ARIADNE: DISAMBIGUATION-GUIDED ACTIVE PERCEPTION FOR ZERO-SHOT 3D VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot 3D visual grounding aims to locate an object in a 3D scene from a natural-language expression without task-specific training. Object-centric local views preserve photographic color and texture detail from captured RGB images, but may omit the anchor objects needed to judge spatial relations. Rendered global perspective views provide wider context, but small object scale, occlusion, and reconstruction artifacts can obscure the details needed for **candidate disambiguation**. We introduce ARIADNE, a disambiguation-guided active perception framework that adapts its observations to the ambiguity remaining among candidates without fine-tuning or test-time training. We first decompose the expression into **visual claims** about target attributes and relations to anchor objects and track which claims remain unresolved for each candidate. Our **Relation-Centric Observation Construction** (RCOC) provides complementary captured, stitched, and rendered views for comparing candidates and anchor objects in shared visual context. When several views are available for a comparison, **Online Evidence Transport** (OET) jointly matches the candidates' evidence needs to these views and selects the next observation. OET balances visibility and visual quality within a transport objective, with graph and history regularization coordinating view selection across candidates and successive comparisons. Together, these components form an adaptive observation and decision loop, with each new view guiding the next comparison toward the distinctions still needed to identify the target. Experiments on ScanRefer and Nr3D show gains of **3.5** percentage points in overall Acc@0.25 and **4.2** percentage points in overall accuracy, respectively, over the current state-of-the-art zero-shot baselines.

1 INTRODUCTION

3D visual grounding (3DVG) asks a model to identify and localize the target object in a 3D scene that is referred to by a natural-language description. It could connect language to physical objects, supporting embodied AI and indoor scene understanding applications. Supervised methods learn this mapping from annotated 3D grounding data (Zhao et al., 2021; Jain et al., 2022; Chen et al., 2022; Zhu et al., 2023). To reduce dependence on costly 3D grounding annotations, *zero-shot 3DVG* composes pretrained perception and language models at inference time, without training a grounding model on task-specific annotations (Yang et al., 2024; Yuan et al., 2024; Fang et al., 2024; Mi et al., 2025; Xu et al., 2025; Li et al., 2025). Image-based approaches present scene views to a pretrained 2D multimodal large language model (MLLM), which compares candidate objects to select the target (Xu et al., 2025; Li et al., 2025). How candidate objects are presented to the MLLM therefore determines **whether the relations needed to distinguish them are visible**.

Recent work has improved how candidate objects are presented to the MLLM. SeeGround renders an elevated global perspective view facing an anchor object (Li et al., 2025), whereas SeqVLM assembles object-centric local views into multi-view sequences and compares candidates in batches (Lin et al., 2025). These object-centric local views use directly captured RGB images, preserving photographic color and texture detail. SPAZER combines global perspective views with object-centric local views of each candidate (Jin et al., 2025). View-on-Graph explores the scene iteratively by traversing a pre-built 3D scene graph (Liu et al., 2026). However, **clear views of individual objects do not ensure that the relation distinguishing them is clearly visible**.

For example, consider a room with several similar chairs and the description "the red chair between the two tables and facing the whiteboard." As illustrated in Figure 1 (left), object-centric local views

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.

