POSE SHORTCUTS AND FRAME REDUCTION: WHEN DO CANONICAL VIEWS HELP SPATIAL VERIFICATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Rendering a relation-conditioned canonical view of a reconstructed scene is a natural way to make a spatial relation visually verifiable. We ask when such a view actually helps a verifier, using fixed-reference relations in 3D Gaussian splatting reconstructions of ScanNet rooms, and find three answers. First, a canonical camera is a function of the queried geometry: for left–right and front–back the camera rotation alone determines the room-frame label in 100% of our views, so a trained verifier that receives the pose need not read the image. A pose-and-mask probe without RGB reaches 99.6% and 91.1% on canonical views of ten previously unused rooms and inverts to 0.2% and 9.6% when only the rotation is replaced by the mirrored camera's; the ResNet verifier itself inverts from 66.2% to 33.9% under the same mirrored rotation, so its front–back training-by-view interaction ($+11.05$ points) is a pose effect. Second, an analytic ceiling shows that a mask-conditioned verifier built on canonical views cannot beat the noisy geometry that placed its camera except when the noise reverses the pair direction, whereas gravity-aligned cameras keep the up–down frame conversion exact; with 1 m center noise and correctly marked objects, GPT-4o verifying up–down in such cameras beats the noisy geometry by 9 points. Third, for vision–language models the value of a canonical view is *frame reduction*: asking the question in the image's own frame and converting the answer with the known camera axes makes canonical views beat default views on left–right for all four models we test, by 6 to 23 points (90.1% for GPT-4o), while the room-frame interface stays at 50–58% in every view. Left–right pairs a model tells apart in the image frame are almost always right; depending on the model, the canonical gain comes from fewer undecided pairs, an exact frame conversion, or both. Canonical views help when they turn a frame-composition problem into a perception problem the verifier can solve; reported gains must be checked against pose shortcuts and geometry ceilings.

1 INTRODUCTION

An agent with a reconstructed scene can render any view before answering a spatial question. A side view separates two objects horizontally, an axial view exposes their depth order, and an eye-level view makes height differences vertical in the image. Analytic, relation-conditioned camera rules that exploit this are appealing components of compositional spatial reasoning, and earlier experiments of ours reported double-digit gains from them. This paper asks a more careful question: *when* does a canonical view help a verifier, and what exactly is it helping with?

Three observations structure the answer. (i) A canonical camera is computed from the geometry of the queried pair, so its pose encodes the queried direction. Any verifier that receives the pose can exploit this shortcut, and whether it does depends on which *identity* cue the architecture can read from the image, not on the spatial relation itself. (ii) If the camera is placed from estimated geometry, the verifier's evidence inherits the estimate's errors; how much a visual step can add is an analytic question with a relation-dependent answer. (iii) For a verifier that grounds entity names itself, such as a vision–language model (VLM), a canonical view changes the task from composing a perceived relation with a coordinate transform into a perception problem in the image's own frame.

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.

We study these observations in a controlled diagnostic: fixed-reference relations (left–right, front–back, up–down) on annotated entity pairs in 3D Gaussian splatting reconstructions of 40 development and ten previously unused ScanNet rooms, with balanced roles, retained failures and room-level uncertainty. Our contributions are:

- **A pose-shortcut analysis.** For canonical left–right and front–back cameras the rotation input determines the room-frame label in all views. A gradient-boosted probe on rotation and projected-mask statistics, without RGB, exceeds every cell of a 63-model ResNet-50 verifier matrix on canonical views and inverts under the mirrored camera's rotation; and the frozen front–back verifier itself inverts from 66.2% to 33.9% under the mirrored rotation with unchanged pixels, so its training-by-view interaction is a pose effect. The matrix's left–right null and up–down "counterexample" are the same mechanism: the pooled network reads vertical but not horizontal mask position, and the up–down models barely respond to the rotation.
- **An analytic ceiling.** An ideal image-frame verifier on cameras placed from noisy centers equals the noisy geometry solver for the horizontal relations except in the direction-reversal regime ($+3$ points at 0.5 m noise), while any horizontal camera keeps the up–down conversion exact ($+29$ points at 0.5 m for the analytic ideal). With real renders, 1 m per-coordinate center noise and correctly marked objects, GPT-4o verifying up–down beats the noisy geometry that placed its camera by about 9 points, while GPT-5.5 does not resolve a gain. Default and broad views cap converted accuracy at 74–77% for the horizontal relations.
- **Frame reduction for VLMs.** On the ten unused rooms, an image-frame prompt with analytic conversion makes canonical views beat default views on left–right for GPT-4o, GPT-5.5, Qwen3-VL-Plus and the open-weight Qwen2.5-VL-7B ($+22.8$, $+16.0$, $+11.3$ and $+6.3$ points), while the room-frame interface stays at 50 to 58%. A role analysis locates the gain: left–right pairs a model tells apart in the image frame are 92–100% correct, so the gain combines fewer pairs answered with one word regardless of role (substantially for GPT-4o) with an exact frame conversion (dominant for GPT-5.5 and Qwen3-VL-Plus). Front–back shows no gain because no model perceives depth order in these renders, and up–down, which every horizontal camera converts exactly, shows only model-specific perceptual differences.
- **A corrected protocol and audit.** Our earlier canonical-view comparisons had camera-dependent label distributions (e.g. 154 left versus 0 right targets under canonical cameras); the fixed-reference protocol removes this, and a two-hop executor study separates valid endpoints from valid chains (Appendix C).

The picture is conditional but mechanistic: canonical views help when they convert a frame-composition task into a perception task the verifier can solve, and apparent gains must be checked against the pose shortcut and the geometry that placed the camera.

2 Related work

Spatial understanding and frames of reference. ScanQA studies question answering over reconstructed scenes, while SQA3D adds a described situation and orientation to the question (Azuma et al., 2022; Ma et al., 2023); ReferIt3D separates view-dependent from view-independent references (Achlioptas et al., 2020). Rel3D and What'sUp change relations while holding object identity fixed, to reduce shortcuts and to document limitations of vision-language models (Goyal et al., 2020; Kamath et al., 2023). SpatialVLM studies training with large-scale spatial supervision, and VSI-Bench evaluates spatial understanding from video (Chen et al., 2024; Yang et al., 2025). VLMs are neither robust nor flexible across frames of reference (Zhang et al., 2025); APC and Allocentric Perceiver move the frame change outside the model, through a simulated viewpoint or explicit target-frame geometry (Lee et al., 2025; Wang et al., 2026). Frame reduction is the special case in which the target frame is the rendered camera's own (Appendix H). Our diagnostic isolates paired evidence-camera changes under a fixed geometric question and does not replace these broader tasks.

Shortcuts and position information. Shortcut learning describes decision rules that succeed on the training distribution for reasons unrelated to the intended task (Geirhos et al., 2020), and partial-input baselines expose such rules in visual question answering and natural language inference

(Agrawal et al., 2018; Poliak et al., 2018). The shortcut we document enters through a non-image input, the camera pose of a query-conditioned view, and is therefore invisible to image-only controls. Which identity cue a network can read from a mask channel is a question of position information: convolutional networks encode absolute position mainly through zero padding and boundary effects and struggle to map coordinates (Islam et al., 2020; Kayhan & van Gemert, 2020; Liu et al., 2018), and representations trained with horizontal-flip augmentation are largely insensitive to chirality (Lin et al., 2020). These properties are consistent with our finding that a globally pooled backbone reads vertical but not horizontal mask order, although we do not isolate their contributions.

Canonical views and active views. Canonical views were studied as preferred object views in human perception (Palmer et al., 1981), and graphics selects informative views by viewpoint entropy (Vázquez et al., 2003); our cameras are instead conditioned on a queried relation. 3D Gaussian splatting provides an explicit representation for novel-view rendering (Kerbl et al., 2023; Ye et al., 2025). ActiveNeRF and FisherRF select reconstruction views by uncertainty or parameter information (Pan et al., 2022; Jiang et al., 2024), objectives that differ from the correctness of a frozen relation verifier. Think3D and SplatReasoner use 3D manipulation or query-conditioned novel views within reasoning and typically pass camera parameters with the rendered view (Zhang et al., 2026; Kim et al., 2026); our analysis shows that for query-conditioned cameras those parameters can carry the answer. We do not evaluate those systems or infer their performance from our results.

3 Task, view policies, and the pose shortcut

Fixed-reference relations. For an entity i, let $c_i \in \mathbb{R}^3$ be the center of its annotated axis-aligned bounding box in the aligned ScanNet room frame (Dai et al., 2017). We define reference right as $+X$, away from the reference observer as $+Y$, and up as $+Z$. Thus $\mathrm{left}(i, j)$ means $c_{i,x} < c_{j,x}$, $\mathrm{front}(i, j)$ means $c_{i,y} < c_{j,y}$, and $\mathrm{above}(i, j)$ means $c_{i,z} > c_{j,z}$. Opposite directions complete three binary relation families. These are reference-observer center relations, not intrinsic object-facing directions or image-axis labels. Eligible pairs have distinct IDs, center distance in $[0.2, 3.0]$ m, and an absolute gap greater than 0.05 m on the queried axis.

Canonical and default evidence. Camera construction uses an unordered entity pair, consistently ordered by ID, so reversing the verbal subject and anchor leaves the physical evidence unchanged. Let h be the unit horizontal direction from the lower-ID entity i to the higher-ID entity j. Canonical left–right views look along $e_z \times h$, perpendicular to the pair; front–back views look along h with a 0.5 m height offset; up–down views look along h from eye level. Our default samples a deterministic azimuth while matching the canonical camera's radius and height offset; a broad-pose control additionally varies elevation. All cameras look at the pair midpoint. Exact rules are in Appendix A.1.

The pose shortcut. Because the canonical forward axis is a fixed function of h (Figure 1), the camera rotation determines h and therefore the room-frame label for the two horizontal relations: $\mathrm{left}(i, j)$ iff $h_x > 0$ and $\mathrm{front}(i, j)$ iff $h_y > 0$. In our evidence this holds for 100% of clean primary canonical views (1,490 left–right and 1,505 front–back views on 16 rooms); under default and broad cameras the same rule agrees with the label in 49.6–50.4% of views, and it carries no information for up–down because h has no height (Table 14, Appendix F). A verifier that receives the rotation therefore only needs to identify *which mask is entity* i: in canonical left–right views entity i is the image-left box (99.7%), in canonical front–back views it is the nearer box, which is larger (78.1%) and lower in the image (74.3%), and for up–down the higher box is the "above" entity under the horizontal canonical and default cameras (99.5% and 99.2%) and in 82.5% of the tilted broad views. Section 5 shows that the trained verifier's canonical-view effects follow this structure.

