HELP OR HINDER? VIEW PRESELECTION IN VLM-BASED ZERO-SHOT 3D GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot 3D visual grounding often restricts a scanned scene to a small set of query-relevant views before a vision–language model (VLM) makes its grounding decision. This reduces computation, but it also removes visual evidence before the grounding model can assess it. We ask whether the accuracy effect of this restriction changes with the grounding VLM. We introduce *Observe-Then-Ground* (OTG), a controlled evaluation protocol in which the VLM selects one image and localizes the target in 2D, and a fixed downstream pathway performs segmentation, RGB-D lifting, and 3D instance assignment. Only the observations supplied to the VLM are varied. Across the same nested CLIP-ranked Top-3/6/12/24 inputs, dense Qwen3-VL checkpoints are less accurate as the input expands beyond their best tested compact condition, whereas GPT-5.6 and Gemini improve through Top-24. Paired-query analysis shows why aggregate accuracy alone is insufficient: more views can recover targets absent from a smaller input, but can also change grounding when the target was already visible. From Top-6 to Top-24, target recovery accounts for most of the net gain for GPT and Gemini, while all dense Qwen checkpoints lose final 3D accuracy on the same already-covered queries. Thus, under this controlled CLIP-ranked interface, query-dependent preselection is an accuracy-relevant component of the grounding system, and its effect differs across the VLM configurations evaluated here.

1 INTRODUCTION

3D visual grounding identifies the object in a scene described by a natural-language referring expression (Chen et al., 2020; Achlioptas et al., 2020). Recent zero-shot approaches use general-purpose vision–language models (VLMs) for this task (Xu et al., 2025a; Drozdov et al., 2026). A scanned scene is observed from many viewpoints, and grounding often requires distinguishing among multiple plausible instances using appearance, spatial relations, and broader scene context. The input must therefore help the model identify the particular instance, not merely recognize its category.

Recent zero-shot systems commonly select, retrieve, or restructure visual evidence before the VLM makes its grounding decision (Xu et al., 2025a; Lin et al., 2025; Jin et al., 2025; Li et al., 2025; Drozdov et al., 2026). A small set of relevant views reduces computation and may make the target easier to find. Yet selecting those views also limits the evidence available for grounding. A view that appears weakly related to the query in isolation may become useful when comparing similar objects or resolving a spatial relation. Early preselection removes views before the grounding VLM can assess them in that context.

This tension suggests a more basic question than choosing a single "best" number of views: *is pre-inference evidence restriction a model-independent preprocessing operation?* A smaller input can improve the signal-to-clutter ratio, while a larger input can preserve observations needed to find the target or distinguish it from alternatives. These effects need not balance in the same way for different VLMs.

We therefore study the response to evidence restriction. Every query starts from the same sampled RGB candidate pool. A fixed CLIP ranking defines nested Top-k inputs, $S_3 \subseteq S_6 \subseteq S_{12} \subseteq S_{24} \subseteq S_{\text{Full}}$. Smaller budgets remove more observations before VLM inference; larger budgets preserve more of the sampled scene. "Full" denotes the complete sampled candidate pool, not every original ScanNet 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.

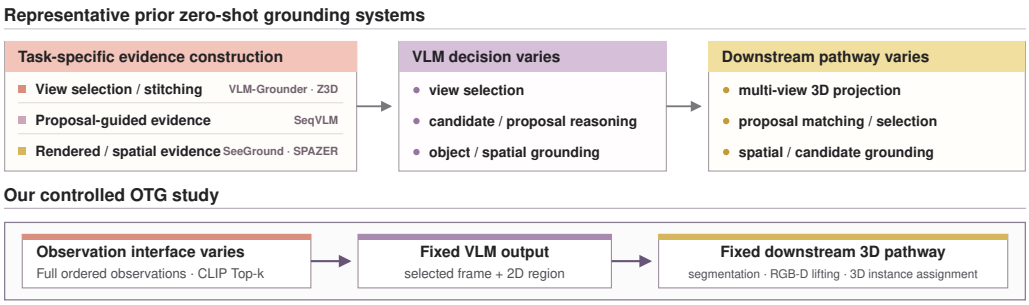

Figure 1: Comparison between representative prior zero-shot 3D grounding pipelines and OTG. Prior systems often co-design the scene-to-VLM evidence, VLM decision format, and downstream 3D pathway. OTG varies only the observations presented to the VLM while fixing the output format and subsequent 2D-to-3D pathway.

Existing zero-shot methods vary both the evidence presented to the VLM and the procedure that converts its decision into a 3D prediction. To separate these choices, we introduce *Observe-Then-Ground* (OTG), a controlled evaluation protocol. The VLM selects one image and localizes the target in 2D; a fixed downstream procedure then performs segmentation, RGB-D lifting, and 3D instance assignment. The purpose of OTG here is to make changes in the supplied observations measurable while holding the grounding contract and downstream 3D pathway fixed (Figure 1).

We evaluate this model $\times$ evidence-budget interaction using four dense Qwen3-VL models (2B, 4B, 8B, and 32B) (Bai et al., 2025), GPT-5.6 (OpenAI, 2026), and Gemini (Google, 2026). Each receives the same CLIP-selected Top-3/6/12/24 views (Radford et al., 2021); selected backbones are also evaluated with full observation.

Across the dense Qwen3-VL checkpoints, the highest observed final 3D accuracy among the tested restricted budgets occurs at Top-3 for 2B/4B and Top-6 for 8B/32B; all four are lower at Top-24 despite 100% target-visible recall. GPT-5.6 instead rises from 57.2 at Top-3 to 62.0 at Top-24 and reaches 62.4 with the complete sampled pool. Gemini rises from 56.4 to 62.4 through Top-24 and then falls to 60.0 with the complete sampled pool. These budgets characterize model-specific responses, not a universal optimal view count. For GPT, Top-6 remains less accurate than Full when PE-Core is replaced with SigLIP2 (Section 4.3).

More importantly, paired analysis separates two effects that aggregate accuracy conflates. Expanding the input can (i) recover targets that were not visible in the smaller input and (ii) alter grounding on queries for which the target was already available. From Top-3 to Top-24, GPT and Gemini gain through both effects. From Top-6 to Top-24, most of their net gain comes from target recovery, whereas every dense Qwen checkpoint loses final 3D accuracy on the same 235 already-covered queries. This model-dependent conditional response is the central observation of the paper.

Our contributions are threefold:

- **We show that view preselection can be an accuracy-relevant design choice, not merely a computational shortcut.** Under the same CLIP-ranked inputs and shared downstream pathway, the accuracy effect of preselection differs across grounding VLMs, motivating joint evaluation of the evidence budget and downstream model.
- **We introduce OTG, a controlled evaluation protocol for this interaction.** OTG fixes the VLM output contract and the 2D-to-3D instantiation pathway while varying the observations supplied before VLM inference. The resulting implementation is also competitive with prior zero-shot systems on ScanRefer-250.
- **We separate target recovery from grounding changes on already-covered queries.** Paired-query analysis relates these two effects to selected-frame visibility, direct 2D localization, and final 3D accuracy. We additionally report the SeqVLM/Z3D-compatible Unique/Multiple class-multiplicity split as a descriptive diagnostic, an alternative-selector check, and field-of-view and observation-order controls.

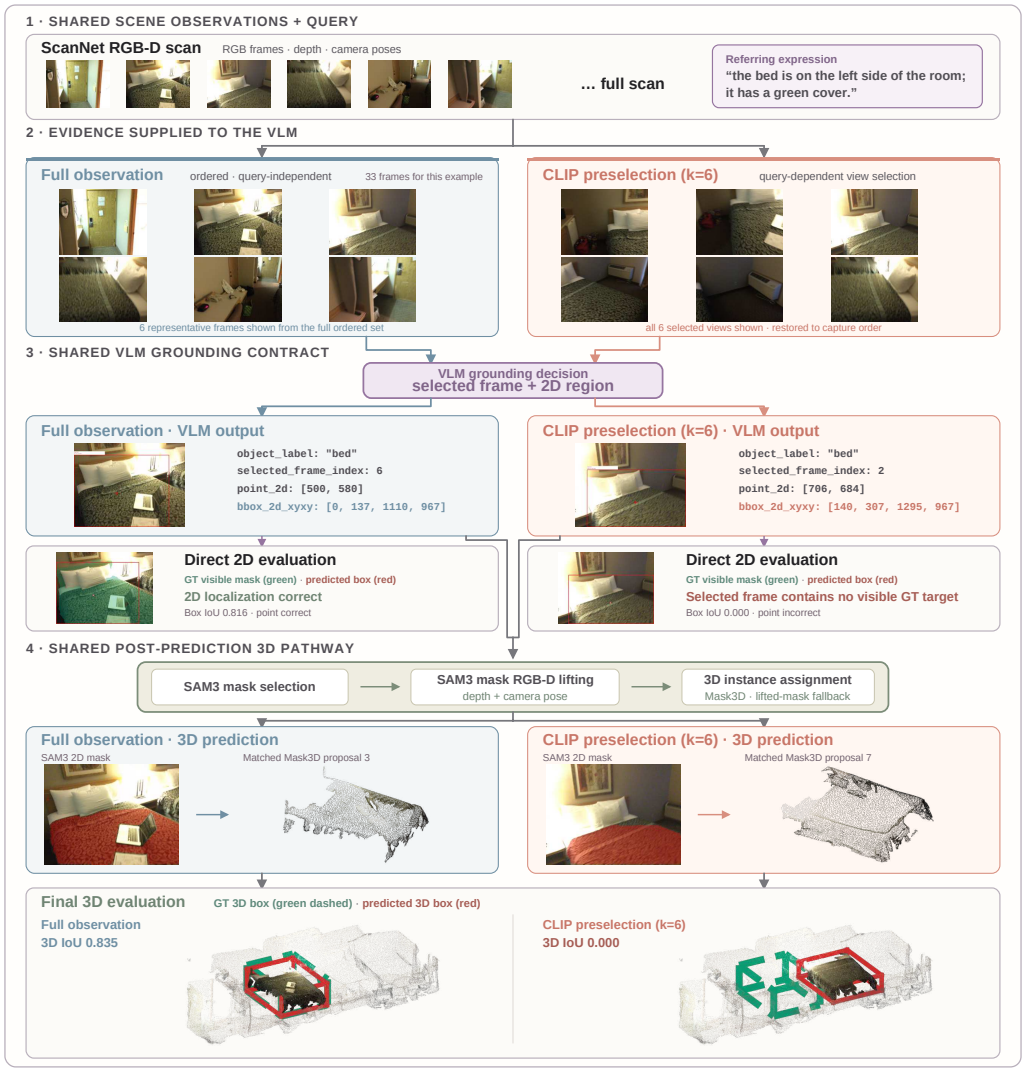

Figure 2: **Concrete OTG comparison using GPT-5.6.** Full observation and CLIP Top-6 differ only in the evidence supplied before VLM inference; they share the grounding contract and post-prediction 3D pathway. Box IoU is computed against the tight bounding rectangle of the GT mask.