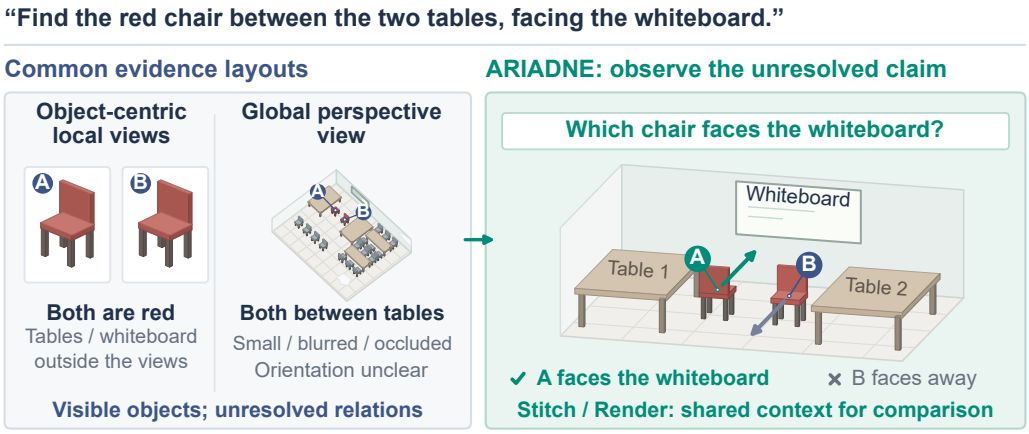


Figure 1: Disambiguation-guided active perception. Object-centric local views and global perspective views can leave the relations distinguishing similar candidates unclear. ARIADNE constructs shared visual context around the unresolved claim to compare candidates against the same anchor objects.

show candidate appearance but omit relevant anchor objects, while global perspective views show the candidates' positions but leave their orientation unclear. In rendered global perspective views, small object scale, occlusion, and reconstruction artifacts can obscure the appearance and spatial relations needed to distinguish candidates. Iterative exploration could provide additional views while still possibly leaving these relations unresolved. The central challenge here is **candidate disambiguation**: distinguishing among objects that remain plausible under the available evidence. Ideally, the next observation should address **the claim that remains unresolved for each candidate**, showing the candidates and anchor objects in the shared visual context needed to judge that claim.

Building on these observations, we propose ARIADNE[1], a disambiguation-guided active perception framework for zero-shot 3DVG, without fine-tuning or test-time training. We use an MLLM to decompose the referring expression into a fixed set of **visual claims** describing the target's attributes and relations to anchor objects. Across successive observations, ARIADNE records a **judgment** for each candidate on each claim: supported, contradicted, or unresolved. We use these judgments to guide the MLLM in formulating a **comparison request** that identifies the next claim to inspect and the shared visual context needed to judge it. The resulting observations in turn are sent to the MLLM to refine these judgments, forming an **adaptive observation and decision loop** that focuses on the remaining ambiguity among candidates. In the same "chair" example, several candidates may satisfy the color and position claims, leaving their orientation toward the whiteboard unresolved. As illustrated in Figure 1 (right), ARIADNE brings the remaining chairs and anchor objects into shared visual context to compare their orientation and distinguish the target.

To obtain the visual context needed for each comparison, we introduce **Relation-Centric Observation Construction** (RCOC), providing three types of observation tool for ARIADNE. Object-centric local views preserve appearance details. Stitching posed RGB-D views combines captured images into wider, continuous views, while rendering from reconstructed 3D mesh offers new perspectives on relations that remain unclear in captured or stitched views. These complementary sources let us bring candidates and anchor objects into compact and shared visual context suited to the current comparison.

These sources can provide several views for the same comparison, but a view that clearly shows one candidate may hide another or its anchor objects. We introduce **Online Evidence Transport** (OET) to select the next view by considering the evidence needs of competing candidates together. OET formulates this selection as a transport problem, matching each candidate's need for evidence on the selected claim to the available views according to visibility and visual quality. We further introduce graph and history regularization to the transport objective to coordinate view selection across candidates and successive observations. Graph regularization favors shared views for candidates that need be observed together, while history regularization lets repeated comparisons build on earlier view preferences as the candidate set evolves. Together, these components form an adaptive observation and decision loop, with each new view guiding the next comparison toward the distinctions still needed to identify the target. Overall, our contributions are summarized as follows:

[1]Like Ariadne's thread through the labyrinth, unresolved claims guide our observations.

