-model conclusions compare conditions within a row of Table 5. Cross-training conclusions use the matched siblings only. The native no-pose evaluation is 58.69 EM; the shared four-cell harness gives 58.66 EM, differing on one question. We retain the shared-harness value for the interaction and the native-path value for the three-model deployment comparison.

The original checkpoint obtains 64.21 EM with GT-relative geometry, 55.23 with a world-frame substitution, and 60.04 when supplied a Loc3R prediction at inference. This is a fixed-checkpoint

Table 5: Checkpoint lineage. Hashes are SHA-256 prefixes recorded with the corresponding runs. A dash indicates that a hash prefix was not included in the source result summary.

Family and role	Training geometry	Seed	Hash prefix
Original: direction and channel diagnostics	GT-relative	42	`884f10c8`
Matched: pose-trained sibling	GT-relative	42	`25a3b339`
Matched: no-pose sibling	World frame	42	`14023870`
Matched: predicted-pose sibling	Loc3R predictions	42	`cc7d2f6e`
Matched: second four-cell crossing	GT / world	43	–

input-substitution experiment. Its 53.6% recovery fraction uses the 8.98-point substitution gap. Table 2 instead compares independently trained native-input models and recovers 31.7% of a 5.52-point gap. The latter is the deployment comparison used in the main text.

B.1 Raw-pose checkpoint evaluation

Two raw-pose checkpoints were trained with the same data, recipe, and schedule as the matched siblings and evaluated on the same 3,261 SQA3D official-validation questions and 65 scenes using the annotated pose. Their 11D layout is $[x_{obj}, y_{obj}, z_{obj}, x_{agent}, y_{agent}, z_{agent}, size_{xyz}, \sin\psi, \cos\psi]$: the model receives the annotated observer pose as raw values next to world-frame object positions and has to construct observer-relative relations itself, whereas the matched GT-relative model receives relations built by the external frame builder. We therefore report these runs separately rather than as cells of Table 4; Section 5.2 compares their scores with the matched models.

Table 6: Raw-pose checkpoint evaluations on SQA3D official validation. Both use the annotated pose and the same frozen evaluation protocol.

Seed	Checkpoint (epoch/step)	EM1	EM1 refined	Correct (EM1/refined)
42	2 / 26,346	60.38	63.11	1,969 / 2,058
43	2 / 26,346	59.67	62.37	1,946 / 2,034

The seed-43 point estimate is 0.71 pp lower in EM1 and 0.74 pp lower in refined EM than seed 42. Checkpoint SHA-256 prefixes are `fd1fc571` (seed 42) and `8b712698` (seed 43); independent row-level score verification passed for both runs.

B.2 Five-benchmark context

The testbed supports grounding, captioning, and question answering through the same object-centric architecture. Table 7 gives the original checkpoint's official-validation results. These scores set the operating context for the audit; the controlled comparisons use common inputs and evaluators within each task.

Table 7: Original multitask checkpoint on five official validation splits. ScanQA includes the legacy and symmetric-normalization EM scores on the same saved predictions.

Benchmark	Metric	Score
ScanRefer	Acc@0.5	58.19
Multi3DRefer	F1@0.5	62.95
Scan2Cap	CIDEr@0.5	87.40
ScanQA	EM, legacy / symmetric normalization	24.49 / 26.60
SQA3D	EM, supplied frame / world substitution	64.21 / 55.23

B.3 Motivating diagnostic: a geometry-free baseline

Before geometry injection was added to the testbed, we diagnosed a Chat-Scene-style CLASP baseline with a Vicuna-7B language model on SQA3D. It is a different checkpoint family from the au-

Table 8: External reference values from Chat-Scene++ Table IV, not reruns under our audit protocol. SR: ScanRefer Acc@0.5; M3D: Multi3DRefer F1@0.5; S2C: Scan2Cap CIDEr@0.5; SQA: SQA3D. All rows use Vicuna-7B.

Published setting	SR	M3D	S2C	ScanQA CIDEr	SQA EM	SQA EM-R
Chat-Scene	50.2	52.4	77.1	87.7	54.6	57.5
Chat-Scene++ w/o 2D	47.3	51.6	72.8	82.5	54.2	57.3
Chat-Scene++ w/ 2D	53.4	58.7	83.0	93.5	55.8	58.5

dited Qwen3-VL-8B models and is reported only as motivation. The baseline's object tokens carried CLASP 3D features and DINOv2 2D features but no positional information: the positional embedding was disabled, the object location tensor was loaded but not used, and the situation entered the prompt only as second-person text, so no observer position or heading reached the model. Trained jointly with SQA3D, this geometry-free baseline reached 54.2 EM1 on the SQA3D validation split, close to the 54.6 that Chat-Scene reports on the test split (Huang et al., 2024a) (a different split), with changes below 0.5 points on ScanRefer, Scan2Cap, and Multi3DRefer.

Its errors frequently involved relations that require an observer frame. Among the 1,495 incorrect predictions, keyword incidence was 890 for front/behind, 512 for right, 474 for left, 285 for counting, and 257 for above/below (a question can match several categories), against 120 for color; these relations are also frequent among all questions, so the counts describe where errors occur. A manual review of 20 failures from one scene attributed 9 to counting or parity, 3 to left/right or clock-face direction, 4 to object or attribute perception, 3 to answer formatting, and 1 to an ambiguous reference. It motivates the audit by showing that situated-QA scores in the range of published text-only results are attainable with no observer position or heading as input, while many residual errors involve relations that an observer frame would resolve.

B.4 Published Chat-Scene++ context

Table 8 records external values from Table IV of Huang et al. (2026). They follow the publication's task-specific evaluation protocols and are not pooled with our official-validation audit. The same paper's Table VIII reports SQA3D EM of 55.8, 56.1, and 56.8 for Vicuna-7B, Vicuna-13B, and LLaMA3-8B, respectively.

C Matched Pose-Representation Training and Direction Trajectories

We compared four 11D geometry conditions in a separate matched training matrix: world-frame geometry (Global), raw annotation position and heading (Raw 11D), continuous observer-relative geometry, and quantized relation geometry. Each condition used two seeds (42 and 43), a global batch size of 32, and 2,000 updates. The three pose-bearing conditions all used annotation agent pose; none is a predicted-pose system. This matrix is distinct from both the raw-pose checkpoints in Appendix B.1 and the epoch-2, step-26,346 siblings used in Table 4.

Full validation endpoint. All eight step-2,000 checkpoints were evaluated on the same 3,261 SQA3D official-validation questions and 65 scenes. Global obtained 49.19 and 51.43 EM1 for seeds 42 and 43, respectively. Table 9 reports within-seed contrasts. The quantized-relation contrast changed from positive for seed 42 to negative for seed 43, with both scene intervals excluding zero. No representation shows a stable advantage across the two seeds.

Fixed horizontal subset across checkpoints. We followed the same 1,216 official-validation questions from 65 scenes at updates 500, 1,000, 1,500, and 2,000 (Figure 5). The first three checkpoints required 24 evaluations (four conditions $\times$ two seeds $\times$ three checkpoints). The eight step-2,000 subset records were taken by question identity from the full-validation predictions. All 32 records passed independent checks of checkpoint identity, question support, prediction hashes, and recomputed EM1 and refined EM.

Table 9: Matched step-2,000 full-validation results. Scores are EM1 percentages; differences and paired 95% scene-bootstrap intervals are in percentage points, relative to the same seed's Global condition. Seeds are reported separately.

Seed	Geometry	EM1	Δ vs. Global	95% CI
42	Raw 11D	50.66	$+1.47$	$[+0.50, +2.46]$
42	Continuous relative	49.56	$+0.37$	$[-0.98, +1.68]$
42	Quantized relation	51.06	$+1.87$	$[+0.57, +3.15]$
43	Raw 11D	50.97	-0.46	$[-1.71, +0.82]$
43	Continuous relative	51.40	-0.03	$[-1.30, +1.31]$
43	Quantized relation	49.83	-1.59	$[-2.61, -0.56]$

On this subset at step 2,000, Global obtained 53.95 and 55.02 EM1 for seeds 42 and 43. All six representation-versus-Global EM1 intervals included zero (Table 10); the six corresponding refined-EM intervals also included zero. The trajectories therefore provide no stable horizontal-direction advantage for a particular pose representation over this training budget.

Table 10: Horizontal-subset endpoint at step 2,000, using the same 1,216 questions for every condition. Differences and paired 95% scene-bootstrap intervals are in percentage points.

Seed	Geometry	EM1	Δ vs. Global	95% CI
42	Raw 11D	54.52	$+0.58$	$[-1.05, +2.25]$
42	Continuous relative	53.13	-0.82	$[-2.55, +0.95]$
42	Quantized relation	53.04	-0.90	$[-2.66, +0.83]$
43	Raw 11D	53.87	-1.15	$[-2.73, +0.41]$
43	Continuous relative	53.54	-1.48	$[-3.12, +0.23]$
43	Quantized relation	54.19	-0.82	$[-2.17, +0.52]$

Uncertainty and scope. Contrasts use 10,000 paired scene-cluster bootstrap draws per seed (resampling seed 20260924). The intervals quantify scene-sampling uncertainty conditional on the saved checkpoints, not training-seed uncertainty. The 48 direction-trajectory contrasts (three comparisons $\times$ two seeds $\times$ four checkpoints $\times$ two metrics) have no multiplicity correction. In particular, an isolated positive seed-43 continuous-relative contrast at step 500 ($+1.89$ pp; pointwise 95% CI $[+0.08, +3.74]$) is exploratory. The comparison covers the tested budget of 2,000 updates.

D Interventions and Statistical Protocol

Fixed-format viewpoint interventions. All main corruptions act before the Geometry MLP and recompute observer-relative features from the intervened pose. They retain object identities, token layout, tensor dimensions, and geometry-validity flags. The translations and heading rotations are separate perturbations; their magnitudes describe physical interventions, with no assumed common difficulty scale. Constant and random viewpoints test dependence on the supplied values while preserving the interface. The within-scene permutation is a situation-level bijection with no self-match. Questions sharing a situation receive the same donor viewpoint. The set of situation-level poses is preserved; because situations have unequal question counts, the induced question-level pose frequencies differ slightly (total variation 0.025). These interventions retain the original situation text and original QA labels and measure susceptibility to inconsistent viewpoint input.

D.1 Position and heading component interventions

We replace predicted XY and heading independently on the predicted-pose-trained checkpoint. P denotes the cached predicted component and G the annotated component: PP is predicted position/predicted heading, GP annotated position/predicted heading, PG predicted position/annotated heading, and GG both annotated. Observer height is fixed at $z = 1.5$ m in every cell. The horizontal subset is frozen from the original prompts and contains 1,216 questions; it is reported separately from all 3,261 questions.

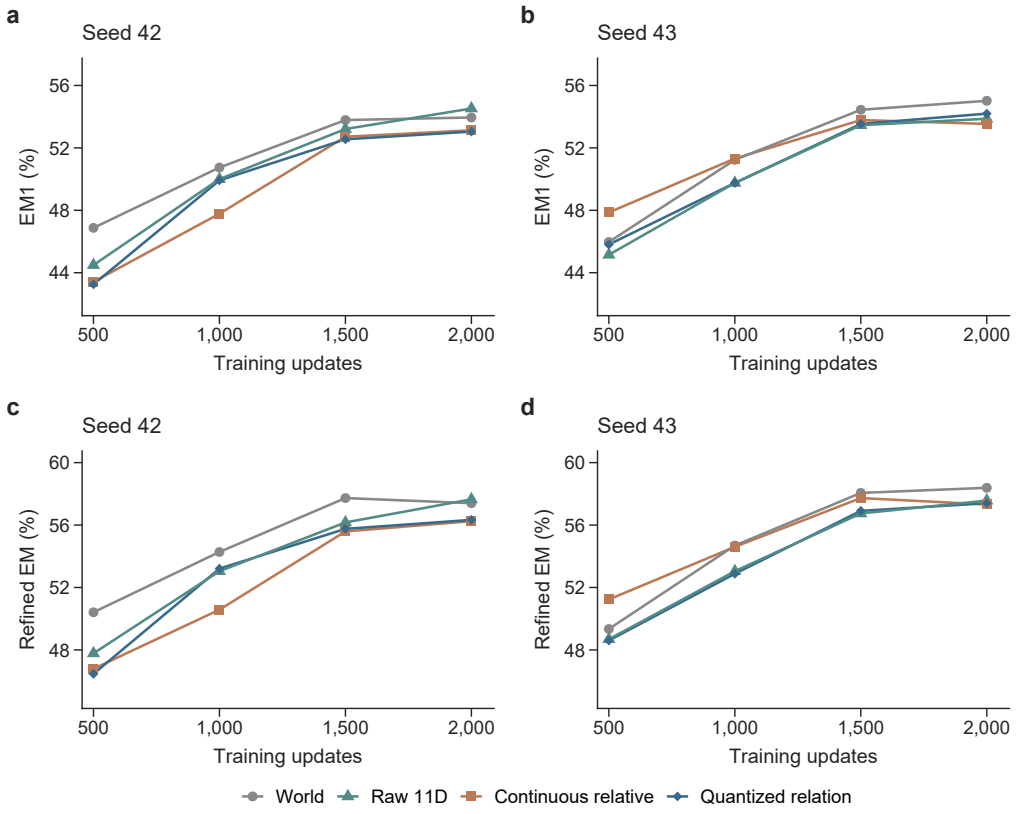


Figure 5: **Direction accuracy over the matched training trajectory.** (a,b) EM1 and (c,d) refined EM on the fixed 1,216-question horizontal subset from 65 scenes, with seeds shown separately. Step-2,000 values are taken from the full-validation predictions. Curves connect fixed-checkpoint point estimates. Paired scene-bootstrap contrasts are reported separately. No representation shows a stable advantage across seeds.