2 Related Work

3D visual grounding and evidence construction. ScanRefer and ReferIt3D established language grounding in reconstructed indoor scenes (Chen et al., 2020; Achlioptas et al., 2020). Supervised methods learn correspondences between language and point clouds, object proposals, or multimodal scene tokens (Wu et al., 2023; Zhu et al., 2023; Jain et al., 2025). Recent zero-shot systems instead use general-purpose VLMs, but differ in the evidence constructed before grounding. VLM-Grounder dynamically stitches image sequences and aggregates multi-view projections (Xu et al., 2025a); SeqVLM filters proposals and constructs proposal-guided sequences (Lin et al., 2025); SPAZER selects a holistic rendered view (Jin et al., 2025); and SeeGround builds a query-conditioned rendered interface (Li et al., 2025). These systems make selecting and organizing visual evidence an integral part of the grounding pipeline.

Spatial understanding from multi-view context. Recent studies evaluate spatial reasoning and multi-view understanding in VLMs (Yang et al., 2025; Yeh et al., 2026). Complementary work develops position-aware and spatially grounded video models, as well as hierarchical scene represen-

tations (Zheng et al., 2025; Yang et al., 2026b;a; Xu et al., 2025b). We focus on which observations are made available to the grounding model before it predicts the target.

View selection and evidence budgets. Z3D combines CLIP-based preselection with VLM view selection and reports ablations on view usage and backbone choices (Drozdov et al., 2026). SPAZER examines candidate counts and rendered-view budgets (Jin et al., 2025), while broader multimodal work develops retrieval, frame selection, and adaptive evidence acquisition (Zou et al., 2026; Xu et al., 2026). We focus on how the accuracy effect of view preselection changes across grounding VLMs. Using the same ranked observations and a shared downstream 3D pathway, we compare each model's response to the same evidence expansion. Paired-query analysis then separates recovery of omitted targets from grounding changes when the target was already visible.

3 OBSERVE-THEN-GROUND

3.1 TASK AND CONTROLLED PROTOCOL

Let a scene provide an RGB-D observation sequence $\mathcal{O} = \{(I_t, D_t, P_t)\}_{t=1}^{T}$, where I_t is an RGB frame, D_t is its depth map, and P_t is the camera pose. Given a referring expression q, the goal is to predict a 3D axis-aligned bounding box B^{3D} for the referred object instance. For an evidence condition γ, the observation interface constructs the RGB input

$$\mathcal{I}_\gamma = g_\gamma\big(\{I_t\}_{t=1}^{T}, q\big)\,. \tag{1}$$

Only $\mathcal{I}_\gamma$ is exposed to the VLM; depth and pose are reserved for post-prediction 3D instantiation.

The VLM receives q and $\mathcal{I}_\gamma$. For each VLM, the grounding prompt is fixed across evidence budgets. The shared output contract is

$$y = (j, c, p, B^{2D}), \tag{2}$$

where j is the selected image index, c is a concise object category, p is a 2D point, and B^{2D} is a 2D bounding box in image j. Thus the VLM first commits to an observation and a 2D region; it never receives depth, pose, point clouds, or 3D proposals. Figure 2 illustrates this shared contract. Across conditions, only the evidence interface changes.

3.2 EVIDENCE INTERFACES

Every query starts from the same RGB candidate pool, sampled every 30 ScanNet (Dai et al., 2017) frames. *Full observation* denotes the complete stride-30 sampled pool in capture order; it is not the set of every original ScanNet frame. For restricted conditions, CLIP ranks this same pool by query–image similarity and retains up to $k \in \{3, 6, 12, 24\}$. Because all restricted conditions use one frozen ranking, the selected sets are nested. The retained frames are then restored to capture order before VLM inference.

The budget k controls the amount of view preselection: smaller k removes more frames, while larger k gives the VLM more images from which to choose and localize the target. The source pool, ranking, query, prompt, output schema, and downstream pathway are fixed across the dense Qwen sweep. GPT-5.6 and Gemini use the same Top-k interfaces. Full is evaluated for Qwen3-VL-2B, Qwen3-VL-30B-A3B, GPT, and Gemini.

CLIP provides a fixed, reproducible ranking with direct precedent in zero-shot 3D grounding (Radford et al., 2021; Drozdov et al., 2026). We hold this selector fixed to isolate how restriction interacts with the downstream model. Target-visible recall is computed from these frozen inputs before VLM inference, so it depends on k rather than the downstream backbone.

3.3 FIXED 2D-TO-3D INSTANTIATION

After the VLM predicts (j, c, p, B^{2D}), category c prompts SAM3 (Carion et al., 2026) on image I_j. We select the candidate mask with the largest pixel-level IoU with B^{2D}; p is retained for evaluation but does not affect mask selection. If no category-prompted mask overlaps the box, the box is used as an explicit SAM3 prompt. The selected mask is eroded, restricted to its dominant valid-depth region, and back-projected using D_j, camera intrinsics, and P_j.

The lifted points are voxelized and matched to Mask3D instance proposals (Schult et al., 2023) by geometric voxel IoU. The highest-overlap proposal defines the final 3D prediction; if none overlaps, the lifted SAM3 mask provides the fallback 3D candidate. This pathway is identical across evidence interfaces and models; Appendix B gives implementation details, and Appendix C gives the full prompts. We evaluate the selected frame and 2D region directly before this pathway and report final 3D accuracy after instantiation.

4 EXPERIMENTS

Setup. We use the fixed 250-expression ScanRefer panel released with VLM-Grounder (Xu et al., 2025a) and the SeqVLM/Z3D-compatible split of 66 *Unique* and 184 *Multiple* queries (Lin et al., 2025; Drozdov et al., 2026). In this split, Unique means that the referred target is the only instance of its NYU40 semantic class in the scene, whereas Multiple means that the scene contains at least one additional instance of the same semantic class. This describes scene-level class multiplicity, not linguistic or perceptual ambiguity. Full observation contains 67.2 stride-30 frames per query on average. We report final axis-aligned 3D box Acc@0.25/0.50; Appendix A defines direct 2D, point, target-visible recall, and official aligned 3D metrics.

Our model $\times$ evidence-budget study evaluates all six VLMs at Top-3/6/12/24. The dense within-family comparison includes Qwen3-VL Thinking checkpoints at 2B, 4B, 8B, and 32B (Bai et al., 2025). A single CLIP ranking, computed with PE-Core-bigG-14-448 (Bolya et al., 2025), is shared across all checkpoints. GPT-5.6 and Gemini 3.8 Flash are evaluated across Top-3/6/12/24/Full. These are independent cross-family comparisons, not points on the dense scaling curve. Qwen3-VL-30B-A3B provides an additional Top-6-versus-Full endpoint. Across evidence budgets, the original images and each model's per-image processing settings are unchanged. Appendix B gives checkpoints, prompts, processing and generation settings, and downstream thresholds.

For direct 2D evaluation, the official filtered ScanNet instance map supplies the visible target mask and its tight box. Point accuracy tests whether the predicted point lies inside this mask. Frame accuracy tests visibility in the selected image, whereas target-visible recall tests the entire presented set. Queries with no target-visible input remain in the denominator and receive zero direct-localization credit.

4.1 GROUNDING PERFORMANCE AGAINST PRIOR SYSTEMS

OTG provides competitive zero-shot grounding performance. With the open-weight Qwen3-VL-32B and Top-6 inputs, OTG reaches 56.4 overall 3D Acc@0.50 on ScanRefer-250, compared with 54.4 for Z3D in Table 1; GPT-5.6 with Full reaches 62.4. Supervised results use the full validation set and provide context only. We treat these cross-system numbers as task-performance context rather than as the basis for the paper's main claim: the evidence-budget conclusions below come from controlled within-model comparisons.

4.2 EVIDENCE-BUDGET RESPONSES ACROSS VLMS AND QUERY REGIMES

Overall response. The same evidence expansion lowers accuracy beyond compact inputs for Qwen, but improves GPT and Gemini through Top-24 (Figure 3(a), Table 2). Within Qwen, the highest observed final 3D accuracy occurs at Top-3 for 2B/4B and Top-6 for 8B/32B. Moving from three to six views lowers 3D Acc@0.50 by 7.2 points for 2B, changes 4B by -0.4, and improves 8B and 32B by 2.4 and 1.6 points, respectively. The adjacent-budget differences are small for the larger checkpoints; all four models show a clearer decline as the budget expands to 12 and 24.

GPT-5.6 instead improves from 57.2 at Top-3 to 60.4, 60.8, and 62.0 at Top-6/12/24, reaching 62.4 with Full. Gemini independently rises from 56.4 to 59.6, 60.8, and 62.4 across the four restricted budgets, then falls to 60.0 with Full.