- **Guiding observation toward candidate disambiguation.** Our **Disambiguation-Guided Active Perception** uses candidate–claim judgments to select the next visual claim, allowing unresolved candidates to receive targeted evidence before final selection.
- **Constructing views for relational comparison.** RCOC organizes captured, stitched, and rendered views around the current comparison, bringing candidates and anchor objects into shared visual context so that their relations can be assessed directly.
- **Coordinating view selection across candidates.** OET jointly selects views through a transport objective with graph and history regularization, so that evidence acquisition considers competing candidates together and carries relevant viewing preferences into later comparisons.
- Extensive experiments further demonstrate the superior performance of our method, surpassing current state-of-the-art zero-shot methods by **3.5%** on ScanRefer and **4.2%** on Nr3D.

2 Related Work

2.1 Supervised 3DVG

Supervised 3DVG learns to associate referring expressions with scene objects from annotated datasets such as ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020). Proposal-based methods learn object–language matching and spatial relations (Zhao et al., 2021; Chen et al., 2022; Chang et al., 2024; Huang et al., 2025), while language-conditioned detectors predict target boxes using both language and object-proposal features (Jain et al., 2022). Vision-language pretraining and instruction-tuning of 3D vision-language models (VLMs) and 3D MLLMs further learn appearance and semantic representations from paired 3D–language data (Zhu et al., 2023; Jia et al., 2025; Huang et al., 2024; Zheng et al., 2025; Qi et al., 2026; Jung et al., 2026). Their dependence on 3D–language supervision motivates zero-shot grounding, which uses pretrained language and multimodal models to reason over scene information without task-specific training. ARIADNE addresses this setting by iteratively acquiring the visual evidence needed for **candidate disambiguation** at inference time.

2.2 Zero-Shot 3DVG

Existing zero-shot approaches follow two main directions: language-based reasoning over structured scene representations and image-based reasoning over scene views. Language-based methods reason over object descriptions and geometric relations (Yang et al., 2024; Fang et al., 2024), executable spatial programs (Yuan et al., 2024; Mi et al., 2025), or constraints (Yuan et al., 2025). These representations support structured reasoning but may omit the visual details needed to distinguish similar candidates. ARIADNE records which **visual claims** remain unresolved for each candidate and directs new observations toward the evidence needed to judge them.

Image-based methods expose scene appearance to underlying MLLMs through captured images or rendered views. SeeGround renders a global perspective view oriented toward an anchor object (Li et al., 2025); SeqVLM assembles object-centric local views from captured images (Lin et al., 2025); and SPAZER combines global perspective views with object-centric local views (Jin et al., 2025). Object-centric local views may omit required anchor objects, while rendered global perspective views can obscure candidate details through small object scale, occlusion, and reconstruction artifacts. Our **Relation-Centric Observation Construction** organizes captured, stitched, and rendered views around the candidates and anchor objects needed for the current comparison, providing shared visual context for judging their relations.

VLM-Grounder selects target images and objects from captured views, using grounding feedback to revise its selections (Xu et al., 2025). More recently, View-on-Graph guides iterative exploration through a pre-built 3D scene graph of views, objects, and spatial relations (Liu et al., 2026). However, both methods present existing views as separate images, limiting direct comparison of candidates and their spatial relations to shared anchor objects. ARIADNE instead uses candidate–claim judgments to identify the unresolved comparison and brings candidates and anchor objects into shared visual context through stitching or rendering. **Online Evidence Transport** (OET) then selects the next view jointly for competing candidates, with graph regularization encouraging views that support their shared comparison.

3 Method

3.1 Method Overview

3D visual grounding (3DVG) localizes the object described by a natural-language referring expression in a 3D scene. ARIADNE addresses this task in the zero-shot setting, without task-specific training.

Formally, given a reconstructed 3D scene $\mathcal{S}$, its source captured RGB-D images $\mathcal{F}$, a referring expression x, and N object proposals $\mathcal{O} = \{o_i\}_{i=1}^{N}$, the goal is to select the target $o_{i^*} \in \mathcal{O}$. The proposals are ground-truth instances or detected objects, depending on the benchmark protocol.

