SELECT, COMPETE, REFINE: TRAINING-FREE 3D VISUAL GROUNDING FROM RGB-D VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free 3D visual grounding from calibrated RGB-D views enables language-conditioned object localization without task-specific model training or preconstructed scene point clouds. However, selecting the object that satisfies a full referring expression remains ambiguous, and geometry reconstructed from partial masks and depth can yield incomplete object boxes. We propose Select, Compete, Refine (SCR), a framework that combines clause-scoped object selection and verification with multi-view boundary recovery using frozen pretrained models. Select organizes entity-level options and clause-bound references. Compete separates challenger discovery under withheld candidate evidence from replacement verification after evidence restoration. Refine recovers observed but underrepresented extent through adaptive per-view bounds and face-wise multi-view support, without additional segmentation or vision-language model calls. We evaluate SCR on SeqAlign3DVG, EmbodiedScan and ScanRefer. On SeqAlign3DVG, SCR achieves 62.96% Acc@0.25 and 38.07% Acc@0.50, improving over our VLM-Grounder reproduction using the same Qwen model by 11.38 and 12.45 percentage points, respectively. Ablation studies further support the benefits of object selection, verification, and boundary recovery.

1 INTRODUCTION

3D visual grounding aims to localize an object in a 3D scene from a natural-language expression (Wang et al., 2024). By connecting linguistic intent to physical objects and their locations, it provides a perceptual basis for language-guided interaction with an environment (Xu et al., 2025). In this work, we study *training-free, image-based 3D visual grounding*: given a referring expression and a fixed set of calibrated RGB-D views, the system predicts a 3D bounding box while keeping all pretrained model weights frozen.

This setting combines complementary advantages. RGB images preserve appearance and relational context for pretrained vision-language models (VLMs), while depth and camera calibration connect those observations to metric 3D geometry. Unlike supervised image-centric approaches that learn task-specific alignment (Lin et al., 2025b; Zheng et al., 2025), frozen inference avoids training a new grounding model. Constructing candidates directly from the available views also removes the input requirement for a preconstructed scene point cloud or external 3D proposals. The central challenge is to use frozen models and the available observations to select the correct object and estimate its 3D extent.

Training-free approaches have explored reasoning over 3D scene representations or object proposals (Yuan et al., 2024; Li et al., 2025b), as well as grounding directly from RGB-D observations. Among the latter, VLM-Grounder (Xu et al., 2025) combines grounding feedback with multi-view ensemble projection. Think, Act, Build (Wang et al., 2026) couples semantic localization with geometric expansion across views, while SceneGraphGrounder (Sun et al., 2026) resolves queries through structured scene-graph matching. These methods establish the value of combining visual reasoning, relational context, and geometry. Building on these advances, we focus on two challenges that can still limit localization: ambiguity in matching the full expression to an object, and incomplete geometry recovered from multi-view observations.

Object-selection ambiguity. Even when relevant reference objects are available, their presence alone does not clarify which entities each clause describes. In an unstructured collection of targets and

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.

references, a VLM must infer these associations while selecting the target. For example, in "the chair beside the second table," "second" orders tables, not chairs. Even with these roles clarified, visually similar candidates can remain ambiguous: a salient appearance match may be selected without satisfying the relation that distinguishes the intended object. Accurate selection therefore requires both explicit language–object associations and a check that an ambiguous initial choice is supported by the full expression.

Incomplete object geometry. Selecting the correct object does not guarantee an accurate 3D box. Back-projecting 2D masks with depth and aggregating observations across views produces geometry whose coverage depends on visible object regions and mask completeness. Partial observations and robust point filtering can omit genuine extremities, leaving the resulting box smaller than the object's supported extent. This geometric error persists even when the semantic choice is correct. Recovering boundaries supported by the available observations thus requires explicit geometric treatment rather than treating the selected candidate's initial box as the final localization.

We introduce *Select, Compete, Refine (SCR)*, a training-free framework that combines *clause-scoped object selection and verification* with *multi-view boundary recovery*. Select organizes duplicate candidate hypotheses into distinct object options and binds reference objects to the clauses they help interpret. Compete selectively tests ambiguous initial choices through *withhold–restore verification*. It temporarily withholds the initial choice's dedicated crop and annotations to elicit an alternative, while leaving raw context RGB unchanged. Both candidates' evidence is then restored for anonymous comparison and a replacement is accepted only when this comparison supports the alternative and provides query-specific counter-evidence against the initial choice.

Refine improves the selected object's box without changing the object choice. It revisits the existing source masks and depth, estimates point-count-adaptive bounds within each view, and updates box faces independently. Each outward face displacement requires support from at least two retained views, and geometric checks limit excessive expansion. This recovers observed but underrepresented extent without additional segmentation or VLM calls.

We evaluate SCR on SeqAlign3DVG (Zhang et al., 2026c), the EmbodiedScan Mini validation set (Wang et al., 2024), and ScanRefer Chen et al. (2020) under their specified input and evaluation protocols. On SeqAlign3DVG, SCR achieves 62.96% Acc@0.25 and 38.07% Acc@0.50. Compared with our VLM-Grounder reproduction using the same Qwen model, these results improve accuracy by 11.38 and 12.45 percentage points, respectively. It also achieves the best performance among the training-free RGB-D methods on EmbodiedScan and ScanRefer. Ablation studies further support the effectiveness of the selection, verification, and refinement components. Our contributions are summarized as follows:

- **Clause-scoped selection and withhold–restore verification.** We make target and reference roles explicit during object selection, and separate proposing an alternative from accepting a replacement through evidence restoration and discriminative comparison.
- **Multi-view boundary recovery.** We introduce point-count-adaptive per-view bounds and independently supported box-face updates to recover observed but underrepresented extent without additional segmentation or VLM calls.

2 Related Work

Learning-based 3D visual grounding. ScanRefer (Chen et al., 2020) established an early learning-based paradigm for grounding free-form referring expressions in reconstructed 3D scenes. Subsequent methods improve cross-modal alignment and contextual reasoning: BUTD-DETR (Jain et al., 2022) combines language and objectness guidance, EDA (Wu et al., 2023) densely aligns textual components with 3D objects, and MiKASA (Chang et al., 2024) models scene context and multiple relational anchors. Learning-based grounding also extends to RGB and RGB-D inputs. Refer-it-in-RGBD (Liu et al., 2021) and Mono3DVG (Zhan et al., 2024) study single-view grounding from RGB-D and monocular images, respectively, while EmbodiedScan (Wang et al., 2024) provides a benchmark for grounding from posed multi-view RGB-D observations. BIP3D (Lin et al., 2025b) and DenseGrounding (Zheng et al., 2025) explore image-centric perception and dense language–vision supervision, respectively. TriCLIP-3D (Li et al., 2025a) introduces parameter-efficient tri-modal adaptation, while DEGround (Zhang et al., 2026b) unifies detection and grounding within a

homogeneous architecture. SeqAlign3DVG (Zhang et al., 2026c) further couples sequence-aligned evaluation with learned voxel reasoning from RGB observations. To reduce reliance on 3D supervision, Scene-R1 (Yuan et al., 2025) trains a video-grounded model through reinforcement learning without 3D grounding annotations. While differing in input modality and supervision, these learning-based approaches require task-specific training, whereas SCR keeps all pretrained model weights frozen.

Training-free 3D grounding from reconstructed scenes. Without task-specific training, LLM-Grounder (Yang et al., 2024) combines language-agent reasoning with open-vocabulary 3D grounding tools, while ZSVG3D (Yuan et al., 2024) and Language-to-Space Programming (Mi et al., 2025) express spatial reasoning as executable programs or constraints over 3D objects. SeeGround (Li et al., 2025b) selects query-relevant rendering viewpoints and combines the resulting images with 3D spatial descriptions, while SeqVLM (Lin et al., 2025a) projects candidate proposals into multi-view image sequences for VLM reasoning. SPAZER (Jin et al., 2025) uses anchor-guided candidate screening followed by joint reasoning over rendered views and retrieved camera images. Agent-Grounder (Huynh et al., 2026) uses adaptive visual tool calls over a point-cloud scene. Although visual observations enrich their reasoning, these methods still depend on preconstructed scene geometry or object proposals. Recent approaches relax the proposal requirement: UniGround (Zhang et al., 2026a) constructs proposals through training-free scene parsing, and OpenGround (Huang et al., 2026) expands object candidates during plan execution rather than relying exclusively on a fixed initial list. Both nevertheless take a scene point cloud as input. SCR instead constructs source-linked candidates from calibrated RGB-D observations, without requiring a preconstructed scene cloud or external 3D proposals.

Training-free 3D grounding from RGB-D views. OVIR-3D (Lu et al., 2023) fuses image-derived object information for open-vocabulary 3D instance retrieval, while ConceptGraphs (Gu et al., 2024) builds object-centric scene graphs from foundation-model predictions. These works provide foundations for multi-view object association and language-accessible scene representations. Among grounding pipelines, VLM-Grounder (Xu et al., 2025) combines image stitching, iterative grounding feedback, and multi-view ensemble projection to estimate 3D boxes without preconstructed scene point clouds. Think, Act, Build (TAB) (Wang et al., 2026) couples semantic reasoning and geometric reconstruction through an agentic tool loop, using Semantic-Anchored Geometric Expansion to acquire additional target masks across views before point-cloud filtering and box estimation. SceneGraphGrounder (Sun et al., 2026) reconstructs a persistent scene graph from RGB-D observations and aligns it with a query graph, using perspective-view VLM reasoning to resolve ambiguity. Mem-ChangingGrounder (Hu et al., 2025) extends grounding to changing scenes through memory retrieval and active exploration, whereas SCR operates on a fixed set of observations. Building on these approaches, SCR focuses on clause-scoped object selection and verification together with multi-view boundary recovery. For object selection, SCR combines entity-level candidates and clause-bound references with a verification procedure that elicits a challenger under withheld evidence and evaluates replacement after restoring that evidence. For boundary recovery, it revisits existing source masks and depth without additional segmentation or VLM calls, using adaptive per-view bounds and independently supported face updates to recover observed but under-represented object extent.