Scene-clustered intervals support this difference in budget response. The Top-3-to-Top-24 change differs by 20.0 points between GPT and Qwen3-VL-2B (95% scene-clustered paired bootstrap interval: 12.4 to 27.6), and by 21.2 points between Gemini and Qwen3-VL-2B (13.6 to 28.7). Appendix G reports the intervals for all six models and the resampling procedure.

Table 1: ScanRefer 3D grounding results. Supervised methods report their original full-validation results; zero-shot methods are evaluated on ScanRefer-250. All values are percentages; "–" denotes an unreported metric.

	Unique		Multiple		Overall	
Method	@0.25	@0.50	@0.25	@0.50	@0.25	@0.50
Supervised						
ScanRefer (Chen et al., 2020)	67.6	46.2	32.1	21.3	39.0	26.1
EDA (Wu et al., 2023)	85.8	68.6	49.1	37.6	54.6	42.3
3D-VisTA (Zhu et al., 2023)	81.6	75.1	43.7	39.1	50.6	45.8
UniVLG (Jain et al., 2025)	–	–	–	–	63.5	56.4
Zero-shot; ScanRefer-250						
VLM-Grounder (Xu et al., 2025a)	66.0	29.8	48.3	33.5	51.6	32.8
SeqVLM (Lin et al., 2025)	77.3	72.7	47.8	41.3	55.6	49.6
SPAZER (Jin et al., 2025)	80.9	72.3	51.7	43.4	57.2	48.8
Z3D (Drozdov et al., 2026)	87.9	81.8	51.7	44.6	61.2	54.4
MCM-VG (Yin et al., 2026)	81.8	74.2	54.9	46.2	62.0	53.6
OTG, Qwen3-VL-32B (Top-6)	**90.9**	**84.8**	51.6	46.2	62.0	56.4
OTG, GPT-5.6 (Full)	87.9	83.3	**60.9**	**54.9**	**68.0**	**62.4**

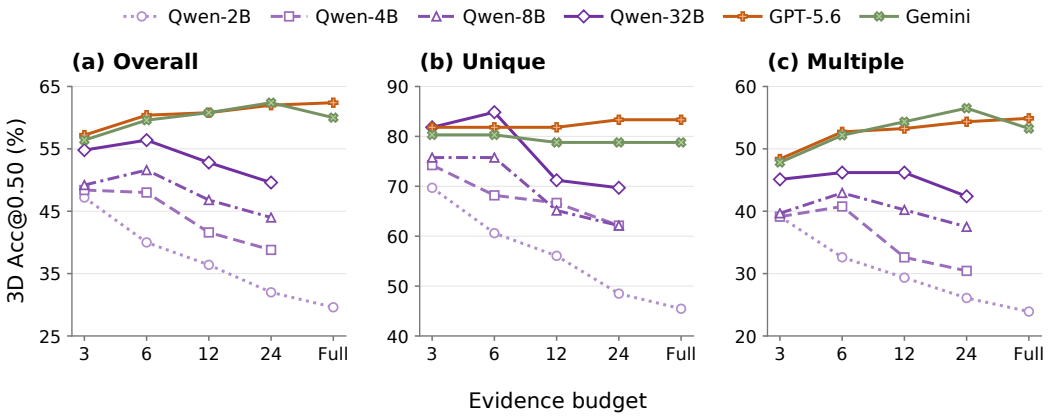

Figure 3: **Evidence-budget responses across VLMs and scene-level class multiplicity.** Final 3D Acc@0.50 on (a) all 250 queries, (b) 66 Unique queries, and (c) 184 Multiple queries. Horizontal-axis numbers denote CLIP Top-k budgets; Full uses the complete stride-30 sampled pool. Dense Qwen checkpoints reach their highest tested accuracy at Top-3 or Top-6. GPT and Gemini improve through Top-24; their aggregate gains are concentrated in the Multiple subset, which also contains every target omitted by Top-6. GPT largely saturates at Full, whereas Gemini declines. Curves show only evaluated conditions. Panels use different vertical ranges for readability.

Target coverage alone does not explain these curves. Target-visible recall rises from 90.0% at Top-3 to 94.0%, 98.4%, and 100.0% at Top-6/12/24. Yet Top-24 is not the best condition for any dense checkpoint.

Unique and Multiple. GPT and Gemini gain mainly on Multiple queries, while Unique changes little (Figure 3(b,c)). For GPT, Unique accuracy is 81.8, 81.8, 81.8, 83.3, and 83.3 across Top-3/6/12/24/Full, while Multiple rises from 48.4 through 52.7, 53.3, and 54.3 to 54.9. Gemini shows the same concentration of aggregate gains in the Multiple subset: 47.8, 52.2, 54.3, and 56.5 across Top-3/6/12/24, whereas Unique is 80.3, 80.3, 78.8, and 78.8. With Full, Gemini's Unique accuracy remains 78.8 and Multiple falls to 53.3.

Table 2: Model × evidence-budget results: final 3D Acc@0.50 (%). Full uses the complete sampled pool; "–" indicates conditions not evaluated. The MoE checkpoint is a separate endpoint comparison.

Backbone	Top-3	Top-6	Top-12	Top-24	Full
Qwen3-VL-2B	47.2	40.0	36.4	32.0	29.6
Qwen3-VL-4B	48.4	48.0	41.6	38.8	–
Qwen3-VL-8B	49.2	51.6	46.8	44.0	–
Qwen3-VL-32B	54.8	56.4	52.8	49.6	–
GPT-5.6	57.2	60.4	60.8	62.0	62.4
Gemini	56.4	59.6	60.8	62.4	60.0
Qwen3-VL-30B-A3B (MoE)	–	50.0	–	–	40.4

Target coverage also differs between these groups: all 15 targets omitted by Top-6 belong to Multiple. Section 4.3 separates recovery of these targets from changes on already-covered queries; recovery explains much of GPT's and Gemini's net improvement beyond Top-6.

Qwen's response also differs between the two groups, but broader input does not consistently help either. From three to six views, 4B loses 6.1 points on Unique but gains 1.6 on Multiple; 8B is unchanged on Unique and gains 3.3 on Multiple.[1] Both splits improve at 32B (Table 7). Further expansion brings substantial Unique losses across the dense family and does not consistently help Multiple.

4.3 What changes when more views are provided?

More views can recover an omitted target, but can also change localization when that target was already visible. We separate these cases using the official visible-instance annotations, and count both wrong-to-correct changes (gains) and correct-to-wrong changes (losses). Correctness uses IoU ≥ 0.50; no queries are removed from the overall evaluation.

Table 3: Correct-query counts at IoU ≥ 0.50 on the same 235 queries whose targets are visible in Top-6. Arrows show Top-6 $\rightarrow$ Top-24.

Backbone	Direct 2D: $6 \rightarrow 24$	Final 3D: $6 \rightarrow 24$
Qwen3-VL-2B	$37 \rightarrow 20$	$100 \rightarrow 80$
Qwen3-VL-4B	$58 \rightarrow 21$	$120 \rightarrow 96$
Qwen3-VL-8B	$60 \rightarrow 33$	$129 \rightarrow 108$
Qwen3-VL-32B	$125 \rightarrow 73$	$141 \rightarrow 123$
GPT-5.6	$165 \rightarrow 163$	$151 \rightarrow 152$
Gemini	$167 \rightarrow 174$	$149 \rightarrow 150$

When the target is already present. From six to 24 views, all four Qwen models lose direct 2D and final 3D accuracy on the same 235 already-covered queries, with 95% scene-clustered paired bootstrap intervals below zero. GPT and Gemini each gain one correct 3D prediction, with intervals spanning zero (Table 3, Appendix G.2). Across all 250 queries, Qwen3-VL-32B selects a target-visible image more often (197 to 203), while correct 2D boxes fall from 125 to 73 and correct final 3D predictions from 141 to 124. Selecting a target-visible view is distinct from correctly localizing the instance. Recovery and localization losses oppose one another here: one recovered 3D success is outweighed by 18 net losses on already-covered queries. Table 15 gives the six-model paired counts; Figures 5–6 show example predictions.

Direct 2D and final 3D can also change in opposite directions. From Top-3 to Top-6, 8B and 32B improve in final 3D while direct 2D Acc@0.50 falls from 28.8 to 24.0 and from 52.8 to 50.0, respectively (Appendix D). Segmentation and instance assignment intervene between the two measurements (Figure 4).

[1] These changes use unrounded accuracies.

GPT-5.6. GPT improves on already-covered queries and recovers omitted targets from Top-3 to Top-24. From Top-6 to Top-24, recovery accounts for most of the net gain. From Top-3 to Top-24, the 225 already-covered queries have 14 gains and seven losses, a net gain of seven. Five further gains come from the 25 omitted-target queries, giving 19 gains and seven losses overall.

From Top-6 to Top-24, the 235 already-covered queries have 12 gains and 11 losses, a net gain of one. Three further gains come from the 15 omitted-target queries, accounting for most of the net improvement.

From Top-6 to Full, the already-covered group has 15 gains and 15 losses, with five gains from omitted targets. Full gives only one more correct prediction than Top-24 (62.4 versus 62.0).

Gemini. Gemini also improves on already-covered queries and recovers omitted targets from Top-3 to Top-24, but loses some of the gain at Full. From Top-3 to Top-24, the 225 already-covered queries have 20 gains and 12 losses, a net gain of eight. Seven further gains come from the 25 omitted-target queries, giving 27 gains and 12 losses overall. Direct 2D localization also improves on the already-covered group (Appendix F).

From Top-6 to Top-24, the 235 already-covered queries have 11 gains and 10 losses, a net gain of one. Six further gains come from the 15 omitted-target queries, again accounting for most of the net improvement.