Table 11: **Fixed-format viewpoint interventions.** Changes in SQA3D EM (percentage points) for the matched seed-42 pose-trained checkpoint, relative to its native GT input (64.21 EM). All rows retain the same geometry dimensions and validity flags. Brackets give paired scene-bootstrap 95% CIs.

Intervention	ΔEM	95% CI
Position translation, 1 m	−0.86	[−1.55, −0.18]
Scene-centroid constant viewpoint	−9.17	[−10.81, −7.53]
Random viewpoint within scene envelope	−9.66	[−11.29, −8.02]
Heading rotation, ±90°	−10.06	[−11.43, −8.70]
Heading reversal, 180°	−17.39	[−19.56, −15.32]
Within-scene viewpoint permutation	−10.49	[−12.14, −8.81]

Paired estimates. Each bootstrap replicate jointly resamples clusters with replacement and computes the question-weighted contrast on the resampled questions. We use 10,000 resamples and bootstrap seed 42 for the four-cell and transfer analyses. The interaction is recomputed from all four cells within each replicate; its confidence interval is not formed by subtracting separate interval endpoints. The SQA3D and MSQA cluster unit is the scene; the Spartun3D unit is the physical location, grouping 144 scans into 44 locations. Intervals are per contrast and are not adjusted for multiple comparisons. They condition on the selected checkpoints and exclude training-seed variation.

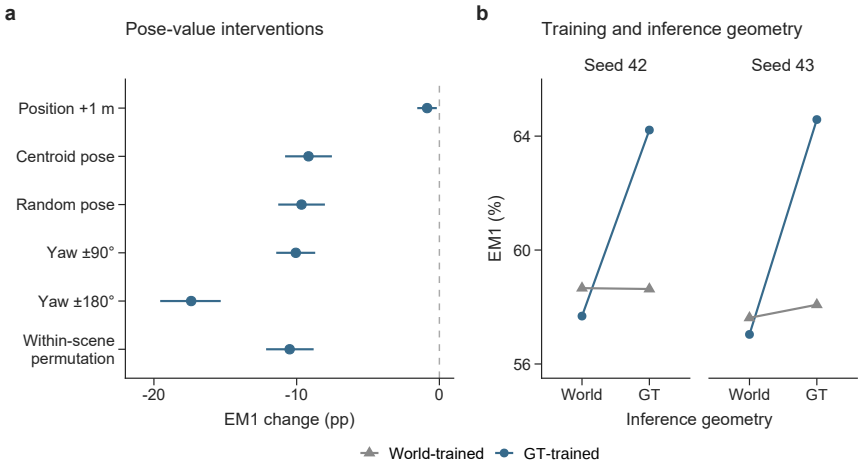


Figure 6: **Viewpoint values and training exposure govern frame use.** (a) SQA3D EM changes under fixed-format viewpoint interventions on the seed-42 matched GT-trained checkpoint, with paired 95% scene-bootstrap intervals (Table 11). (b) The training–inference crossing for each seed; lines connect the same checkpoint across inference geometries (Table 14).

Table 12: EM1 for the fixed-height position-by-heading interventions. Seed 43 PP reuses the shared-pipeline evaluation; the other conditions use the same within-seed pipeline.

Seed	Support	PP	GP	PG	GG
42	All (3,261)	60.44	61.02	62.93	63.35
42	Horizontal (1,216)	63.08	63.65	65.13	64.47
43	All (3,261)	61.24	61.36	63.94	64.12
43	Horizontal (1,216)	63.16	64.39	64.97	64.80

D.2 Cross-system training-by-input interactions

The matched four-cell interaction was also evaluated on a Vicuna model line using continuous relative geometry. All rows use the same 3,261 SQA3D official-validation questions within the row. The two model lines differ in geometry encoding and implementation, so the table tests whether the interaction recurs within each line.

Vicuna wrong-frame donor check. For the Vicuna seed-42 checkpoint, the correct GT-relative aggregate EM1 is 59.58%; a same-scene donor-pose substitution yields 52.10% EM1 and 54.74% refined EM. The world-input score (53.66%) is the absent-frame control. The correct-frame predictions were paired with the donor outputs on all 3,261 question identities and 65 scenes. Re-scoring gives 1,943 versus 1,699 correct answers: donor substitution repairs 138 answers and breaks 382. The paired donor-minus-correct-frame contrast is -7.48 percentage points (95% scene-bootstrap CI $[-9.11, -5.84]$; 10,000 draws, seed 42). Saved configuration and initialization manifests match apart from the donor annotation and output routing.

D.3 Matched predicted-pose comparisons across two seeds

Seed 42 uses its native evaluation path; seed 43 uses the shared deterministic pipeline of the four-cell crossing. Each row is internally matched over the same qids, scenes, and reference order, but the two rows use different evaluation pipelines and are not pooled. Recovery is computed within each row as $100 \times (\text{Predicted} - \text{Global})/(\text{GT} - \text{Global})$. The seed-43 recovery interval is bootstrapped jointly by scene; for seed 42, a post-hoc scene bootstrap (10,000 draws) of the recovery ratio gives $[5.67, 53.54]\%$, computed separately from the $P - G$ interval.

Table 13: Paired component contrasts in percentage points. $PG - GP$ directly compares correcting heading versus correcting position; $I = (GG - GP) - (PG - PP)$. Brackets are 95% scene-cluster bootstrap intervals conditional on each fixed checkpoint.

Seed	Support	$PG - GP$ EM1	$PG - GP$ refined	I EM1	I refined
42	All	$+1.90$ $[0.80, 3.02]$	$+2.05$ $[0.94, 3.15]$	-0.15 $[-1.09, 0.78]$	-0.40 $[-1.38, 0.58]$
42	Horizontal	$+1.48$ $[-0.15, 3.12]$	$+1.32$ $[-0.25, 2.94]$	-1.23 $[-2.61, 0.08]$	-1.40 $[-2.89, 0.00]$
43	All	$+2.58$ $[1.41, 3.74]$	$+2.64$ $[1.45, 3.84]$	$+0.06$ $[-0.85, 0.99]$	-0.18 $[-1.09, 0.77]$
43	Horizontal	$+0.58$ $[-0.99, 2.09]$	$+0.33$ $[-1.26, 1.85]$	-1.40 $[-2.58, -0.25]$	-1.15 $[-2.29, 0.00]$

Table 14: Training-dependent observer-frame use across the Qwen3-VL-8B and Vicuna model lines. W is world-frame input and G is GT-relative input; $I = (A_{11} - A_{10}) - (A_{01} - A_{00})$. Intervals are paired by scene within each checkpoint.

Model line	Seed	$W \to W$	$W \to G$	$G \to W$	$G \to G$	I [95% CI]
Qwen3-VL-8B	42	58.66	58.63	57.68	64.21	$+6.56$ [5.06, 8.08]
Qwen3-VL-8B	43	57.62	58.08	57.04	64.58	$+7.08$ [5.57, 8.54]
Qwen3-VL-8B	44	58.39	58.30	58.39	63.85	$+5.55$ [4.18, 6.93]
Vicuna	42	53.94	53.94	53.66	59.58	$+5.92$ [4.54, 7.28]
Vicuna	43	53.82	53.73	51.73	59.40	$+7.76$ [6.05, 9.44]
Vicuna	44	54.03	53.97	52.16	59.34	$+7.24$ [5.50, 9.03]

The native seed-42 global score is 58.6937%, whereas the shared-harness score for the corresponding row is 58.6630%; the two evaluation pipelines differ on one question (qid `220602012875`). We retain each score only within its own pipeline. The intervals condition on selected checkpoints and are not estimates of training-seed variability.

Injection equivalence and individual answers. The prespecified equivalence margin for SQA3D injection into the no-pose sibling is ± 1.0 EM point. Seed 42 gives -0.03 points, CI $[-0.67, +0.58]$, satisfying this criterion. Seed 43 gives $+0.46$, CI $[-0.19, +1.13]$, whose upper bound exceeds the ± 1.0-point margin. Small mean changes can coexist with changes on individual questions: for seed 42, 6.2% of answer strings change, including 58 repairs and 59 newly incorrect answers. Within-scene pose permutation also reduces the world-trained model under injected geometry by 0.77 points, CI $[-1.40, -0.12]$.

Lexical diagnostic groups. For descriptive analysis, the original checkpoint's questions are grouped by lexical cues. These groups overlap. Table 18 records the exact grouping; the controlled direction diagnostic in Appendix E provides a separate binary task with geometry-defined labels.

D.4 Two-seed exploratory lexical candidate analysis

The paired comparison holds each GT-trained checkpoint fixed and contrasts its world-frame input (A_{10}) with native GT-relative input (A_{11}). Lexical groups are overlapping search candidates generated from question text; they have not been manually validated as semantic classes or as questions that truly depend on the observer frame. Repair/damage counts compare per-question correctness across the two inputs.

The candidate groups contain 6,180 memberships across 3,261 questions, with 1,711 questions appearing in multiple non-full groups. In particular, the residual group mixes attribute words with questions not captured by other rules, and spatial constructions can cross group boundaries. A separate human-reviewed observer-dependence analysis is reported in Appendix D.5.

D.5 Human-reviewed observer-frame dependence

Review scope and definition. We reviewed question semantics using AI-assisted initial labeling followed by human review and adjudication; the final labels follow the human review. A question is observer-dependent if interpreting it or resolving its referent requires the observer's position and/or horizontal heading. An explicit observer-relative selector is sufficient: for example, asking about the color of a chair to the observer's right remains observer-dependent even though the requested answer is an attribute. In nested expressions, an explicit outer observer selector determines this label

Table 15: Vicuna seed-42 aggregate scores for correct, absent, and donor-supplied frames. All rows cover 3,261 official-validation questions; the donor row is an inference-only same-scene substitution.

Inference geometry	EM1 (%)	Questions
World frame (no observer-relative fields)	53.66	3,261
Correct GT-relative frame	59.58	3,261
Same-scene donor pose (incorrect frame)	52.10	3,261

Table 16: Native-input SQA3D EM1 for the matched global, predicted-pose, and GT-pose models. Values and GT gaps are percentages/percentage points.

Seed / pipeline	Global	Predicted	GT	GT−Global
42 / native	58.6937	60.4416	64.2134	5.5198
43 / shared	57.6204	61.2389	64.5814	6.9611

without resolving the inner object's directional axis. Object–object relations and fixed room/map directions alone do not imply observer dependence, and sitting versus standing alone does not imply position/heading dependence under this rubric. Room membership and fixed room directions are distinct: the former can locate the observer within a space, whereas the latter need not use the observer as a reference.

Of the 3,261 official-validation questions, 24 earlier rubric-pilot items were kept separate. The remaining 3,237 comprise 2,998 previously settled items and 239 adjudicated items. The latter contain 153 dependent, 49 non-dependent, and 37 unresolved labels. Merging by question identity gives **2,892 dependent and 308 non-dependent (observer-independent) questions**, or **3,200 questions in the primary labeled analysis**. The 37 semantically unresolved items are listed in the manifest to be released and are excluded from the dependent/non-dependent comparison: 28 omit the reference frame, four have ambiguous action/path readings, four have incomplete wording, and one leaves the historical entry frame unspecified. These are exclusions for semantic uncertainty, not prediction correctness. Original question text, identities, and fixed-review-subset membership are unchanged.

This is a post-hoc, human-reviewed textual annotation without double annotation. Labels were frozen before joining the per-question predictions for this analysis, but the study was not prospectively registered or blinded to earlier aggregate findings.