Information available to the verifier. The trained verifier receives rendered RGB, two subject/anchor projected-AABB mask channels, and the nine entries of the camera-from-aligned-world rotation. It does not receive entity centers, camera translation, depth, relation labels, visibility flags, or a camera-policy identifier. The masks and rotation are privileged geometric information, and by the argument above they are sufficient. A retrained boxes-only model receives the same masks and rotation with constant normalized RGB. All visual comparisons retain blank renders and missing boxes.

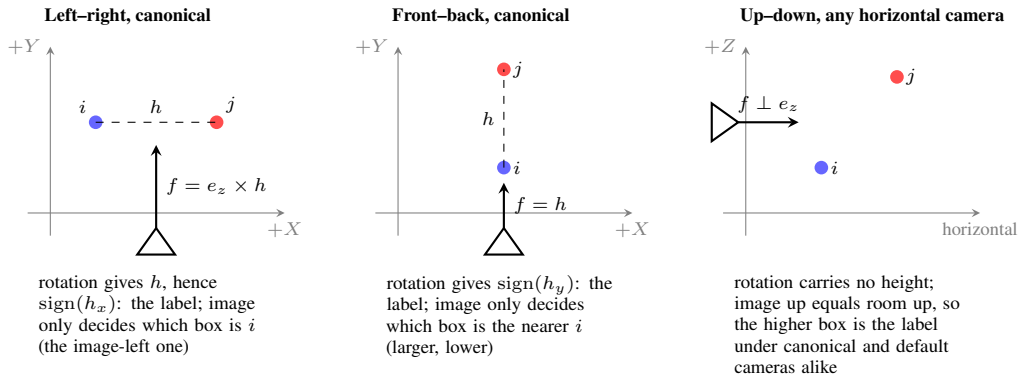


Figure 1: Why a query-conditioned camera leaks the answer. The canonical left–right and front–back cameras look along a fixed function of the pair direction h (from the lower-ID entity i to j), so the rotation input determines $\mathrm{sign}(h_x)$ or $\mathrm{sign}(h_y)$, which is the room-frame label up to the identity of the subject; the image only has to reveal which box is entity i. Up–down cameras are horizontal, so the rotation carries no height and the vertical image order is the label itself.

Geometric sufficiency. Because camera selection already has the centers, an explicit geometry solver answers the clean task; Section 6 quantifies what a visual step could add once the centers are noisy. A two-hop extension of the task is defined and analyzed in Appendix C.

4 Experimental Design

Cohorts and evidence. All 40 previously examined ScanNet rooms are development data, split into 28 training, six validation, and six development-evaluation rooms. Ten additional physical rooms were selected by a frozen availability rule before predictive outcomes, with no scene replacement by reconstruction quality. Their reconstructions use 400 photographs sampled across each trajectory and 30,000 optimization steps. The confirmation evidence run contains 26,130 records across the six development-evaluation rooms and ten unused rooms, with no reconstruction failure; the unused-room primary single-hop cohorts contain 923 left–right, 935 front–back, and 894 up–down unordered pair–relation units, each evaluated in both role orders. All primary results below concern the ten unused rooms unless stated otherwise.

Trained verifier matrix. We use an ImageNet-initialized ResNet-50 (He et al., 2016) with two additional mask channels, global average pooling, and a nonlinear head that combines pooled features with the rotation. Three training policies (canonical, default, equal-budget mixed), two modalities, three relations, and three seeds give 54 primary models; nine additional up–down presentation models give 63. All conditions use ten epochs of 5,000 ordered pair exposures with identical pair schedules across camera policies, validation-selected checkpoints, and no selection on the unused rooms (Appendix B).

Pose-shortcut probe and rotation interventions. A gradient-boosted tree (Pedregosa et al., 2011) receives only the rotation and twelve projected-mask statistics (centroid, width, height, area, missing flag for subject and anchor), trained on the 28 training rooms under the same three policies and evaluated on the same views as the matrix. At test time we keep the mask statistics fixed and replace the rotation by that of the camera mirrored through the pair midpoint ($R\,\mathrm{diag}(-1,-1,1)$) or by zeros; masks-only and centroid-only or size-only feature sets isolate the identity cues. The same two rotation interventions are applied to all 63 frozen checkpoints on the same views (Appendix F).

Analytic ceiling. For all eligible pairs of the 50 rooms, ten shared per-entity Gaussian draws scaled by $\sigma \in \{0, \ldots, 1.0\}$ m perturb the centers that place the cameras and feed a geometry solver. An ideal image-frame verifier reads the true relative placement inside the selected camera and converts it with the known camera axes (Appendix G). Labels and eligibility stay clean.

Table 1: Single-hop accuracy (%) and paired canonical-minus-default effect (percentage points) on ten unused rooms. RGB plus AABB masks and rotation; clean primary pairs, overlays off. Intervals are 95% room-bootstrap intervals.

Relation	Training views	Canonical	Default	Difference [interval]
Left–right	Canonical	48.88	49.53	-0.65 $[-2.22, 0.56]$
	Default	51.45	50.82	$+0.63$ $[-0.25, 1.44]$
	Mixed	50.74	50.65	$+0.10$ $[-0.66, 0.69]$
Front–back	Canonical	66.19	55.79	$+10.41$ $[8.13, 12.92]$
	Default	49.74	50.39	-0.64 $[-1.32, -0.06]$
	Mixed	53.45	50.99	$+2.46$ $[1.53, 3.37]$
Up–down	Canonical	94.27	94.92	-0.65 $[-1.38, 0.20]$
	Default	92.04	94.60	-2.56 $[-4.44, -1.07]$
	Mixed	87.65	91.30	-3.66 $[-4.71, -2.56]$

Prompt frame by evidence view for a VLM. On the ten unused rooms, 40 pairs per room and relation (first frozen pair identifiers), both roles, three evidence views and two prompt frames on the same pixels: the previously reviewed room-frame interface (header image with projected room axes and the numeric camera-from-room rotation in the prompt) and an image-frame prompt whose answer is converted outside the model by the sign of the cosine between the questioned camera axis and the room axis (Appendix D). GPT-4o (`gpt-4o-2024-08-06`, temperature 0), GPT-5.5 (`gpt-5.5-2026-04-23`, reasoning effort none), Qwen3-VL-Plus (`qwen3-vl-plus-2025-12-19`, temperature 0, thinking disabled) and the open-weight Qwen2.5-VL-7B-Instruct (local, greedy) each answered at least 14,443 of 14,448 planned cases; Gemini could not be evaluated because the project key is limited to 250 requests per model per day.

Estimands and uncertainty. For training policy t and relation r, the paired view effect is

$$\Delta_{r,t} = \frac{1}{S} \sum_{s=1}^{S} \left[\bar{A}_{s,r,t}(\mathrm{C}) - \bar{A}_{s,r,t}(\mathrm{D}) \right], \tag{1}$$

where each room mean averages paired cases and, for the matrix, the three trained seeds. Intervals use 2,000 paired room-bootstrap resamples (Field & Welsh, 2007). Repeated roles, camera conditions, seeds, noise draws and prompts remain within rooms; they are not independent samples. Intervals are exploratory, unadjusted for multiplicity, and conditional on the trained models.

5 The trained verifier's canonical-view effects are pose shortcuts

What the matrix shows. Table 1 reports the frozen RGB-and-box verifiers on the ten unused rooms (Figure 2 in Appendix B adds the six development-evaluation rooms). Canonical-trained front–back models reach 66.19% under canonical views versus 55.79% under default views ($+10.41$ points), while default-trained models give -0.64 points; the directly bootstrapped training-by-view interaction is $+11.05$ points $[9.08, 13.33]$. Up–down models attain 87.65–94.92% with no positive canonical effect, and every left–right RGB model stays at chance even on its own training data (all nine conditions, and still after forty epochs in a development control, Appendix B). Read on their own, these three results would suggest a training-distribution effect, a competent relation without a canonical benefit, and a learning failure. The probe shows they are one mechanism.

A probe without pixels exceeds every cell. Table 2 gives a gradient-boosted tree the verifier's rotation input and twelve projected-mask statistics and nothing else. Trained on canonical development views, it reaches 99.60% left–right, 91.10% front–back and 99.27% up–down on canonical views of the unused rooms, above every RGB verifier cell, with canonical-minus-default effects of $+25.11$ and $+26.99$ points for the horizontal relations. Replacing only the rotation by that of the camera mirrored through the pair midpoint inverts the left–right probe to 0.20% and the front–back probe to 9.59% while leaving up–down at 99.27%. Removing the rotation drops both horizontal

Table 2: Pose-shortcut probe on ten unused rooms: room-macro accuracy (%) of a gradient-boosted tree that receives only the verifier's rotation input and twelve projected-mask statistics (no RGB), trained on canonical development views. *Mirrored* replaces the rotation by that of the camera mirrored through the pair midpoint while the mask statistics stay fixed. For comparison, the canonical-trained RGB verifier of Table 1 scores 48.88 / 49.53 (left–right), 66.19 / 55.79 (front–back) and 94.27 / 94.92 (up–down) on canonical / default views. Intervals are 95% room bootstraps.

Relation	Inputs	Canonical	Default	Mirrored (canon.)	Canonical−default
Left–right	rotation + masks	99.60	74.49	0.20	$+25.11$ $[19.73, 30.89]$
	masks only	47.74	50.44	–	-2.70 $[-5.90, 0.50]$
Front–back	rotation + masks	91.10	64.10	9.59	$+26.99$ $[23.06, 31.37]$
	masks only	53.67	50.38	–	$+3.29$ $[-1.24, 8.71]$
Up–down	rotation + masks	99.27	99.13	99.27	$+0.14$ $[-0.51, 0.99]$
	masks only	99.27	99.13	–	$+0.14$ $[-0.66, 1.14]$

relations to near chance (49.72% and 54.52%), and the masks alone give 47.74% and 53.67%, but 99.27% for up–down. Closed-form rules make the cues explicit: "pair direction from the rotation XOR subject mask is image-left" scores 99.14% for left–right and 0.86% under the mirrored rotation; "...XOR subject mask is larger" scores 75.26% for front–back; "subject mask is higher" scores 95.01% (canonical) and 96.41% (default) for up–down.

Reading the matrix through the probe. For the horizontal relations the room-frame label is available from the pose and the only visual task is an identity cue. The globally pooled ResNet reads the front–back cue (size and vertical position of the nearer target) but not the left–right cue (pure horizontal position, which is exactly the information that flip-augmented, globally pooled features are least likely to retain; Islam et al., 2020; Lin et al., 2020); on default views, where a default-trained probe with the same inputs reaches 87% for left–right, the network stays at 50%. The front–back interaction is reproduced in sign and pattern by the probe without RGB ($+26.99$ canonical-trained versus -6.20 default-trained, against $+10.41$ versus -0.64 for the RGB verifier) because canonical poses encode the pair axis and default poses do not; it is a property of the information contract, not of appearance. Up–down needs no rotation at all: vertical mask order is the label under any horizontal camera, which is why the verifier is accurate in both views and no canonical benefit is possible.