From Top-6 to Full, the already-covered group has 11 gains and 15 losses, offset by five gains from omitted targets, leaving one net gain. Unlike GPT, Gemini declines from Top-24 to Full: eight gains and 14 losses leave six fewer correct predictions (60.0 versus 62.4), despite complete target-visible recall in both conditions.

Relation to the Multiple split. The Multiple curves combine target recovery with changes on already-covered queries. All 15 targets omitted by Top-6 belong to Multiple. For GPT, the 169 already-covered Multiple queries have 11 gains and 11 losses at Top-24, and 14 gains and 15 losses at Full, relative to Top-6. Its net improvement on Multiple therefore mainly reflects target recovery, rather than a net gain on already-covered Multiple queries. Gemini's eight gains and 14 losses from Top-24 to Full all occur on Multiple.

View-selector ablation. Replacing PE-Core with SigLIP2 (Tschannen et al., 2025) at GPT Top-6 retains only 63.1% shared frames on average. With the same prompt, decoding, and chronological presentation, target-visible recall rises from 94.0% to 95.6%, but neither direct 2D nor final 3D accuracy improves (Table 4). This limited check reinforces the distinction between coverage and grounding quality: higher target-visible recall alone does not imply better grounding.

Table 4: GPT-5.6 Top-6 selector comparison: all 250 queries, IoU ≥ 0.50 (%).

Selector	Target recall	2D accuracy	3D accuracy
PE-Core	94.0	66.0	60.4
SigLIP2	95.6	65.2	58.8

4.4 Broader evidence and efficiency

For GPT and Gemini, broader selected evidence approaches or exceeds Full accuracy at lower image count. Top-24 uses 23.8 rather than Full's 67.2 images on average—64.6% fewer— with complete target-visible recall. It finishes one correct query below Full for GPT (62.0 versus 62.4) and six above for Gemini (62.4 versus 60.0). Direct 2D accuracy also shows no gain from Full (Appendix D). Conversely, compact Top-6 improves over Full for Qwen3-VL-2B and Qwen3-VL-30B-A3B (Table 2), reinforcing the model-dependent value of preselection.

4.5 FIELD-OF-VIEW AND ORDER CONTROLS

Both controls reduce GPT's final 3D accuracy under the same OTG pathway. Center-50 crops halve image area: Full retains higher target recall than Top-6 (99.6% versus 92.4%), while 3D Acc@0.50 drops similarly from uncropped inputs (4.4 versus 4.0 points), to 58.0 versus 56.4.

The order control changes the organization of the same images. Shuffling identical Full frames changes 3D accuracy from 62.4 to 59.6 with timestamps and 58.4 without them for the single tested shuffle. Direct 2D changes are smaller and non-monotonic; full metrics are in Appendix E and Table 12.

5 DISCUSSION

Evidence restriction is part of the model–system interaction. The same nested evidence expansion produces different accuracy responses: dense Qwen declines beyond compact tested inputs, GPT and Gemini improve through Top-24, and Gemini declines at Full. Preselection is therefore an accuracy-relevant system choice, not only an efficiency choice.

Coverage and grounding on already-visible targets are distinct. GPT and Gemini gain through both effects from Top-3 to Top-24, but mainly through recovery from Top-6 to Top-24. All four dense Qwen models instead lose final 3D accuracy on already-covered queries in the latter expansion. Coverage alone cannot explain these curves.

Finding a view and localizing the object are distinct. Qwen3-VL-32B selects a target-visible frame more often at Top-24 than at Top-6, yet produces fewer correct 2D boxes. Direct 2D and final 3D can also diverge because segmentation and instance assignment intervene between them. OTG exposes these intermediate stages under a shared downstream pathway.

Evidence expansion changes the observation interface, not only evidence quantity. Increasing the budget changes the views and their organization within a longer sequence. The single GPT shuffle also changes accuracy without changing image identity. These curves therefore characterize the observation interface, not an isolated mechanism such as visual clutter or attention dilution.

Interpreting Unique and Multiple. GPT and Gemini gains concentrate on Multiple queries, which also contain every target omitted by Top-6. The split identifies where gains occur; paired analysis distinguishes target recovery from grounding changes within that group.

Scope and limitations. Our ScanRefer-250 sweep fixes the PE-Core ranking and downstream pathway; SigLIP2 tests only GPT at Top-6. Broader cross-selector and cross-dataset validation remains necessary. Each model–budget condition uses one saved generation per query; bootstrap intervals measure scene-resampling uncertainty, not run-to-run generation variability.

Implications for multi-view VLMs. Evidence budgets and selectors should be evaluated with the chosen backbone, not inherited as model-independent preprocessing constants. Evaluations should report target coverage and direct grounding separately from final downstream accuracy. Whether fine-tuning, alternative selectors, or spatial evidence organization reduce the observed sensitivity remains open.

6 CONCLUSION

We introduced OTG to study view preselection under a fixed grounding contract and downstream 3D pathway. The same CLIP-ranked evidence expansion produces different accuracy responses across VLMs: dense Qwen checkpoints decline beyond compact tested inputs, while GPT and Gemini improve through Top-24. Paired analysis separates target recovery from grounding changes on already-visible targets, whose balance depends on the model and starting budget. Under the controlled interfaces studied here, these results show that view preselection can be part of the VLM–evidence interaction rather than merely a model-independent computational shortcut. Evidence budgets should therefore be evaluated together with the grounding VLM.

REPRODUCIBILITY STATEMENT

Section 4 specifies the fixed query panel, evidence interfaces, model comparisons, and evaluation protocol. Appendices A–G provide the exact checkpoints, downstream settings, prompt, detailed metrics, diagnostic definitions, and scene-clustered uncertainty analysis. The released pipeline records the ranked and presented frames, target visibility, VLM prediction, and final 3D output needed to reproduce each evaluation and the paired-query counts. Code and reproduction materials are available in the anonymous repository at `https://anonymous.4open.science/r/otg-4463`.

ETHICS STATEMENT

We use the established ScanNet and ScanRefer benchmarks and collect no new human-subject data. Our evaluation inherits the privacy, geographic, and representational limitations of these datasets; released artifacts will follow their original licenses and access conditions.

AI USE STATEMENT

The authors conceived the research question, designed the experiments, interpreted the results, and determined all claims and conclusions. Under author direction, generative AI assistants supported hypothesis and method refinement, code implementation and editing, experiment orchestration, literature and result summarization, figure preparation, and manuscript drafting and editing. The authors reviewed all AI-assisted outputs, verified experimental settings and executions, numerical results, citations, and claims, and take full responsibility for the paper.

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 *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 422–440. Springer, 2020.

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

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Niessner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *Computer Vision – ECCV 2020*, volume 12365 of *Lecture Notes in Computer Science*, pp. 202–221. Springer, 2020.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2432–2443, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Nikita Drozdov, Andrey Lemeshko, Nikita Gavrilov, Anton Konushin, Danila Rukhovich, and Maksim Kolodiazhnyi. Z3D: Zero-shot 3D visual grounding from images. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 147–154. Association for Computational Linguistics, 2026.

Google. Gemini 3.8 Flash, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash/. Accessed September 26, 2026.

Ayush Jain, Alexander Swerdlow, Yuzhou Wang, Sergio Arnaud, Ada Martin, Alexander Sax, Franziska Meier, and Katerina Fragkiadaki. Unifying 2D and 3D vision-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26717–26739. PMLR, 2025.

Zhao Jin, Rong-Cheng Tu, Jingyi Liao, Wenhao Sun, Xiao Luo, Shunyu Liu, and Dacheng Tao. SPAZER: Spatial-semantic progressive reasoning agent for zero-shot 3D visual grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model

serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. URL https://arxiv.org/abs/2309.06180.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. SeeGround: See and ground for zero-shot open-vocabulary 3D visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3707–3717, 2025.

Jiawen Lin, Shiran Bian, Yihang Zhu, Wenbin Tan, Yachao Zhang, Yuan Xie, and Yanyun Qu. SeqVLM: Proposal-guided multi-view sequences reasoning via VLM for zero-shot 3D visual grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3094–3103. ACM, 2025.

OpenAI. Previewing GPT-5.6 Sol: A next-generation model, June 2026. URL https://openai.com/index/previewing-gpt-5-6-sol/. Accessed September 26, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation*, pp. 8216–8223. IEEE, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. EDA: Explicit text-decoupling and dense alignment for 3D visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19231–19242, 2023.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long video understanding with learnable retrieval in video-language models. *IEEE Transactions on Multimedia*, 28:5254–5261, 2026.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. VLM-Grounder: A VLM agent for zero-shot 3D visual grounding. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3961–3985. PMLR, 2025a.

Wenting Xu, Viorela Ila, Luping Zhou, and Craig T. Jin. TB-HSU: Hierarchical 3D scene understanding with contextual affordances. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):8960–8968, 2025b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025.

Jihan Yang, Zifan Zhao, Xichen Pan, Shusheng Yang, Junyi Zhang, Bingyi Kang, Hu Xu, Shang-Wen Li, and Saining Xie. Cambrian-P: Pose-grounded video understanding. In *Computer Vision – ECCV 2026*, 2026a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. In *International Conference on Learning Representations*, 2026b.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in MLLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):12000–12008, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/38188.

Yufei Yin, Jie Zheng, Qianke Meng, Zhou Yu, Minghao Chen, Jiajun Ding, Min Tan, Yuling Xi, Zhiwen Chen, and Chengfei Lv. Multiple consistent 2D–3D mappings for robust zero-shot 3D visual grounding. *arXiv preprint arXiv:2604.26261*, 2026.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2911–2921, 2023.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, 2026.

A Detailed Evaluation Definitions

Visible 2D ground truth. The official ScanNet filtered 2D instance map provides a pixel mask for each visible object instance in a frame. We convert the target mask to its tight enclosing rectangle for box IoU and test point accuracy directly against its pixels; the mask itself is not supplied to the VLM. Frames in which the target has no annotated visible pixel cannot produce a correct direct 2D localization.