Paired analysis. Within each existing GT-trained Qwen checkpoint, we compare world-frame input A_{10} with native GT-relative input A_{11} on exactly the same questions. A repair is an incorrect-to-correct transition and a damage is a correct-to-incorrect transition. EM1 is the primary metric; refined EM and the four-cell interaction are supplementary. All intervals below use 10,000 paired scene-family bootstrap draws with seed 20260925. Each draw resamples the 65 source scene families and recomputes question-weighted scores on the selected group; all groups and cells share the same draws. The non-dependent group occurs in 63 of these families. Group-difference intervals are computed directly from the paired replicate differences, not by comparing separate confidence intervals. Intervals are pointwise and conditional on the recorded checkpoints, without training-seed or annotation uncertainty; seeds are reported separately.

The dependent group gains 7.30 and 8.37 points for seeds 42 and 43, whereas the non-dependent group's changes are -1.62 and -0.32 points, with both intervals containing zero (Table 20). The directly paired gain difference between dependent and non-dependent groups is $+8.92$ points (95% CI $[+5.39, +12.40]$) and $+8.69$ points ($[+5.67, +11.75]$), respectively. Thus, the aggregate advantage is concentrated in the human-labeled observer-dependent group. The matched world-trained model shows the same pattern: $A_{11} - A_{00}$ is $+6.09$ and $+7.57$ points on dependent questions, versus -0.32 and $+0.97$ on non-dependent ones. This association is consistent with use of the supplied frame; the groups also differ in question content and size.

Supplementary metrics and training interaction. The same grouping also permits the original matched four-cell sensitivity analysis, with $I = (A_{11} - A_{10}) - (A_{01} - A_{00})$. It uses the 2D-on

Table 17: Matched predicted-versus-global gains and within-pipeline recovery. Scene-paired 95% intervals; *post-hoc scene bootstrap of the ratio, computed separately from the $P - G$ interval.

Seed / pipeline	$P - G$ EM1 [95% CI]	Recovery [95% CI]
42 / native	$+1.7479$ $[+0.2432, +3.2738]$	31.67% $[5.67, 53.54]^*$
43 / shared	$+3.6185$ $[+1.9308, +5.4518]$	51.98% $[29.86, 76.27]$

Table 18: Overlapping lexical groups on SQA3D validation, original diagnostic checkpoint. Gaps use unrounded scores.

Group	Questions	GT EM	World EM	Gap (pp)
All questions	3,261	64.21	55.23	8.98
Left/right	2,027	64.97	56.14	8.83
Front/behind	1,998	63.51	54.80	8.71
Above/below	250	63.60	52.80	10.80
Counting	732	55.87	52.19	3.69

predictions of the same checkpoints. Table 21 reports refined EM gains and the EM1 interaction. The stratification covers SQA3D only.

Support and sensitivity checks. The existing fixed 653-question review subset contains the 24 pilot items. Of its remaining 629 items, six remain unresolved, leaving 623 labeled questions (563 dependent and 60 non-dependent). On this subset, dependent-group EM1 gains are $+5.51$ and $+7.64$ points, versus -6.67 and -11.67 for the non-dependent group. This is a sensitivity check within the same data. The 37 unresolved questions have a $+5.41$-point EM1 change in each seed and remain excluded regardless of that change; their intervals include zero. The analysis to be released provides all support-specific scores, four-cell values, repair/damage counts, conditional repair/damage rates, frozen labels, input hashes, and bootstrap settings.

E Direction and Channel Diagnostics

Question construction and metric. The training-source diagnostic contains 512 explicit-object direction questions from 16 scenes: four situations, four object proposals, and two horizontal axes per scene. Target identities are given, and labels are computed from ground-truth geometry. Donor viewpoint substitution keeps scene, target, prompt, and weights fixed. Of 512 questions, 455 have unambiguous donor labels; 227 of these require a different answer under the donor viewpoint. Donor-label agreement on the changed-label subset is the proportion of predictions matching the new geometry-derived answer. It is computed without selecting for original correctness.

The held-out diagnostic uses 40 scene families selected under a frozen rule from 65 eligible SQA3D validation families. Scene-set intersection checks exclude every family appearing in the nine training annotation sets. It contains 1,280 questions and 517 unambiguous changed-label items. Substituted-viewpoint donor-label agreement is 88.0% on the changed-label items, compared with 14.1% with the original frame retained. Re-scoring all 3,840 saved outputs with the hash-matched direction parser confirms 455/517 donor-label agreements under substituted geometry and 73/517 under original geometry. Post-hoc, question-weighted bootstrap resampling of the 40 scenes gives 95% intervals of $[84.73, 91.09]\%$ and $[11.08, 17.38]\%$, respectively, and a paired increase of 73.89 points, CI $[68.38, 79.28]$ (10,000 draws, seed 42). The 14.1% value is agreement with the donor label under the original input, not original-label accuracy. The scene separation applies to the audited QA training annotations. The question construction uses explicit identifiers and ground-truth geometry to control referent and label ambiguity.

What information does the frame contain? Table 23 removes selected channels at inference on the original checkpoint. Removing the explicit bearing pair reduces accuracy from 84.2% to 78.1%, while preserving most of the improvement over world-frame inputs. Observer-relative displacement still contains heading information through its horizontal components. The result therefore characterizes channel dependence; the bearing pair and displacement are redundant carriers of

Table 19: Exploratory lexical candidate groups on SQA3D official validation. Entries give ΔEM1 in percentage points [95% paired scene-bootstrap CI], followed by repair/damage counts.

Rule-based candidate group	n	Seed 42	Seed 43
All questions	3,261	$+6.53$ $[5.20, 7.92]$; 372/159	$+7.54$ $[6.12, 8.93]$; 381/135
Horizontal words	2,093	$+4.01$ $[2.53, 5.61]$; 168/84	$+4.30$ $[3.00, 5.58]$; 170/80
Vertical words	198	$+2.53$ $[-2.68, 7.98]$; 13/8	$+0.51$ $[-3.27, 4.41]$; 8/7
Distance / proximity words	124	$+4.03$ $[-4.65, 13.13]$; 17/12	$+10.48$ $[2.50, 19.13]$; 19/6
Object-relation cues	837	$+3.58$ $[0.94, 6.29]$; 67/37	$+3.35$ $[1.21, 5.55]$; 55/27
Counting words	731	$+0.41$ $[-2.55, 3.46]$; 54/51	$+1.09$ $[-1.51, 3.66]$; 48/40
Direction-conditioned counting	533	$+2.06$ $[-1.70, 6.17]$; 45/34	$+1.31$ $[-1.85, 4.55]$; 38/31
Attribute / residual words	1,664	$+9.50$ $[7.32, 11.78]$; 235/77	$+10.34$ $[7.78, 12.93]$; 238/66

Table 20: Human-reviewed question groups on SQA3D official validation. Scores are EM1 percentages; A_{00} is the matched world-trained model with world-frame input, and $\Delta = A_{11} - A_{10}$ is in percentage points with a paired 95% scene-bootstrap interval. R/D gives repair/damage counts. The combined row contains the same 3,200 questions for both seeds.

Group	Seed	n	A_{00}	A_{10}	A_{11}	Δ [95% CI]	R/D
Dependent	42	2,892	58.78	57.57	64.87	$+7.30$ $[+5.87, +8.76]$	351/140
Dependent	43	2,892	57.81	57.02	65.39	$+8.37$ $[+6.84, +9.92]$	363/121
Non-dependent	42	308	61.36	62.66	61.04	-1.62 $[-4.90, +1.58]$	12/17
Non-dependent	43	308	59.42	60.71	60.39	-0.32 $[-3.32, +2.54]$	11/12
Combined	42	3,200	59.03	58.06	64.50	$+6.44$ $[+5.11, +7.81]$	363/157
Combined	43	3,200	57.97	57.38	64.91	$+7.53$ $[+6.08, +8.96]$	374/133

observer-relative direction. The raw-pose condition inserts position and heading into the same slots at inference, obtaining 54.5%. It documents the behavior of this checkpoint under a changed representation. The separate matched-training matrix and checkpoint trajectory in Appendix C address representation choice under a fixed budget.

A traceable qualitative example. For scene0038_00, situation QID 220602010683 and target `OBJ000` (chair, GT box row 10), the explicit front/behind question yields *Behind* with world geometry, *Front* with the original GT-relative geometry, and *Behind* with the same-scene donor geometry. The original geometric label is front and the donor label is behind. This illustration, used in Figure 1, comes from a four-question, two-scene training-source pilot.

F Cross-Dataset Protocol and Results

Evaluation domains and inputs. The fixed seed-42 matched checkpoints are evaluated on MSQA validation questions from 64 ScanNet scenes (Dai et al., 2017), 46 3RScan scenes (Wald et al., 2019), and 72 ARKitScenes scenes (Baruch et al., 2021), and on Spartun3D object validation in 44 physical locations (144 scans). These sets have zero exposure in the nine QA training configurations; 3RScan-family separation follows dataset provenance. The evaluated MSQA sets contain 890, 622, and 579 questions, respectively; Spartun3D contains 9,751. The 3RScan, ARKitScenes, and Spartun3D protocols use zero placeholders for unavailable offline 2D features, and mask that branch identically in every compared cell. The transferred 3D features and geometry remain available. Reported effects are within-domain contrasts under this feature protocol.

Viewpoint conventions. MSQA viewpoints follow the dataset generator's $z = 0$ convention and are injected as provided. For Spartun3D we reconstruct each pose from its situation construction, at the anchor box's height and with one of four axis-aligned headings. Frame checks report 100% pose-in-AABB rates on the 3RScan and ARKit domains, 92.1%/92.3% bearing reverse-check rates, and 96.2% yaw–label agreement on Spartun3D. These conventions determine the geometry under which each within-dataset intervention is interpreted.

Identical support for the four-cell interaction. The main transfer interaction is a post-hoc reanalysis of saved predictions. For MSQA, a common-support question has a canonicalizable reference

Table 21: Supplementary sensitivity analyses on the same human-reviewed groups. Entries are percentage points with paired 95% scene-bootstrap intervals.

Group	Seed	Refined-EM Δ	EM1 interaction I
Dependent	42	$+7.92$ $[+6.50, +9.36]$	$+7.47$ $[+5.92, +9.06]$
Dependent	43	$+8.89$ $[+7.43, +10.43]$	$+7.81$ $[+6.18, +9.44]$
Non-dependent	42	-1.95 $[-5.23, +1.31]$	-1.62 $[-5.34, +2.17]$
Non-dependent	43	$+0.32$ $[-2.85, +3.48]$	-0.32 $[-3.48, +2.67]$

Table 22: Training-source direction diagnostic. The last column uses only questions whose two valid labels differ.

Input geometry	Original-label accuracy $n = 512$	Donor-label accuracy $n = 455$	Changed-label agreement $n = 227$
World frame	50.2	48.6	47.1
Original GT viewpoint	84.2	49.9	15.4
Substituted viewpoint	49.4	85.1	85.9

and parseable predictions under the frozen scorer in all four cells. The same question identities are used for every cell mean and bootstrap replicate. Table 24 gives the cells and Table 25 their joint interaction intervals. Spartun3D includes its full evaluated set. All means and intervals are question-weighted; repeated scans remain grouped by physical location during resampling.

Reference-defined sensitivity check. Because common support also selects using model predictions, we separately retain every canonicalizable reference and count unparseable predictions as incorrect. This post-hoc analysis changes the missing-prediction treatment while retaining the frozen canonicalization rules. Table 26 shows positive interaction estimates in all three domains. Its denominators and scoring rule differ from Table 24; both analyses use canonicalizable reference subsets.

Parsing coverage and physical-scene grouping. Table 27 separates reference coverage from prediction parsing on the reference-defined support. Parsing success is not answer correctness. The 46 evaluated MSQA-3RScan scan IDs map one-to-one to 46 official physical reference groups; the reference-defined subset likewise contains 44 scans in 44 groups. Scene and physical-group clustering therefore coincide on these evaluated supports. Spartun3D retains its separate 144-scan/44-location grouping.

Prespecified pairwise contrasts. The initial transfer analysis estimates the pose effect within the pose-trained model and the injection effect within the no-pose model on their respective scoreable pairwise intersections. The MSQA pose/injection denominators are 303/328, 251/255, and 300/307. Table 28 preserves those contrasts separately from the common-support interaction. For example, the Spartun3D pose effect is $+0.77$ points, with a location-clustered CI of $[0.13, 1.42]$; the corresponding injection effect is -0.52, CI $[-0.94, -0.12]$. Direction-question filtering on MSQA-ScanNet gives a pose effect of $+5.62$, CI $[1.15, 10.42]$.