3 METHODOLOGY

Problem formulation. Given a referring expression q and calibrated RGB-D observations $\mathcal{O} = \{(I_v, D_v, K_v, T_v)\}_{v=1}^{V}$, we localize the described object in 3D. Here K_v is the camera intrinsic matrix and T_v is the camera-to-world pose. We predict a 3D bounding box $B = [\mathbf{L}, \mathbf{U}]$ in the shared world frame, where $\mathbf{L} = (x_{\min}, y_{\min}, z_{\min})$ and $\mathbf{U} = (x_{\max}, y_{\max}, z_{\max})$ are its two opposite corners. No pretrained weights are updated, and no preconstructed scene point cloud or external 3D proposals are required.

Overview. Figure 1 presents *Select, Compete, Refine* (SCR). Starting from mask-derived 3D candidates, SCR addresses two high-level challenges through three stages. Select and Compete address object-selection ambiguity: Select organizes entity-level answer options and clause-scoped references, while Compete verifies ambiguous initial choices through evidence withholding and restora-

tion. Refine addresses incomplete geometry through adaptive boundary estimates and independently supported box-face updates.

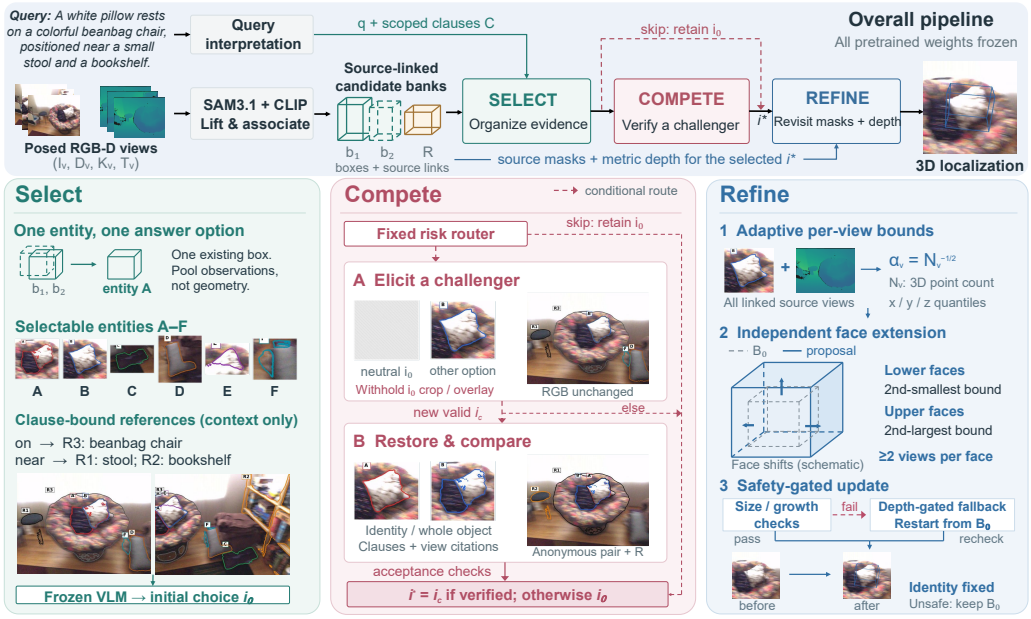


Figure 1: **Select, Compete, Refine.** Select organizes entity-level answer options and clause-bound references. Compete elicits an alternative through evidence withholding and verifies it after restoration. Refine updates individual box faces from multi-view boundary support, with a depth-gated fallback for unsafe proposals. Dashed arrows denote conditional execution. The query is abbreviated in the figure.

3.1 From Query and RGB-D Views to Object Candidates

We convert language into segmentation prompts and reasoning clauses, then associate depth-lifted masks across views. Each candidate retains its source masks and view indices, linking subsequent visual decisions to geometric observations.

Query interpretation. A frozen pretrained language model, prompted with text only, extracts target information and clauses describing appearance, relations, ordinals, viewpoint, and absolute position. We derive a short target phrase and a canonical category from the parsed target information as complementary segmentation prompts, such as *brown chair* and *chair*. Clauses retain the detailed conditions for reasoning and specify whether each condition concerns the target or a reference object. Dataset-specific model choices are given in Section 4.1 and the 27B parser is detailed in Appendix A.2.

Multi-view candidate construction. SAM3.1 (Carion et al., 2026; Meta, 2026) segments the views with these prompts. After mask suppression, we back-project valid depth pixels using K_v and T_v, filter noisy points, and associate spatially compatible fragments across views. Robust boxes fitted to their points form the target bank $\mathcal{B} = \{b_i\}_{i=1}^{N}$. CLIP (Radford et al., 2021) similarity and geometric support provide an initial ranking. Reference-object hypotheses from the same observations form a contextual bank. Both banks retain source observations, not only 3D boxes. Parsing and candidate construction are detailed in Appendix A.2, and deterministic scoring and entity representation in Appendix A.3.

3.2 Select: Scope-Aware Entity Evidence

Visual selection requires both distinct answer options and correctly assigned context. Duplicate fragments can make one object appear to be several choices, while an unstructured collection of

targets and references can obscure which objects a clause describes. For example, in *the chair beside the second table*, the ordinal orders tables rather than chairs. Select addresses both ambiguities before asking the VLM to choose.

Instance-level answer options. We group candidate hypotheses likely to depict the same physical object using center proximity and 3D/2D containment. A new member must be compatible with every existing member. Each group receives one selectable label and an existing representative box, chosen by identity relevance and geometric consistency. Member observations are available for evidence selection, but their boxes are not spatially united. Thus multiple views can support one answer option without creating a larger synthetic object.

Clause-scoped targets and references. We shortlist identity-relevant entities and alternatives supported by soft relation scores, retaining at most six entities, or eight for target ordinals. Each external reference phrase is matched to its bank hypotheses. For an ordinary relation, such as *beside the table*, we propose one reference instance. If the query orders reference objects, as in *beside the second table*, we retain multiple table hypotheses so the VLM can resolve that ordering. References are selected by detection quality, shared visibility, and relation compatibility, within a small global budget. The prompt binds each reference to its clause and marks it with a context-only R label. Target entities serve as context for same-category relations. A missing reference is marked unresolved rather than replaced by an assumed relation. These proposed references identify what to inspect, while the VLM still judges whether the stated relation holds.

Joint visual evidence. A greedy planner selects up to three full-frame views, rewarding visibility of targets and bound references and coverage of additional clauses. For ordinal queries, spatial separation of the relevant instances also contributes to view selection. One source-view detail crop per entity supports appearance comparison, while full frames preserve relational context. The VLM receives both image sets, the full query, and its scoped clauses, and returns the initial choice i_0. Candidate labels are shuffled, and numerical heuristic scores are hidden from the VLM. Invalid choices fall back to deterministic ranking (Appendices A.3.3 and A.4).

3.3 COMPETE: WITHHOLD–RESTORE VERIFICATION

A salient appearance match may dominate the initial decision even when a different instance better satisfies the query. Reconsidering the same evidence can preserve that preference. Withhold–restore verification separates finding an alternative from accepting a replacement, using different evidence conditions for the two judgments.

Selective invocation. A fixed risk rule routes multi-candidate queries using ranking ambiguity, VLM confidence, geometric support, and decision type. Unrouted queries retain i_0. The routing rule and its development protocol are specified in Appendices A.5.1 and A.5.5.

Stage A: elicit a challenger. Let E denote the Select evidence. We replace i_0's detail crop with a neutral placeholder and suppress its label and contour in context panels. Raw context RGB remains unchanged, and i_0 remains selectable. With the same query, clauses $\mathcal{C}$, views, and candidate labels, the frozen VLM F proposes

$$i_c = \text{select}\, F(q, \mathcal{C}, W_{i_0}(E)), \tag{1}$$

where W_{i_0} is the evidence-withholding operation. A different valid choice proceeds to Stage B. Otherwise, we retain i_0.

Stage B: restore, compare, and justify. We restore both candidates' evidence and anonymously relabel the pair A/B. Up to two clause-bound references and three joint views support comparison. The VLM judges target identity and whole-object status and cites visible differences using at most two discriminative clauses. We accept i_c only when the VLM chooses it with confidence at least 0.80, supports its identity, does not contradict its whole-object status, and provides counter-evidence against i_0. Counter-evidence is a contradicted identity or whole-object judgment for i_0, or a clause that supports i_c and contradicts i_0 with valid displayed-view citations. Otherwise, $i^* = i_0$. A replacement is accepted only after comparison with the restored evidence, evidence withholding alone cannot trigger a replacement.

3.4 REFINE: OBSERVATION-SUPPORTED MULTI-VIEW GEOMETRY

Partial views and robust point filtering can omit an object's extremities from its initial box. Refine revisits the selected object's source masks and depth to recover boundaries supported by multiple observations. It preserves the chosen identity and refines box faces independently, allowing nonuniform changes in both box size and center.

Adaptive per-view bounds. For each view linked to the selected entity, we union its source masks and lift all valid depth pixels. Let P_v contain the resulting N_v points. Views with $N_v < 128$ are discarded. Along axis j,

$$\alpha_v = N_v^{-1/2}, \qquad \ell_{vj} = Q_{\alpha_v}(P_{v,j}), \qquad u_{vj} = Q_{1-\alpha_v}(P_{v,j}), \tag{2}$$

where Q denotes an empirical quantile. Denser observations therefore receive less trimming than sparsely supported observations.

Independent face support. For $n \geq 2$ retained views, sort the lower and upper endpoints separately in ascending order. Given the initial box $B_0 = [\mathbf{L}_0, \mathbf{U}_0]$, the global update is

$$L_{g,j} = \min\{L_{0,j}, \ell_{(2)j}\}, \qquad U_{g,j} = \max\{U_{0,j}, u_{(n-1)j}\}. \tag{3}$$

Using the second extreme requires at least two views to support each outward displacement. Different faces can draw support from different views.

Safety-gated update. We retain a global proposal only if its maximum side length is at most 5 m and its volume is at most four times that of B_0. An unsafe update triggers a fallback from B_0: restrict source pixels to the box's camera-depth interval, retain one center- and depth-consistent connected component per view, and recompute the bounds. Each face uses the smaller extension from its tightest agreeing pair of views, or remains unchanged without agreement. If the fallback also fails the safety checks, we retain B_0.