Official aligned 3D coordinates. Predicted lifted points and ScanRefer ground-truth instance points are transformed into the official axis-aligned scene frame exactly once. Axis-aligned boxes are reconstructed and compared using volumetric IoU. Queries with IoU greater than or equal to the threshold count as correct.

Target-visible recall. Target-visible recall is measured before VLM inference. A query is recalled if at least one presented frame contains pixels of the ground-truth target in the official 2D annotation. Failed-recall queries remain in the accuracy denominator because removing them would hide an error introduced by the input interface.

Unique/Multiple class-multiplicity split. We use the frozen SeqVLM/Z3D-compatible partition of 66 Unique and 184 Multiple queries. Following the SeqVLM evaluation code, we exclude wall, floor, and ceiling instances, map the remaining scene-instance labels to NYU40 classes, and count all scene instances in the target's class. A count of one is Unique; a count greater than one is Multiple. This label is determined from scene annotation and is independent of the VLM prediction or evidence budget. It does not specify how many same-class alternatives are visible in a particular Top-k input and is therefore not treated as a direct measure of perceptual ambiguity.

Efficiency accounting. We report mean image count, estimated image tokens, and VLM request time per query. Image tokens are estimated from each input's dimensions using the documented GPT-5.6 Sol high-detail rule,[2] excluding text, output, and reasoning tokens. These are estimates rather than recorded API usage. Request time excludes downstream segmentation, RGB-D lifting, and 3D instance assignment.

[2] https://developers.openai.com/api/docs/guides/images-vision

B Implementation Details

The dense sweep uses the official `Qwen/Qwen3-VL-2B-Thinking`, `Qwen/Qwen3-VL-4B-Thinking`, and `Qwen/Qwen3-VL-8B-Thinking` BF16 checkpoints, together with the official `Qwen/Qwen3-VL-32B-Thinking-FP8` release. The additional Qwen endpoint uses `Qwen3-VL-30B-A3B-Thinking`. All Qwen checkpoints are served with vLLM (Kwon et al., 2023) on one NVIDIA RTX PRO 6000 Blackwell Max-Q GPU (96 GB). They use the checkpoint generation configuration, seed 0, and at most 20,480 new tokens, and return normalized $[0, 1000]$ x, y coordinates. GPT-5.6 denotes `gpt-5.6-sol` (OpenAI, 2026), accessed through the Responses API with low reasoning effort, no explicit temperature, and at most 1,024 output tokens; it returns native x, y image coordinates. Gemini denotes `gemini-3.8-flash`, accessed through the Interactions API with low thinking level, high media resolution, no explicit temperature, top-p, or top-k, and at most 8,192 output tokens; it returns normalized $[0, 1000]$ y, x coordinates. All normalized outputs are converted to native x, y image coordinates before the shared downstream pathway. CLIP ranking uses PE-Core-bigG-14-448 with `meta` pretrained weights.

All budget conditions use the original images without client-side resizing. Image-processing settings are fixed across budgets: Qwen retains the same per-image resolution and visual-token allocation, while GPT and Gemini use their respective high-detail settings.

We use `facebook/sam3`. Masks smaller than 64 pixels are rejected and the selected mask is eroded by 5 pixels before valid-depth filtering. Lifted points are voxelized at 0.02 m. Mask3D ScanNet200 proposals below confidence 0.2 are discarded and matched at 0.05 m voxel resolution. These settings are fixed across evidence interfaces.

Alternative-selector check. We rank the same stride-30 pool using `google/siglip2-so400m-patch14-384`, revision `e8e4872`, with FP32 inference, native image preprocessing, and cosine similarity. Original descriptions use the native 64-token limit; two are truncated. Top-6 views are presented chronologically, and all queries remain in the denominator. Relative to PE-Core, SigLIP2 yields 10 gains and 14 losses in final 3D correctness, and 14 gains and 16 losses in direct 2D correctness at IoU ≥ 0.50. GPT inference uses the same hosted API and decoding settings in both runs. The shared SAM3/Mask3D pipeline runs on an NVIDIA RTX PRO 6000 Blackwell Max-Q (96 GB) for PE-Core and an NVIDIA RTXA6000-24Q (24 GB vGPU) for SigLIP2, with unchanged downstream settings. This check covers one VLM and budget, not the full model–budget interaction.

Target coverage across budgets. Table 5 compares target-visible recall from the same stride-30 pool. SigLIP2 has lower recall at Top-3, higher recall at Top-6 and Top-12, and the same complete coverage at Top-24. Neither selector therefore dominates across all budgets. These are retrieval-only measurements; SigLIP2's downstream grounding is evaluated only at GPT Top-6 (Table 4). At that budget, its higher recall does not improve grounding accuracy, separating target coverage from the usefulness of the selected evidence for grounding.

Table 5: Target-visible recall (%) across selection budgets on all 250 queries. A query is covered when at least one selected view contains the target under the same visibility criterion used for PE-Core. These values measure target coverage, not grounding accuracy.

Selector	Top-3	Top-6	Top-12	Top-24
PE-Core	90.0	94.0	98.4	100.0
SigLIP2	86.4	95.6	99.6	100.0

C Grounding Prompt and Output Schema

Every request uses the system message `Return only the requested JSON object.` The user content contains, for each observation in presentation order, its metadata header followed by its image bytes, and then one complete grounding template below. In the templates, `<<QUERY>>`

is replaced by the ScanRefer expression and <<FRAME_LINES>> by the presented frame metadata. Evidence budgets change only these observation blocks and frame lines; the prompt wording is otherwise fixed.

GPT-5.6 native-pixel prompt. This is the complete user-text template supplied to GPT-5.6.

```
ScanNet color-frame grounding task.

Target description:
"<<QUERY>>"

You will receive sampled original color frames from the ScanNet
    scene. Each image is preceded by image_index, timestamp_sec,
    and ScanNet source_frame.

Please do all of the following carefully:
1. Identify the target object described in the sentence.
2. Infer the object label/category name for the selected target
    object, such as chair, towel, picture, table, or cabinet.
3. Select the single frame/time where the target object is most
    clearly visible and best matches the description.
4. Give a 2D point (x, y) near the visual center of the target
    object in pixel coordinates of the selected image.
5. Give a tight 2D bounding box for the visible extent of the
    target object in pixel coordinates: [x_min, y_min, x_max,
    y_max].
6. In the JSON, `selected_frame_index` must be the `image_index`
    value of the selected frame.

Important instructions:
- The object_label should be the category name of the object you
    selected.
- Return a concise, visually segmentable object category, such as
    chair, table, or cabinet. Do not include attributes, relations
    , or explanatory phrases. The object_label will be used as the
     text prompt for downstream SAM3 segmentation.
- If the label is directly stated in the target description, use
    that label and set label_source="from_description".
- If the label is not directly stated, infer the best label from
    the description and visual evidence and set label_source="
    inferred".
- If there are multiple similar objects, use the relational
    description carefully to choose the correct one.
- The point should be near the visual center of the object.
- The bbox should tightly cover the visible extent of the object.
- The JSON fields must describe the same selected frame, point,
    bbox, and label.

Available frames:
<<FRAME_LINES>>

Output rules:
- Return exactly one valid JSON object.
- Do not output chain-of-thought.
- Do not output Markdown.
- The first character of your response must be {.
- The last character of your response must be }.

JSON format:
```

```
{
  "object_label": "string",
  "label_source": "from_description",
  "selected_timestamp_sec": 0.0,
  "selected_frame_index": 0,
  "point_2d": [0, 0],
  "bbox_2d_xyxy": [0, 0, 0, 0],
  "image_size": [0, 0],
  "confidence": 0.0,
  "reason": "one short sentence"
}
```

Qwen3-VL normalized-coordinate prompt. The following is the complete post-adaptation prompt supplied to every Qwen3-VL checkpoint in the dense sweep.

```
ScanNet color-frame grounding task.

Target description:
"<<QUERY>>"

You will receive sampled original color frames from the ScanNet
    scene. Each image is preceded by image_index, timestamp_sec,
    and ScanNet source_frame.

Please do all of the following carefully:
1. Identify the target object described in the sentence.
2. Infer the object label/category name for the selected target
    object, such as chair, towel, picture, table, or cabinet.
3. Select the single frame/time where the target object is most
    clearly visible and best matches the description.
4. Give a 2D point (x, y) near the visual center of the target
    object on a fixed 0-1000 coordinate canvas: x=0 is the left
    edge, x=1000 is the right edge, y=0 is the top edge, and y
    =1000 is the bottom edge.
5. Give a tight 2D bounding box for the visible extent of the
    target object on the same fixed 0-1000 coordinate canvas: [
    x_min, y_min, x_max, y_max].
6. In the JSON, `selected_frame_index` must be the `image_index`
    value of the selected frame.

Important instructions:
- The object_label should be the category name of the object you
    selected.
- Return a concise, visually segmentable object category, such as
    chair, table, or cabinet. Do not include attributes, relations
    , or explanatory phrases. The object_label will be used as the
     text prompt for downstream SAM3 segmentation.
- If the label is directly stated in the target description, use
    that label and set label_source="from_description".
- If the label is not directly stated, infer the best label from
    the description and visual evidence and set label_source="
    inferred".
- If there are multiple similar objects, use the relational
    description carefully to choose the correct one.
- The point should be near the visual center of the object.
- The bbox should tightly cover the visible extent of the object.
- Treat x and y independently when mapping the selected image onto
     the fixed 0-1000 canvas.
- Every point and bbox coordinate must be between 0 and 1000.
```

```
- Set image_size=[1000, 1000] to describe the coordinate canvas;
    do not guess the JPEG pixel dimensions.
- The JSON fields must describe the same selected frame, point,
    bbox, and label.

Available frames:
<<FRAME_LINES>>

Output rules:
- Return exactly one valid JSON object.
- Do not output chain-of-thought.
- Do not output Markdown.
- The first character of your response must be {.
- The last character of your response must be }.

JSON format:
{
  "object_label": "string",
  "label_source": "from_description",
  "selected_timestamp_sec": 0.0,
  "selected_frame_index": 0,
  "point_2d": [0, 0],
  "bbox_2d_xyxy": [0, 0, 0, 0],
  "image_size": [1000, 1000],
  "confidence": 0.0,
  "reason": "one short sentence"
}
```