The frozen verifiers invert under the mirrored pose. Applying the same intervention to the frozen models themselves (Table 15, Appendix F) confirms the reading: with identical pixels and masks, replacing the rotation by that of the mirrored camera turns the canonical-trained front–back RGB verifier from 66.19% into 33.89% and its canonical-minus-default effect from $+10.43$ into -11.60 points, while zero rotation sends it to 50.50%; the boxes-only front–back model inverts from 80.16% to 20.56%. Left–right RGB models stay at chance under every intervention, and the nine up–down training conditions move by at most 0.2 points after seed averaging (single checkpoints by at most 0.4 points). The matrix's positive canonical-view effects (front–back under canonical and mixed training; the latter inverts from $+2.49$ to -2.51) are therefore pose effects on the real network, not only on the probe.

Consequence. None of the matrix's canonical-view effects measures appearance-based verification. Claims about visual verification must therefore either remove the rotation from the verifier and convert frames analytically, or use a verifier that grounds entity names itself. Sections 6 and 7 examine each route.

6 What a canonical-view visual step can add to noisy geometry

A canonical camera is placed from estimated entity centers. When those centers are noisy the camera is misplaced in the same way as a geometry solver's answer, so the natural question is how much an image-frame verifier could add. We answer it analytically with an *ideal* verifier that reads the true relative placement of the targets inside whatever camera the noisy geometry selected and converts the answer with the known camera axes (Appendix G). Table 3 gives the result on the ten unused rooms (Figure 4 in Appendix G plots all noise levels).

Table 3: Analytic ceiling under shared per-coordinate center noise σ (m), ten unused rooms, room-macro accuracy (%): the noisy geometry solver and an *ideal* image-frame verifier that reads the true placement inside the camera the noisy geometry selected and converts it with the known camera axes. Gain is the paired room difference (95% room bootstrap); reversal is the fraction of cases whose noisy pair direction is reversed.

Relation	σ	Geometry	Ideal, canon.	Ideal, default	Gain canon.−geometry	Reversal
Left–right	0.2	93.06	94.02	73.84	$+0.95$ $[0.61, 1.30]$	1.6%
	0.5	82.42	85.87	73.84	$+3.44$ $[2.57, 4.35]$	7.2%
	1.0	72.20	80.12	73.84	$+7.93$ $[6.19, 9.81]$	16.9%
Front–back	0.2	93.75	93.98	76.65	$+0.23$ $[-0.11, 0.65]$	1.6%
	0.5	84.58	87.42	76.53	$+2.84$ $[1.78, 4.02]$	7.4%
	1.0	73.90	81.69	76.75	$+7.80$ $[5.65, 10.60]$	17.4%
Up–down	0.2	85.44	100.00	100.00	$+14.56$ $[11.69, 17.60]$	1.8%
	0.5	70.97	100.00	100.00	$+29.03$ $[25.86, 32.17]$	7.5%
	1.0	61.76	100.00	100.00	$+38.24$ $[35.35, 40.92]$	17.5%

Table 4: Up–down with cameras placed from noisy centers (ten unused rooms, 400 pairs, both roles, two noise draws; all planned cases, room-macro %). The A/B markers are the true objects' projected boxes (correct 2D identification, noisy 3D localization); the geometry solver uses the same noisy centers. VLM minus geometry is a paired room difference with a 95% room bootstrap. The ideal verifier reads the true image order in the rendered camera, which perspective can reverse when the camera height is noisy.

σ (m)	Camera	Geometry	GPT-4o	GPT-4o−geometry	GPT-5.5	GPT-5.5−geometry	Ideal
0.5	canonical	72.12	76.62	$+4.50$ $[-0.44, 10.13]$	68.50	-3.62 $[-8.88, 1.94]$	90.88
0.5	default	72.12	74.38	$+2.25$ $[-4.12, 9.12]$	68.81	-3.31 $[-7.69, 1.81]$	95.13
1.0	canonical	62.13	70.94	$+8.81$ $[1.19, 15.94]$	65.63	$+3.50$ $[-3.50, 10.50]$	89.75
1.0	default	62.13	71.69	$+9.56$ $[2.75, 15.88]$	65.81	$+3.69$ $[-1.56, 8.06]$	91.50

Horizontal relations inherit the geometry's error. Whenever the noisy pair direction is not reversed, the converted image-frame answer is identical to the noisy geometry solver's answer; the verifier can only help when the noise reverses the pair direction, which requires center errors comparable to the pair distance (7% of cases at $\sigma = 0.5$ m, 17% at 1.0 m). The maximal gain is $+0.95$ points at 0.2 m, $+3.44$ at 0.5 m and $+7.93$ at 1.0 m for left–right, and similar for front–back. Within the non-reversed stratum the two are equal to the decimal (86.68% both at 0.5 m).

Gravity-fixed reductions are exact. For up–down any horizontal camera has image up equal to room up, so the conversion does not depend on the noisy geometry at all: the ideal verifier stays at 100% while the geometry solver falls to 61.76% at 1.0 m. The value of a visual step is therefore relation-specific and large exactly where the image axis used by the question is fixed by gravity rather than by the estimated pair axis.

Measured with real renders and VLMs. We tested this prediction on the 400 up–down pairs of Section 7, placing canonical and default cameras from centers with 0.5 or 1.0 m per-coordinate noise and marking the true objects, so that 2D identification is correct while 3D localization is noisy (Table 4). Perspective from the noisy camera height lowers the ideal verifier to 90–95%. GPT-4o, which keeps most of its clean up–down accuracy (81.5% to 71%), beats the noisy geometry by $+8.81$ $[1.19, 15.94]$ and $+9.56$ $[2.75, 15.88]$ points at 1.0 m, but not resolvably at 0.5 m; GPT-5.5, which perceives up–down less well in these renders, resolves no gain. The benefit predicted by the ceiling is therefore real but limited by perception, and in this test it is resolved only at 1.0 m.

Default and broad views cap converted accuracy. Even a perfect image-frame verifier converts to only 73.84% (left–right) and 76.9% (front–back) from default or broad cameras, because the questioned image axis is not aligned with the room axis. This is the construction-level reason a canonical view can matter for an image-frame verifier: it makes the frame conversion exact. Front–

Table 5: Left–right room-frame accuracy (%) of four VLMs on ten unused rooms, 400 pairs, both roles, all planned cases (failures counted incorrect). The image-frame prompt is converted to the room frame outside the model with the known camera axes; the room-frame prompt is the explicit-rotation interface on the same pixels. Effects are paired canonical-minus-default room differences with 95% room bootstraps. "Same word" is the share of pairs answered identically in both role orders (canonical / default, over the 397–400 pairs with both orders answered), which makes those pairs exactly 50% correct. Prompt-by-view interactions versus default: $+21.38$, $+8.50$ and $+11.38$ points (intervals above zero); for Qwen2.5-VL-7B the room-frame arm is constant, so its $+6.25$ equals the image-frame effect. Conversion ceilings: canonical 100%, default 74%. Front–back and up–down are in Table 10.

Model	Prompt frame	Canonical	Default	Broad	Canonical−default	Same word
GPT-4o	image frame	90.13	67.38	66.50	$+22.75$ $[16.63, 29.00]$	19 / 40
	room frame	51.12	49.75	49.62	$+1.38$ $[0.50, 2.25]$	92 / 98
GPT-5.5	image frame	84.12	68.12	62.88	$+16.00$ $[12.25, 19.88]$	27 / 29
	room frame	58.13	50.62	57.00	$+7.50$ $[1.00, 13.50]$	52 / 51
Qwen3-VL-Plus	image frame	68.25	57.00	55.62	$+11.25$ $[7.63, 14.88]$	63 / 67
	room frame	50.37	50.50	49.88	-0.13 $[-0.88, 0.62]$	95 / 97
Qwen2.5-VL-7B	image frame	59.12	52.87	54.75	$+6.25$ $[3.50, 9.13]$	82 / 88
	room frame	50.00	50.00	50.00	0.00 (degenerate)	100 / 100

back canonical views are only approximately reducing (99.51% at $\sigma = 0$) because the 0.5 m camera elevation tilts the depth axis.

These are ceilings for any verifier whose camera and masks derive from the same geometry; identity or grounding errors that place the camera on the wrong object are invisible to such a verifier by construction and can only be addressed by a verifier that grounds names itself.

7 Frame reduction for vision–language verifiers

A VLM grounds the named entities from the marked image (Yang et al., 2023), so it is a verifier for which visual evidence can in principle add information. Our earlier interface asked the fixed-room-frame question with the numeric camera-from-room rotation in the prompt. We compare it, on the same pixels, cameras and roles of the ten unused rooms, with an image-frame prompt whose answer is converted outside the model, for four models: GPT-4o and GPT-5.5 (reasoning disabled), Qwen3-VL-Plus, and the open-weight Qwen2.5-VL-7B run locally (OpenAI, 2024; 2026; Bai et al., 2025a;b). Table 5 gives left–right; Appendix D gives all relations (Table 10) and role statistics (Table 11).

For left–right, canonical views help exactly when the frame conversion leaves the model. With the image-frame interface the canonical view beats the default view for every model: $+22.75$ $[16.63, 29.00]$ (GPT-4o, 90.13%), $+16.00$ $[12.25, 19.88]$ (GPT-5.5, 84.12%), $+11.25$ $[7.63, 14.88]$ (Qwen3-VL-Plus, 68.25%) and $+6.25$ $[3.50, 9.13]$ points (Qwen2.5-VL-7B, 59.12%); between-model differences were not tested and adjacent intervals overlap. With the room-frame interface three models stay at 50–51% in every view and GPT-5.5 reaches only 58.13%. The prompt-by-view interactions of GPT-4o, GPT-5.5 and Qwen3-VL-Plus are positive with intervals above zero ($+21.38$, $+8.50$, $+11.38$); because the room-frame arms of GPT-4o and Qwen3-VL-Plus give one word for 92–98% of pairs, only GPT-5.5's compares two informative view effects, and for Qwen2.5-VL-7B the room-frame arm is constant, so its interaction equals its image-frame effect.