4 EXPERIMENTS

We evaluate SCR on SeqAlign3DVG, EmbodiedScan, and ScanRefer against supervised and training-free grounding methods. We then examine how evidence organization, challenger verification, and box refinement affect localization, and report inference costs relative to VLM-Grounder.

4.1 EXPERIMENTAL SETUP

Datasets and protocols. SeqAlign3DVG (Zhang et al., 2026c) contains 2,916 evaluation queries from 141 scenes, each paired with 15 posed views. On EmbodiedScan (Wang et al., 2024), we follow the official Mini validation protocol, evaluating 11,776 eligible queries across 651 scenes or regions from ScanNet, 3RScan, and Matterport3D. We fix the 50-view inputs produced by the official sampler. For ScanRefer (Chen et al., 2020), we evaluate all 9,508 validation queries from 141 scans. Originally a point-cloud-based grounding benchmark, ScanRefer does not prescribe a fixed RGB-D input-view list. We select up to 48 views per scan by pose-based farthest-point sampling, shared across queries without target annotations, while retaining the original queries, targets, and 3D evaluation criteria.

Metrics and implementation. We report Top-1 Acc@0.25/0.50 with axis-aligned 3D IoU on SeqAlign3DVG and ScanRefer. EmbodiedScan uses official oriented IoU for Hit@10, Hit@10 checks whether any of the ten ranked boxes localizes the target. EmbodiedScan Top-1 is also reported because native VLM-Grounder outputs only one final box. Failures stay in the denominator. SeqAlign3DVG and EmbodiedScan use Qwen3.6-27B Qwen Team (2026a), including our VLM-Grounder reproductions. ScanRefer uses Qwen3.5-9B Qwen Team (2026b) for text-only query interpretation and image-conditioned selection and verification, retaining SAM3.1 and the SCR stages. All weights are frozen.

Table 1: SeqAlign3DVG results (%). Input denotes test-time modalities. For RGB BIP3D, the two rows differ in depth supervision during training.

Method	Task training	Input	Acc@0.25	Acc@0.50
BIP3D Lin et al. (2025b), w/o depth sup	Supervised	RGB	50.05	20.12
BIP3D Lin et al. (2025b), w/ depth sup	Supervised	RGB	59.60	24.55
SeqAlign3DVG Zhang et al. (2026c)	Supervised	RGB	51.30	22.39
Grounding DINO Liu et al. (2024) + back-projection	None	RGB-D	26.61	5.62
VLM-Grounder Xu et al. (2025), GPT-5.4-mini	None	RGB-D	47.28	21.18
VLM-Grounder Xu et al. (2025), Qwen3.6-27B	None	RGB-D	51.58	25.62
SCR (Ours)	None	RGB-D	**62.96**	**38.07**

4.2 Comparison with Existing Methods

SeqAlign3DVG. Table 1 compares SCR with supervised BIP3D (Lin et al., 2025b) and SeqAlign3DVG (Zhang et al., 2026c), a direct Grounding DINO (Liu et al., 2024) baseline, and training-free VLM-Grounder (Xu et al., 2025). SCR achieves 62.96/38.07, improving over the same-Qwen VLM-Grounder reproduction by 11.38/12.45 percentage points. The ablations below separate improvements in object selection from those in geometric estimation.

Table 2: EmbodiedScan basic Mini validation results (%). Both metrics use official oriented 3D IoU. The EmbodiedScan row is the released basic-prompt baseline; "–" denotes an unreported metric.

		Top-1		Hit@10	
Method	Task training	@0.25	@0.50	@0.25	@0.50
EmbodiedScan Wang et al. (2024)	Supervised	–	–	33.59	14.40
VLM-Grounder Xu et al. (2025)	None	14.25	6.05	–	–
SCR (Ours)	None	15.16	6.73	41.80	15.15

EmbodiedScan. SCR obtains 41.80/15.15 Hit@10, exceeding the released baseline by 8.21/0.75 points (Table 2). Its 15.16/6.73 Top-1 exceeds the completed same-Qwen VLM-Grounder result by 0.91/0.68 points. VLM-Grounder returns one box, so native Hit@10 is unavailable. SCR's Top-1–Hit@10 gap highlights remaining selection ambiguity.

Table 3: ScanRefer Overall accuracy (%). PC denotes a reconstructed scene point cloud, possibly with RGB evidence. Unmarked rows use full validation. †: 14-scan subset results.

Method	Task training	Input	Acc@0.25	Acc@0.50
BUTD-DETR (Jain et al., 2022)	Supervised	PC	52.2	39.8
EDA (Wu et al., 2023)	Supervised	PC	54.59	42.26
ZSVG3D (Yuan et al., 2024)	None	PC	36.4	32.7
SeeGround (Li et al., 2025b)	None	PC	44.1	39.4
SceneGraphGrounder† (Sun et al., 2026)	None	RGB-D	36.8	26.0
SCR (Ours)	None	RGB-D	41.03	24.41
SCR (Ours)†	None	RGB-D	**45.03**	**26.42**

ScanRefer. SCR reaches 41.03/24.41 on full validation (Table 3). Among RGB-D training-free methods, SceneGraphGrounder reports 36.8/26.0 on a subset (the first 14 scans of the full validation set), while our model yields 45.03/26.42 on this subset.

Figure 2 provides representative examples of the three behaviors emphasized in SCR: retaining a correct initial choice while improving geometry, accepting a challenger that better satisfies the query, and refining the selected object's final extent.

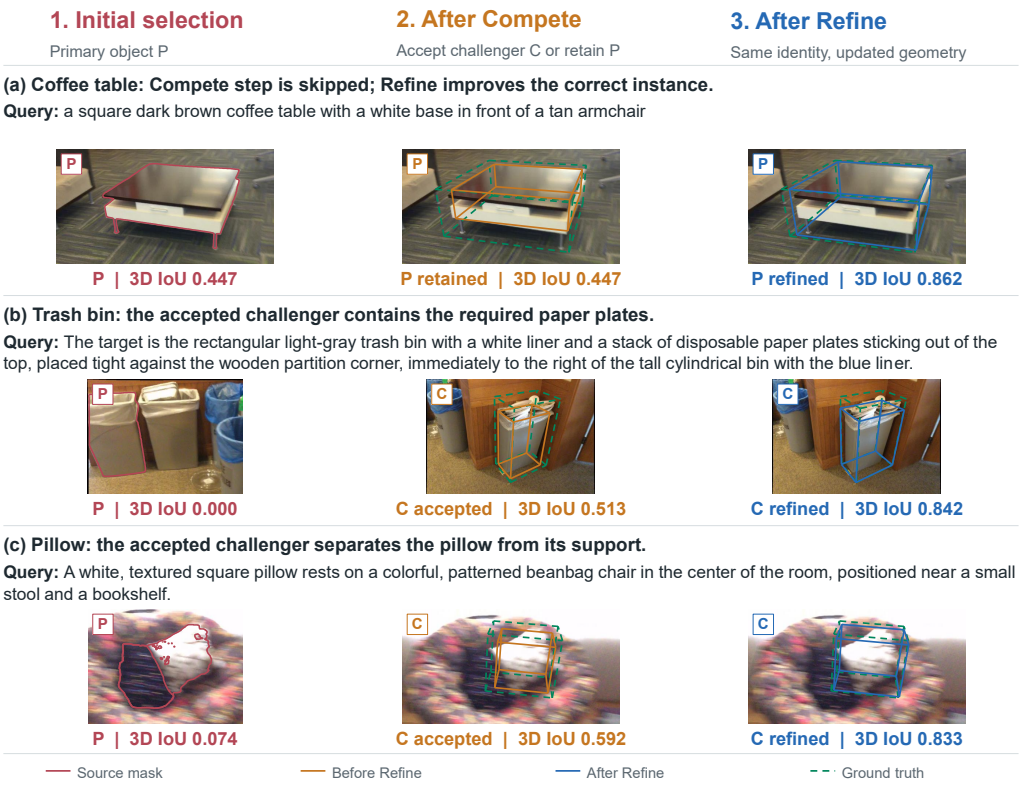


Figure 2: Qualitative results of SCR on SeqAlign3DVG. Each row shows one representative example across the three stages: the initial selection, the verified selection after COMPETE, and the final 3D box after REFINE. In (a), COMPETE is skipped and REFINE recovers the missing extent of the correct instance. In (b), the accepted challenger satisfies the required attribute (*paper plates*), and REFINE further improves the box. In (c), the accepted challenger separates the pillow from its supporting beanbag chair, after which REFINE recovers a tighter 3D localization. Red contours indicate source masks, orange boxes indicate the pre-refinement box, blue boxes indicate the final refined box, and green dashed boxes indicate ground truth.

4.3 ABLATION STUDIES

Table 4: Module combinations on SeqAlign3DVG (%). † denotes the unscoped evidence without clause-bound references. The remaining rows share the frozen primary predictions.

Selection evidence	Compete	Refine	Acc@0.25	Acc@0.50
Unscoped context†	No	No	51.37	20.03
Scope-aware Select	No	No	54.18	21.47
Scope-aware Select	Yes	No	55.62	21.98
Scope-aware Select	No	Yes	61.32	36.80
Scope-aware Select	Yes	Yes	**62.96**	**38.07**

Select: organizing candidates and their context. The unscoped baseline already shows source masks and RGB scene context, but does not explicitly associate references with query clauses. Presenting distinct entity-level options and clause-bound references in joint views improves accuracy from 51.37/20.03 to 54.18/21.47 (Table 4).

Compete: eliciting and verifying alternatives. Withhold–restore verification improves Select + Refine from 61.32/36.80 to 62.96/38.07. With the same 1,337 routed queries, per-call limits, and

restored verification procedure, SCR with withhold–restore verification reaches 62.96/38.07 versus 62.35/37.41 for structured reconsideration with intact evidence.