As illustrated in Figure 2, ARIADNE distinguishes candidate objects (plausible targets) through an adaptive observation and decision loop. A pretrained 2D multimodal large language model (MLLM) first decomposes the referring expression x into **visual claims**: attributes of the target or relations to anchor objects (scene objects used as references). To identify missing evidence, **Disambiguation-Guided Active Perception** assesses whether candidates satisfy each claim and specifies the next comparison. The required visual context guides **Relation-Centric Observation Construction** (RCOC) in proposing captured (LOCAL), stitched (STITCH), and rendered (RENDER) views. Because views can favor different candidates, **Online Evidence Transport** (OET) selects a view by considering candidates together and reusing relevant earlier viewing preferences. **MLLM verification** evaluates the selected images to guide the next comparison. See Algorithm 1 in the supplementary material for the complete procedure.

3.2 DISAMBIGUATION-GUIDED ACTIVE PERCEPTION

To determine what evidence is still needed to distinguish candidates, we first decompose the referring expression x into K fixed visual claims, $\mathcal{K} = \{c_k\}_{k=1}^{K}$ (Section A.2). For example, "the red chair between the two tables and facing the whiteboard" gives c_1: "The target is a red chair," c_2: "The target is between the two tables," and c_3: "The target faces the whiteboard."

We compare candidates over multiple rounds, maintaining a **judgment** for each candidate on each visual claim: supported, contradicted, or unresolved, according to the available images. Missing context or ambiguous identity leaves the judgment unresolved. We index these rounds by t and write $J_t(i, k)$ for the latest judgment of candidate i on visual claim c_k.

In each round t, the MLLM uses these judgments to formulate a **comparison request** specifying which visual claim $c_{k_t} \in \mathcal{K}$ to assess and which anchor objects or other context to show with the candidates. For example, checking whether each of two candidate chairs faces a whiteboard calls for a view of both chairs and that whiteboard. For the current comparison, $\mathcal{I}_t \subseteq \{1, \ldots, N\}$ contains the indices of the candidates being compared. We pair each candidate with the selected visual claim:

$$\mathcal{E}_t = \{(i, k_t) : i \in \mathcal{I}_t\}. \tag{1}$$

For each pair $e = (i, k_t)$, we assess whether candidate i satisfies the selected visual claim.

To obtain the requested visual evidence, **RCOC** proposes views, and **OET** selects one for the current comparison. Once the selected view is obtained, **MLLM verification** assesses its images and updates the corresponding judgments. Clear contradictions exclude candidates from later iterations of the loop; an unresolved judgment alone leaves a candidate under consideration. These updated judgments guide the next round's comparison request and the eventual target selection.

3.3 RELATION-CENTRIC OBSERVATION CONSTRUCTION

Relation-Centric Observation Construction (RCOC) seeks shared visual context for the current candidates (indexed by $\mathcal{I}_t$) together with the requested context objects. To reveal relations that individual captured images may not show clearly, RCOC proposes new views through stitching, which expands the visible context, and rendering, which changes the viewpoint. We use three complementary tools to obtain these views: LOCAL, STITCH, and RENDER.

LOCAL reuses captured frames from $\mathcal{F}$, selected by their coverage of the requested objects, and marks these objects with tight boxes and consistent IDs following prior visual prompting practice (Jin et al., 2025; Li et al., 2025). STITCH uses depth and camera poses to combine overlapping captured frames into a continuous panorama. The requested candidates and anchor objects remain marked with tight boxes in the panorama, using the same object IDs as in the local views. This brings these objects into a wider shared view while retaining photographic appearance detail.