Independent grader checks. Two locally deployed graders, Llama-3.3-70B-Instruct and GLM-4-32B-0414, score the three MSQA domains using the OpenEQA correctness prompt (Majumdar et al., 2024). They see each prediction separately, without knowing which model produced it. A mark of at least 4 on the five-point scale counts as correct. Each grader preserves the signs of all six pose/injection contrasts (Table 29). Agreement with frozen EM on canonical subsets is 96.2%/97.2%/95.6% for Llama and 96.9%/95.3%/94.95% for GLM across ScanNet/3RScan/ARKit; the final GLM value is marginal against the prespecified 95% agreement threshold. After a recovery pass, 1.2% of Llama marks remain unparseable; GLM parses all marks. These checks assess contrast robustness within each domain.

F.1 Source-domain 2D feature-branch control

We tested whether the SQA3D training–inference geometry interaction persisted when the projected 2D object-feature branch was zeroed at inference. Each seed used the same four check-

Table 23: Channel interventions on the 512-question direction diagnostic. The final column compares each condition's outputs under its own supplied input with the changed-label donor targets. Raw-pose substitution is inference-only.

Supplied information	Accuracy	Changed-label donor agreement
World frame	50.2	47.1
Raw viewpoint values	54.5	43.6
Relative displacement, bearing pair removed	78.1	21.6
Full relation geometry	84.2	15.4

Table 24: Transfer EM on identical four-cell support. A_{ti} indexes training and inference geometry as in Equation 6. MSQA support is conditional on reference and prediction canonicalizability.

Domain	Questions	Scenes/scans	A_{00}	A_{01}	A_{10}	A_{11}
MSQA-ScanNet	293	64	39.25	38.57	35.49	39.93
MSQA-3RScan	243	43	33.74	30.86	33.74	36.21
MSQA-ARKit	291	66	37.80	34.02	37.11	39.18
Spartun3D	9,751	144	29.71	29.19	30.73	31.49

point/geometry cells as its 2D-on references, with 3,261 official-validation questions from 65 scenes. Question identities, prompts, deterministic decoding, and the answer scorer were matched within each seed. Seeds are analyzed separately.

The intervention zeroed the projected 2D feature after projection, positional addition, and validity masking. The 2D token slot remained in the three-token object layout; 3D features, pose inputs, and geometry were unchanged. Actual-forward checks verified nonzero 2D embeddings before intervention and exactly zero packed 2D vectors afterward. This is a fixed-checkpoint inference ablation, not token removal or training without images.

The interaction remained positive with the branch zeroed for both seeds (Table 30). Its change was inconclusive for seed 42 and negative for seed 43, in both metrics. Thus, the source-domain interaction persisted under this intervention, but its change was not stable across seeds. The intervals condition on the fixed checkpoints and quantify scene-sampling uncertainty only.

G White-Box Representation Diagnostics

Readout probes complement the behavioral interventions by testing whether selected geometric variables are accessible from hidden states. They use 64 training-source scene families, split into 48 fitting and 16 development families, with scene-family bootstrap intervals. Ridge readouts recover world XYZ from the 3D token with 0.44 m mean error under world-frame geometry, versus 1.10 m after removing geometry and 2.14 m for a constant predictor. Two-layer MLP readouts ($4096 \to 256$) agree with ridge within 0.01–0.06 m, while label-shuffled controls produce 1.99–2.11 m error.

A prompt-conditioned study uses 2,048 object–question probes with global-input, situation-swap, question-swap, object-binding-swap, and GT-input conditions. Table 31 reports the development sign readouts. World positions remain readable, while the supplied frame makes horizontal observer-relative signs much more accessible. Vertical relative signs are accessible in both conditions, consistent with their weaker dependence on heading. These outcomes characterize the tested probe families and hidden-state selections.

A restricted activation-patching analysis uses FP32 projection into a rank-three world-XYZ subspace at transformer block 18. It adds the projected clean-minus-corrupted difference to the corrupted run and measures agreement with the clean answer. Among 641 changed answers, the coordinate-subspace, norm-matched random, and full-hidden patches restore 15, 13, and 113 answers, respectively. Across all 3,261 questions, coordinate-minus-random agreement is -0.03 points, CI $[-0.42, 0.37]$. The selected world-coordinate subspace therefore does not provide a stronger repair than this random control under the tested patch. Behavioral frame-use claims rely on the direct

Table 25: Four-cell transfer interactions on common support. Spartun3D resamples 44 physical locations; MSQA resamples scenes. All analyses use 10,000 joint bootstrap draws.

Domain	Interaction (pp)	95% CI
MSQA-ScanNet	$+5.12$	$[1.21, 9.33]$
MSQA-3RScan	$+5.35$	$[-1.26, 12.31]$
MSQA-ARKit	$+5.84$	$[-0.98, 13.19]$
Spartun3D	$+1.29$	$[0.48, 2.14]$

Table 26: Reference-defined support, with unparseable predictions counted as incorrect. The ARKit lower bound rounds to 0.00.

Domain	Questions	Interaction (pp)	95% CI
MSQA-ScanNet	397	$+4.79$	$[1.52, 8.44]$
MSQA-3RScan	282	$+6.74$	$[1.17, 12.55]$
MSQA-ARKit	343	$+5.83$	$[0.00, 12.09]$

input interventions; the representation analysis characterizes selected readable features and their intervention effects.

G.1 Full decoder-layer readout scan

We extended the internal readout diagnostic to all 36 decoder blocks of the original Qwen3-VL-8B diagnostic checkpoint (SHA-256 prefix `884f10c8`; Table 5). The frozen selection contained 2,048 proposal–question records from 256 questions and 64 scene families. Five input conditions (Global, situation swap, question swap, binding swap, and GT) were extracted at three object-token sites per block: the projected-3D-feature token, pre-query identifier, and post-query identifier. These are hidden states at the specified token positions in every decoder block.

Readout and controls. A fixed ridge readout ($\alpha = 1$) was fitted on 1,536 records from 48 scene families and evaluated on 512 records from 16 held-out probe scene families. The primary metric was Euclidean error in the observer-forward/left plane, in metres. We defined each contrast as control error minus readout error, so a positive value favored the named readout. Using 5,000 bootstrap resamples of development scene families (seed 42), we formed 95% max-statistic intervals separately for three prespecified contrast families: 108 Global-versus-training-mean comparisons (36 layers $\times$ three sites), 36 post-query-versus-pre-query comparisons, and 36 post-query-versus-within-scene-situation-label-shuffle comparisons. The GT condition remained a separate ground-truth-conditioned positive control.

Results and interpretation. No Global layer/site showed a familywise-supported improvement over the training-mean predictor (Table 32, Figure 8). Neither the pre/post-query comparison nor the situation-label-shuffle control supported a post-query advantage after correction. The full scan thus supplies no positive evidence for these specific linear-readout advantages. The 16 development families were held out from fitting the readout; the frozen checkpoint may have encountered the source scenes. This is an internal accessibility diagnostic.

Extraction and statistical verification. All 15 condition/site layer-18 feature slices matched an earlier independent extraction exactly, and separate three-dimensional relative-coordinate ridge readouts reproduced its layer-18 errors within 1.11×10^{-9} m; those XYZ errors are verification targets, not the primary forward/left metric above.

H Exploratory Semantic-Pair and Object-Binding Pilots

These inference-only pilots used the two existing Qwen3-VL-8B checkpoints at seeds 42 and 43. The model inputs excluded reference answers and geometry-derived labels; those labels were used only for scoring after generation. The pilots are separate from the primary held-out direction bench-

Table 27: Reference coverage (Ref.%) uses the full evaluated set; prediction parsing percentages use the reference-defined subset in each cell. Unparseable predictions remain incorrect in the reference-defined EM analysis.

Domain	Ref./Full	Ref.%	A_{00}	A_{01}	A_{10}	A_{11}
MSQA-scannet	397/890	44.61	83.88	83.12	78.59	80.60
MSQA-3RScan	282/622	45.34	91.13	91.49	91.49	91.49
MSQA-ARKit	343/579	59.24	90.38	90.38	88.92	90.09

Table 28: Original pairwise transfer contrasts, in percentage points with paired 95% CIs. Pose effect is $A_{11} - A_{10}$; injection effect is $A_{01} - A_{00}$, each on its own scoreable pairwise intersection.

Domain	Pose effect	Injection effect
MSQA-ScanNet	$+4.29\ [0.64, 8.45]$	$-0.61\ [-1.81, 0.58]$
MSQA-3RScan	$+3.59\ [-1.18, 8.66]$	$-2.75\ [-5.99, 0.73]$
MSQA-ARKit	$+2.00\ [-3.56, 7.79]$	$-3.58\ [-7.12, -0.30]$
Spartun3D	$+0.77\ [0.13, 1.42]$	$-0.52\ [-0.94, -0.12]$

mark and are too small to support population-level inference. The earlier R04 and R05 Qwen batches used inherited generation defaults, whose base configuration enabled sampling. We distinguish those exploratory batches from the deterministic R04 expansion and R05 replays below.

H.1 Single-scene object-binding intervention (R04)

The R04 pilot used the human-approved pair of distinct tables in scene0064_00, assigned to proposal slots 51 and 23. The approved descriptions were "the small side table between the brown armchair and the black sofa" and "the large rectangular coffee table beside the brown armchair" (the prompts used the reviewed Chinese-language versions of these descriptions). We used pose QID 220602000227 only as a geometric anchor and removed its original situation text. Each target was queried using either its explicit object token or its reviewed natural-language description, under aligned features or a bilateral swap of the complete 3D and 2D feature vectors for slots 51 and 23, and under the correct heading or a 180° counterfactual heading. The feature swap retained geometry, object identifiers, masks, prompts, and all other scene inputs. This yielded 16 answers per checkpoint, with two target answers per table cell.

Both seeds answered the object-token conditions correctly for both tables under both headings. Under natural-language references, correct-heading accuracy was 0/2 and 1/2 for seeds 42 and 43. Swapping both visual feature channels changed none of the 16 matched outputs across targets, references, and headings. This pilot contains a single reviewed pair in one scene.

H.2 Six-scene object-binding expansion (R04)

We expanded the constructed R04 diagnostic to six additional scenes using the same two Qwen3-VL-8B checkpoints and an existing Vicuna-7B checkpoint. Of ten visually reviewed candidate pairs, eight were retained, one remained unresolved, and one was rejected. Six retained pairs admitted an existing validation pose with opposite left/right target labels and an absolute lateral margin of at least 0.30 m for each target. These six pairs were frozen before inference; proposal-center and ground-truth-center labels agreed for all twelve targets. The two remaining visually retained pairs lacked an eligible pose under this contrast and were not included. Selection did not use model predictions.

The reviewed English descriptions were used verbatim. The original situation text and reference answers were omitted; the human record that situation leakage could not be assessed from the visual cards was retained separately from the audit of the actual model input. Each checkpoint generated 96 answers: two targets, two reference forms, two feature-binding conditions, and two supplied headings in each scene. All conditions disabled sampling and used five-beam decoding, with repeated-output and token-map checks. The earlier Chinese-description single-scene pilot retained inherited generation defaults and is reported separately. The new runs use supplied geometry. The Qwen and

Table 29: MSQA contrasts under two independent graders, in percentage points with scene-paired 95% CIs.

Grader	Domain	Pose effect	Injection effect
Llama-3.3-70B	ScanNet	$+3.44\ [1.44, 5.48]$	$-0.34\ [-1.44, 0.78]$
	3RScan	$+2.64\ [-0.17, 5.88]$	$-1.14\ [-3.02, 0.68]$
	ARKit	$+2.99\ [-0.95, 7.26]$	$-1.22\ [-3.35, 0.76]$
GLM-4-32B	ScanNet	$+2.02\ [0.23, 3.96]$	$-0.56\ [-1.64, 0.47]$
	3RScan	$+2.25\ [0.00, 4.70]$	$-0.96\ [-2.37, 0.37]$
	ARKit	$+1.73\ [-1.45, 5.01]$	$-1.04\ [-3.12, 0.89]$

Table 30: Source-domain geometry interaction with the 2D feature branch on and zeroed. All estimates and paired 95% scene-bootstrap intervals are in percentage points. $\Delta I = I_{\text{off}} - I_{\text{on}}$, with I defined in Equation 6. Intervals use the same 10,000 scene draws across the eight cells within each seed; these jointly recomputed intervals can differ slightly from the original four-cell analysis.

Seed	Metric	I_{on}	I_{off}	ΔI
42	EM1	6.56	7.24	$+0.67$
	95% CI	$[5.06, 8.05]$	$[5.89, 8.64]$	$[-0.95, +2.35]$
43	EM1	7.08	3.43	-3.65
	95% CI	$[5.59, 8.52]$	$[1.64, 5.32]$	$[-5.37, -1.87]$
42	Refined EM	7.39	7.45	$+0.06$
	95% CI	$[5.86, 8.93]$	$[6.10, 8.92]$	$[-1.70, +1.87]$
43	Refined EM	7.67	3.86	-3.80
	95% CI	$[6.19, 9.08]$	$[2.07, 5.76]$	$[-5.51, -2.05]$

Vicuna checkpoints keep their own geometry modes and training histories, so this is a cross-system check under the same query and intervention interface.