Refine: separating point recovery from boundary estimation. With selections from the Compete stage fixed, Table 5 pairs each initial estimator with the same Refine policy, without new SAM/Qwen calls. Using all valid source-mask depth points, dense raw min/max yields 47.67/19.48, whereas pooled 2–98% quantiles yield 59.40/26.54. With object selections fixed, dense raw min/max underperforms pooled quantiles, indicating that adding points alone is insufficient for accurate box estimation in this setting. Refine further improves raw-quantile boxes to 62.86/37.45 and boxes initialized with statistical outlier removal (SOR) and pooled quantiles from 55.52/21.64 to 63.58/38.48. These paired controls support observation-supported per-view boundary estimation beyond pooled quantiles; raw min/max remains unchanged. The main results retain the frozen original-fragment-box initialization, which improves from 55.62/21.98 to 62.96/38.07 with Refine.

Table 5: Paired geometry controls on SeqAlign3DVG (%). Object selections and observations are fixed. Dense estimators use all valid source-mask depth points. Each row uses the same Refine policy. $*$ marks the main model.

	Estimator only		+ Refine	
Initial box estimator	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
Original fragment boxes	55.62	21.98	62.96*	38.07*
Dense raw min/max	47.67	19.48	47.67	19.48
Dense raw pooled 2–98%	59.40	26.54	62.86	37.45
Dense SOR pooled 2–98%	55.52	21.64	**63.58**	**38.48**

4.4 Inference Efficiency

On SeqAlign3DVG, recorded costs are 33.31 s/query for SCR and 72.26 s/query for the same-Qwen VLM-Grounder reproduction. Within the withhold–restore controls, selective verification uses 11.63 versus 21.49 extra Qwen seconds per query for verification of all eligible queries, while reaching 62.96/38.07 versus 62.38/37.62.

5 Conclusion

We presented SCR, a training-free framework for 3D visual grounding from calibrated RGB-D views. SCR addresses object-selection ambiguity through clause-scoped target and reference organization and withhold–restore verification, separating the discovery of an alternative from the decision to replace an initial choice. Complementarily, multi-view boundary recovery refines the selected object's extent using adaptive per-view bounds and face-wise support, without additional segmentation or VLM calls. Experiments on SeqAlign3DVG, EmbodiedScan Mini, and ScanRefer assess SCR under different observation protocols and model configurations. Reported results exceed our same-Qwen VLM-Grounder reproductions on SeqAlign3DVG and EmbodiedScan, while ablations support the selection, verification, and boundary-recovery components. SCR remains dependent on candidate quality and the available calibrated observations, and its outward-only updates cannot correct already oversized boxes or guarantee recovery of unobserved extent. Future work will investigate broader candidate coverage and boundary estimation under sparse or uncertain observations.

AI use statement

We used generative AI tools to assist with research ideation and manuscript preparation. Specifically, these tools supported literature retrieval and summarization; discussion of research ideas and experimental design; drafting of the manuscript; and language polishing. We did not use generative AI to generate synthetic datasets or to prove mathematical claims.

We have reviewed and revised the AI-assisted material included in the final manuscript. AI suggestions were treated as inputs to author judgment, not as substitutes for original literature or ex-

perimental evidence. We are responsible for verifying literature attributions and ensuring that the technical descriptions, reported results, and visualizations are consistent with the implementation and experimental records. The use of pretrained language and vision-language models within the proposed grounding framework is described separately in the Methodology and Experiments sections. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

Ethics statement

This work studies language-conditioned 3D object localization, with evaluation on the existing SeqAlign3DVG, EmbodiedScan, and ScanRefer benchmarks. The proposed method is intended for object-level scene understanding, not for identifying individuals or inferring sensitive personal attributes.

Reproducibility statement

Section 3 defines SCR and its principal equations, while Section 4.1 specifies the datasets, observation protocols, metrics, and dataset-specific model choices. Appendix A.2 details the 27B parser and candidate construction; Appendices A.3–A.4 specify deterministic scoring, entity grouping, clause binding, and visual-evidence construction. Appendix A.5 provides the routing and replacement-acceptance rules, and Appendices A.6–A.7 detail boundary recovery, fallback processing, and safety checks. Appendix A.8 documents dataset-specific adaptations, box conversion, ranked outputs, and execution-accounting conventions.

References

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, 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, 2026. URL https://arxiv.org/abs/2511.16719.

Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. MiKASA: Multi-key-anchor & scene-aware transformer for 3D visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14131–14140, 2024.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, pp. 202–221. Springer, 2020. doi: 10.1007/978-3-030-58565-5_13.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Miao Hu, Zhiwei Huang, Tai Wang, Jiangmiao Pang, Dahua Lin, Nanning Zheng, and Runsen Xu. Changinggrounding: 3d visual grounding in changing scenes. arXiv preprint arXiv:2510.14965, 2025. URL https://arxiv.org/abs/2510.14965.

Wenyuan Huang, Zhenyu Zhang, Zhao Wang, Zhou Wei, Ting Huang, Fang Zhao, and Jian Yang. Openground: Planning-based online perception for open-world 3d visual grounding, 2026. URL https://arxiv.org/abs/2512.23020.

Cuong Huynh, Maxim Popov, Denis Gridusov, and Sergey Kolyubin. Agentgrounder: Zero-shot 3d visual pointcloud grounding using multimodal language models, 2026. URL https://arxiv.org/abs/2605.25901.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pp. 417–433. Springer, 2022.

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. arXiv preprint arXiv:2506.21924, 2025. URL https://arxiv.org/abs/2506.21924.

Fan Li, Zanyi Wang, Zeyi Huang, Guang Dai, Jingdong Wang, and Mengmeng Wang. TriCLIP-3D: A unified parameter-efficient framework for tri-modal 3D visual grounding based on CLIP. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5257–5266. Association for Computing Machinery, 2025a. doi: 10.1145/3746027.3755836.

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, 2025b.

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. Association for Computing Machinery, 2025a. doi: 10.1145/3746027.3754985.

Xuewu Lin, Tianwei Lin, Lichao Huang, Hongyu Xie, and Zhizhong Su. BIP3D: Bridging 2D images and 3D perception for embodied intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9007–9016, 2025b. doi: 10.1109/CVPR52734.2025.00842.

Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6032–6041, 2021.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pp. 38–55, 2024.

Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. OVIR-3D: Open-Vocabulary 3D Instance Retrieval Without Training on 3D Data. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1610–1620. PMLR, 2023.

Meta. SAM 3.1 release notes. Official software release, March 27, 2026, March 2026. URL https://github.com/facebookresearch/sam3/blob/main/RELEASE_SAM3p1.md.

Boyu Mi, Hanqing Wang, Tai Wang, Yilun Chen, and Jiangmiao Pang. Language-to-Space Programming for Training-Free 3D Visual Grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3844–3864. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.191/.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5.

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. URL https://proceedings.mlr.press/v139/radford21a.html.

Xuefei Sun, Xujia Zhang, Brendan Crowe, Doncey Albin, and Christoffer Heckman. Scenegraph-grounder: Zero-shot 3d visual grounding via structured scene graph matching, 2026. URL https://arxiv.org/abs/2605.21788.

Haibo Wang, Zihao Lin, Zhiyang Xu, and Lifu Huang. Think, act, build: An agentic framework with vision language models for zero-shot 3d visual grounding, 2026. URL https://arxiv.org/abs/2604.00528.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19757–19767, 2024.

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 (CVPR)*, pp. 19231–19242, 2023.

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 Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3961–3985. PMLR, 2025. URL https://proceedings.mlr.press/v270/xu25c.html.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7694–7701, 2024.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20623–20633, 2024.

Zhihao Yuan, Shuyi Jiang, Chun-Mei Feng, Yaolun Zhang, Shuguang Cui, Zhen Li, and Na Zhao. Scene-R1: Video-Grounded Large Language Models for 3D Scene Reasoning without 3D Annotations. arXiv preprint arXiv:2506.17545, 2025. URL https://arxiv.org/abs/2506.17545.

Yang Zhan, Yuan Yuan, and Zhitong Xiong. Mono3dvg: 3d visual grounding in monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6988–6996, 2024. doi: 10.1609/aaai.v38i7.28525.

Jiaxi Zhang, Yunheng Wang, Wei Lu, Taowen Wang, Shuning Zhang, Yixiao Feng, Junzhe Xu, Shunwang Sun, Weisheng Xu, Yuetong Fang, and Renjing Xu. Uniground: Universal 3d visual grounding via training-free scene parsing, 2026a. URL https://arxiv.org/abs/2603.08131.

Yani Zhang, Dongming Wu, Hao Shi, Yingfei Liu, Tiancai Wang, and Xingping Dong. DEGround: An effective baseline for ego-centric 3D visual grounding with a homogeneous framework. *arXiv preprint arXiv:2506.05199v3*, 2026b. URL https://arxiv.org/abs/2506.05199v3.

Yi Zhang, Yi Wang, Yueting Wu, Kaiyue Yang, Yuejiao Su, and Lap-Pui Chau. SeqAlign3DVG: A sequence-aligned benchmark and voxel reasoning framework for 3D visual grounding. In *Proceedings of the 34th ACM International Conference on Multimedia*. Association for Computing Machinery, 2026c. doi: 10.1145/3767308.3835644.

Henry Zheng, Hao Shi, Qihang Peng, Yong Xien Chng, Rui Huang, Yepeng Weng, Zhongchao Shi, and Gao Huang. DenseGrounding: Improving dense language-vision semantics for ego-centric 3D visual grounding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iGafR0hSln.

A APPENDIX

A.1 QUERY INTERPRETATION AND CANDIDATE CONSTRUCTION

This appendix specifies the implementation of Select, Compete, Refine (SCR). We distinguish a *fragment*, obtained from one source mask, a *candidate*, obtained by associating fragments across views, and an *entity*, which groups likely duplicate candidates into one answer option. Candidate construction associates observed geometry; entity grouping organizes the answer space without replacing member boxes by their union. All distances below are in meters unless otherwise indicated.

A.2 FROZEN MODELS AND QUERY INTERPRETATION

Models. The 27B configuration uses SAM3.1 for text-conditioned masks, CLIP ViT-B/32 for appearance retrieval, and Qwen3.6-27B in BF16 for query interpretation and visual judgments. Visual calls use greedy decoding, a maximum of 1,003,520 pixels per input image, and at most 450 output tokens. The segmentation resolution is 1,008 pixels. Table 6 summarizes the principal settings. These are inference-time components with frozen pretrained weights; none is fine-tuned for grounding.