Why the gain arises. The two role orders of a pair always have opposite correct answers, so a pair answered with the same word in both orders is exactly 50% correct. In the left–right image frame, pairs that a model does tell apart are almost always right, 92–100% for every model in canonical and default views alike (Table 11); perception fails mainly by same-word answers rather than wrong answers. The canonical view reduces same-word pairs substantially for GPT-4o (19% versus 40%), modestly for the Qwen models (by 3.7 and 6.5 points) and not for GPT-5.5 (27% versus 29%); these rates are descriptive. The canonical gain therefore decomposes into a perceptual part (+10.75, +4.00, about +2.4 and +3.25 points of image-frame accuracy) and the exact conversion, since the default

view caps converted accuracy at 74% (losing 12.0, 12.0, about 8.8 and 3.0 points); for GPT-5.5 and Qwen3-VL-Plus the conversion dominates. For left–right, the room-frame interface fails the same way from the other side: GPT-4o and both Qwen models give one word for 92–100% of pairs, which holds their accuracy near 50% by construction, while GPT-5.5 answers about half the pairs role-sensitively but is right on only 67% and 52% of them, consistent with attempting the coordinate composition and performing it poorly.

A canonical view only converts what is perceived. Front–back is at 47–54% for every model, frame and view, with no resolved canonical effect: no model perceives depth order in these renders. Up–down is not a frame-reduction case, because every horizontal camera converts it exactly; its effects are perceptual and model-specific (GPT-4o $+6.00$ $[2.62, 9.00]$ in the image frame, GPT-5.5 $+5.75$ $[1.25, 10.25]$ in the room frame, the Qwen models at 53–61% with no resolved positive effect and Qwen2.5-VL-7B reversing to -6.75).

Limits. Each run answered at least 14,443 of 14,448 planned cases; decoding settings, parsing rules and the render style are given in Appendix D, and all cases remain in the denominator.

8 Discussion and limitations

Three mechanisms now account for what looked like a collection of unrelated canonical-view results. A query-conditioned camera encodes the queried direction, so a verifier that receives the pose can succeed without reading the image; which relations it "solves" depends on the identity cue its architecture can extract, and the resulting training-by-view interactions are properties of the information contract. A mask-conditioned visual step cannot add more than the reversal regime of the geometry that placed the camera, except where the image axis used by the question is fixed by gravity. And for a verifier that grounds names itself, the canonical view's value is frame reduction, which we could measure only after taking the coordinate transform out of the model.

The scope is deliberately limited. We use annotation-derived geometry and masks, oracle parsing and entity identities, three center-relation families, one trained architecture, four VLMs (three proprietary models from two providers and one open-weight 7B model), and ten unused rooms. Room-bootstrap intervals do not measure uncertainty across architectures or providers. The pose shortcut we document applies to any pipeline that passes query-conditioned camera parameters to its reasoning model; we have not measured its prevalence in complete systems. The ceiling analysis and its up–down test cover center noise only, with correct 2D markers; identity errors need a name-grounding verifier by construction. The two-hop executor study (Appendix C) was built on the shortcut-affected verifiers and is reported as an executor-level diagnostic. Rendering quality is imperfect.

For future work, the results suggest a concrete recipe rather than an adaptive selector: choose the camera that makes the questioned room axis coincide with an image axis, ask the verifier in the image's frame, convert analytically, and audit any trained verifier with a pose-only probe and a mirrored-camera intervention before attributing gains to vision.

9 Conclusion

Canonical camera construction does not by itself establish a benefit for spatial verification. The pose of a query-conditioned camera carries the answer for the horizontal relations, so trained verifiers that receive it can shortcut; a visual step on cameras placed from noisy geometry adds little beyond the direction-reversal regime, except for gravity-fixed reductions; and a vision–language verifier benefits from canonical views substantially only when the question is asked in the image's own frame and converted analytically (residual room-frame effects are at most $+7.5$ points across four models). The appropriate claim is conditional and mechanistic: canonical views help by reducing frame composition to perception, and reported gains should be audited against pose shortcuts and geometry ceilings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting sections of this paper and to improve the clarity, organization, and readability of the writing. We did not use generative AI tools to run the experiments. We reviewed and revised all AI-assisted text and checked technical claims, citations, and descriptions of experiments and results against the underlying work and original sources. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 defines the relations, the camera rules and the information available to the trained verifier, and Section 4 describes the cohorts, the verifier matrix, the probe and rotation interventions, the ceiling analysis, the vision–language study and the uncertainty estimates. Appendix A lists the room splits and the exact camera construction; Appendix B the architecture, optimization and checkpoint selection; Appendix C the two-hop task, executor and statistics; Appendix D the prompts, model identifiers, decoding settings, parsing rules and the image-frame conversion; Appendix F the probe features and interventions; and Appendix G the ceiling protocol. All experiments use public ScanNet scans; the main seeds are stated in Appendices A–C, and uncertainty is estimated with paired room-level bootstraps (Section 4).

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pp. 422–440, 2020. doi: 10.1007/978-3-030-58452-8_25.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3d question answering for spatial scene understanding. In *CVPR*, 2022. URL https://arxiv.org/abs/2112.10482.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

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

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. URL https://arxiv.org/abs/1702.04405.

C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. arXiv:2004.07780.

Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3D: A minimally contrastive benchmark for grounding spatial relations in 3d. In *Advances in Neural Information Processing*

Systems, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/76dc611d6ebaafc66cc0879c71b5db5c-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. URL https://arxiv.org/abs/1512.03385.

Md Amirul Islam, Sen Jia, and Neil D. B. Bruce. How much position information do convolutional neural networks encode? In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:2001.08248.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and mapping with radiance fields using Fisher information. In *ECCV*, pp. 422–440, 2024. doi: 10.1007/978-3-031-72624-8_24. URL https://arxiv.org/abs/2311.17874.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023. URL https://aclanthology.org/2023.emnlp-main.568/.

Osman Semih Kayhan and Jan C. van Gemert. On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2003.07064.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Yu-Ji Kim, Dahye Lee, Jun-Seong Kim, Hyeon-Woo Nam, GeonU Kim, Yongjin Kwon, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. SplatReasoner: Enhancing embodied reasoning and grounding by novel view synthesis. In *ECCV*, 2026. URL https://arxiv.org/abs/2601.13132v2.

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. doi: 10.1109/ICCV51701.2025.00863. URL https://arxiv.org/abs/2504.17207.

Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2006.09512.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1807.03247.

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. URL https://openreview.net/forum?id=IDJx97BC38.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. GPT-5.5 system card, 2026. URL https://deploymentsafety.openai.com/gpt-5-5. Model snapshot gpt-5.5-2026-04-23.

Stephen Palmer, Eleanor Rosch, and Paul Chase. Canonical perspective and the perception of objects. In John Long and Alan Baddeley (eds.), *Attention and Performance IX*, pp. 135–151. 1981.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. ActiveNeRF: Learning where to see with uncertainty estimation. In *ECCV*, 2022. URL https://arxiv.org/abs/2209.08546.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In **SEM*, pp. 180–191, 2018. doi: 10.18653/v1/S18-2023.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH Conference Proceedings*, 2023. URL https://arxiv.org/abs/2302.04264.

Pere-Pau Vázquez, Miquel Feixas, Mateu Sbert, and Wolfgang Heidrich. Automatic view selection using viewpoint entropy and its application to image-based modelling. *Computer Graphics Forum*, 22(4):689–700, 2003. doi: 10.1111/j.1467-8659.2003.00717.x.

Hengyi Wang, Ruiqiang Zhang, Chang Liu, Guanjie Wang, Zehua Ma, Han Fang, and Weiming Zhang. Allocentric perceiver: Disentangling allocentric reasoning from egocentric visual priors via frame instantiation. *arXiv preprint arXiv:2602.05789*, 2026. URL https://arxiv.org/abs/2602.05789.

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

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. URL http://jmlr.org/papers/v26/24-1476.html.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Yizhuang Peng, Zhenfei Yin, Lijun Wang, and Huchuan Lu. Think3D: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026. URL https://arxiv.org/abs/2601.13029v4.

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*, 2025. URL https://openreview.net/forum?id=84pDoCD4lH.

A DATA, GEOMETRY, AND CAMERA CONSTRUCTION

Entity construction and splits. Vertices are transformed by the scan's recorded rigid axis-alignment matrix. Each retained annotated entity has at least 50 vertices and an AABB diagonal between 0.1 and 6.0 m. The AABB midpoint defines its center. Pair sampling uses the frozen seed 20260923, a maximum of 100 unordered pairs per room and relation, and the distance/gap rules in Section 3. Both ordered roles are retained, so the binary targets are balanced without relabeling a question after moving the camera.

The following scene identifiers omit only the common `scene` prefix. Physical room groups, not scan suffixes, define the unused-room exclusion. The train/validation/evaluation partitions are fixed before model training.

> Training (28): 0000_00, 0001_00, 0009_00, 0011_00, 0013_00, 0020_00, 0024_00, 0027_01, 0031_00, 0032_00, 0050_00, 0084_00, 0144_00, 0354_00, 0034_00, 0035_00, 0051_00, 0052_00, 0054_00, 0057_00, 0068_00, 0072_00, 0082_00, 0087_00, 0090_00, 0098_00, 0099_00, 0106_00.
>
> Validation (6): 0002_00, 0017_00, 0028_00, 0038_00, 0056_00, 0083_00.
>
> Development evaluation (6): 0003_00, 0022_00, 0029_01, 0044_00, 0060_00, 0091_00.
>
> Unused-room confirmation (10): 0039_00, 0059_02, 0062_02, 0065_01, 0067_01, 0073_01, 0074_01, 0078_02, 0079_01, 0080_01.

The confirmation rooms are the first ten lexicographically available physical room groups outside development with pre-existing RGB/pose/intrinsic exports and annotation geometry. This is availability-conditioned selection, not random sampling from all indoor scenes. Each room has 400 equally spaced samples from finite, rigid valid poses across the full trajectory. Reconstruction uses Nerfstudio tooling (Tancik et al., 2023) with gsplat rasterization (Ye et al., 2025), 30,000 iterations, seed 42, degree-three spherical harmonics, native-resolution images, no camera optimizer, and random Gaussian initialization without a seed point cloud. Whole-trajectory sampling differs from the historical prefix-biased reconstruction procedure.

A.1 EXACT CAMERA RULES

Let (i, j) be the pair in increasing entity-ID order, let $m = (c_i + c_j)/2$, and let h be the normalized horizontal projection of $c_j - c_i$. If its norm is below 10^{-8}, set $h = (1, 0, 0)$. Define $\rho = \max(1.5, 1.5\|c_j - c_i\|_2)$, in meters, and $e_z = (0, 0, 1)$. The camera position is

$$p_r = m - \rho f_r + \eta_r e_z, \qquad f_{\mathrm{LR}} = e_z \times h, \quad f_{\mathrm{FB}} = f_{\mathrm{UD}} = h, \tag{2}$$

with $\eta_{\mathrm{FB}} = 0.5$ m and zero otherwise. A look-at transform points at m with gravity-aligned up. The default replaces $-f_r$ by $(\cos\theta, \sin\theta, 0)$ for a deterministically seeded uniform azimuth, retaining ρ and η_r. The broad policy samples azimuth uniformly and elevation uniformly from $[-35, 35]$ degrees at radial distance ρ. The camera seed is 2026092501, with deterministic scene/pair/relation keys.