Under aligned features and the correct heading, object-token versus description matches were Qwen3-VL-8B seed 42: 8/12 versus 4/12; Qwen3-VL-8B seed 43: 10/12 versus 7/12; Vicuna-7B seed 42: 10/12 versus 5/12. Natural-description answers changed after the visual-feature swap on Qwen3-VL-8B seed 42: 2/24; Qwen3-VL-8B seed 43: 3/24; Vicuna-7B seed 42: 1/24 target-by-heading comparisons. These are repeated comparisons within the same six scenes.

The natural descriptions often identify objects through scene anchors, such as a bed or desk. Exchanging target feature bundles while retaining anchor geometry need not move the referent to the other slot. We therefore report original-slot matches, opposite-slot matches, and paired answer changes separately. Opposite-slot and original-slot matches are complementary for valid binary answers.

H.3 Complete-object sequence-order control (R04)

We added the planned sequence-order control on the same six reviewed scene pairs. After geometry fusion and image masking, we reversed the order of the complete three-token object packages: identifier, geometry-fused 3D feature, and masked 2D feature. Object identities and their feature/geometry bindings, prompts, target labels, and supplied poses remained unchanged. Padding was excluded by the original packing function. The original ordering was rerun alongside the reversed ordering under fixed decoding. All 144 outputs passed input and per-object package-identity checks; the 72 original-order outputs exactly reproduced their counterparts in the preceding six-scene experiment.

Object-ID answers did not change under this reversal in any of the three checkpoints. Natural-description answers changed on 2/12 targets for Qwen seed 42 and on 0/12 for Qwen seed 43 and Vicuna. For Qwen seed 42, the two correct-heading feature-swap changes and the two order-control changes overlap on one target. Sequence sensitivity therefore remains an alternative contributor to some description-conditioned answer changes.

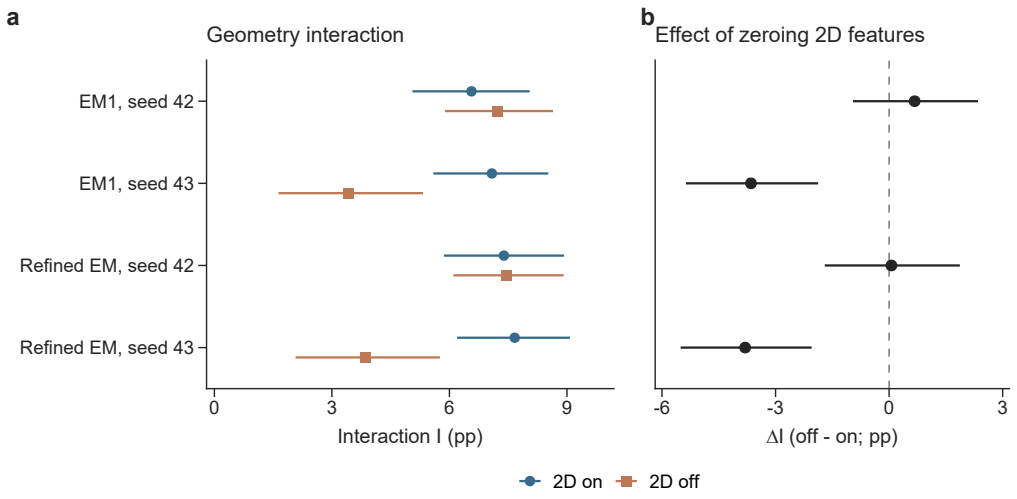


Figure 7: **Geometry interaction under a source-domain modality intervention.** (a) Training–inference interaction I with the projected 2D branch enabled or zeroed. (b) The paired change $I_{\text{off}} - I_{\text{on}}$. Each seed uses the same 3,261 official-validation questions from 65 scenes. Bars are paired 95% scene-bootstrap intervals from 10,000 shared draws, conditional on the fixed checkpoints. Zeroing retains the 2D token slots and is applied only at inference. The change is inconclusive for seed 42 and negative for seed 43.

Table 31: Balanced sign accuracy for prompt-conditioned readouts on the development probes.

Input	Forward/backward	Left/right	Above/below
World frame	54.4	48.5	97.4
GT-relative frame	89.5	90.0	92.0

H.4 R04: ADDITIONAL REVIEWED OBJECT PAIRS

Of eleven new candidate pairs, we adopted nine after an AI-assisted visual review; two remained unresolved because of competing table references or an unclear natural cabinet-unit boundary. Descriptions were revised when needed to use visually supported distinguishing attributes and were frozen before inference. The original situation text was omitted, as in the earlier R04 protocol. All nine pairs passed the existing-QID, opposite-label, 0.30 m lateral-margin, and proposal/ground-truth label-consistency checks.

Eight scenes form the new R04 diagnostic batch; one scene previously used in R05 is reported separately as development support. All are from existing validation scenes. Three existing checkpoints each produced 144 answers across two targets, two reference forms, two visual bindings, and two supplied headings. All 432 outputs passed the frozen identity and intervention checks.

Qwen, seed 42 obtained 15/16 ID and 13/16 description matches under the original heading and aligned features. Qwen, seed 43 obtained 15/16 ID and 13/16 description matches under the original heading and aligned features. Vicuna, seed 42 obtained 14/16 ID and 10/16 description matches under the original heading and aligned features. These counts are reported without pooling with the earlier six-scene batch. Many descriptions identify a target through another scene object. Swapping target feature bundles while retaining anchor geometry need not move the referent. This expansion tests supplied-frame use and reference sensitivity.

H.5 SEMANTIC VIEWPOINT-PAIR PILOT (R05)

The R05 manifest contains three changed-label and three invariant cases over four scenes and eight QIDs. Each prompt was evaluated with its own pose and its paired QID pose. A and B denote the

Table 32: Full-layer Global readout diagnostics. Counts refer to familywise 95% intervals excluding zero. Positive reductions favor the readout over the named control. Layer-18 values in the final two rows are examples from the complete scan, not selected best layers.

Contrast family	Tests	Result
Readout vs. training-mean predictor	108	All estimates negative (-0.2241 to -0.0134 m); 33 corrected negative, 0 corrected positive.
Post-query vs. pre-query	36	0 exclude zero; layer 18: $+0.0357$ m, $[-0.0511, +0.1226]$.
Post-query vs. situation-label shuffle	36	0 exclude zero; layer 18: $+0.0101$ m, $[-0.0892, +0.1094]$.

Table 33: R04 single-scene table-pair diagnostic. Entries give correct answers / two target answers. The 180° condition reverses the geometry-derived left/right labels.

Seed	Reference	Visual binding	Correct heading	180° heading
42	Object token	Aligned	2/2	2/2
42	Object token	3D+2D swapped	2/2	2/2
42	Reviewed description	Aligned	0/2	1/2
42	Reviewed description	3D+2D swapped	0/2	1/2
43	Object token	Aligned	2/2	2/2
43	Object token	3D+2D swapped	2/2	2/2
43	Reviewed description	Aligned	1/2	1/2
43	Reviewed description	3D+2D swapped	1/2	1/2

two QID endpoints of a changed pair; joint accuracy counts a pair only when both endpoints are correct. The invariant cases reuse two changed-case QID pairs, so the six case pairs share scenes.

The changed-label joint score differed between the two checkpoints, while neither seed changed its answer on any of the three invariant cases. Only four scenes were evaluated, and two QID pairs contribute to both case types. This batch did not include Vicuna. Fixed-decoding replays, including Vicuna, are reported in Appendix H.7.

H.6 Pose-verified semantic pairs (R05 pass 5)

After human review of 16 observer poses (13 matches, three unresolved semantic anchors, and no contradictions), we retained five pairs with both A/B pose anchors confirmed: RR10, RB01, RB02, RB04, and RB08. Three B-side anchors remain unresolved (RR02, RB05, RB06) and are excluded; RB03 and RB07 remain dropped. The resulting inference-only diagnostic has six changed-label horizontal cases, ten QIDs, and four scenes. RB02 contributes two axes. No below/above questions or invariant controls are included. The reviewed cases use the same frozen situation, target, and question while replacing only the complete observer pose; references and labels remain outside model inputs.

Aggregated over the six cases, seed 42 obtains own-pose A/B/joint accuracy of 5/6, 5/6, and 5/6, with paired-pose accuracy of 10/12; seed 43 obtains 5/6, 4/6, and 4/6, with paired-pose accuracy of 9/12. Outputs change under pose substitution on 10/12 endpoint prompts for each seed; A/B self-pose answers switch on 6/6 and 5/6 cases, respectively. These small results vary across checkpoints. They cover only four scenes, omit invariant controls, and support no scene-bootstrap interval or generalization claim. The earlier three-changed/three-invariant pilot remains a separate, non-overlapping batch.

H.7 Fixed-decoding semantic-pair replays (R05)

We reran both previously reviewed batches with sampling explicitly disabled and inference seed 20260925, using the two existing Qwen checkpoints and the existing Vicuna checkpoint. The scene inputs, QIDs, situation texts, questions, target mappings, poses, and labels remained frozen. Each checkpoint–batch combination passed an exact repeated-output and token-map smoke check before

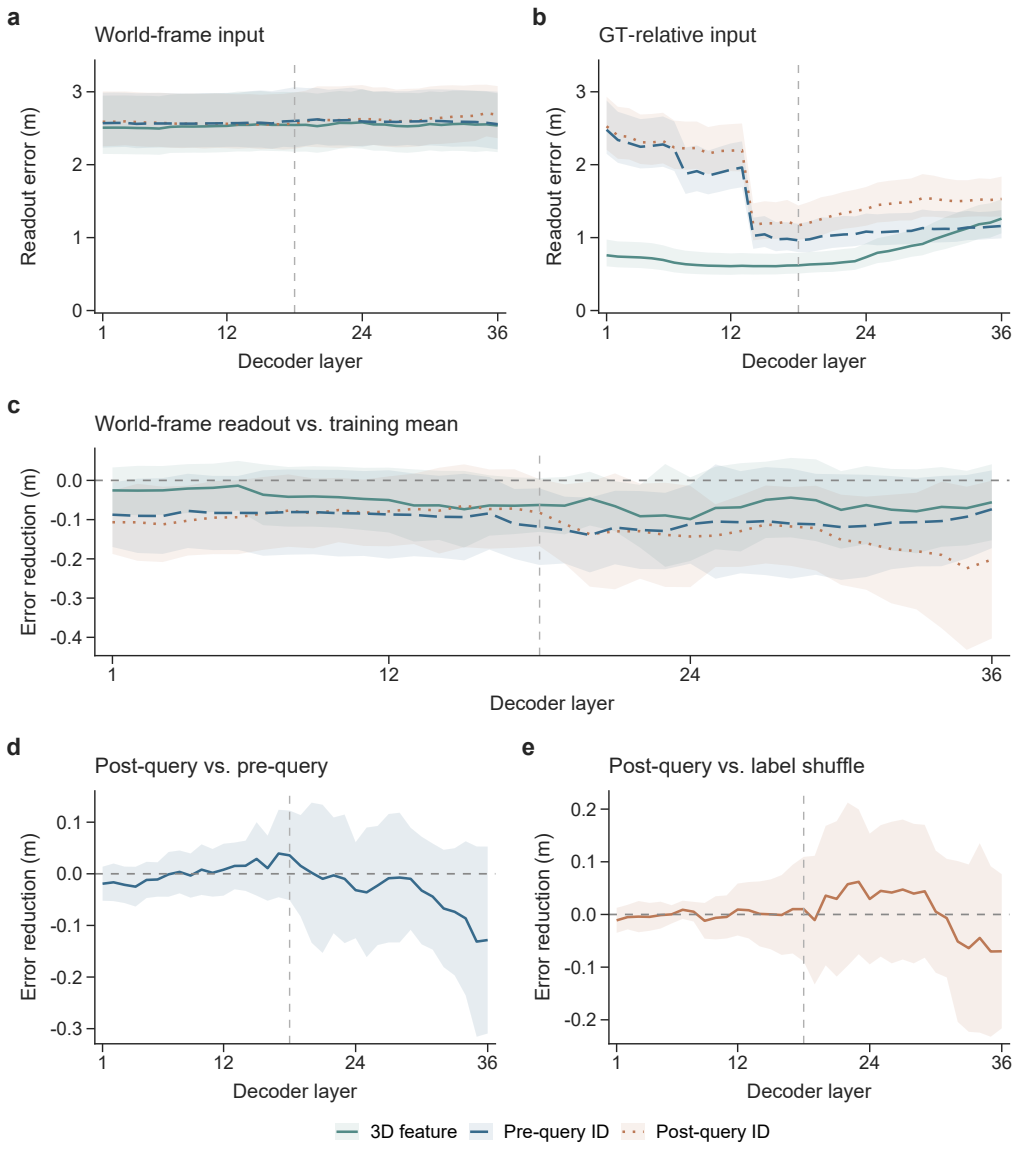


Figure 8: **Full decoder-layer scan of observer-relative linear readouts.** (a,b) World-frame and GT-relative inputs: mean forward/left Euclidean error with pointwise 95% intervals, on a common vertical scale. (c) World-frame readout versus the training-mean predictor; (d) post-query versus pre-query; (e) post-query versus within-scene situation-label shuffle. Contrast bands use familywise 95% max-statistic intervals within the respective 108-, 36-, and 36-test families; positive values favor the named readout. Vertical dashed lines mark the earlier layer-18 anchor. The 16 evaluation scene families are held out from probe fitting. No corrected Global improvement or post-query advantage is supported.

its full run. All 144 generated answers and their actual object-token mappings passed coverage and identity checks. The earlier sampled-decoding batches above are not pooled with these replays.