Identity and clauses. Query interpretation has two cached text-only passes using the same Qwen model. The first returns the target core, category head, aliases, attributes, reference phrases, relations, and forbidden target terms. The second receives the query and an identity contract containing the first pass's core, head, and aliases. It must preserve that identity while extracting appearance, relation, ordinal, viewpoint, and absolute-position clauses. Each clause records its query span and subject or scope, so an ordinal of an anchor group is distinguished from an ordinal of the target. Identity-copy checks and deterministic normalization validate the record, standardize predicate names and explicit directions, and remove duplicate clauses. The second pass is not an independent replacement for the first-pass identity.

The first output budget is 512 tokens. The clause budget is 900 tokens; six truncated responses in the frozen SeqAlign3DVG cache were regenerated with 1,600 tokens. Both caches contain 2,916 Qwen3.6-27B records. The implementation can inherit an incorrect first-pass identity even if the clause decomposition is internally consistent.

Text sent to SAM. The target prompt generator extracts a canonical noun from the category head, falling back to the core. It shortens the core after that noun where possible; otherwise, it removes trailing descriptions introduced by terms such as *with*, *featuring*, or *which*. Articles are removed. The distinct short target phrase and category are sent as *separate* SAM prompts. For example, *brown chair with high backrests* yields *brown chair* and *chair*, not their concatenation. The first preserves a useful appearance cue; the second avoids requiring the segmenter to satisfy it. Identical strings are evaluated once and forbidden terms are filtered. Alias expansion is disabled in this configuration. Fixed proposal score offsets are 0.030 for the short phrase and 0.020 for the category. Downstream visual reasoning still receives the complete original query.

A.2.1 FROM MASKS TO THE TARGET BANK

Mask retention. The SAM confidence threshold is 0.25. Masks smaller than 80 pixels or larger than 70% of an image are excluded; mask-IoU suppression uses threshold 0.82. At most eight masks per prompt and eight masks per view survive, with a global target-mask budget of 120. These limits can discard valid observations and are part of the reported configuration.

Back-projection and compact fragments. Write the camera-to-world pose as $(R_v, \mathbf{t}_v)$. For homogeneous pixel $\tilde{\mathbf{u}} = (u_x, u_y, 1)^\top$, back-projection is

$$\mathbf{x}_v(\mathbf{u}) = R_v\big[D_v(\mathbf{u})K_v^{-1}\tilde{\mathbf{u}}\big] + \mathbf{t}_v. \tag{4}$$

Depth is metric camera-axis depth. Masks are resized to the depth grid by nearest-neighbor interpolation, with correspondingly scaled intrinsics. Dataset alignment is included in the common coordinate frame. A world-to-camera extrinsic matrix must be inverted before use in Eq. equation 4.

A fragment needs at least 40 valid depth points. Up to 8,192 points per mask are retained. Target-fragment statistical outlier removal uses 20 neighbors and a standard-deviation ratio of 1.0, with at least 30 points retained after filtering. Compact boxes use coordinatewise 2% and 98% quantiles. Initial fragment dimensions are checked against 0.01 and 5.0 m. The original masks and their source links are retained separately from the compact point sample.

Cross-view association. For fragments i, j from different views, let d_{ij} be centroid distance and J_{ij} their axis-aligned 3D box IoU. An association edge requires

$$d_{ij} \le 0.45, \qquad 0.55 \exp(-d_{ij}/0.22) + 0.35 J_{ij} \ge 0.45, \tag{5}$$

and a pairwise enclosing box with no dimension above 4 m. Appearance and projection terms do not contribute in this configuration. Connected components of the accepted edges are obtained by union–find. This is *not* complete-link clustering: two same-view fragments have no direct edge but can enter the same component through a third view.

Associated point sets are pooled, capped at 8,192 points using a fixed seed, and fitted again with coordinatewise $2\%/98\%$ bounds. A union of fragment boxes is used only when point data are unavailable. Each candidate stores its source view IDs, fragment and mask IDs, source rectangles, and fragment qualities. Candidate quality is the mean fragment quality. Before semantic ranking, candidates are ordered lexicographically by view count, quality, source-point count, and volume. Box NMS suppresses candidates with 3D IoU at least 0.85 with an earlier retained candidate; at most 80 remain. The strict bank does not add projection variants or single-fragment backfill proposals.

A.2.2 Reference-Object Bank

An independent reference bank is built from the first parser's anchor phrases and relation references. Up to four concepts are retained after deduplication and filtering of generic or collective same-category terms. For each concept, a canonical noun and a short anchor phrase are sent to SAM, with score offsets 0.015 and 0.025, respectively. Lifting and cross-view association use the same coordinate convention and quantile-box construction as the target branch; this reference builder does not run the target branch's additional statistical-outlier-removal step.

After fusion, the reference builder keeps up to three instances per concept according to

$$s_{\text{anchor}} = q + 0.16 \min(1, n_v/5) + 0.05 \min\left(1, \frac{\log(1+n_f)}{\log 6}\right) + 0.03 \operatorname{clip}_{[0,1]}\left(\frac{\log(1+V)}{\log 9}\right), \tag{6}$$

where q, n_v, n_f, V denote quality, source-view count, fragment count, and volume. CLIP is not used in this reference-instance ranking. The bank is a set of hypotheses, not a set of identified ground-truth anchors. Its legacy noun canonicalizer and first-pass concept list can omit a reference present in the later clause record. Such a reference is left unresolved rather than assumed to exist or regenerated from annotations.

A.3 Deterministic Ranking and Entity Representation

The deterministic scores serve different purposes: initial bank ordering, identity-based entity ordering, and selection of an existing representative box. They are fixed heuristics, not calibrated probabilities or learned predictions of ground-truth IoU.

A.3.1 Appearance and Identity Scores

For each candidate, up to four source views are selected by maximum fragment quality per view. CLIP computes the maximum crop–text cosine similarity over those views and three forms of the parsed target phrase: the phrase itself, *a photo of a [phrase]*, and *a cropped photo of a [phrase]*. This parsed phrase can be longer than the shortened SAM prompt. The crop extraction retains context rather than replacing the RGB background with a binary mask. The initial score and subsequent identity score are

$$s_{\text{pre}} = s_{\text{CLIP}} + 0.04v + 0.02s + 0.04f - 0.03t + 0.12h, \tag{7}$$

$$s_{\text{id}} = s_{\text{pre}} + 0.04q + 0.03v + 0.04h - 0.08t. \tag{8}$$

Component	Frozen setting
SAM3.1 resolution; confidence	1,008; 0.25
Mask area; mask-IoU NMS	$[80,\ 0.70WH]$ pixels; 0.82
Target masks: per view / total	8 / 120
Point cap: per mask / candidate	8,192 / 8,192
Target SOR: neighbors / ratio	20 / 1.0
Compact box quantiles	0.02, 0.98
Association: distance cap / scale / score	0.45 m / 0.22 m / 0.45
Candidate NMS; target-bank limit	0.85; 80
Reference concepts / instances per concept	4 / 3
CLIP views / image batch size	4 / 256
Selectable entities: ordinary / target ordinal	6 / 8
References: ordinary / external ordinal group	3 / 4
Primary joint views; detail crops per entity	At most 3; 1
Qwen precision; visual decoding	BF16; greedy
Visual pixels per image; output tokens	1,003,520; 450
Stage-B references / joint views	At most 2 / 3
Stage-B confidence threshold	0.80
Refine minimum points / distinct views	128 / 2
Refine maximum dimension / volume ratio	5 m / 4

Table 6: Principal inference settings. Source observations may be fewer than the stated maxima. The generic reranker setting of top-12 does not override the scope-aware six/eight-entity display budget.

There is no anchor-relation bonus in s_{pre}. For dimensions $\mathbf{d}$, $d_{\min} = \min_j d_j$, and $V = \prod_j \max(d_j, 10^{-3})$, the terms are

$$v = \frac{\log(1+n_v)}{\log 16}, \qquad f = \min\left(1, \frac{\log(1+n_f)}{\log 64}\right), \tag{9}$$

$$s = \tanh(V/0.08)\min(1, d_{\min}/0.12), \qquad t = \mathbf{1}[d_{\min} < 0.08]. \tag{10}$$

The view term v is not clipped, so it can exceed one with more than 15 source views. The repeated terms in Eq. equation 8 are additive. Quality q averages fragment scores; a fragment's score is its stored SAM confidence, including its prompt offset, multiplied by $\min(1, N/40)$ for the valid lifted-point count N.

Shape preference. The shape term compares dimensions with a fixed category prototype:

$$h = \exp\left[-\frac{1}{3}\sum_{j=1}^{3}\left|\log\frac{\tilde{d}_j}{\tilde{d}_j^{\text{ref}}}\right|\right]. \tag{11}$$

The two horizontal dimensions are sorted independently for the candidate and prototype; height remains the third dimension. Dimensions are floored at 10^{-3} m. The longest matching category key in the target phrase selects the prototype; an unmatched category receives zero. These manually specified preferences are not a learned completeness classifier.

A.3.2 Consistency Features and Representative Score

For a group $\mathcal{G}$, the representative is an existing member maximizing

$$s_{\text{rep}}(i) = s_{\text{id}}(i) + 0.05p_i + 0.04d_i^{\text{dep}} + 0.03\beta_i + 0.02m_i. \tag{12}$$

Ties favor higher identity score and then earlier original rank. Consistency features are recomputed from the compact candidate geometry and source observations, before selected-only Refine.