We render at 256×256 pixels with a 60-degree field of view and full exported spherical-harmonic appearance. The trained model uses 224×224 inputs and nearest pixel-center mapping for the two AABB masks. Reversing roles swaps masks and target labels, never the physical camera. The reference rotation maps aligned-world displacements into OpenGL camera axes: image right, image up, and toward the camera. It conveys orientation but no entity translation.

A.2 SHARED NOISY GEOMETRY

For each scene, entity, and fixed noise seed, draw one $\epsilon_i \sim \mathcal{N}(0, I_3)$ and use $c_i + \sigma\epsilon_i$ as the method center. Translate the AABB corners by the same vector. Both camera selection and masks use this perturbed geometry; the geometry solver uses the same centers. The reconstructed RGB scene, clean labels, and clean eligibility stay fixed. The visual follow-up uses $\sigma \in \{0.05, 0.20\}$ m, three noise seeds (20260925011–20260925013), and up to ten unordered primary pairs per scene/relation. Its 480 pair–relation units span all 16 rooms; the ten unused rooms contribute the matched confirmation subset. This is a direction-only test on a fixed cohort, not the earlier experiment that also recomputed noisy distance-gated chains.

Table 6: Development-only left–right budget control, seed 41. Percentages compare separately selected ten- and forty-epoch runs; no confirmation inference.

Training	Modality	Last training, $10 \to 40$	Selected validation, $10 \to 40$
Canonical	Boxes only	$49.00 \to 99.32$	$49.77 \to 77.45$
Canonical	RGB + boxes	$50.86 \to 51.50$	$51.13 \to 49.83$
Default	Boxes only	$54.92 \to 83.62$	$58.52 \to 83.48$
Default	RGB + boxes	$49.96 \to 50.00$	$50.03 \to 52.14$
Mixed	Boxes only	$78.20 \to 87.50$	$71.42 \to 81.91$
Mixed	RGB + boxes	$50.08 \to 50.00$	$50.20 \to 51.07$

B Training and learning diagnostics

The ResNet-50 first convolution is expanded to five channels. Its RGB weights are copied from the pinned ImageNet initialization, while the two mask-channel weights start at zero. The pooled 2,048-dimensional feature is concatenated with rotation9 and passed through a 256-unit linear layer, ReLU, and a two-logit output layer. The backbone and head are both optimized with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, and bfloat16 autocast. Seeds are 41, 42, and 43. Training uses ten epochs and 5,000 ordered exposures per epoch, drawn from paired permutations with both roles included.

Mixed training alternates canonical/default views within the same exposure budget; it does not double the number of examples. Validation always pools canonical, default, and broad evidence with no gravity overlay and selects the earliest minimum scene-macro cross-entropy. Boxes-only retraining zeros the normalized RGB channels throughout training and evaluation. Frozen RGB suppression applies that change only to an already trained RGB model and is a distinct input intervention.

The nine additional up–down RGB conditions cross three seeds with canonical training plus a gravity overlay, broad training without it, and broad training with it. Rotation9 already conveys gravity; the overlay changes presentation, not access to gravity information. Without an evaluation overlay, the broad up–down model's confirmation canonical-minus-default effect is -2.83 points $[-3.86, -1.82]$.

B.1 Extended development training

An audit retained every condition and identified 22 of 63 models in a descriptive band of 45–55% for both last online training accuracy and selected pooled validation accuracy. This is not a formal equivalence criterion. A separate float32 CPU control fit four unordered training pairs in both roles with each modality, reaching a ten-step perfect-fit criterion. This verifies tiny-set fitting capacity only.

The longer-budget diagnostic trains all three left–right camera policies with both modalities at seed 41 for forty epochs. Each uses the same 2,438 training pairs and 200,000 ordered exposures, compared with 50,000 in the ten-epoch run. The first ten pair-exposure schedules match, but the runs are separately restarted, not bitwise continuations. No forty-epoch model is selected or evaluated on the unused-room confirmation cohort. Table 6 reports accuracy at the checkpoint chosen by each budget's validation-loss rule, not maximum validation accuracy. Longer training also increases the number of selection opportunities.

Figure 3 shows all six learning curves, and Figure 2 the canonical-minus-default effects of the RGB matrix on both evaluation cohorts.

C Two-hop execution: results and statistical analysis

The two-hop study reuses the frozen matrix verifiers of Section 5. Because those verifiers are pose-shortcut verifiers (left–right hops are at chance and front–back hops depend on the pose), the results below are an executor-level diagnostic of how validity criteria and search depth interact, not evidence about visual composition. They are retained in full for completeness and comparability with the historical study.

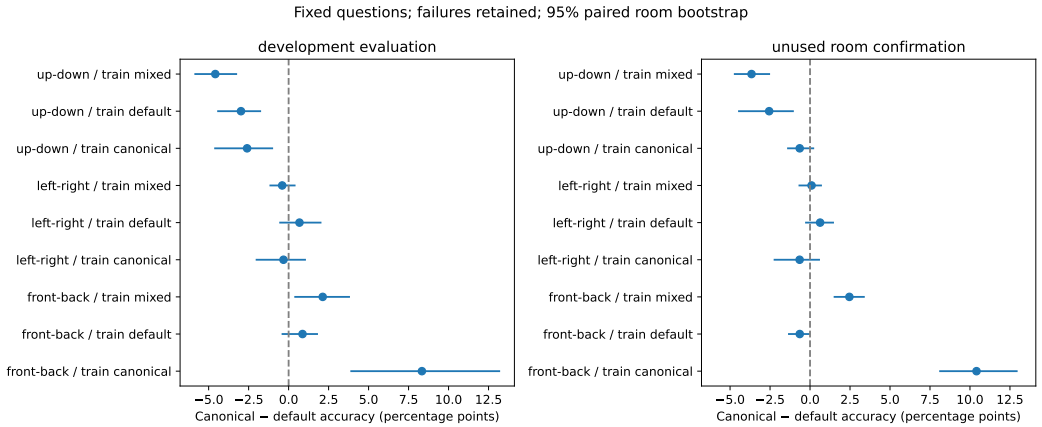


Figure 2: Canonical-minus-default effects for RGB-and-box models on six development-evaluation rooms (left) and ten unused rooms (right). Intervals resample rooms after averaging trained seeds. Table 2 shows that the front–back effects and the absence of left–right effects are both explained by the pose shortcut and the identity cue the network can read.

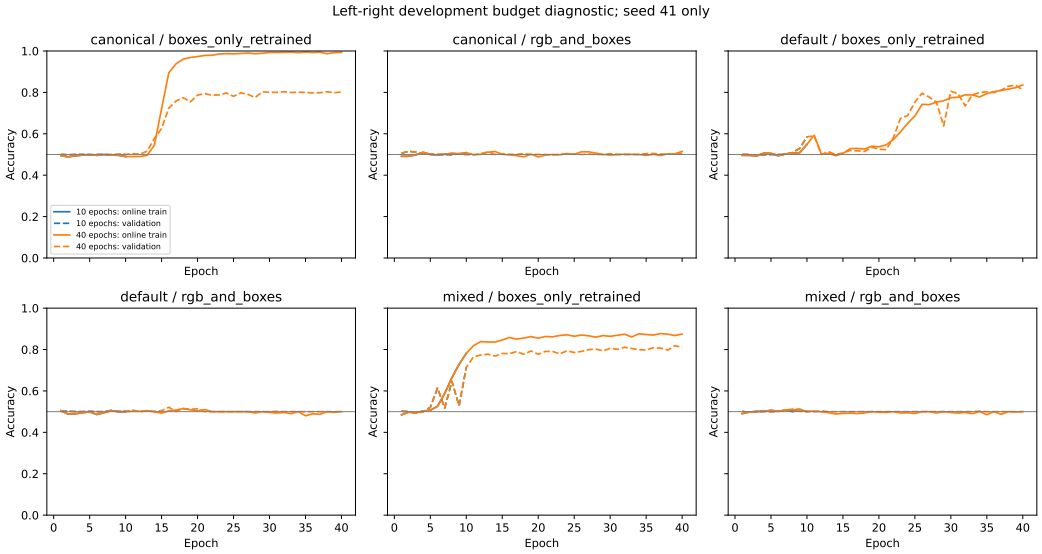


Figure 3: All six development learning-budget conditions. Online training accuracy is measured while weights change; validation curves concern the fixed six development validation rooms. The RGB plateau persists at forty epochs, while mask models improve with remaining generalization gaps.

C.1 Two-hop questions

A query specifies an anchor a and directions r_1, r_2. Its valid witness set is

$$\mathcal{V}_q = \{(z, b) : r_1(z, a) \wedge r_2(b, z), \quad a, z, b \text{ distinct}\}, \tag{3}$$

where each edge also satisfies the geometric eligibility rule of Section 3. Parsing, entity IDs, and eligibility are given; only edge directions are visually verified. Multiple-choice candidates contain one valid endpoint and invalid distractors. Open-set evaluation uses all non-anchor entities on the same query cohort. We score both whether the returned endpoint belongs to some pair in $\mathcal{V}_q$ and whether the returned pair itself belongs to $\mathcal{V}_q$.

Table 7: Both canonical cameras minus both default cameras on the same 656 unused-room queries, with full predicted-intermediate search. Effects and 95% room-bootstrap intervals are in percentage points.

Training	Answer format	Final valid	Selected chain valid
Canonical	Multiple choice	$+2.55$ $[0.54, 4.49]$	$+3.28$ $[1.10, 5.40]$
Default	Multiple choice	-1.88 $[-4.52, 0.30]$	-1.28 $[-3.68, 1.17]$
Mixed	Multiple choice	-2.85 $[-5.11, -0.48]$	-1.32 $[-3.27, 0.67]$
Canonical	Open set	-1.45 $[-3.73, 1.07]$	$+2.71$ $[0.07, 5.68]$
Default	Open set	-2.09 $[-3.98, 0.11]$	-0.33 $[-3.71, 2.50]$
Mixed	Open set	-2.07 $[-4.70, 0.59]$	-0.81 $[-2.63, 1.11]$

C.2 Compositional results

We reuse the frozen RGB verifier predictions in a common two-hop executor. An edge is positive when the requested-direction probability exceeds 0.5. The executor ranks positive prefix intermediates by probability, counts their positive suffix witnesses for each candidate, and returns the endpoint with the most witnesses and its first ranked supporting intermediate. It abstains when there is no positive path. Gold endpoints and witnesses are accessed only after inference. Full, top-one, and top-five intermediate search share the same edge predictions; oracle controls replace whole hop predicates.