RENDER further samples camera positions around the requested objects and aims each camera toward their center. Each viewpoint is assessed by how much of each object is visible, how large it appears in the image, and whether candidates and anchor objects can be shown together. Changing the viewpoint can reveal spatial arrangements obscured by occlusion or unsuitable viewing angles in captured or stitched images. Section A.5 describes the full construction procedure.

Together, these tools define a finite set $\mathcal{A}_t$ of **2D view proposals**. Each $a \in \mathcal{A}_t$ specifies captured images, a stitched panorama, or a rendered view from a particular camera position. A single tool can contribute multiple alternatives, and OET selects among these views for the next observation.

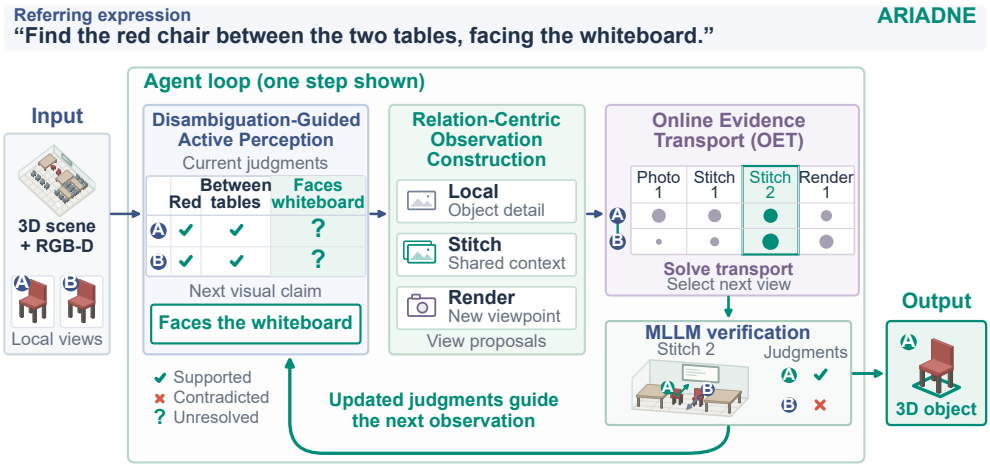


Figure 2: ARIADNE overview. The four components form an agentic loop, with updated judgments guiding each successive comparison. **Disambiguation-Guided Active Perception** selects the next claim to verify. **Relation-Centric Observation Construction** proposes views, and **Online Evidence Transport** solves a transport problem to jointly select a view for competing candidates. **MLLM verification** updates the candidate–claim judgments from the acquired images. Circle areas indicate allocation weights.

To evaluate these view proposals for the next observation, we estimate how clearly each can show the candidates being compared and the context objects specified in the MLLM's comparison request. Let $\mathcal{R}_t \subseteq \{1, \ldots, N\}$ index these context objects. For a candidate–claim pair $e = (i, k_t)$, the set $\mathcal{M}_t(e) = \{i\} \cup \mathcal{R}_t$ contains candidate i and its requested context objects. The geometric observability score $v_t(j, a) \in [0, 1]$ estimates how much of object j is visible in view proposal a and how large it appears in the image. The fidelity score $f_t(a) \in [0, 1]$ estimates how much appearance detail a captured, stitched, or rendered image retains. We combine these estimates into the observation-quality score $\widehat{q}_t(e, a)$ and its corresponding cost $C_t(e, a)$:

$$\widehat{q}_t(e, a) = f_t(a) \min_{j \in \mathcal{M}_t(e)} v_t(j, a), \qquad C_t(e, a) = 1 - \widehat{q}_t(e, a). \tag{2}$$

These scores estimate how well each view can reveal the evidence needed to assess a candidate–claim pair. Online Evidence Transport uses the corresponding costs to select the next view by considering competing candidates together.