The pass-5 changed-label joint counts are 5/6 and 4/6 for Qwen seeds 42 and 43, compared with 0/6 for Vicuna. On this batch, answers change after pose substitution for 10/12 endpoint prompts in each Qwen checkpoint and 0/12 in Vicuna. Results vary substantially across the tested systems. In the pilot, all three checkpoints have zero invariant answer flips, but invariant joint correctness is 3/3, 2/3, and 2/3, respectively, so stable answers can still be wrong.

Table 34: R04 six-scene expansion. Entries count matches to the original target-slot geometry label out of 12 target queries per condition. The 180° column uses recomputed labels. Six scene pairs, rather than 96 generations, determine the sampling support.

Model	Seed	Reference	Features	Correct	180°
Qwen3-VL-8B	42	Object token	Aligned	8/12	10/12
Qwen3-VL-8B	42	Object token	3D+2D swapped	7/12	9/12
Qwen3-VL-8B	42	Description	Aligned	4/12	5/12
Qwen3-VL-8B	42	Description	3D+2D swapped	6/12	5/12
Qwen3-VL-8B	43	Object token	Aligned	10/12	11/12
Qwen3-VL-8B	43	Object token	3D+2D swapped	8/12	10/12
Qwen3-VL-8B	43	Description	Aligned	7/12	4/12
Qwen3-VL-8B	43	Description	3D+2D swapped	7/12	5/12
Vicuna-7B	42	Object token	Aligned	10/12	9/12
Vicuna-7B	42	Object token	3D+2D swapped	8/12	10/12
Vicuna-7B	42	Description	Aligned	5/12	6/12
Vicuna-7B	42	Description	3D+2D swapped	5/12	5/12

Table 35: Complete-package sequence reversal under aligned features and the correct supplied frame. Accuracy entries and answer changes each use 12 target queries over six scenes. Reversal changes sequence position while preserving every object's identity and evidence bundle.

Model	Seed	ID orig.	ID rev.	ID changes	Desc. orig.	Desc. rev.	Desc. changes
Qwen3-VL-8B	42	8/12	8/12	0/12	4/12	6/12	2/12
Qwen3-VL-8B	43	10/12	10/12	0/12	7/12	7/12	0/12
Vicuna-7B	42	10/12	10/12	0/12	5/12	5/12	0/12

Conditional correctness at B given a correct A is the joint count divided by the A-correct count (undefined if that count is zero).

These are reused, small diagnostic supports. The pass-5 cases include two axes from one QID pair and omit invariant cases. Situation A/B may differ in both observer position and heading, so the intervention is a complete-viewpoint substitution rather than a pure-yaw comparison. The model lines retain their different geometry implementations and training histories. No population interval or backbone-only causal conclusion is inferred from this table.

H.8 Native world-frame sibling control (R05)

We evaluated the corresponding existing world-frame-trained siblings on the same two reviewed semantic-pair supports. Within each model line, the world and GT training/validation annotations matched on QIDs, scene identities, question text, reference answers, and target IDs; geometry availability differed. We retained the matched epoch-2, step-26,346 endpoints. The world control supplies each original situation and question once, without an observer-pose input. Actual forward checks verified the native world geometry mode, zero observer-relative channels, unchanged world positions and sizes, and the absence of pose fields from model inputs. Loading checks verified every saved model tensor and coverage of all runtime trainable parameters, without asserting equality for unsaved base-model state.

Each checkpoint generated 12 answers per batch, yielding 72 new control outputs. Their prompt hashes exactly matched the corresponding GT replay prompts; decoding and visual-input contracts were held fixed. We did not duplicate the world prompts into nominal paired-pose conditions, since no pose is provided.

The world-trained checkpoints obtain zero changed-label joint successes in both batches. On pass 5, the GT-trained own-pose checkpoints obtain 5/6 and 4/6 for Qwen, whereas Vicuna remains at 0/6. Thus the reviewed pairs distinguish the tested Qwen conditions but not the Vicuna ones. This comparison jointly changes training exposure and native inference geometry.

H.9 R05: seven additional reviewed viewpoint pairs

We evaluated seven additional scene blocks, each containing two original situations, one changed-label question, and one invariant question about the same target. The resulting 14 question pairs use

Model	Reference	Aligned	Swapped	Aligned 180°	Swapped 180°
New R04 batch: eight scenes					
Qwen, seed 42	ID	15/16	15/16	15/16	15/16
Qwen, seed 42	Description	13/16	12/16	11/16	9/16
Qwen, seed 43	ID	15/16	15/16	16/16	16/16
Qwen, seed 43	Description	13/16	12/16	12/16	10/16
Vicuna, seed 42	ID	14/16	15/16	14/16	14/16
Vicuna, seed 42	Description	10/16	10/16	10/16	9/16
Development: one previously used scene					
Qwen, seed 42	ID	2/2	2/2	1/2	1/2
Qwen, seed 42	Description	2/2	1/2	1/2	1/2
Qwen, seed 43	ID	2/2	2/2	2/2	2/2
Qwen, seed 43	Description	2/2	0/2	2/2	1/2
Vicuna, seed 42	ID	1/2	1/2	2/2	2/2
Vicuna, seed 42	Description	1/2	1/2	1/2	1/2

Table 36: Additional R04 original-slot label matches. The eight-scene batch has 16 targets per condition; development has two. Labels are recomputed for the substituted heading. The full nine-scene counts and paired changes are listed in the per-query results to be released.

Table 37: R05 Qwen semantic-pair pilot with sampled decoding. Changed-label A/B/joint accuracies use three cases each; invariant accuracy and pose-change rates use three cases. Paired-pose accuracy uses 12 prompts.

Seed	Changed A/B/joint	Invariant accuracy	Paired-pose accuracy	Changed / spurious answer changes
42	3/3, 3/3, 3/3	3/3	12/12	3/3 / 0/3
43	1/3, 3/3, 1/3	2/3	9/12	1/3 / 0/3

horizontal relations only. We adopted an AI-assisted visual review of target uniqueness, situation–pose compatibility, and target-relation leakage. Original situation text was preserved. Two other candidate blocks did not meet the situation-anchor review criteria and were excluded before inference. The retained pairs can differ in both position and heading, so they test complete-viewpoint changes. These are new diagnostic questions from existing validation scenes.

We reused the two Qwen checkpoints and the Vicuna checkpoint with the frozen deterministic decoding protocol. Each GT-trained model received both the situation's own pose and its paired pose, giving 56 outputs per model. Each matched world-trained model received the same 28 situation–question prompts without observer-pose inputs. All 252 outputs passed identity, coverage, and direction-format checks. Table 41 reports the native-condition results separately from the earlier batches.

Both Qwen seeds answered both endpoints correctly in 6/7 changed-label pairs and all 7/7 invariant pairs. On the 14 changed-label endpoints evaluated under the substituted paired pose, each matched the supplied-frame label in 14/14 cases; Vicuna matched in 7/14. Vicuna obtained 0/7 changed-pair joint accuracy under both native GT and native world conditions. Thus this expansion supports supplied-frame use for the evaluated Qwen checkpoints. Native world versus native GT comparisons jointly change training exposure and inference geometry. The seven selected scene blocks form a small diagnostic sample.

H.10 Expanded reviewed reference and viewpoint diagnostics

A further review batch retained ten R04 object pairs from eight scenes and fourteen R05 viewpoint pairs from fourteen scenes. The R04 phrases incorporate the three author-specified anchor clarifications before inference. The R05 review retained twelve original unique-category cases and two qualified-reference cases (C09 and C10); three cases remained unresolved and two were excluded for directional leakage, including leakage through a target–anchor relation. Unresolved cases and both questions of any excluded pair are omitted from these results. Review decisions were made by the authors. All cases derive from existing official-validation assets.

The three existing GT-trained checkpoints and the matched world-trained R05 controls were reused without training. Source, target, prompt, checkpoint, and output-coverage checks passed for 480

Table 38: R05 pass-5 endpoint accuracy (sampled decoding) on the six human-approved changed-label cases. Each A/B entry is correct / one endpoint under its own pose or the paired endpoint's pose. Seeds are separate existing Qwen3-VL-8B checkpoints; no scores are pooled.

Pair / axis	Seed 42		Seed 43	
	Own pose (A/B)	Paired pose (A/B)	Own pose (A/B)	Paired pose (A/B)
RR10 / left–right	1/1, 1/1	1/1, 1/1	1/1, 1/1	1/1, 1/1
RB01 / front–back	0/1, 0/1	0/1, 1/1	0/1, 0/1	0/1, 0/1
RB02 / front–back	1/1, 1/1	1/1, 1/1	1/1, 1/1	1/1, 1/1
RB02 / left–right	1/1, 1/1	1/1, 1/1	1/1, 1/1	1/1, 1/1
RB04 / front–back	1/1, 1/1	0/1, 1/1	1/1, 0/1	0/1, 1/1
RB08 / left–right	1/1, 1/1	1/1, 1/1	1/1, 1/1	1/1, 1/1

Table 39: R05 fixed-decoding replay, keeping the two batches separate. A/B and joint count changed-label cases under their own situation–pose pairing. Invariant both counts cases with both endpoints correct; invariant flip counts different answers between those endpoints. Paired match is agreement with the substituted-pose label on 12 endpoint prompts, including invariant prompts where present. Off-diagonal situation–pose inputs can conflict, so these entries are label-match diagnostics. Each batch covers only four scenes.

Batch	Model	Seed	A / B	Joint	Invariant both	Invariant flip	Paired match
Pilot	Qwen3-VL-8B	42	3/3, 3/3	3/3	3/3	0/3	12/12
Pilot	Qwen3-VL-8B	43	1/3, 3/3	1/3	2/3	0/3	9/12
Pilot	Vicuna-7B	42	1/3, 2/3	0/3	2/3	0/3	8/12
Pass 5	Qwen3-VL-8B	42	5/6, 5/6	5/6	–	–	10/12
Pass 5	Qwen3-VL-8B	43	5/6, 4/6	4/6	–	–	9/12
Pass 5	Vicuna-7B	42	2/6, 4/6	0/6	–	–	6/12

R04 and 504 R05 predictions. This batch is reported separately from the earlier reviewed sets; two R04 scenes contribute multiple pairs. Decoding remains fixed, with sampling disabled and inference RNG seed 20260925, distinct from the checkpoint training seeds.

Under a 180-degree intervention on the supplied heading, aligned natural-description accuracy against the new frame's labels is 15/20, 18/20, and 13/20 for Qwen seed 42, Qwen seed 43, and Vicuna, respectively. Joint correctness across the original and intervened frames is 12/20, 16/20, and 6/20. This is a geometry-input intervention. Feature permutation also changes correct-frame ID answers on 4/20, 0/20, and 2/20 targets. The effects differ across checkpoints.

Both Qwen checkpoints often follow the supplied frame on changed-label cases, but invariant consistency is imperfect and differs between seeds (Table 43). Vicuna's two changed-pair successes occur in the qualified-reference stratum; its original unique-category stratum is 0/12. The two qualified-reference scenes are insufficient to infer an effect of description type. A low invariant-flip rate is also insufficient for correctness: a model may give the same wrong answer to both endpoints. These complete-viewpoint pairs may change position and heading together. They strengthen the behavioral audit while bounding claims of robust perspective reasoning, uniform cross-system behavior, or autonomous frame acquisition. The world comparison jointly changes the trained checkpoint and inference geometry and is interpreted at that scope.