A valid answer and a valid chain are different outcomes. On the 656 unused-room queries, canonical-trained full-search models gain 2.55 points in multiple-choice final validity and 3.28 points in chain validity. In open set, however, their final-valid accuracy decreases from 70.99% to 69.55%, while chain validity increases from 39.91% to 42.62%. The final-valid effect has an interval spanning zero; the chain-valid effect is $+2.71$ points $[0.07, 5.68]$. Default- and mixed-trained families show other effect patterns (Table 7). Per-family intervals alone are not a direct interaction test between chain-training conditions.

Account for the available path universe. For the canonical-trained open-set family, uniform sampling from each policy's own predicted positive paths yields final validity of 64.36%/64.44% and chain validity of 36.97%/34.66% under canonical/default views. Deterministic selection exceeds these matched random chain baselines by 5.65 $[3.58, 7.73]$ and 5.25 $[3.31, 7.35]$ points, respectively. The baselines condition on different predicted path sets; their difference is not a common policy-independent chance level. Reporting both metrics and their matched comparators distinguishes endpoint permissiveness from path quality.

Executor controls expose additional constraints. For canonical-trained multiple-choice models with canonical cameras, final-valid accuracy is 47.94% with full search, 44.14% with top five, and 24.61% with top one. Replacing the prefix with an oracle yields 55.21%; replacing the suffix yields 80.42%; replacing both yields 100%. The next subsection gives chain validity and exact scoring. These interventions alter whole predicates and path scores, so their gaps are diagnostic ceilings, not additive causal allocations of error. They also do not validate a near-chance relation verifier merely because it appears within a successful chain.

C.3 Executor definition and statistics

For each of the 36 ordered direction sequences, the query generator retains up to two queries per scene, subject to finite availability. Each multiple-choice query has one valid final and three invalid distractors, with frozen randomized option order. The open-set version retains the same anchor and directions but supplies all non-anchor IDs in a deterministically shuffled order. It is therefore an MCQ-conditioned anchor cohort, not an unrestricted sample of open-world language questions.

Let $P(a, z)$ and $Q(z, b)$ be the requested-direction probabilities of the first and second hop. Both are set to zero on edges outside the clean eligibility graph defined in Appendix A; noisy inputs never change this graph. The eligible prefix set is $Z = \{z \neq a : P(a, z) > 0.5\}$, ordered by decreasing probability with lower ID breaking ties. Top-K keeps its first K entries. For candidate b, the score is

Table 8: Executor controls for canonical-trained RGB models, both cameras canonical, multiple-choice unused-room queries. Descriptive room-mean accuracy (%); oracle gaps are not additive causal error contributions.

Executor	Final valid	Selected chain valid
Actual, full search	47.94	28.89
Actual, top five	44.14	26.37
Actual, top one	24.61	15.27
Oracle prefix, actual suffix	55.21	48.71
Actual prefix, oracle suffix	80.42	58.61
Both oracle, full search	100.00	100.00

the count of distinct positive witnesses,

$$w(b) = \sum_{z \in Z_K} \mathbf{1}[b \notin \{a, z\}] \mathbf{1}[Q(z, b) > 0.5]. \tag{4}$$

The executor selects the first maximizer in the public candidate order, then the first ranked supporting intermediate. If every score is zero it abstains, which is scored as incorrect. Probabilities for reverse roles are obtained by separate model inference with swapped masks, not manufactured as one minus the forward prediction.

Gold witness sets are separate from public inputs and from the actual-model inference function. Oracle controls replace a complete hop graph using the clean predicate and probability one on positive edges. Every full-search, both-oracle query copy was asserted to return a valid final and selected chain. Nine camera-training/seed verifier families, four prefix/suffix camera crosses, two answer formats, and six controls produce 457,920 evaluations of the 1,060 underlying questions. Table 8 concerns only the 656 unused-room questions and the specified canonical-trained family.

For each condition, the matched random baseline samples uniformly from its own predicted positive (z, b) paths after prefix truncation and candidate restriction. Its final-valid and chain-valid probabilities are calculated exactly; both are zero when no path exists. The candidate/path universe is identical to that condition's deterministic executor, but may differ between camera policies. This baseline is not uniform over all scene entities.

Aggregation and uncertainty. Case means are first computed within scene and trained seed. Three seed means are then averaged within each scene, and scenes receive equal weight. Paired camera effects and interactions resample the same room indices across all contrasted cells, using 2,000 bootstrap draws and seed 20260925099. Role reversals, shared-noise draws, and repeated query controls remain clustered within their room. Intervals are percentile intervals conditional on the trained models. Cohort comparisons resample the two room cohorts separately. Quality-conditioned subsets and direction-sequence tables are descriptive sensitivities and do not replace the all-case primary estimate.

Frozen verifiers under shared geometry noise. Table 9 compares the canonical-trained RGB verifiers with the direction-only geometry solver on a matched unused-room subset with $0.20\,\mathrm{m}$ per-coordinate shared center noise (same perturbation for cameras, masks and solver; clean labels and eligibility). The verifiers trail the solver by 41.33, 31.17 and 3.56 points; given Section 5, this reflects the verifiers' inability to read the identity cue rather than a property of the visual evidence, and Section 6 bounds what any such verifier could add.

D Vision–language interfaces and rendering diagnostics

D.1 Prompt frame by evidence view (Section 7)

Cases are the first 40 frozen pair identifiers per unused room and relation whose canonical, default and broad views all exist (1,200 pairs, no exclusions), both role orders, one colour assignment (subject red, anchor blue), 14,400 scene cases and 48 synthetic units. The image-frame prompt shows the plain

Table 9: Shared 0.20 m per-coordinate geometry noise, matched unused-room pairs. Visual column uses canonical-trained RGB models and canonical views; contrast intervals resample rooms. Clean labels and eligibility remain fixed.

Relation	Geometry accuracy (%)	Visual minus geometry (points)
Left–right	92.00	-41.33 $[-46.28, -37.00]$
Front–back	95.33	-31.17 $[-33.44, -28.78]$
Up–down	82.00	-3.56 $[-6.39, -0.89]$

render with the two labelled outlines and asks, for left–right, "Answer using THIS IMAGE's own left and right only. Is the center of object A to the LEFT or to the RIGHT of the center of object B in this image?"; for front–back, whether A is in FRONT of B (closer to the camera) or BEHIND B; for up–down, whether A is ABOVE or BELOW B in this image; each with "Reply with exactly one word". The room-frame arm is the interface of Section D.2 unchanged: the header image with projected room-axis arrows and the numeric camera-from-room rotation in the prompt. The image-frame label is defined by perspective projection of the true centers (pixel x, depth, pixel y); the room-frame prediction of the image-frame arm is the model's answer if the cosine between the questioned camera axis (right, forward, up) and the room axis ($+X$, $+Y$, $+Z$) is positive and its complement otherwise. Ideal-conversion ceilings on the planned cases are 100 / 99.5 / 100% for canonical, 74 / 80 / 100% for default and 74 / 78 / 82% for broad cameras (left–right / front–back / up–down). GPT-4o (temperature 0, 128 output tokens, high-detail image input) answered 14,443 cases with 14,448 HTTP calls and five unparseable answers; GPT-5.5 (`reasoning_effort` none, 128 completion tokens, no temperature parameter, zero reasoning tokens recorded) answered 14,443 cases with 14,449 HTTP calls, its five unparseable answers being left–right room-frame responses of "FRONT". Both runs used 5.77 M and 3.37 M prompt tokens respectively (GPT-5.5 bills images by patch) and completed in about 38 minutes each. Qwen3-VL-Plus (`qwen3-vl-plus-2025-12-19`) ran on the DashScope Singapore endpoint with temperature 0, thinking disabled and 128 output tokens; the endpoint rate-limits near 60 requests per minute, so it was paced at two workers with a 1.05-second admission interval, answered 14,446 cases with 14,448 HTTP calls and used 3.32 M prompt tokens (a first attempt at six workers was stopped after 204 rows and is unused). The open-weight Qwen2.5-VL-7B-Instruct ran locally (bf16, greedy decoding, 8 new tokens, the snapshot's repetition penalty of 1.05, 14.3 minutes on one RTX 4090) and answered all 14,448 cases, always with exactly one listed word. GPT-4o accepts only the whole cleaned answer (its 5 unparseable replies are refusals on front-back image-frame questions, counted incorrect); GPT-5.5 and Qwen3-VL-Plus accept the whole answer with a flagged first-word fallback, used 0 times; the local model takes the unflagged first word, and since all its replies are exactly one listed word the whole-answer rule would give identical predictions. The all-planned and answered-only room estimates agree within 0.22 points for every model. Gemini 3.1 Pro (preview) was limited by a project quota of 250 requests per model per day and is not reported.

Table 10 reports all relations for the four models and Table 11 the role statistics for left–right. Because both role orders of a pair are shown on the same render with the A/B marks swapped and always have opposite correct answers, a pair answered with the same word in both orders is exactly 50% correct, which holds the left–right room-frame cells of GPT-4o and both Qwen models near 50% by construction (exactly 50% for Qwen2.5-VL-7B); the zero-width 50% cells of Qwen2.5-VL-7B are this degenerate case and carry no information about views. Synthetic two-sphere units (eight per relation and arm; four left–right and front–back image-frame units are undefined at azimuth 0) are interface checks only: GPT-4o answered LEFT on seven of eight image-frame left–right units and FRONT on six of eight front–back units, and was correct on all up–down units.

D.2 Development-room room-frame evaluation

The completed original VLM evaluation uses `gpt-4o-2024-08-06` and `gemini-3.1-pro-preview`, temperature zero, and output budgets of 128 and 2,048 tokens, respectively. Per provider it contains 2,400 primary explicit-rotation cases, 480 matched header-only cases, 480 current-camera cases, and 96 synthetic controls. Primary cases use ten unordered pairs per relation in each of the ten cached development rooms, both evidence policies,

Table 10: All relations: room-frame accuracy (%) of the four VLMs on ten unused rooms, all planned cases, with paired canonical-minus-default effects and 95% room bootstraps. IF = image frame with analytic conversion; RF = room frame with explicit rotation. Conversion ceilings for IF are 100 / 99.5 / 100% (canonical) and 74 / 80 / 100% (default) for left–right / front–back / up–down.