Gemini normalized-coordinate prompt. This is the complete post-adaptation user-text template supplied to Gemini. Relative to the GPT template, only the coordinate instructions and corresponding JSON fields change. The adapter reorders the returned y, x coordinates and converts them to native pixels before the shared downstream pathway.

```
ScanNet color-frame grounding task.

Target description:
"<<QUERY>>"

You will receive sampled original color frames from the ScanNet
    scene. Each image is preceded by image_index, timestamp_sec,
    and ScanNet source_frame.

Please do all of the following carefully:
1. Identify the target object described in the sentence.
2. Infer the object label/category name for the selected target
    object, such as chair, towel, picture, table, or cabinet.
3. Select the single frame/time where the target object is most
    clearly visible and best matches the description.
4. Give a 2D point near the visual center of the target object as
    [y, x], with both coordinates normalized to integers from 0 to
     1000.
5. Give a tight 2D bounding box for the visible extent of the
    target object as [y_min, x_min, y_max, x_max], with all
    coordinates normalized to integers from 0 to 1000.
6. In the JSON, `selected_frame_index` must be the `image_index`
    value of the selected frame.

Important instructions:
```

```
- The object_label should be the category name of the object you
    selected.
- Return a concise, visually segmentable object category, such as
    chair, table, or cabinet. Do not include attributes, relations
    , or explanatory phrases. The object_label will be used as the
     text prompt for downstream SAM3 segmentation.
- If the label is directly stated in the target description, use
    that label and set label_source="from_description".
- If the label is not directly stated, infer the best label from
    the description and visual evidence and set label_source="
    inferred".
- If there are multiple similar objects, use the relational
    description carefully to choose the correct one.
- The point should be near the visual center of the object.
- The bbox should tightly cover the visible extent of the object.
- The JSON fields must describe the same selected frame, point,
    bbox, and label.

Available frames:
<<FRAME_LINES>>

Output rules:
- Return exactly one valid JSON object.
- Do not output chain-of-thought.
- Do not output Markdown.
- The first character of your response must be {.
- The last character of your response must be }.

JSON format:
{
  "object_label": "string",
  "label_source": "from_description",
  "selected_timestamp_sec": 0.0,
  "selected_frame_index": 0,
  "point_2d_yx": [0, 0],
  "bbox_2d_yxyx": [0, 0, 0, 0],
  "image_size": [1000, 1000],
  "confidence": 0.0,
  "reason": "one short sentence"
}
```

D Detailed Evidence-Budget Results

Table 6: GPT-5.6 final 3D Acc@0.50 and target-visible recall on ScanRefer-250. Images is the mean input count; other values are percentages. Top-24 uses 64.6% fewer images than Full while retaining a target-visible view for every query.

Input	Images	Recall	Overall	Unique	Multiple
Top-3	3.0	90.0	57.2	81.8	48.4
Top-6	6.0	94.0	60.4	81.8	52.7
Top-12	12.0	98.4	60.8	81.8	53.3
Top-24	23.8	100.0	62.0	**83.3**	54.3
Full	67.2	100.0	**62.4**	**83.3**	**54.9**

Table 7: Effect of relaxing Top-3 to Top-6 by scene-level class multiplicity. Values are final 3D Acc@0.50.

	Unique		Multiple	
Model	Top-3	Top-6	Top-3	Top-6
2B	69.7	60.6	39.1	32.6
4B	74.2	68.2	39.1	40.8
8B	75.8	75.8	39.7	42.9
32B	81.8	84.8	45.1	46.2

Table 8: Detailed Qwen3-VL metrics across all four restricted budgets, with the 2B Full condition included for reference. Latency is mean VLM request time in seconds.

Model	Budget	Recall	2D@.25	2D@.50	Point	3D@.25	3D@.50	Latency
2B	3	90.0	38.0	13.2	32.4	51.6	47.2	13.2
2B	6	94.0	36.4	14.8	28.4	44.8	40.0	13.9
2B	12	98.4	29.2	8.0	23.2	40.8	36.4	14.5
2B	24	100.0	22.4	8.4	20.4	36.0	32.0	14.5
2B	Full	100.0	22.4	6.8	17.2	32.0	29.6	23.5
4B	3	90.0	53.2	31.2	40.8	54.8	48.4	23.6
4B	6	94.0	48.4	23.2	40.4	53.6	48.0	27.9
4B	12	98.4	34.4	14.8	31.2	46.4	41.6	29.2
4B	24	100.0	30.4	8.4	24.0	43.6	38.8	34.6
8B	3	90.0	50.8	28.8	42.8	55.2	49.2	37.3
8B	6	94.0	50.4	24.0	47.2	56.8	51.6	44.1
8B	12	98.4	44.8	16.4	38.0	52.4	46.8	43.8
8B	24	100.0	35.2	13.2	32.8	50.4	44.0	48.7
32B	3	90.0	67.6	52.8	55.6	60.8	54.8	50.4
32B	6	94.0	64.4	50.0	54.0	62.0	56.4	65.9
32B	12	98.4	63.6	42.4	51.2	58.0	52.8	65.9
32B	24	100.0	54.0	29.2	42.4	55.2	49.6	78.0

Table 9: Restricted-budget results by scene-level class multiplicity. Values are final 3D Acc@0.50; each sequence follows Top-3/6/12/24. Full-observation results are listed separately below.

Model	Budget	Unique	Multiple
2B	3/6/12/24	69.7/60.6/56.1/48.5	39.1/32.6/29.3/26.1
4B	3/6/12/24	74.2/68.2/66.7/62.1	39.1/40.8/32.6/30.4
8B	3/6/12/24	75.8/75.8/65.2/62.1	39.7/42.9/40.2/37.5
32B	3/6/12/24	81.8/84.8/71.2/69.7	45.1/46.2/46.2/42.4
GPT-5.6	3/6/12/24	81.8/81.8/81.8/83.3	48.4/52.7/53.3/54.3
Gemini	3/6/12/24	80.3/80.3/78.8/78.8	47.8/52.2/54.3/56.5
2B	Full	45.5	23.9
GPT-5.6	Full	83.3	54.9
Gemini	Full	78.8	53.3

Table 10: GPT-5.6 evidence-budget metrics.

Condition	Images	Recall	2D@.50	Point	3D@.25	3D@.50
Top-3	3.0	90.0	62.0	60.4	63.2	57.2
Top-6	6.0	94.0	66.0	64.0	67.2	60.4
Top-12	12.0	98.4	66.0	66.8	67.6	60.8
Top-24	23.8	100.0	66.0	68.0	68.4	62.0
Full	67.2	100.0	65.2	65.6	68.0	62.4

Table 11: Gemini 3.8 Flash metrics on all 250 queries. Frame accuracy is the fraction of VLM-selected images containing a visible target pixel, not input target-visible recall. All values are percentages.

	Final 3D			Direct 2D box			Frame
Input	mIoU	@.25	@.50	mIoU	@.25	@.50	Acc
Top-3	54.21	62.8	56.4	56.71	70.4	64.4	81.6
Top-6	57.01	66.4	59.6	59.50	72.4	66.8	81.6
Top-12	58.15	67.2	60.8	61.43	72.4	69.2	86.0
Top-24	59.78	68.8	62.4	62.56	75.2	72.0	86.4
Full	57.40	66.4	60.0	61.54	74.4	70.4	84.4

Table 12: GPT-5.6 diagnostic metrics for Full/Top-6, field-of-view, and order controls. Image count, estimated image tokens, and VLM request time are averaged over queries. Token estimates follow Appendix A.

Condition	Images	Est. image tokens	VLM sec.	Recall
Full, ordered	67.2	101.4k	14.9	100.0
Top-6, ordered	6.0	9.0k	9.1	94.0
Full, shuffled, timestamps	67.2	101.4k	14.5	100.0
Full, shuffled, no timestamps	67.2	101.4k	15.3	100.0
Full, center-50	67.2	50.9k	31.2	99.6
Top-6, center-50	6.0	4.5k	11.2	92.4

Condition	2D@.25	2D@.50	Point	3D@.25	3D@.50
Full, ordered	74.0	65.2	65.6	68.0	62.4
Top-6, ordered	72.8	66.0	64.0	67.2	60.4
Full, shuffled, timestamps	74.0	66.8	66.8	66.4	59.6
Full, shuffled, no timestamps	72.4	64.0	64.0	65.2	58.4
Full, center-50	71.2	64.4	64.0	64.0	58.0
Top-6, center-50	69.6	64.0	63.2	62.4	56.4

D.1 Qualitative 2D–3D disagreement

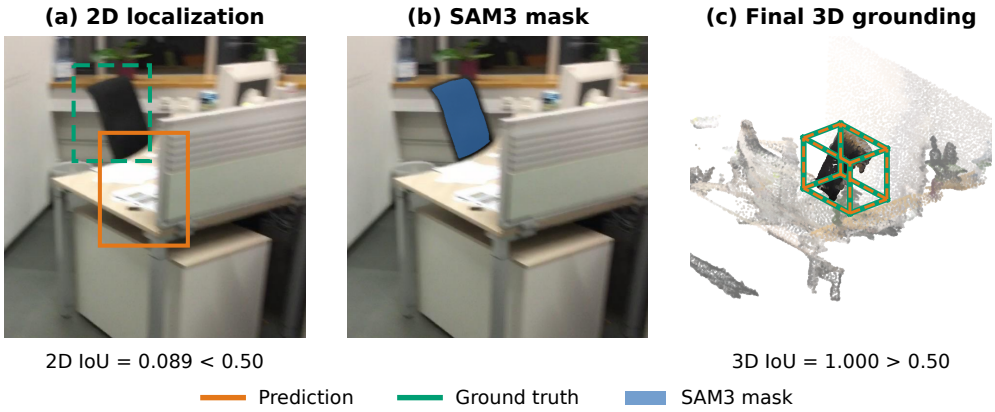

Figure 4: A below-threshold 2D box yielding correct final 3D grounding with Qwen3-VL-32B (FP8) and PE-Core Top-6. The query is: *"the chair is pulled into the desk. the chair has a wall behind it. the couch is in the background past it."* (a) The predicted box overlaps only a small part of the target's visible GT box (2D IoU 0.089). (b) The filtered SAM3 mask used for lifting covers the visible chair back. (c) The matched proposal's official aligned 3D box coincides with the GT box (3D IoU ≈ 1.000). Panels (a, b) use the same display-only image crop; panel (c) shows the local scene point cloud.