H.11 Matched prompt-order and reference diagnostics

We applied the same prompt diagnostic to the Vicuna checkpoint and both existing Qwen checkpoints, crossing natural-language versus explicit object-token references with original versus reversed direction-option order. Each cell contains the same fourteen validation-derived scenes, with fourteen changed and fourteen invariant cases, two situation endpoints, and self/paired supplied poses (112 correlated outputs per cell). All three systems use identical case manifests, while retaining their original checkpoints, geometry modes, and native prompt templates. The original Qwen conditions exactly reproduced all 224 previously recorded prediction records. No training or outcome-based case selection was performed.

Table 40: Native world-trained control versus the GT-trained, own-pose replay on the same questions. G and W denote the GT-trained and world-trained checkpoints. Joint/both counts require both A and B endpoints to be correct. World endpoint accuracy includes changed and invariant cases when present. Batches and model lines remain separate.

Batch	Model	Seed	Changed joint G	Changed joint W	Invariant both G	Invariant both W	W endpoints
Pilot	Qwen3-VL-8B	42	3/3	0/3	3/3	0/3	5/12
Pilot	Qwen3-VL-8B	43	1/3	0/3	2/3	0/3	4/12
Pilot	Vicuna-7B	42	0/3	0/3	2/3	2/3	7/12
Pass 5	Qwen3-VL-8B	42	5/6	0/6	–	–	5/12
Pass 5	Qwen3-VL-8B	43	4/6	0/6	–	–	6/12
Pass 5	Vicuna-7B	42	0/6	0/6	–	–	6/12

Model	Geometry	Endpoint	Changed	Invariant	Inv. flip
Qwen, seed 42	GT	27/28	6/7	7/7	0/7
	World	14/28	1/7	3/7	1/7
Qwen, seed 43	GT	27/28	6/7	7/7	0/7
	World	17/28	1/7	4/7	1/7
Vicuna, seed 42	GT	16/28	0/7	4/7	1/7
	World	12/28	0/7	3/7	0/7

Table 41: Additional R05 diagnostics on seven scene blocks. Endpoint counts correct individual answers; Changed and Invariant count pairs with both endpoints correct; Inv. flip counts answer changes on invariant pairs (lower is better). Each row uses 28 endpoint answers, not 28 independent scenes. GT and World denote each model's native training and inference geometry. Review provenance and selection limits are described in the text.

For Qwen seed 42, option reversal changed 9/112 answers with natural references and 3/112 with object tokens. Changed-pair joint correctness held under both orders for 11/14 natural-reference cases and 12/14 object-token cases.

For Qwen seed 43, option reversal changed 12/112 answers with natural references and 16/112 with object tokens. Changed-pair joint correctness held under both orders for 9/14 natural-reference cases and 8/14 object-token cases.

For Vicuna seed 42, option reversal changed 97/112 answers with natural references and 20/112 with object tokens. Changed-pair joint correctness held under both orders for 0/14 natural-reference cases and 4/14 object-token cases.

These controlled wording interventions assess prompt sensitivity and referential representation. The twelve unique-category and two qualified-description scenes are reported separately in the results to be released.

Natural-reference option reversal changed 9/112 and 12/112 answers for the two Qwen checkpoints, compared with 97/112 for Vicuna. Qwen changed-pair joint correctness was 12/14 versus 12/14 for seed 42 and 11/14 versus 10/14 for seed 43, whereas Vicuna remained at 2/14 under both orders. Thus aggregate accuracy concealed substantial Vicuna instability, while the Qwen results constrain a first-option-following explanation for their original changed-pair performance. Residual invariant errors and prompt sensitivity remain visible.

This robustness is conditional: for Qwen seed 43 with object-token references, option reversal reduced changed-pair joint correctness from 12/14 to 8/14 and invariant joint correctness from 14/14 to 11/14. All sixteen changed answers were correct-to-incorrect transitions. Explicit identifiers therefore do not uniformly improve performance across orders and checkpoints.

H.12 Human-reviewed semantics under the 2D-branch intervention

We joined the sixteen 2D-on/off prediction files to the author-reviewed labels in Appendix D.5. The primary support contains 2,892 observer-dependent and 308 non-dependent questions; 37 unresolved questions and 24 earlier rubric-pilot questions remain separate. No questions were relabeled or selected using prediction outcomes, and no new model inference was needed. All 52,176 predic-

Table 42: **Expanded R04 diagnostic: ten pairs in eight scenes.** Each cell counts correct answers among twenty targets. ID and natural descriptions use aligned features and the correct frame; natural-swapped uses the same geometry and target-slot labels after pairwise visual-feature permutation. The last column counts changes from the aligned natural-description answers. With anchor-dependent descriptions, the intended referent after permutation can be ambiguous.

Checkpoint	ID	Natural	Natural-swapped	Answer changes
Qwen seed 42	16/20	16/20	15/20	1/20
Qwen seed 43	20/20	17/20	14/20	5/20
Vicuna seed 42	19/20	13/20	13/20	2/20

Table 43: **Expanded R05 diagnostic: fourteen reviewed scene pairs.** Changed and invariant columns require both self-pose endpoints to be correct. Invariant flips count an answer change although the geometric label is unchanged. Paired-frame agreement uses the substituted frame's label on the twenty-eight changed-case endpoint prompts. World joint is changed-pair joint correctness for the separately trained world-input checkpoint.

Checkpoint	Changed	Invariant	Inv. flips	Paired frame	World joint
Qwen seed 42	12/14	9/14	5/14	25/28	3/14
Qwen seed 43	11/14	13/14	1/14	25/28	1/14
Vicuna seed 42	2/14	7/14	1/14	17/28	0/14

tion records were rescored using the recorded answer scorer; question, scene, and reference identities matched, and the eight 2D-on cells exactly reproduced the records used in the semantic analysis.

For each group, $G = A_{11} - A_{10}$ measures the supplied-frame gain within the GT-trained checkpoint, and $I = (A_{11} - A_{10}) - (A_{01} - A_{00})$ is the training–inference interaction. We reuse the same 10,000 scene-bootstrap draws across all groups, branches, and cells, and compute $\Delta G = G_{\text{off}} - G_{\text{on}}$ and $\Delta I = I_{\text{off}} - I_{\text{on}}$. Intervals are post-hoc, pointwise, and conditional on the fixed checkpoints. Figure 9 shows the primary EM1 contrasts.

With the 2D branch zeroed, the dependent-group EM1 frame gain remains positive: $+7.19$ points (95% CI $[+5.76, +8.68]$) for seed 42 and $+4.60$ ($[+2.96, +6.31]$) for seed 43. The corresponding non-dependent gains are $+0.32$ ($[-3.53, +3.84]$) and -1.95 ($[-5.25, +1.32]$). The paired dependent-minus-non-dependent gain differences remain positive at $+6.87$ ($[+2.73, +11.12]$) and $+6.55$ ($[+3.32, +9.84]$).

Zeroing the branch changes the dependent-group gain by -0.10 points ($[-1.62, +1.47]$) for seed 42 and -3.77 ($[-5.27, -2.34]$) for seed 43. The analogous interaction changes are $+0.52$ ($[-1.24, +2.32]$) and -3.63 ($[-5.50, -1.87]$). Both seeds' between-group differences in these changes have intervals containing zero. Thus, semantic concentration of frame usefulness persists under this particular 2D intervention, while its magnitude can be reduced. The retained 3D features, learned interfaces, and text still supply information when the 2D branch is zeroed.

H.13 Same-Qwen language-adapter ablation and world-trained controls

To reduce reliance on a comparison between different backbone systems, we evaluated the same two GT-trained Qwen checkpoints with their language LoRA contributions switched off, alongside the two existing world-trained Qwen siblings. We retained all fourteen reviewed scene blocks and crossed the same natural/object-token references with original/reversed direction-option order. Each GT condition contains 112 outputs (two endpoints, changed/invariant cases, and self/paired poses); each native world condition contains 56, because no pose is supplied and duplicating self/paired inputs would add no intervention. The eight LoRA-off cells and eight world cells add 1,344 predictions. The full-model cells are reused unchanged from the prompt diagnostic.

The ablation sets the active scaling of all 252 unmerged language LoRA linear layers to zero. It preserves the learned token embeddings, object and image projectors, geometry/fusion parameters, and pretrained base weights. Actual-forward hooks verified that each affected layer's output exactly equaled its base-layer output on its first invocation, with all parameter identities and versions un-

Table 44: Matched prompt diagnostics. Joint correctness requires both self-pose endpoints to be correct. First-option counts include all 112 self/paired-pose outputs per cell.

System	Reference	Order	Changed joint	Invariant joint	First option
Qwen seed 42	Natural	Original	12/14	9/14	49/112
Qwen seed 42	Natural	Reversed	12/14	12/14	64/112
Qwen seed 42	Object token	Original	13/14	13/14	55/112
Qwen seed 42	Object token	Reversed	12/14	13/14	58/112
Qwen seed 43	Natural	Original	11/14	13/14	57/112
Qwen seed 43	Natural	Reversed	10/14	12/14	67/112
Qwen seed 43	Object token	Original	12/14	14/14	52/112
Qwen seed 43	Object token	Reversed	8/14	11/14	76/112
Vicuna seed 42	Natural	Original	2/14	7/14	104/112
Vicuna seed 42	Natural	Reversed	2/14	7/14	105/112
Vicuna seed 42	Object token	Original	8/14	11/14	41/112
Vicuna seed 42	Object token	Reversed	4/14	9/14	91/112

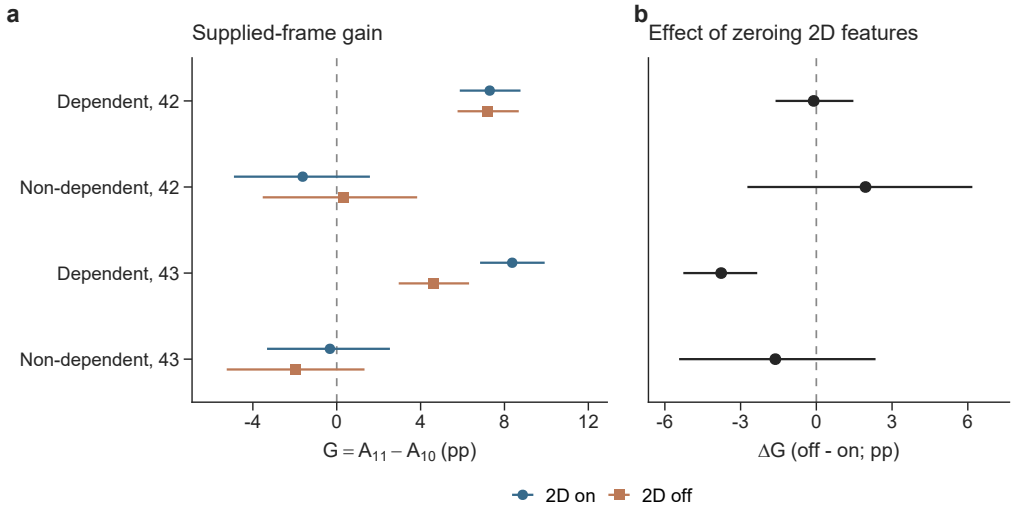


Figure 9: **Human-reviewed semantics under the 2D-branch intervention.** (a) Within-checkpoint EM1 frame gain $G = A_{11} - A_{10}$ with 2D features enabled or zeroed; (b) its paired change $G_{\text{off}} - G_{\text{on}}$. Row labels identify the semantic group and training seed. The unchanged human-reviewed support has 2,892 dependent and 308 non-dependent questions; 37 unresolved and 24 pilot questions remain separate. Bars show post-hoc, pointwise 95% intervals using 10,000 shared scene-bootstrap draws, conditional on fixed checkpoints. Dependent-group gains remain positive with the branch zeroed, while the reduction is clear only for seed 43.

changed during inference. Every saved checkpoint tensor was checked for exact loading. Thus, this is a language-adapter ablation *conditional on the trained input interface*. Turning off one co-adapted component can also induce an input–model mismatch.

The world-trained controls use their native world geometry, with observer-relative fields verified to be zero and agent-pose inputs absent. Their original natural-reference predictions exactly reproduce the previously recorded 112 outputs across both seeds. World-versus-GT comparisons jointly change the trained checkpoint and the inference geometry. All conditions retain the fixed beam-search settings, complete object-token input, and frozen question manifests. No new training, outcome-based selection, or label revision was performed. Figure 10 summarizes all diagnostic cells.

Table 45: Within-checkpoint supplied-frame gain under the 2D intervention, in percentage points. Dep. and Non-dep. denote the human-reviewed groups; group sizes are 2,892 and 308. EM-R is refined EM.