Relation	Model	Prompt	Canonical	Default	Broad	Canonical − default
Left–right	GPT-4o	IF	90.13	67.38	66.50	$+22.75$ $[16.63, 29.00]$
		RF	51.12	49.75	49.62	$+1.38$ $[0.50, 2.25]$
	GPT-5.5	IF	84.12	68.12	62.88	$+16.00$ $[12.25, 19.88]$
		RF	58.13	50.62	57.00	$+7.50$ $[1.00, 13.50]$
	Qwen3-VL-Plus	IF	68.25	57.00	55.62	$+11.25$ $[7.63, 14.88]$
		RF	50.37	50.50	49.88	-0.13 $[-0.88, 0.62]$
	Qwen2.5-VL-7B	IF	59.12	52.87	54.75	$+6.25$ $[3.50, 9.13]$
		RF	50.00	50.00	50.00	0.00 (degenerate)
Front–back	GPT-4o	IF	54.12	51.88	50.50	$+2.25$ $[-4.38, 9.75]$
		RF	50.00	50.25	50.25	-0.25 $[-0.75, 0.25]$
	GPT-5.5	IF	47.00	49.88	52.38	-2.87 $[-8.75, 2.88]$
		RF	53.00	52.25	52.50	$+0.75$ $[-4.25, 5.75]$
	Qwen3-VL-Plus	IF	49.88	49.88	49.38	0.00 $[-1.12, 1.38]$
		RF	49.88	50.00	50.12	-0.13 $[-0.38, 0.00]$
	Qwen2.5-VL-7B	IF	49.50	49.12	49.50	$+0.37$ $[-1.37, 1.87]$
		RF	51.12	49.88	50.37	$+1.25$ $[-0.62, 3.25]$
Up–down	GPT-4o	IF	81.50	75.50	71.50	$+6.00$ $[2.62, 9.00]$
		RF	71.50	71.62	70.88	-0.12 $[-3.25, 3.12]$
	GPT-5.5	IF	72.25	70.00	67.50	$+2.25$ $[-2.00, 6.13]$
		RF	77.50	71.75	69.25	$+5.75$ $[1.25, 10.25]$
	Qwen3-VL-Plus	IF	57.12	56.88	55.63	$+0.25$ $[-1.63, 2.00]$
		RF	54.00	54.62	53.12	-0.62 $[-2.12, 0.75]$
	Qwen2.5-VL-7B	IF	53.75	60.50	57.50	-6.75 $[-9.12, -4.62]$
		RF	50.00	50.00	50.00	0.00 (degenerate)

Table 11: Role invariance on left–right, ten unused rooms, scene pairs with both role orders answered (400 per cell, except GPT-5.5 room frame 398 / 397 and Qwen3-VL-Plus image-frame default 399); synthetic controls are excluded. "Same word" is the share of pairs answered identically in both role orders (exactly 50% correct by construction). "Accuracy when distinguished" is the prompt-frame accuracy on the remaining pairs, which are either fully right or fully wrong; n is their number. It covers only pairs the model told apart and is not a bias-corrected accuracy.

		Same word (%)		Accuracy when distinguished (%, n)	
Model	Prompt	Canonical	Default	Canonical	Default
GPT-4o	image frame	19.25	39.75	99.7 (323)	98.8 (241)
	room frame	92.25	98.00	64.5 (31)	37.5 (8)
GPT-5.5	image frame	27.25	28.75	96.9 (291)	92.3 (285)
	room frame	51.51	51.39	67.4 (193)	51.8 (193)
Qwen3-VL-Plus	image frame	63.00	66.67	99.3 (148)	97.7 (133)
	room frame	95.25	96.50	57.9 (19)	64.3 (14)
Qwen2.5-VL-7B	image frame	81.75	88.25	100.0 (73)	100.0 (47)
	room frame	100.00	100.00	– (0)	– (0)

both roles, and both red/blue assignments. Thus each relation has 400 paired role/color cases per policy from 100 physical pairs, not 400 independent rooms or objects.

The explicit-reference prompt defines left/right using room X, front/behind using room Y, and above/below using room Z, and gives the rotation matrix from room to OpenGL camera coordinates. It states that moving the evidence camera does not change the answer and asks for exactly one direction word. Images mark subject A and anchor B and show projected positive-axis arrows with depth-direction cues. The header-only variant omits the numeric matrix. Synthetic images contain equal-radius targets differing only on the queried room axis, rendered at four azimuths with both role/color assignments. Their horizontal failures (8/16 for each horizontal relation, LEFT on every explicit left–right unit) mean that this interface is not a validated route from image evidence to room-frame answers; Section 7 shows the image-frame alternative on unused rooms.

Table 12 reports the development-room results of this interface. Current-camera diagnostic questions use their own camera-relative labels. GPT-4o canonical/default accuracies on those cases are 78.75/71.25% (left–right), 58.75/56.25% (front–back), and 65.00/70.00% (up–down). All paired room-bootstrap intervals include zero. These are 80 paired role/color cases per relation and a differ-

Table 12: GPT-4o explicit-reference development evaluation. Accuracy (%) and canonical-minus-default points, with room-bootstrap intervals. Horizontal synthetic controls fail to validate this interface; these are not confirmation results for the trained verifier.

Relation	Canonical	Default	Difference [interval]
Left–right	50.25	50.50	-0.25 $[-2.50, 2.25]$
Front–back	49.75	50.00	-0.25 $[-0.75, 0.00]$
Up–down	72.00	74.75	-2.75 $[-10.75, 6.75]$

Table 13: Unused-room rendering diagnostics (%, canonical/default). These proxies are descriptive, with no primary-case exclusion.

Relation	Blank views	Missing boxes	Both targets mesh-visible
Left–right	1.65 / 0.65	1.42 / 3.13	64.31 / 58.12
Front–back	0.40 / 1.31	5.10 / 2.42	49.64 / 61.99
Up–down	0.73 / 1.61	5.93 / 3.53	50.92 / 59.38

ent task, not corrected labels for the primary fixed-reference comparison. On the matched prompt subset, the up–down difference between explicit-matrix and header-only view effects is -17.5 points $[-30.0, -5.0]$, an exploratory interface sensitivity.

Response coverage. GPT-4o produced 3,456 valid answers. The original Gemini run produced 242: 66 synthetic, 118 primary, 31 header-only, and 27 current-camera answers. Seven further HTTP-success responses were unparseable, and 3,207 cases failed at the transport layer. A first paced recovery received three further HTTP429 responses and recovered no answers. A second unchanged-input recovery also stopped under its access/rate-limit guard, with 488 valid answers including the 242 originals. Its terminal response coverage remains incomplete; its accuracy analysis is pending and is excluded from this draft's quantitative comparisons. Failure-inclusive request success is not verifier accuracy, and the small completed-pair intersection cannot establish a Gemini null effect. Recovery preserves original valid and HTTP-success unparseable answers rather than resampling them for a preferred outcome.

D.3 Retained rendering diagnostics

Table 13 equally weights the ten unused rooms on matched clean primary pairs. Mesh visibility requires each annotated target to cover at least 16 first-hit pixels; it is not a segmentation of recognizable RGB content. A render is blank if its uint8 RGB mean or standard deviation is below 5, and a box is missing if either target's projected box is undefined. Fifty checks at observed reconstruction-input cameras, five per unused room, all completed, but their 19.74–33.65 dB room-mean PSNR range is an input-fit diagnostic. Novel analytic views can still contain severe artifacts. We do not exclude these cases from the primary camera comparison or treat a visibility-conditioned subset as a randomized intervention.

E Historical results and comparability audit

Earlier exploratory experiments used a different question/reference protocol, camera default, rendering pipeline, training distribution, and executor. Their numbers are retained as motivation and audit targets, not pooled with the corrected fixed-reference results. All previously inspected rooms are now development data, including those formerly called held out.

The expanded boundary-focused two-hop cohort contains 958 queries from 12 yielding scenes and gives a canonical-minus-default effect of $+16.493$ points. A second cohort contains 806 queries from nine yielding scenes and gives $+1.241$ points. A direct scene-clustered bootstrap of the first effect minus the second gives $+15.252$ points $[1.072, 29.001]$ with query weighting, but $+13.298$ $[-0.963, 28.318]$ with equal scene weighting. These are different estimands. They support sensitivity to weighting and scene heterogeneity, not a causal attribution to scene distribution. First hops are also skewed: 942 of 958 in the first cohort are up–down, and all 806 are up–down in the second.

Table 14: Pose leak and identity cues in the clean primary evidence of all 16 evaluation rooms (six development-evaluation and ten unused rooms), pooled rates (%). *Rule* is the agreement between the stored room-frame label and the label implied by the pair direction recovered from the camera rotation, over all views (n); the cue columns describe the projected boxes of the lower-ID entity i over views with both boxes (n_{box}); the last column is the agreement between "the higher box is the subject" and "above". Ties in box coordinates or areas count as "not left / lower / larger / higher".

Relation	View	n	n_{box}	Rule	i image-left	i lower	i larger	higher = above
Left–right	canonical	1,490	1,463	100.00	99.73	50.58	58.92	46.55
	default	1,490	1,448	50.07	49.24	51.10	58.49	46.27
	broad	1,490	1,439	49.93	49.13	52.54	60.39	48.99
Front–back	canonical	1,505	1,419	100.00	48.27	74.28	78.08	48.48
	default	1,505	1,460	49.57	49.66	52.88	59.38	45.96
	broad	1,505	1,458	50.43	49.79	53.16	60.84	45.13
Up–down	canonical	1,463	1,375	44.70	46.18	52.36	77.45	99.49
	default	1,463	1,415	49.08	50.74	53.36	58.45	99.15
	broad	1,463	1,409	50.92	48.47	52.66	58.62	82.47

An audit of 438 complete-case archived VLM pairs reproduces the old means but reveals camera-dependent target distributions. Canonical left–right has 154 left and zero right targets, compared with default's 73 left and 81 right. Canonical front–back has 150 front and one behind, compared with default's 73 front and 78 behind. The aggregate model differences therefore do not constitute a paired fixed-question transfer test. Reweighting default front–back class recalls to the canonical label proportions yields 29.120%, near the observed 28.477% canonical accuracy; this is descriptive and does not establish a causal correction. Historical render failures were omitted from that complete-case analysis, unlike the new primary denominator.

For 400 historical open-set queries, accounting for all valid answers gives 91.0% chain validity for canonical and 91.5% for default. Correcting the metric does not establish a canonical advantage. That cohort has a mean 26.37 valid finals among 31.75 scene objects, so exact recovery of one sampled endpoint is a different target from relation satisfaction. The new fixed-reference local-graph study changes eligibility, query construction, and execution; its 42.62/39.91% canonical-trained chain-valid rates are not a drop measured on the same task. Likewise, the new 656-query unused-room comparison is not an exact replication of the historical 958/806-query study.