D.2 Qualitative evidence-budget comparisons

Figures 5 and 6 show the complete Top-6 and Top-24 inputs for two queries, followed by the selected-frame predictions of Qwen3-VL-32B (FP8) and GPT-5.6. The models receive identical inputs at each budget, and Top-6 is a subset of Top-24. Thumbnails follow the actual presentation order, left to right and then top to bottom; labels `fN` identify ScanNet source frames. Q and G mark the frames selected by Qwen and GPT, respectively. These markers and the prediction/GT overlays are for visualization only. Correctness uses IoU ≥ 0.50.



Figure 5: Opposite evidence-budget changes for the query *"this is a rectangular flat panel monitor. it is the last monitor on the right in the front row of monitors."* Qwen changes from correct to incorrect final 3D grounding, while GPT changes from incorrect to correct. The target is already clearly visible in Top-6: frame 1980, selected by Qwen at Top-6 and GPT at Top-24, belongs to both inputs. The target is absent from Qwen's selected Top-24 frame; only 15 target pixels occur at the far right edge of GPT's selected Top-6 frame.

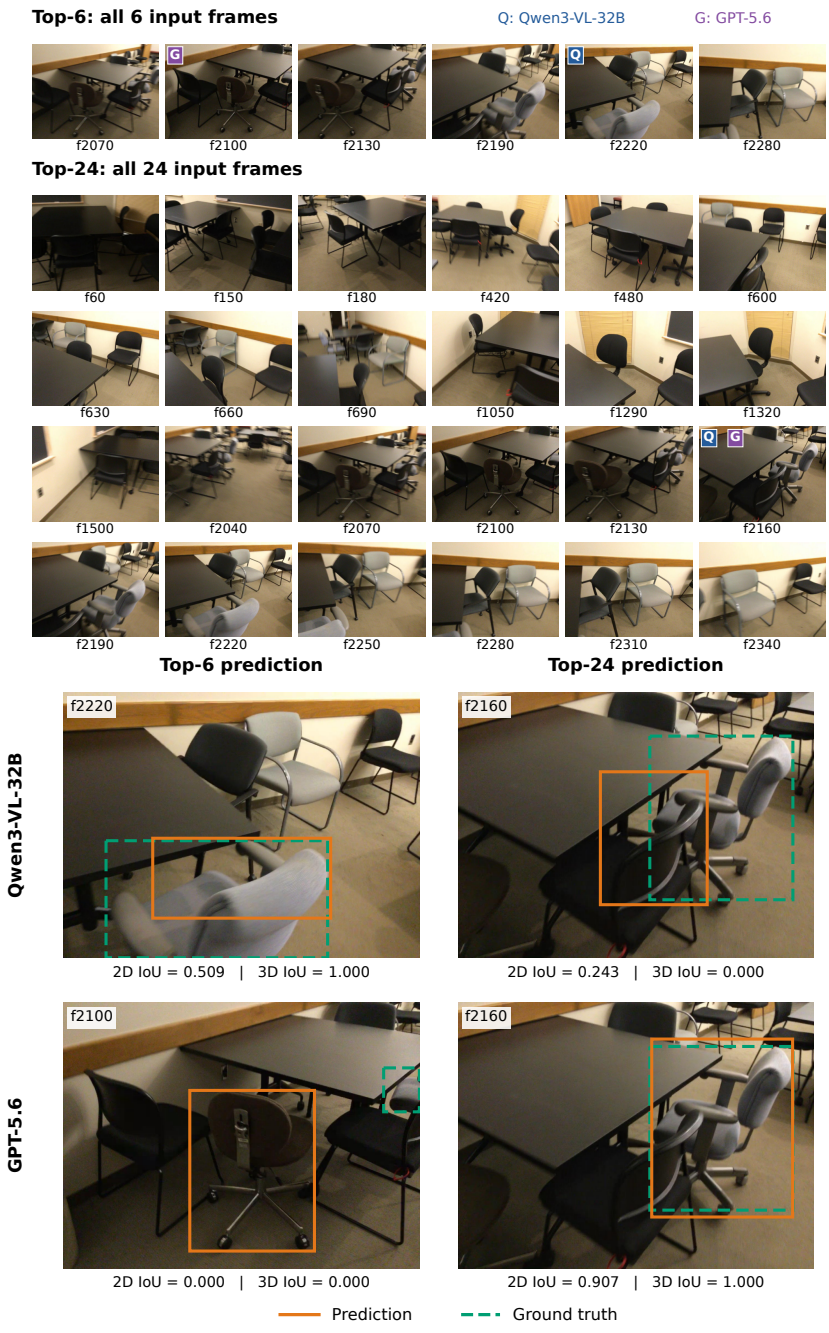

Figure 6: Opposite evidence-budget changes for the query *"the gray padded chair is on rollers. it is the only rolling chair at the table."* The target is visible in every Top-6 input frame. Qwen changes from correct to incorrect final 3D grounding, while GPT changes from incorrect to correct. At Top-24, both models select the same frame (2160), but their predicted boxes locate different regions. The complete inputs and the selected-frame predictions separate the change in available observations from the change in localization.

E FIELD-OF-VIEW AND ORDER STRESS TESTS

Center-crop stress test. We center-crop every image in the stride-30 pool to 50% of its original area. For Top-6, PE-Core ranking is recomputed on the cropped images, so the selected frames can differ from the uncropped Top-6 set. After VLM inference, predicted coordinates are mapped back to the original image, and SAM3 operates on that original image before the shared RGB-D lifting and Mask3D matching stages. This gives GPT-5.6 target-visible recall of 99.6% for full observation and 92.4% for Top-6. Final 3D Acc@0.25/0.50 is 64.0/58.0 for full observation and 62.4/56.4 for Top-6. Center cropping lowers both conditions similarly at 3D Acc@0.50: -4.4 points for full observation and -4.0 points for Top-6. Full retains higher target-visible recall and final accuracy under the narrower per-view field of view.

Observation order. Using the same full image set but one shuffled ordering changes GPT-5.6 final 3D Acc@0.50 from 62.4% to 59.6% when timestamps are retained and to 58.4% when timestamps are hidden. Because each shuffled ordering requires a separate GPT-5.6 API evaluation over all 250 queries, API cost limited this control to one ordering.

F PAIRED QUERY ANALYSIS

These tables provide the detailed paired-query counts underlying Section 4.3. Tables 13 and 14 compare GPT-5.6 and Gemini over five budget expansions. Table 15 compares all six VLMs on Top-3/6 → Top-24, separating initially present and absent targets. Table 16 gives the GPT/Gemini Unique/Multiple breakdown; Table 17 supplements main-text Table 3 with frame and point counts. We pair predictions by scene, target object, and annotation ID. Present/absent denotes target visibility in the smaller input; each starting budget defines its own group. A gain is an incorrect final 3D prediction becoming correct at IoU ≥ 0.50; a loss is the reverse, and Net is gains minus losses. All entries are query counts, not percentages. Each model–budget condition uses one saved run, so paired changes can include run-to-run variation.

Table 13: GPT-5.6 and Gemini: final 3D gains and losses underlying Section 4.3. Unlike the before/after totals in main-text Table 3, this table records changes in both directions and includes initially omitted targets.

			GPT-5.6			Gemini		
Expansion	Initial target	Queries	Gains	Losses	Net	Gains	Losses	Net
Top-3 → Top-24	Present	225	14	7	+7	20	12	+8
	Absent	25	5	0	+5	7	0	+7
Top-6 → Top-24	Present	235	12	11	+1	11	10	+1
	Absent	15	3	0	+3	6	0	+6
Top-3 → Full	Present	225	15	9	+6	17	16	+1
	Absent	25	7	0	+7	8	0	+8
Top-6 → Full	Present	235	15	15	0	11	15	−4
	Absent	15	5	0	+5	5	0	+5
Top-24 → Full	Present	250	11	10	+1	8	14	−6

Table 14: Stage-by-stage counts for the same groups in Table 13. Arrows show smaller → larger input; Frame counts target-visible selections. Box and final 3D correctness use IoU ≥ 0.50. The Top-6 → Top-24 Present rows include the totals in main-text Table 3 for completeness.