3.4 Online Evidence Transport

Competing candidates need comparable evidence. Selecting a clear view independently for each candidate can separate them from the common context needed to judge a relation. Online Evidence Transport (OET) allocates observation weight from candidate–claim pairs to view proposals, jointly selecting the next view for the candidates. Inspired by entropic transport with relaxed mass (Chizat et al., 2018) and online transport optimization (Ballu & Berthet, 2023), we update the allocation as new observations change the evidence needed to distinguish candidates.

An allocation has one row per candidate–claim pair in Equation (1) and one column per view proposal; an entry gives the weight allocated from that pair to that proposal. We write this allocation as $P \in \mathbb{R}_{\geq 0}^{m_t \times n_t}$, with $m_t = |\mathcal{E}_t|$ candidate–claim pairs and $n_t = |\mathcal{A}_t|$ view proposals. To give each pair the same reference total weight in the comparison, we set $r_t(e) = 1$ for every $e \in \mathcal{E}_t$. These values form the vector r_t, and the optimization encourages the total weight for each candidate–claim pair (its row sum) to remain near its reference value. To treat view proposals equally in the entropy term, we use a uniform reference matrix U_t, with $U_t(e, a) = 1/n_t$. For each candidate–claim pair, this reference divides one unit equally among the n_t view proposals; the optimized weights then reflect current observation costs and the regularizers below.

When a candidate–claim pair is assessed again, some previously considered views may still be useful. To carry earlier viewing preferences into the new decision, **history regularization** compares the new allocation with a history reference $B_t \in \mathbb{R}_{>0}^{m_t \times n_t}$. The entry $B_t(e, a)$ gives the reference weight for showing view a when assessing pair e. We form B_t from the most recent allocation for each pair: weights for views still available are normalized separately for each candidate–claim pair and mixed with U_t so that new views also receive positive reference weight. For a pair without matching

history, we use its uniform reference weights from U_t. Current observation costs can still move the new allocation away from this reference.

When competing candidates can be observed together, coordinating their view allocations helps compare them in shared visual context. **Graph regularization** therefore links candidate–claim pairs more strongly when a view can provide clear evidence for both. We record these link strengths in a symmetric nonnegative matrix $W_t \in \mathbb{R}_{\geq 0}^{m_t \times m_t}$ with zero diagonal. For distinct candidates on the selected claim, $W_t(e, e')$ measures the best shared-view quality for pairs e and e', estimated from the observation-quality scores $\widehat{q}_t$ in Equation (2). A larger weight imposes a stronger penalty on differences between the two pairs' view allocations. To express this pairwise penalty compactly, we use the graph Laplacian $L_t = D_t - W_t$, where the diagonal matrix D_t collects each candidate–claim pair's total link strength, $D_t(e, e) = \sum_{e'} W_t(e, e')$. Section A.6 gives the construction of W_t and B_t.

With the observation costs C_t from Equation (2) and the references above fixed, we optimize the current allocation P:

$$\begin{aligned} P_t = \arg\min_{P \geq 0} \quad & \langle C_t, P \rangle + \epsilon \, \mathrm{KL}(P \| U_t) \\ & + \lambda \, \mathrm{KL}(P\mathbf{1} \| r_t) + \eta \, \mathrm{KL}(P \| B_t) \\ & + \gamma \, \mathrm{tr}(P^\top L_t P), \end{aligned} \tag{3}$$

where $\epsilon, \lambda > 0$ weight the entropy and relaxed-demand terms, while $\eta, \gamma \geq 0$ control history and graph regularization, respectively. Here $\langle C_t, P \rangle = \sum_{e,a} C_t(e, a) P(e, a)$, and $\mathbf{1} \in \mathbb{R}^{n_t}$ is the all-ones vector, so $P\mathbf{1}$ collects the total weight for each candidate–claim pair (the row sums). The generalized divergence $\mathrm{KL}(X \| Y) = \sum_\ell [X_\ell \log(X_\ell / Y_\ell) - X_\ell + Y_\ell]$ compares nonnegative arrays of possibly different total mass, with $0 \log(0/Y_\ell) = 0$; ℓ ranges over entries and Y is positive. The symbols $\top$ and tr denote transpose and trace.