F POSE-SHORTCUT PROBE AND ROTATION INTERVENTIONS

Pose determines the horizontal labels. For each clean primary view of the 16 evaluation rooms we recover the camera forward axis in room coordinates from the stored camera-from-room rotation ($f = -R_{3,\cdot}$) and the implied pair direction h ($h = (f_y, -f_x, 0)$ for canonical left–right cameras, $h = (f_x, f_y, 0)$ otherwise). Table 14 reports the agreement of the rule "label 0 iff $h_x > 0$" (left–right) or "$h_y > 0$" (front–back) with the stored label: 1,490 of 1,490 and 1,505 of 1,505 canonical views, 49.6–50.4% of default and broad views, and 45–51% for up–down (42–52% across the individual cohorts), where the rule is a no-information control. Over views with both projected boxes, entity i (lower ID) is the image-left box in 99.73% of canonical left–right views and the larger (78.08%) or lower (74.28%) box in canonical front–back views, while the higher box is the room-frame "above" entity in 99.49% of canonical, 99.15% of default and 82.47% of broad up–down views; on the ten unused rooms alone these rates are 99.78, 78.67, 72.35, 99.64, 99.19 and 80.95%. Ties in box coordinates or areas count as "not left", "not lower", "not larger" or "not higher". These are statistics of the evidence, computed by a reviewed audit run, not model outputs.

Probe. Features are the nine rotation entries and, for subject and anchor, the projected-box centroid (x, y divided by 256), width, height, area and a missing-box flag. A scikit-learn `HistGradientBoostingClassifier` (200 iterations, learning rate 0.1, 31 leaves, no early stopping, no validation-based selection) is fitted per relation, training policy (canonical, default, or their union) and feature set (all; masks only; rotation with centroids; rotation with sizes) on the 28 training rooms, both roles. Evaluation covers the same clean primary views, both roles and all three policies of the six development-evaluation and ten unused rooms. The mirrored-rotation intervention

Table 15: Rotation interventions on the frozen verifiers, ten unused rooms, canonical views, room-macro accuracy (%) with 95% room bootstraps where shown. Pixels and masks are unchanged; only the rotation input is replaced by that of the camera mirrored through the pair midpoint or by zeros. The last column gives the canonical-minus-default effect under identity / mirrored rotation.

Relation	Frozen model	Identity	Mirrored	Zero	Changed	Effect identity / mirrored
Front–back	canonical-trained RGB+boxes	66.19 [63.36, 69.15]	33.89 [31.48, 36.55]	50.50	55.5%	$+10.43$ / -11.60
	canonical-trained boxes only	80.16 [75.38, 84.54]	20.56 [16.07, 25.36]	51.59	92.3%	$+20.91$ / -20.61
	mixed-trained boxes only	71.69 [67.69, 75.60]	28.33 [24.47, 32.41]	50.62	73.9%	$+13.81$ / -14.56
	default-trained RGB+boxes	49.71	49.33	49.29	1.2%	-0.68 / -0.98
Left–right	canonical-trained RGB+boxes	48.96	48.86	48.90	2.1%	-0.56 / -0.62
	mixed-trained boxes only	62.74 [61.74, 63.62]	37.42 [36.47, 38.57]	49.89	31.4%	$+6.40$ / -5.81
Up–down	canonical-trained RGB+boxes	94.27 [93.01, 95.63]	94.24	94.23	0.1%	-0.65 / -0.65
	default-trained boxes only	94.99	94.96	94.98	0.1%	-1.65 / -1.72

replaces R by $R\,\mathrm{diag}(-1,-1,1)$, the camera-from-room rotation of the camera rotated 180 degrees about the vertical axis through the pair midpoint at the same elevation; the zero intervention sets all nine entries to zero. Rooms are the independent unit.

Frozen-model interventions. The 63 frozen confirmation checkpoints were re-evaluated on the same clean primary views with the rotation input replaced by the mirrored camera's rotation or by zeros, pixels and masks unchanged (Table 15; 0.135 lock-held GPU hours in total). Every front–back model that had learned inverts under the mirror and falls to chance under zero rotation; the default-trained front–back RGB models, which never learned, change about 1% of their predictions. Left–right RGB models stay at chance: the canonical- and mixed-trained groups change 1–2% of their predictions, the default-trained group 11.2% (mirrored) and 6.6% (zero) through one seed that weakly reads the pose, and eight of the nine checkpoints put 74–99% of their predictions in one class; the boxes-only models that had partially learned (mixed training, 62.74%) invert to 37.42%. All 27 up–down conditions, including the gravity-overlay and broad-pose variants, change at most 0.6% of their predictions after seed averaging (at most 1.2% for a single checkpoint) and keep their canonical-minus-default effects to within 0.2 point on the unused rooms (0.25 point on the development rooms), so those verifiers barely respond to the rotation. Zero rotation is an out-of-support ablation reported for completeness; the mirrored rotation stays in distribution. Identity re-inference reproduces the confirmation estimates to within 0.10 point for the RGB cells of Table 1 (e.g. $+10.43$ here versus $+10.41$) and within 0.19 points of accuracy and 0.29 points of effect for boxes-only cells, because batch composition differs. Aggregation follows Appendix C: case means within seed and room, seeds averaged within room, 2,000-replicate room bootstraps.

Additional cells. Default-trained probes with rotation and masks reach 87.00% (left–right) and 90.35% (front–back) on default views and 84–90% on canonical views, so the verifier's non-RGB inputs determine the answer well beyond canonical cameras. Rotation with mask centroids reproduces the full probe for left–right (99.55% canonical) and up–down, and for front–back both reduced feature sets carry most of the signal (centroids 85.71%, sizes 80.17% on canonical views, against 91.10% for the full set). Development-evaluation rooms show the same pattern. Probe training accuracies are 87–100%.

G Analytic ceiling protocol

For every room of both cohorts and every relation, all pairs satisfying the eligibility rule are enumerated and capped at 100 per room and relation by a seeded permutation; both role orders are evaluated. Ten standard-normal draws per room and entity are shared across $\sigma \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0\}$ m. Noisy centers $\hat{c} = c + \sigma\epsilon$ place the canonical, default and broad cameras with the rules of Appendix A.1 (default azimuth and broad azimuth and elevation seeded per pair) and feed the geometry solver $\mathrm{label}(\hat{c}_A, \hat{c}_B)$. The ideal verifier uses the true centers: with camera right, up and forward axes r, u, f and $d = c_A - c_B$, it predicts the sign of d_x as $\mathrm{sign}(d\cdot r)\ \mathrm{sign}(r_x)$ for left–right, of d_y as $\mathrm{sign}(d\cdot f)\ \mathrm{sign}(f_y)$ for front–back, and of d_z as $\mathrm{sign}(d\cdot u)\ \mathrm{sign}(u_z)$ for up–down. For canonical left–right cameras $r = \hat{h}$, so the converted answer equals the noisy solver's answer whenever $(c_B - c_A)\cdot\hat{h} > 0$; a reversal is recorded when the noisy

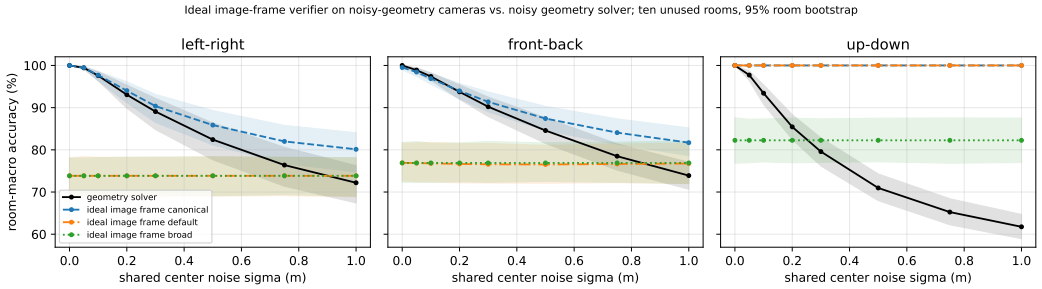


Figure 4: Room-macro accuracy of the noisy geometry solver and of an ideal image-frame verifier on canonical, default and broad cameras placed from the same noisy centers, ten unused rooms, 95% room bootstraps. Left–right and front–back canonical curves track the geometry solver; up–down is exact under horizontal cameras for this perspective-free ideal; default and broad cameras are capped by the conversion.

and true horizontal pair directions have negative dot product. Accuracies are room-macro; gains are paired room differences with 2,000-replicate percentile bootstraps. The 40 development rooms give the same pattern, with accuracies within 3.3 points and gains within 2.3 points of the unused-room values. Identity errors, occlusion and rendering quality are outside this bound.

H Extended related work

Canonical views. Canonical views entered vision science as the preferred, most typical views of objects (Palmer et al., 1981), and graphics selects informative views by maximizing viewpoint entropy (Vázquez et al., 2003). Both notions are properties of an object or a scene. Our canonical cameras are properties of a queried relation: they place the pair so that the relation becomes a displacement along one image axis.

Frames of reference and frame conversion. COMFORT shows that VLMs are neither robust nor consistent across frames of reference and default to the camera-relative conventions of English (Zhang et al., 2025). APC builds an abstract scene representation and simulates the requested viewpoint, so that other-perspective questions become camera-perspective questions (Lee et al., 2025). Allocentric Perceiver recovers metric 3D structure, transforms it into a query-conditioned target frame and passes it to the model as text (Wang et al., 2026). These methods and our frame reduction share the principle that the frame change is computed outside the model. Frame reduction is the special case in which the target frame is the rendered camera's own: the model answers from pixels in the image frame, and one sign per relation, fixed by the known camera axes, converts the answer. Our contribution is not this principle but a controlled test of when canonical views help. On identical pixels, cameras and roles, the canonical gain for left–right appears under the image-frame prompt, while the room-frame prompt with the numeric camera rotation stays near chance (Section 7).

View-dependent grounding and visual marks. ReferIt3D separates view-dependent from view-independent referring expressions in ScanNet scenes (Achlioptas et al., 2020). Our fixed-reference relations are view-independent by construction, so the evidence camera changes only what the verifier sees. We mark the two entities with labelled coloured outlines, as in visual mark prompting (Yang et al., 2023).

Partial-input baselines and position information. Question-prior and hypothesis-only analyses show that a model can succeed without the evidence a benchmark intends to test (Agrawal et al., 2018; Poliak et al., 2018). Our pose-and-mask probe applies the same test to a non-image input: it drops the RGB image and keeps the camera rotation and the projected mask statistics, and the mirrored-rotation intervention then shows which input determines the answer (Appendix F). Position information in convolutional networks (Islam et al., 2020; Kayhan & van Gemert, 2020; Liu et al., 2018; Lin et al., 2020) explains which identity cues the pooled verifier can read.

Models and software. We evaluate GPT-4o (OpenAI, 2024), GPT-5.5 (OpenAI, 2026), Qwen3-VL-Plus, an API model of the Qwen3-VL family (Bai et al., 2025a), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025b). Reconstructions use Nerfstudio (Tancik et al., 2023), rendering uses gsplat (Ye et al., 2025), the probe uses scikit-learn (Pedregosa et al., 2011), and intervals use a room-level cluster bootstrap (Field & Welsh, 2007).