Seed	Group	Metric	G_{on}	G_{off}	ΔG	95% CI for ΔG
42	Dep.	EM1	7.30	7.19	-0.10	$[-1.62, +1.47]$
42	Non-dep.	EM1	-1.62	0.32	$+1.95$	$[-2.73, +6.18]$
42	Dep.	EM-R	7.92	7.61	-0.31	$[-1.91, +1.31]$
42	Non-dep.	EM-R	-1.95	0.32	$+2.27$	$[-2.44, +6.43]$
43	Dep.	EM1	8.37	4.60	-3.77	$[-5.27, -2.34]$
43	Non-dep.	EM1	-0.32	-1.95	-1.62	$[-5.43, +2.35]$
43	Dep.	EM-R	8.89	4.84	-4.05	$[-5.49, -2.66]$
43	Non-dep.	EM-R	0.32	-1.30	-1.62	$[-5.60, +2.53]$

Table 46: Training–inference interaction under the 2D intervention, in percentage points. Dep. and Non-dep. denote the human-reviewed groups; group sizes are 2,892 and 308. EM-R is refined EM.

Seed	Group	Metric	I_{on}	I_{off}	ΔI	95% CI for ΔI
42	Dep.	EM1	7.47	7.99	$+0.52$	$[-1.24, +2.32]$
42	Non-dep.	EM1	-1.62	0.97	$+2.60$	$[-2.19, +7.12]$
42	Dep.	EM-R	8.26	8.20	-0.07	$[-2.01, +1.89]$
42	Non-dep.	EM-R	-1.95	0.97	$+2.92$	$[-1.86, +7.50]$
43	Dep.	EM1	7.81	4.18	-3.63	$[-5.50, -1.87]$
43	Non-dep.	EM1	-0.32	-1.95	-1.62	$[-6.67, +3.78]$
43	Dep.	EM-R	8.37	4.60	-3.77	$[-5.59, -2.05]$
43	Non-dep.	EM-R	-0.00	-1.30	-1.30	$[-6.85, +4.32]$

For the original natural-reference prompts, switching off language LoRA reduces changed-pair joint correctness from 12/14 to 4/14 for seed 42 and from 11/14 to 3/14 for seed 43. The native world-trained controls obtain 3/14 and 1/14. The full grid is in Tables 48 and 49.

All 1,344 new inputs passed the complete-input and deterministic-generation checks. All 1,344 outputs contained exactly one distinct valid direction term, so the decline is not due to unparseable or conflicting direction answers. Four LoRA-off outputs included additional wording; the remaining 1,340 were single direction words after punctuation normalization. All outputs remain in the denominator under the unchanged parser. The results to be released also report first-option counts, invariant flips, fixed-text pose-pair correctness, and separate twelve-scene unique-category and two-scene qualified-reference strata. The controls reveal distinct failure patterns. Reversing natural-reference options changes 36/56 and 28/56 native world answers for seeds 42/43; both reversed conditions select the first option on all 56 outputs and obtain 0/14 changed-pair joint correctness. In contrast, the LoRA-off models change only 8/112 natural-reference answers under option reversal in each seed, despite their poor fixed-text pose-pair correctness (Table 51). Low option-order sensitivity alone is therefore insufficient evidence of frame use. The audit distinguishes order-driven answering from a loss of responsiveness to supplied poses without attributing either pattern to a unique internal cause. The within-checkpoint ablation supports dependence of this trained system's diagnostic behavior on the learned language adapters. The weaker and order-sensitive Vicuna results are reported in Appendix H.11.

I Qualitative cases with native predicted viewpoints

We selected 24 distinct official-validation scenes from the native world-frame, predicted, and GT prediction files, with one question per scene. Selection was post hoc and purposive across correct/incorrect outcome combinations, restricted to human-reviewed observer-dependent questions whose reference category maps to a single annotated instance. No new inference or label revision was performed. Figure 4 in the main text shows a repair, a GT-only success, a failure of the predicted-pose model, and a failure of all three models; Table 52 supplies the unchanged situations.

Table 47: **Same-model comparison on the 14 reviewed scene pairs** (original prompt). Joint correctness requires both endpoints of a pair to be correct; full and LoRA-off share each seed's checkpoint, the world row is a separately trained control.

Condition	Geometry at inference	Seed 42		Seed 43	
		Changed	Invariant	Changed	Invariant
Full Qwen	GT-relative	**12/14**	9/14	**11/14**	**13/14**
Language LoRA off	GT-relative	4/14	8/14	3/14	6/14
World-trained Qwen	Native world	3/14	5/14	1/14	5/14

Table 48: Changed-pair joint correctness. Each entry is out of 14 scene pairs and requires both situation endpoints to be correct. Full and LoRA-off use self-pose endpoints; World uses native no-pose inputs.

Seed	Reference	Order	Full model	LoRA-off	World
42	Natural	Original	12/14	4/14	3/14
42	Natural	Reversed	12/14	5/14	0/14
42	Object token	Original	13/14	2/14	7/14
42	Object token	Reversed	12/14	1/14	1/14
43	Natural	Original	11/14	3/14	1/14
43	Natural	Reversed	10/14	2/14	0/14
43	Object token	Original	12/14	1/14	0/14
43	Object token	Reversed	8/14	2/14	0/14

The mapped instance boxes agree with boxes recomputed from their segmented point-cloud vertices (IoU above 0.99999 for all 24 candidates). This verifies the visual annotation identity. The 24-case gallery, to be released with the code, includes the full situations, raw model answers, original exact-match scores, two views per scene, and source identities.

I.1 Proposal grounding on the same 24 scenes

For each of the 24 gallery scenes we issued one additional grounding query to the three native-input checkpoints: the original situation and question were retained and the model was asked to output the object identifier of the referent. The generated identifier maps to a fixed proposal box, which is compared with the annotated target box by 3D IoU; the target category, instance, box, reference answer, and QA prediction are not given to the model, and the GT-trained model receives the observer pose, not the target box. Figure 3 shows three cases for the predicted-pose model. The three models are separately trained checkpoints, and the 24 scenes were selected for illustration.

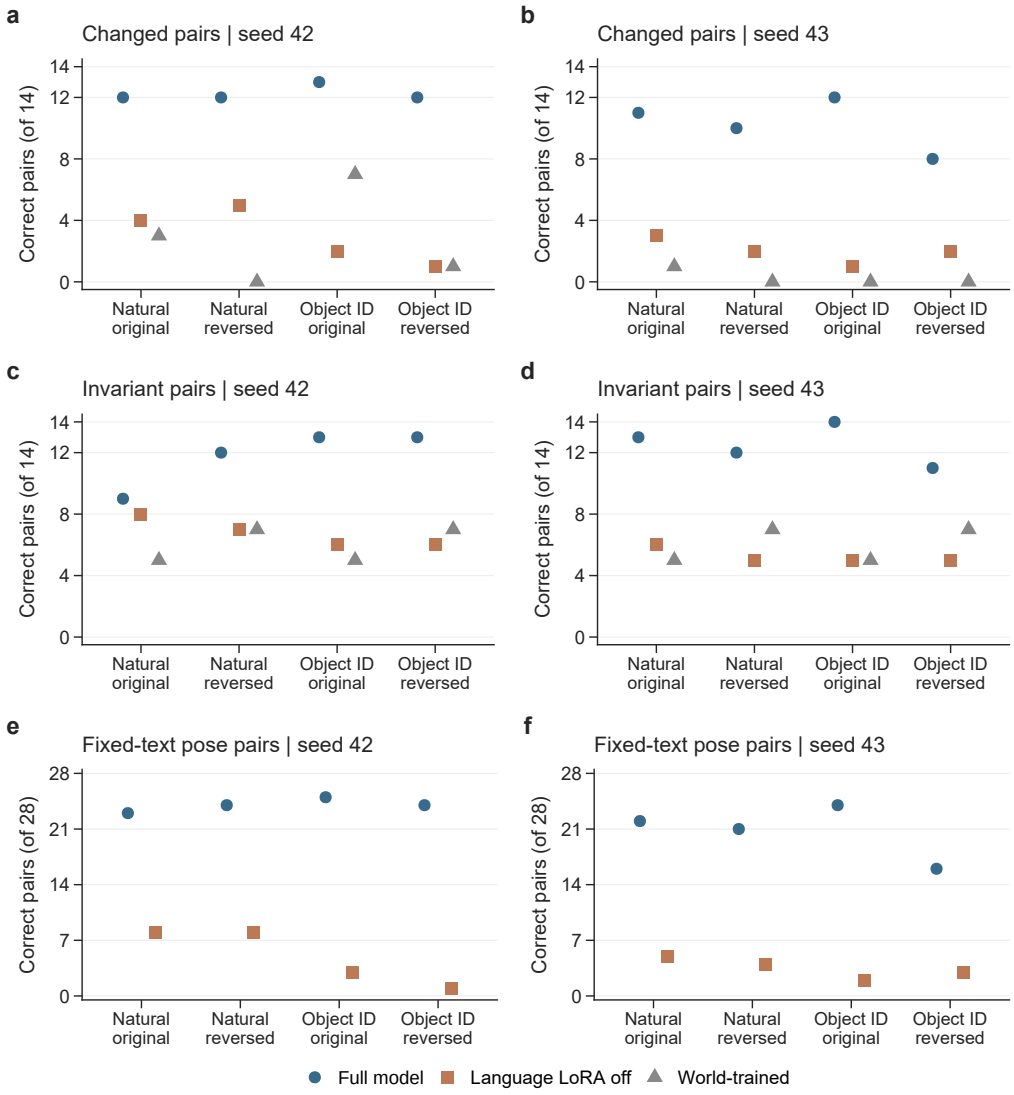


Figure 10: **Language-adapter dependence across the complete Qwen diagnostic grid.** (a,b) Changed-pair and (c,d) invariant-pair joint correctness: both situation endpoints have to be correct in each of 14 reviewed scene pairs. (e,f) Fixed-text pose-pair correctness: both supplied-pose conditions have to be correct for each of 28 prompts; World is not applicable because it has no pose input. Counts are correlated within fourteen selected validation-derived scenes. LoRA-off preserves the trained input interface and is not pristine pretrained Qwen. World changes both the trained checkpoint and inference geometry. Exact counts and paired transitions appear in Tables 48–51.

Table 49: Invariant-pair joint correctness. Each entry is out of 14 scene pairs and requires both situation endpoints to be correct. Full and LoRA-off use self-pose endpoints; World uses native no-pose inputs.

Seed	Reference	Order	Full model	LoRA-off	World
42	Natural	Original	9/14	8/14	5/14
42	Natural	Reversed	12/14	7/14	7/14
42	Object token	Original	13/14	6/14	5/14
42	Object token	Reversed	13/14	6/14	7/14
43	Natural	Original	13/14	6/14	5/14
43	Natural	Reversed	12/14	5/14	7/14
43	Object token	Original	14/14	5/14	5/14
43	Object token	Reversed	11/14	5/14	7/14

Table 50: Paired output changes when switching language LoRA from on to off. Each row matches all 112 self/paired outputs by question, case, condition, and exact prompt. Repair and damage mean incorrect-to-correct and correct-to-incorrect transitions. These counts are correlated within scenes.

Seed	Reference	Order	Changed answers	Repairs	Damages
42	Natural	Original	33/112	3	30
42	Natural	Reversed	32/112	0	32
42	Object token	Original	46/112	2	44
42	Object token	Reversed	51/112	2	49
43	Natural	Original	37/112	1	36
43	Natural	Reversed	39/112	0	39
43	Object token	Original	47/112	0	47
43	Object token	Reversed	34/112	2	32

Table 51: Fixed-text pose-pair correctness on changed cases. For each of 28 fixed prompts (two endpoints in each of 14 scenes), both the self-pose and paired-pose predictions have to match their respective labels. Only the supplied pose changes within each test; situation text, question, visual inputs, and target identity remain fixed.

Seed	Reference	Order	Full model	LoRA-off
42	Natural	Original	23/28	8/28
42	Natural	Reversed	24/28	8/28
42	Object token	Original	25/28	3/28
42	Object token	Reversed	24/28	1/28
43	Natural	Original	22/28	5/28
43	Natural	Reversed	21/28	4/28
43	Object token	Original	24/28	2/28
43	Object token	Reversed	16/28	3/28

Table 52: Unchanged situations for the four illustrated cases. Questions and answers appear in Figure 4; question identifiers link to the validation predictions.

Panel	Situation and question identifier
a	I am flattening the surface of my bed with the chair directly behind me. QID 220602008182
b	I sat at the table eating my lunch and to my right was another table. QID 220602003726
c	Stepping off the ladder at the head of the bed and going to head in between the cabinet and the desk. QID 220602029533
d	I am preparing my food on the kitchen counter. QID 220602016789