Projection consistency p. Up to five views are chosen by maximum fragment quality per view. Let P_v be the projected 3D-box rectangle and S_v the enclosing source-mask rectangle, expressed on

Category	Dimensions (m)	Category	Dimensions (m)
trash/garbage can	.30,.30,.60	coffee table	.60,1.00,.45
book shelf/bookshelf	.35,1.00,1.70	whiteboard	.06,1.20,.90
keyboard	.15,.45,.05	monitor	.08,.50,.35
picture/painting	.05,.80,.60	cabinet	.45,1.00,1.00
dresser	.45,1.10,.90	pillow	.25,.50,.18
bottle	.08,.08,.25	toilet	.45,.70,.75
window	.08,1.10,1.10	chair	.55,.65,.85
stool	.35,.35,.55	bench	.45,1.20,.50
table	.75,1.35,.75	desk	.70,1.20,.75
sofa/couch	.90,1.80,.80	shelf	.35,1.00,1.50
sink	.45,.60,.25	door	.08,.90,2.00
lamp	.20,.20,.80	plant	.40,.40,.80
box	.40,.50,.40	bag	.30,.40,.40
bed	1.40,2.00,.60		

Table 7: All category-size prototypes in the frozen identity score.

the same image grid. Then

$$p = 0.75\,\mathrm{med}_v\,\mathrm{IoU}_{2D}(P_v, S_v) + 0.25\,\mathrm{med}_v \exp\left[-\frac{\|c(P_v) - c(S_v)\|_2}{0.20\sqrt{W_v^2 + H_v^2}}\right]. \tag{13}$$

Only valid projections contribute; no valid values yield zero. This is rectangle agreement, not mask IoU. The projected rectangles are not drawn as the visual evidence shown to Qwen.

Depth consistency d^{dep}. Project the box center and eight corners into each selected view. For finite positive-depth probes inside the image, sample depth at rounded, resolution-adjusted pixel coordinates. With at least two valid probes,

$$d_v^{\mathrm{dep}} = \exp[-\,\mathrm{med}_k\,|D_v(u_k) - z_k|/0.08]\,, \qquad d^{\mathrm{dep}} = \mathrm{med}_v\, d_v^{\mathrm{dep}}. \tag{14}$$

An empty valid set yields zero. These probes can fall on occluders or background; they provide a coarse compatibility feature, not a dense surface-consistency certificate.

Source-box stability β. Using all source views, let a_v and r_v be the normalized area and aspect ratio of S_v. With at least two observations,

$$\beta = \exp[-\tfrac{1}{2}(\mathrm{CV}(a) + \mathrm{CV}(r))], \qquad \mathrm{CV}(x) = \frac{\mathrm{std}(x)}{\max(\mathrm{mean}(x), 10^{-6})}. \tag{15}$$

Widths and heights are floored at one pixel. Fewer than two observations give $\beta = 0.5$. This measures variation of 2D support, not uncertainty from independently fitted 3D boxes.

Multi-view support m. The support feature is $m = \frac{1}{2}\min(1, n_v/6) + \frac{1}{2}\min(1, n_f/8)$. Repeated support is useful but does not guarantee that every associated fragment belongs to the same physical object.

A.3.3 Grouping, Geometry, and Anonymous Labels

Candidates are visited in decreasing s_{id}, with original order breaking ties. A candidate joins the first group for which it is compatible with *every* member; otherwise, it starts a new group. This complete-link rule is distinct from the earlier union–find association. Define 3D containment as intersection volume divided by the smaller volume. For each shared source view, define 2D containment analogously using enclosing source rectangles. Ordinary compatibility requires:

1. center distance at most 0.18 m and 3D containment at least 0.80;
2. median 2D containment at least 0.90;
3. at least two shared views with 2D containment at least 0.70.

If only one shared view exists, it must have containment at least 0.95. An inherited same-fusion-family exception uses distance 0.25 m, 3D containment 0.65, and median 2D containment 0.82, retaining the shared-view requirement. The strict front end does not generate additional family variants.

The group score is $\max_{i \in \mathcal{G}} s_{\mathrm{id}}(i)$, while its geometry is the member maximizing Eq. equation 12. These need not be the same member. Member source records are pooled and deduplicated by view, fragment, and mask identifiers. The representative box is never replaced by a spatial union at this stage. Thus two hypotheses of one chair can provide observations to one selectable entity without inventing a larger chair. Conversely, an incorrect grouping can hide a genuinely independent object from the answer list; retaining the raw bank does not prevent that selection error.

After shortlist construction, labels are shuffled using a NumPy random state seeded by the integer value of the first eight hexadecimal digits of the record-identifier MD5 hash. Labels therefore do not encode score rank. One entity receives one selectable letter and retains the same mapping across the primary call and Compete Stage A. Reference labels begin with `R` and are not valid answers. Stage B uses a separate fixed salted shuffle for its two alternatives.

A.4 Select: Clause-Scoped Visual Evidence

Select first builds a compact target shortlist, then binds reference hypotheses to individual clauses, and finally selects views that expose these targets and references together. Binding and visibility are planning signals: they do not establish that a relation is true.

A.4.1 Target Shortlist and Soft Relation Heuristics

The top three entities by identity score are protected. For each supported target-subject relation, a preliminary text binder selects an available anchor concept; up to three instances are considered. The relation score of a target is its best score over those instances. The two highest-scoring targets per relation are added to the protected set, then remaining slots are filled in identity order. If the union exceeds the budget, it is pruned lexicographically by protected status, best relation score, and entity score, then restored to identity order. The budget is six, or eight if a target ordinal exists. An anchor ordinal alone does not raise the target budget.

The preliminary binder accepts an exact normalized phrase match or token overlap $|A \cap B| / \min(|A|, |B|) \geq 0.75$ with a runner-up margin of 0.15. Normalization removes articles and common relation function words and normalizes plurals. This is separate from the more permissive scoped binder below. Unsupported relations add no ranking bonus. No intersection of hard relation predicates is used to exclude targets.

Geometric terms. For target and reference centers $\mathbf{c}_T, \mathbf{c}_R$ and dimensions $\mathbf{d}_T, \mathbf{d}_R$, set $\Delta = \mathbf{c}_T - \mathbf{c}_R$ and

$$g = \left\| \max\left(|\Delta_{xy}| - \tfrac{1}{2}(\mathbf{d}_{T,xy} + \mathbf{d}_{R,xy}), 0\right) \right\|_2, \tag{16}$$

$$a = \max\left(0.25, \tfrac{1}{4}[\mathrm{mean}(\mathbf{d}_{T,xy}) + \mathrm{mean}(\mathbf{d}_{R,xy})]\right), \tag{17}$$

$$P = \exp(-g/a), \qquad P_b = \exp[-g/\max(0.75, 2a)], \tag{18}$$

$$Z = \exp[-\max(0, |\Delta_z| - \tfrac{1}{2}(d_{T,z} + d_{R,z}))/0.30], \tag{19}$$

$$w = \min(1, n_{\mathrm{shared}}/2). \tag{20}$$

Let O be the product of the per-axis horizontal intersection lengths divided by the smaller target/reference dimension, clipped to $[0, 1]$. Let C_{3D} denote 3D containment, σ the logistic function, and $z_T^-, z_T^+, z_R^-, z_R^+$ the lower and upper heights. The soft scores are

$$s_{\text{next to}} = 0.76P + 0.14Z + 0.10w, \tag{21}$$

$$s_{\text{near}} = 0.88P_b + 0.12w, \tag{22}$$

$$s_{\text{adjacent}} = 0.72\exp[-g/\max(0.16, 0.60a)] + 0.18Z + 0.10w, \tag{23}$$

$$s_{\text{on}} = 0.55\exp(-|z_T^- - z_R^+|/0.20) + 0.35O + 0.10w, \tag{24}$$

$$s_{\text{under}} = 0.50\sigma(-6\Delta_z) + 0.25\exp(-|z_T^+ - z_R^-|/0.30) + 0.15O + 0.10w, \tag{25}$$

$$s_{\text{closest}} = 0.90P + 0.10w, \tag{26}$$

$$s_{\text{above}} = 0.62\sigma(6\Delta_z) + 0.28O + 0.10w, \tag{27}$$

$$s_{\text{in}} = (0.72C_{3D} + 0.18O + 0.10w)\mathbf{1}[V_T \le V_R], \tag{28}$$

$$s_{\text{against}} = 0.62P + 0.20Z + 0.10\left(1 - \frac{\min_j d_{R,j}}{\max_j d_{R,j}}\right) + 0.08w. \tag{29}$$

Dimensions are floored at 10^{-4} m when needed. For left/right, use the horizontal center offsets of normalized source rectangles in shared views. If A is the fraction with the requested sign, $E = \min(1, \text{med}\,|\Delta x_v|/0.08)$ and $W = \min(1, n_{\text{valid}}/2)$, the score is $A(0.65 + 0.20E + 0.15W)$. It is image-relative, not a world-coordinate left/right estimator. All relation scores are clipped to $[0, 1]$.

A.4.2 Clause Binding and Reference Selection

Which clauses need references? Each relation requests its object phrase. An anchor ordinal requests its group concept and, when present, its own reference; a target ordinal requests its explicit reference. Viewpoint and absolute-position clauses can likewise request a concrete reference. Generic notions such as the room or viewer are not treated as segmentable objects. Same-category group descriptions use other target entities as context rather than creating an additional selectable copy. Reference requirements retain their clause IDs and roles.

Matching a phrase to the available bank. The scoped binder lowercases text, removes punctuation and selected function words, normalizes plurals, and applies a fixed alias map, including couch/settee to sofa, refrigerator to fridge, television to TV, monitor to screen, and wastebasket/trashcan to bin. It compares a requested phrase's token set D with the bank concept's phrase, short phrase, and category. For one such variant S, let $p = |D \cap S|/|S|$, $r = |D \cap S|/|D|$, H indicate a matching canonical head, and X indicate an exact canonical phrase match. The concept score is the maximum across its variants of

$$b(D, S) = \max\{X,\ 0.50p + 0.25H + 0.20\min(1, 2r) + 0.05\min(1, |S|/4)\}. \tag{30}$$

The best concept must score at least 0.67 and exceed the runner-up by 0.08; the margin is waived for a score of at least 0.999. Otherwise, the requirement remains unresolved. Matching operates only on the cached bank, not on ground-truth classes or boxes.