Expansion	Initial target	Queries	Frame	2D box	Final 3D
GPT-5.6					
Top-3 → Top-24	Present	225	199 → 201	155 → 161	143 → 150
	Absent	25	0 → 9	0 → 4	0 → 5
Top-6 → Top-24	Present	235	206 → 203	165 → 163	151 → 152
	Absent	15	0 → 7	0 → 2	0 → 3
Top-3 → Full	Present	225	199 → 197	155 → 157	143 → 149
	Absent	25	0 → 11	0 → 6	0 → 7
Top-6 → Full	Present	235	206 → 199	165 → 159	151 → 151
	Absent	15	0 → 9	0 → 4	0 → 5
Top-24 → Full	Present	250	210 → 208	165 → 163	155 → 156
Gemini					
Top-3 → Top-24	Present	225	204 → 205	161 → 172	141 → 149
	Absent	25	0 → 11	0 → 8	0 → 7
Top-6 → Top-24	Present	235	204 → 208	167 → 174	149 → 150
	Absent	15	0 → 8	0 → 6	0 → 6
Top-3 → Full	Present	225	204 → 198	161 → 168	141 → 142
	Absent	25	0 → 13	0 → 8	0 → 8
Top-6 → Full	Present	235	204 → 204	167 → 172	149 → 145
	Absent	15	0 → 7	0 → 4	0 → 5
Top-24 → Full	Present	250	216 → 211	180 → 176	156 → 150

Table 15: Final 3D gains and losses across all six VLMs for **(a)** Top-3 → Top-24 and **(b)** Top-6 → Top-24. Present/absent refers to target visibility in the starting input: 225/25 queries in (a), and 235/15 in (b). Each part uses identical groups across models, and every target is visible in Top-24. Present net changes in (b) match the final 3D differences in main-text Table 3.

Backbone	Initial target	Queries	Gains	Losses	Net
(a) Top-3 → Top-24					
Qwen3-VL-2B	Present	225	15	54	−39
	Absent	25	1	0	+1
Qwen3-VL-4B	Present	225	13	41	−28
	Absent	25	4	0	+4
Qwen3-VL-8B	Present	225	21	37	−16
	Absent	25	3	0	+3
Qwen3-VL-32B	Present	225	11	27	−16
	Absent	25	3	0	+3
GPT-5.6	Present	225	14	7	+7
	Absent	25	5	0	+5
Gemini	Present	225	20	12	+8
	Absent	25	7	0	+7
(b) Top-6 → Top-24					
Qwen3-VL-2B	Present	235	22	42	−20
	Absent	15	0	0	0
Qwen3-VL-4B	Present	235	15	39	−24
	Absent	15	1	0	+1
Qwen3-VL-8B	Present	235	18	39	−21
	Absent	15	2	0	+2
Qwen3-VL-32B	Present	235	16	34	−18
	Absent	15	1	0	+1
GPT-5.6	Present	235	12	11	+1
	Absent	15	3	0	+3
Gemini	Present	235	11	10	+1
	Absent	15	6	0	+6

Table 16: Unique/Multiple breakdown of the already-covered queries in Table 13. Model columns report final 3D gains/losses (wrong-to-correct / correct-to-wrong). Both models use the same query groups.

	Unique			Multiple		
Expansion	Queries	GPT-5.6	Gemini	Queries	GPT-5.6	Gemini
Top-3 → Top-24	65	1/0	1/2	160	13/7	19/10
Top-6 → Top-24	66	1/0	0/1	169	11/11	11/9
Top-3 → Full	65	1/0	1/2	160	14/9	16/14
Top-6 → Full	66	1/0	0/1	169	14/15	11/14
Top-24 → Full	66	0/0	0/0	184	11/10	8/14

Table 17: Companion to main-text Table 3: frame and point counts on the same 235 already-covered queries. Arrows show Top-6 → Top-24. A target-visible selection contains a GT target pixel; a correct point falls inside the GT target mask.

Backbone	Selected-visible: $6 \rightarrow 24$	Correct point: $6 \rightarrow 24$
Qwen3-VL-2B	$184 \rightarrow 157$	$71 \rightarrow 50$
Qwen3-VL-4B	$196 \rightarrow 183$	$101 \rightarrow 60$
Qwen3-VL-8B	$199 \rightarrow 191$	$118 \rightarrow 82$
Qwen3-VL-32B	$197 \rightarrow 201$	$135 \rightarrow 105$
GPT-5.6	$206 \rightarrow 203$	$160 \rightarrow 167$
Gemini	$204 \rightarrow 208$	$167 \rightarrow 164$

Which selected frames change? Among GPT's 12 Top-6-to-Top-24 gains with targets already present, seven select a newly available frame, four select a different frame already in Top-6, and one retains the same frame. The 11 losses comprise five selections of new frames and six switches to other old frames. Ten gains also change direct 2D box correctness from wrong to correct; nine losses change it from correct to wrong. These 2D counts refer only to queries whose final 3D correctness flips, not to all 235 queries.

G SCENE-CLUSTERED PAIRED UNCERTAINTY

G.1 FULL-PANEL RESPONSE: TOP-3 TO TOP-24

The 250 queries come from 115 ScanNet scenes, with one to eight expressions per scene. We estimate uncertainty for the change in final 3D Acc@0.50 from Top-3 to Top-24 using 10,000 paired scene-cluster bootstrap resamples (seed 20260926). Each resample draws 115 scenes uniformly with replacement and includes every query from each drawn scene, including repeated copies when a scene is drawn more than once. The same scene draws are used for all models and both budgets, preserving their query-level correspondence.

For each resample, we compute query-weighted accuracy at both budgets and their difference, using the actual number of resampled queries as the denominator. This retains the reported query-level metric rather than averaging scene accuracies. We also subtract Qwen3-VL-2B's budget effect from GPT's and Gemini's within each resample. The 2.5th and 97.5th percentiles of each resampled statistic form its 95% interval (Table 18).

Table 18: Scene-clustered paired uncertainty for Top-3 $\rightarrow$ Top-24. Changes and intervals are in percentage points of final 3D Acc@0.50. The bottom rows compare budget effects, not model accuracies at one budget.

Model or difference in effects	Change	95% interval
Qwen3-VL-2B	-15.2	$[-22.0, -8.6]$
Qwen3-VL-4B	-9.6	$[-15.5, -3.5]$
Qwen3-VL-8B	-5.2	$[-10.9, +0.8]$
Qwen3-VL-32B	-5.2	$[-10.4, -0.4]$
GPT-5.6	$+4.8$	$[+1.2, +8.7]$
Gemini	$+6.0$	$[+1.2, +10.9]$
GPT-5.6 minus Qwen3-VL-2B	$+20.0$	$[+12.4, +27.6]$
Gemini minus Qwen3-VL-2B	$+21.2$	$[+13.6, +28.7]$

Both cross-model intervals exclude zero, indicating different Top-3-to-Top-24 responses from Qwen3-VL-2B for GPT-5.6 and Gemini under scene-level resampling. Among the individual model effects, Qwen3-VL-8B is the only 95% interval that includes zero. These intervals quantify scene-resampling uncertainty for the 250-query panel.

G.2 ALREADY-COVERED QUERIES: TOP-6 TO TOP-24

We apply the same procedure to the fixed 235 queries whose targets are already visible in Top-6 (Table 3). These queries span 111 scenes. Each of 10,000 resamples (seed 20260926) draws 111 scenes with replacement and includes all retained queries from each drawn scene, repeated with that scene's multiplicity. The same draws are shared across all six VLMs, both budgets, and both metrics. Accuracy remains query-weighted, with the actual number of sampled queries as the denominator. Point estimates here use 235 queries, not the full 250-query panel.

Table 19: Top-6 → Top-24 on the same 235 already-covered queries. Changes and 95% scene-clustered paired intervals are in percentage points of Acc@0.50.

	Final 3D		Direct 2D	
Model	Change	95% interval	Change	95% interval
Qwen3-VL-2B	-8.51	$[-15.07, -2.02]$	-7.23	$[-12.40, -2.16]$
Qwen3-VL-4B	-10.21	$[-16.94, -3.73]$	-15.74	$[-22.03, -9.77]$
Qwen3-VL-8B	-8.94	$[-15.13, -2.31]$	-11.49	$[-17.89, -5.29]$
Qwen3-VL-32B	-7.66	$[-13.02, -2.18]$	-22.13	$[-29.33, -15.04]$
GPT-5.6	$+0.43$	$[-3.98, +4.70]$	-0.85	$[-5.08, +3.49]$
Gemini	$+0.43$	$[-3.64, +4.78]$	$+2.98$	$[-2.21, +8.30]$

All four Qwen declines have intervals below zero in both metrics. GPT's and Gemini's intervals include zero; their small observed net changes do not establish equivalence between budgets. To directly compare model responses, we subtract each Qwen model's final 3D change from GPT's or Gemini's within each shared resample (Table 20). All eight pointwise intervals are above zero, supporting different budget responses on the same already-covered queries.

Table 20: Cross-model differences in the Top-6 → Top-24 effect on final 3D Acc@0.50, using the same 235 already-covered queries. Values are percentage points; these compare accuracy changes, not accuracies at a single budget.

Difference in budget effects	Change	95% interval
GPT-5.6 minus Qwen3-VL-2B	$+8.94$	$[+0.85, +17.02]$
GPT-5.6 minus Qwen3-VL-4B	$+10.64$	$[+3.50, +17.81]$
GPT-5.6 minus Qwen3-VL-8B	$+9.36$	$[+1.84, +16.43]$
GPT-5.6 minus Qwen3-VL-32B	$+8.09$	$[+1.62, +14.72]$
Gemini minus Qwen3-VL-2B	$+8.94$	$[+1.30, +16.67]$
Gemini minus Qwen3-VL-4B	$+10.64$	$[+3.24, +18.10]$
Gemini minus Qwen3-VL-8B	$+9.36$	$[+2.10, +16.45]$
Gemini minus Qwen3-VL-32B	$+8.09$	$[+1.34, +14.86]$

Interpretation of the intervals. Each model–budget condition in these analyses uses one saved generation per query. The intervals quantify scene-resampling uncertainty in those predictions, not variation across repeated model generations. They are pointwise percentile intervals without multiple-comparison adjustment; the cross-model contrasts are descriptive rather than preregistered.