Choosing instances. For up to three available instances of a bound concept, compute

$$s_R = 0.40q_R + 0.25\min(1, n_{\text{shared}}/3) + 0.20s_{\text{rel}} + 0.15/\operatorname{rank}_R, \tag{31}$$

where shared views are with any shortlisted target, s_{rel} is the largest applicable relation score over those targets, and rank_R is the one-based bank rank from Eq. equation 6. An ordinary requirement proposes one instance. An external anchor-ordinal group proposes two or three, according to the requested group size clipped to $[2, 3]$, with three used if size is unspecified, subject to availability.

Repeated object IDs are deduplicated while pooling clause memberships and roles. References are then ordered by number of bound clauses, external ordinal-group priority, and s_R. The total budget is three, raised to four when an external ordinal-group requirement exists. A same-category target group alone does not increase it. Global pruning can leave a clause without a reference. This is a compact hypothesis planner, not a joint optimization that certifies which physical anchor the query denotes.

A.4.3 Selecting the Three Joint Views

Within each source view, the planner encloses all pooled source-mask rectangles of an entity and normalizes their coordinates to $[0, 1]$. A rectangle's presence defines observed visibility; its area and center give size and horizontal-position statistics. References are treated analogously. Visibility means that source evidence exists, not that an object is fully visible or unoccluded.

For a clause with external references, each target is paired with the clause-bound reference sharing the most source views. The shared views are eligible to inspect that target–reference clause. Same-category context uses shared views with other targets. A clause with no external requirement uses the target's own views; an unresolved external requirement provides no joint coverage. A view

covers a clause if it is eligible for at least one target. This predicts inspectability, not satisfaction of the predicate.

Let T_v, R_v, C_v be the targets, references, and observable clauses in view v, with global counts N_T, N_R, N_C. Define

$$t_v = \frac{|T_v|}{\max(1, N_T)}, \qquad a_v = \frac{|R_v|}{\max(1, N_R)}, \tag{32}$$

$$c_v = \frac{|C_v|}{\max(1, N_C)}, \qquad u_v = \min(1, \mathrm{mean}_{i \in T_v} A_{iv}/0.16), \tag{33}$$

where A_{iv} is normalized rectangle area and the empty mean is zero. The clause denominator includes clauses with no eligible views. For target ordinals, $o_v = t_v \min(1, \mathrm{span}_{T,v}/0.45)$; otherwise, for anchor ordinals, $o_v = a_v \min(1, \mathrm{span}_{R,v}/0.45)$; otherwise $o_v = 0$. Each span is the range of horizontal rectangle-center coordinates, zero for fewer than two instances. This spread proxy remains horizontal even for vertical or distance ordinals; it is not an ordinal solver.

The fixed base score and greedy incremental score are

$$b_v = 0.34t_v + 0.27c_v + 0.14a_v + 0.15o_v + 0.10u_v, \tag{34}$$

$$g(v \mid \mathcal{V}) = b_v + 0.18\frac{|T_v \setminus T_{\mathrm{seen}}|}{\max(1, N_T)} + 0.20\frac{|C_v \setminus C_{\mathrm{seen}}|}{\max(1, N_C)}. \tag{35}$$

Starting with no selected views, choose the maximum-scoring unused view, update the covered sets, and repeat for at most three views. Ties favor smaller view indices. The coverage bonuses discourage repeatedly selecting the same visible entities or clauses; this is not simply the three highest static scores. All parsed clauses remain in the prompt even if no selected view supports inspecting them.

A.4.4 Rendering and the Primary Judgment

The input consists of a joint-view sheet and a source-mask detail sheet. For each entity, a crop view is scored by the enclosing source rectangle's normalized area plus 0.10 times the highest source-record score, with a 0.35 penalty if the rectangle touches an image border within six pixels. The best view gives one detail crop, using context scale 1.35 and minimum size 180 pixels. All pooled source masks in that view can contribute contours of one color under the entity's single label. Thus the frozen renderer can show multiple contours for one entity, rather than enforcing one cleaned mask per entity and view.

The joint views retain RGB appearance and use source-mask contours and labels, not projected 3D boxes. Reference labels are context-only. The prompt includes the full query, typed clauses, selectable labels, and clause-bound reference descriptions. It requests a selected label, runner-up, decision type, identity judgment, clause and anchor statuses, view citations, confidence, and a short rationale. Candidate ranks and numerical heuristic scores are not supplied. The label is mapped back to its representative object ID; an unusable choice falls back to the stored deterministic ranking. The detail and joint sheets preserve context, but contour clutter or erroneous grouping can still affect the visual judgment.

A.5 Compete: Routing, Intervention, and Acceptance

Compete separates eliciting an alternative from deciding whether it should replace the primary choice. Evidence withholding alone is not treated as proof that the primary prediction is wrong.

A.5.1 Selective Invocation

Eligible queries have a valid primary choice and at least two displayed target entities. Let n be their number, r_0 the primary's rank among them, Δ_s its deterministic score minus the best alternative score, and Δ_e the analogous entity-score gap. Deterministic scores use the stored `pre_consistency_final_score`, falling back to `final_score`; ties use object ID. Let γ_0 be primary confidence, and p_0, h_0, m_0 its projection consistency, stored shape prior, and multi-view support. The router reads `shape_prior`, not the separate `vlm_shape_prior` field used

for the identity score; an absent router field uses the default below rather than recomputing it. For $\mathrm{clip}(x) = \min(1, \max(0, x))$, define

$$z_1 = \mathrm{clip}\left(\frac{r_0 - 1}{\max(1, n-1)}\right), \qquad z_2 = 1 - \mathrm{clip}\left(\frac{\max(0, \Delta_s)}{0.12}\right), \tag{36}$$

$$z_3 = 1 - \mathrm{clip}\left(\frac{\max(0, \Delta_e)}{0.12}\right), \qquad z_4 = 1 - \mathrm{clip}(\gamma_0), \tag{37}$$

$$z_5 = 1 - \mathrm{clip}(p_0), \qquad z_6 = 1 - \mathrm{clip}(h_0), \tag{38}$$

$$z_7 = 1 - \mathrm{clip}(m_0), \qquad z_8 = \mathrm{clip}((n-2)/6). \tag{39}$$

Missing confidence defaults to 0.9; missing projection, shape, and support features default to 0.5. Confidence is a self-reported score, not a calibrated probability. The risk is

$$r = \mathbf{w}^\top \mathbf{z}, \qquad \mathbf{w} = (0.24, 0.20, 0.12, 0.14, 0.12, 0.08, 0.05, 0.05). \tag{40}$$

Stage A is invoked when $r \geq 0.33806227953614815$ and the *primary response's* decision type is identity, appearance, relation, or mixed. Ordinal, viewpoint, position, and unknown decision types skip Compete and retain their primary choice. This decision type is not simply the list of clause types present in the query. Skipped rows remain in the evaluation denominator.

A.5.2 Stage A: Withholding Primary Presentation Evidence

Stage A retains the query, clause record, candidate set, label mapping, and joint-view plan. It replaces the primary entity's detail crop by a neutral placeholder and suppresses its label and contours in the joint views. Raw scene RGB is unchanged, and the primary remains a selectable option. Its appearance may therefore remain visible in the scene or in overlapping crops. This is a candidate-specific presentation intervention, not removal of the physical object from every image.

The model is not forced to change its answer. Stage B is reached only if Stage A returns a valid, different candidate in the displayed set. A repeated primary, invalid choice, or unavailable challenger retains the primary prediction. All visual calls use the same frozen model and decoding settings.

A.5.3 Stage B: Pair-Specific References and Views

Stage B restores evidence and compares only the primary i_0 and challenger i_c, under newly shuffled anonymous pair labels. Their source-view sets V_0, V_c include the pooled observations of their entity groups. Put $B = V_0 \cap V_c$ and $U = V_0 \cup V_c$.

Reference selection. Only references already retained by Select are considered; Stage B does not run another detector. Let S_R be a reference's stored shared-view set with the primary shortlist, C_R its bound clause IDs, and s_R its score from Eq. equation 31. References with no intersection with U are removed. The remaining score is

$$b_R = 4\mathbf{1}[S_R \cap B \neq \varnothing] + 0.50|S_R \cap B| + 0.20|S_R \cap U| + 0.25|C_R| + 0.05 s_R. \tag{41}$$

At most two references are retained, with original reference order and object ID breaking ties. This prioritizes a reference that can be inspected with both alternatives rather than independently inventing a new anchor.

View selection. Consider views in U. Let $n_v \in \{1, 2\}$ be the number of pair members visible, P the primary's selected context-view set, a_v the number of retained references whose S_R contains v, and C_v the stored observable-clause set from the primary plan. Rank views once by

$$b_v^{\mathrm{pair}} = 8\mathbf{1}[n_v = 2] + 2\mathbf{1}[v \in P] + 1.25 a_v + 0.35 \min(6, |C_v|) + 0.20 n_v. \tag{42}$$

Ties favor earlier primary-plan position and then smaller view index. Retain at most three views. If none contains the primary, replace the last selected view by its smallest source-view index; apply the analogous check to the challenger and deduplicate. Unlike Eq. equation 35, this is a pair-specific static ranking with coverage checks, not a greedy novelty objective.

The prompt supplies the original query, clauses, pair evidence, and retained context-only references, without exposing which pair label was the primary. It requests identity and whole-object judgments for both options, a final choice and confidence, and at most two discriminative clause comparisons with view citations. These judgments come from Qwen, not a separate ground-truth-based identity or completeness classifier.

A.5.4 EXACT ACCEPTANCE GATE

The challenger replaces the primary only if all of the following hold:

1. The Stage-B response has a valid selectable choice.
2. That choice maps to the challenger, not the primary.
3. Both pair labels have valid object-ID mappings.
4. The reported confidence is at least 0.80.
5. Challenger identity is `supported`.
6. Challenger whole-object status is not `contradicted`.
7. There is explicit counter-evidence against the primary, as defined below.
8. Every view citation in a decisive clause is a valid rendered context-view ID.

Counter-evidence is satisfied if the primary's identity or whole-object status is `contradicted`, or if at least one decisive clause exists. A decisive clause must be listed among the response's distinguishing clause IDs, marked discriminative, judged supported for the challenger and contradicted for the primary, and cite a nonempty set of view IDs. Malformed or out-of-plan citations in such clauses reject the update. If primary identity or whole-object contradiction already supplies counter-evidence, a decisive clause and its citations are not required.

The distinction between *not contradicted* and *supported* is important: a weak or unobservable challenger whole-object judgment does not itself fail this gate. The gate checks structured judgments and citation membership, not whether the pixels independently entail the claims. Any failed acceptance condition retains the primary. Thus an incorrectly proposed challenger does not automatically change the output, but a confident, internally consistent model error can still pass.

A.5.5 POLICY DEVELOPMENT AND COMPUTATION ACCOUNTING

The invocation policy was selected on 500 declared development rows using accuracy and recorded computational cost. The search considered decision-type families (appearance/relation/mixed; those types plus identity; relation/mixed; and all types) and eligible-row risk budgets. Thresholds were empirical risk quantiles. The retained policy used the identity-inclusive family and development budget 0.65, followed by checks on separate row subsets. Deployment uses the fixed threshold above; it does not force 65% of every new dataset to invoke Compete.

Inference does not read target annotations, but policy selection did use development annotations.

Let a_i indicate invocation of Stage A and b_i invocation of Stage B. Excluding retries, the number of visual calls for a valid primary row is $1 + a_i + b_i$, with $b_i \le a_i$. Token and latency accounting includes all invoked calls, including rejected challengers. The efficient SeqAlign3DVG result is an offline replay of these cached deterministic calls on the router-selected rows, not a newly measured cold end-to-end run. Cached replay time, summed model latency, and parallel wall-clock throughput must be reported separately.

A.6 REFINE: ADAPTIVE MULTI-VIEW BOUNDS AND THE FALLBACK

Refine acts only after the final entity has been selected. Its input box $B_0 = [\mathbf{L}_0, \mathbf{U}_0]$ is that entity's existing representative, with center $\mathbf{c}_0$ and dimensions $\mathbf{d}_0$. All linked source views, not just the three visual-evidence views, can contribute. There are no additional SAM or Qwen calls, and refinement does not change identity or feed revised geometry back into ranking.

A.6.1 DENSE SOURCE SUPPORT AND GLOBAL BOUNDS

Within each view, union the selected entity's source masks and lift every finite positive-depth pixel. Unlike compact bank construction, this stage uses neither point subsampling nor statistical outlier removal. Discard views with fewer than 128 valid points and require at least two distinct views. With N_v points in view v, set

$$\alpha_v = N_v^{-1/2}, \qquad \ell_{vj} = Q_{\alpha_v}(\{x_j\}), \qquad u_{vj} = Q_{1-\alpha_v}(\{x_j\}), \tag{43}$$

using linear-interpolated empirical quantiles for each world-coordinate axis j. This trims approximately $\sqrt{N_v}$ values per tail, not a fixed percentage. Sparse views have a larger fractional trim but must still meet the minimum-point rule. The construction is a robust support heuristic, not a distribution-free confidence interval.

For each axis, let $\ell_{(2)j}$ be the second-smallest lower endpoint and $u_{(V-1)j}$ the second-largest upper endpoint across valid views. Propose

$$L_j^{\mathrm{glob}} = \min(L_{0j}, \ell_{(2)j}), \qquad U_j^{\mathrm{glob}} = \max(U_{0j}, u_{(V-1)j}). \tag{44}$$

An outward face location therefore has support from at least two views; it is not a majority vote, and the supporting pair can differ by face. Quantile trimming does not itself expand a box: expansion comes from revisiting denser, per-view source support that the compact candidate construction may have omitted. The update can move the center as well as dimensions and is not uniform scaling.

If the global result passes the safety checks below, it is returned, including when the original box already encloses the supported intervals and no face moves. Every other global-stage failure status triggers the fallback from B_0, not from a rejected enlarged box.

A.6.2 Depth-Gated, Single-Component Source Selection

Depth gating. Transform the eight corners of B_0 to camera coordinates. Reject a view with nonfinite corners, or with no corner depth above 10^{-6} m. Over the positive-depth corners, obtain $z_{\min}, z_{\max}$ and set

$$\delta = \max\{0.05,\ 0.25\max(z_{\max} - z_{\min}, 10^{-6})\}. \tag{45}$$

For each original source mask, retain only its pixels with finite positive depth in $[z_{\min} - \delta, z_{\max} + \delta]$. The original center must also have positive camera depth and a valid finite image projection. This removes depth-incompatible observations without spatially cropping the mask to the projected box.

Component choice. Find eight-connected image components after depth gating. For component C, let $(u_c, v_c), z_c$ be the projected original box center and its camera depth. Define

$$s(C) = \sqrt{\left(\frac{|\operatorname{med} D_C - z_c|}{r}\right)^2 + \left(\frac{\|(\operatorname{med} u_C, \operatorname{med} v_C) - (u_c, v_c)\|_2}{h}\right)^2} + |C|^{-1/2}, \tag{46}$$

where $r = \max\{(z_{\max} - z_{\min} + 2\delta)/2, \delta\}$ and $h = \max(1, \sqrt{w_M^2 + h_M^2})$. Here w_M, h_M are the inclusive pixel spans of the full resized source mask *before* gating. Retain the minimum-score component, breaking ties toward more pixels. It must yield at least 128 points and satisfy

$$\rho = \sqrt{\frac{1}{3}\sum_{j=1}^{3}\left(\frac{\operatorname{med} x_j - c_{0j}}{\max(d_{0j}/2, 0.05)}\right)^2} \le 1.5. \tag{47}$$

If several source masks in one view yield valid hypotheses, choose lexicographically by $(s(C), \rho, -N)$. Only one hypothesis per view can vote. Components are 2D connected regions after gating, not independently segmented 3D objects. A mask can become disconnected because of occlusion, invalid depths, depth gating, or segmentation errors; disconnectedness alone does not prove multiple object identities.

A.7 Independent Face Consensus and Safety

Recompute Eq. equation 43 on the selected components. For each lower or upper face, the outward proposals are

$$e_{vj}^{-} = L_{0j} - \ell_{vj}, \qquad e_{vj}^{+} = u_{vj} - U_{0j}. \tag{48}$$

Retain finite proposals of at least 0.01 m, sort them, and inspect adjacent pairs. Choose the pair with the smallest spread, breaking ties toward larger minimum displacement. It is accepted only if its spread is at most $\max(0.05, 0.35d_{0j})$; move the face by the *smaller* displacement of that pair. A face without an accepted pair remains unchanged. This tightest-pair rule differs from the global second-order-statistic rule.

Both stages reject nonfinite boxes, dimensions above 5 m, or volume growth above 4 times the original volume; original dimensions must be strictly positive. Safety violations reject updates rather than clamp them. If the fallback's combined box fails, all its proposed face changes are rejected and B_0 is returned.

The estimator is outward-only: it cannot repair an oversized initial box by shrinking it, and it cannot reconstruct surfaces absent from all source observations. Its benefit depends partly on how much support was removed when the initial candidate box was fitted. Comparisons against denser point fitting and weaker filtering are therefore necessary to distinguish the value of multi-view estimation from recovery of previously discarded observations.

A.8 Evaluation and Reproducibility Conventions

Dataset-specific inputs. SeqAlign3DVG uses its provided 15-view query sequences. The reported EmbodiedScan Mini run uses a frozen, seeded 50-view scene/region manifest following the benchmark's ordered sampling configuration, rather than query-specific ground-truth visibility selection. Different queries in a scene can share the same images. This reproduces the sampling rule, not an unavailable historical random draw from another method. The retained Mini evaluation has 11,776 queries across ScanNet, 3RScan, and Matterport3D after the official validity filtering; prediction failures are not an additional exclusion criterion.

EmbodiedScan distance adapter. The EmbodiedScan run additionally applies a dataset-specific language contract: ordinary positive near/close-to and far-from clauses are normalized to nearest/farthest, rank-one, same-category anchor-distance ordinals. This follows the benchmark's distance-based description construction, not a general claim about English semantics. Negations, far-end position descriptions, and explicit second/third ranks are not overwritten. Raw queries and parser responses are preserved. This rule is enabled only for the EmbodiedScan adapter and is not applied to SeqAlign3DVG. The frozen bank, scope-aware planner, Qwen model, and withhold–restore acceptance gate otherwise retain their settings.

Geometry and metrics. SCR predicts a 3D bounding box in the aligned dataset frame. For an oriented-box evaluator, its center is $(\mathbf{L} + \mathbf{U})/2$, dimensions are $\mathbf{U} - \mathbf{L}$. EmbodiedScan ground truth retains its original nine-parameter oriented boxes and rotation convention; it is not replaced by an enclosing axis-aligned box. A prediction being axis-aligned does not change the benchmark's oriented-IoU definition.

Top-1 evaluates the final selected prediction. Ranked top-10 evaluates whether any of the first ten saved predictions passes the IoU threshold, with the final winner placed first and the remaining predictions retaining the method's saved ordering. It is not ten independent VLM judgments or an oracle sort by ground truth. Bank oracle instead evaluates the best IoU over the specified bank. Raw-bank oracle, displayed-entity oracle, and a bank with only its selected prediction refined are different diagnostics and must be named separately. Empty banks and unrecoverable inference failures remain in the denominator.

Execution records and efficiency. Predictions are keyed by row index and a stable record identifier rather than assuming that sample names are unique. Resumed runs reuse completed records and retain failure diagnostics. Efficiency accounting distinguishes query parsing, target/reference bank construction, primary VLM selection, both Compete stages, and CPU refinement. Report the model precision, number and type of GPUs, actual call and token counts, cache reuse, and whether time is summed per-query latency or parallel throughput. Model loading, retries, and data preparation should be stated explicitly when included or excluded. Neither offline geometry replay nor cached Compete replay measures cold end-to-end inference speed.

Scope of the configuration. Unless stated otherwise, the detailed parser and inference settings above describe the Qwen3.6-27B configuration used on SeqAlign3DVG and EmbodiedScan. ScanRefer uses Qwen3.5-9B for text-only query interpretation and image-conditioned selection and verification, with up to 48 pose-sampled RGB-D views per scan, as specified in Section 4.1. The implementation does not use scene meshes, ground-truth object proposals, or target boxes to generate evidence or choose an object.