The cost favors views that clearly show the candidates and requested context while preserving visual detail (see Section A.7 for the numerical solution). We select the view proposal receiving the most weight across the current candidate–claim pairs (the largest column sum), $a_t^* = \arg\max_{a \in \mathcal{A}_t} \sum_{e \in \mathcal{E}_t} P_t(e, a)$.

MLLM verification. As illustrated in Figure 2, the MLLM uses the images obtained from the selected view proposal a_t^* to assess the selected visual claim c_{k_t} for each candidate $i \in \mathcal{I}_t$, in the context of the referring expression x. The resulting assessments update the judgments, $J_t(i, k_t) \to J_{t+1}(i, k_t)$. In the Stitch 2 view shown in the lower-right MLLM verification panel of Figure 2, A faces the whiteboard and B does not, supporting A and contradicting B on that claim. The updated judgments guide the next comparison request and the final selection of o_{i^*}. The complete procedure and final selection policy are given in Algorithm 1 and section A.8.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate on **ScanRefer** (Chen et al., 2020) and **Nr3D** (Achlioptas et al., 2020), both built on ScanNet (Dai et al., 2017). Following established evaluation practice, we use a fixed 250-example subset of each benchmark (Xu et al., 2025). For ScanRefer, Acc@0.25 and Acc@0.5 measure the fraction of predicted boxes whose 3D IoU with the ground-truth box is at least 0.25 or 0.5, respectively. Results are reported on *Unique* (only one object of the target class) and *Multiple* (same-class distractors) subsets. For Nr3D, we report instance selection accuracy with given object boxes, on *Easy*/*Hard* subsets with at most two/more than two same-class objects (including the target), and *View-Dependent*/*View-Independent* subsets according to viewpoint dependence. We also report overall results on both benchmarks and separate Nr3D settings with and without ground-truth object class information for fair comparison.

Implementation details. Following standard practice, we use Mask3D (Schult et al., 2023) proposals for ScanRefer and ground-truth object boxes for Nr3D. RGB-D frames are taken directly from ScanNet. Seed 2.0 Lite (ByteDance Seed Team, 2026) serves as the MLLM backend without task-specific training. Further implementation details are provided in Section B.1.

4.2 Comparison with State-of-the-Art Methods

Tables 1 and 2 present quantitative comparisons with representative supervised and zero-shot 3DVG methods on ScanRefer and Nr3D. Following recent zero-shot evaluations, we separate methods trained

Table 1: ScanRefer quantitative comparison. Results on ScanRefer (Chen et al., 2020) are reported for Unique (no same-class distractors) and Multiple (with same-class distractors) subsets, along with overall performance. Bold denotes the best results among zero-shot methods.

Method	Unique		Multiple		Overall	
	@0.25↑	**@0.5↑**	**@0.25↑**	**@0.5↑**	**@0.25↑**	**@0.5↑**
Fully-Supervised						
ScanRefer (Chen et al., 2020)	67.6	46.2	32.1	21.3	39.0	26.1
ChatScene (Huang et al., 2024)	89.6	82.5	47.8	42.9	55.5	50.2
Video-3D LLM (Zheng et al., 2025)	88.0	78.3	50.9	45.3	58.1	51.7
GPT4Scene (Qi et al., 2026)	90.3	83.7	56.4	50.9	62.6	57.0
Zero-Shot						
OpenScene (Peng et al., 2023)	20.1	13.1	11.1	4.4	13.2	6.5
LLM-Grounder (Yang et al., 2024)	—	—	—	—	17.1	5.3
ZSVG3D (Yuan et al., 2024)	63.8	58.4	27.7	24.6	36.4	32.7
VLM-Grounder (Xu et al., 2025)	66.0	29.8	48.3	33.5	51.6	32.8
SeeGround (Li et al., 2025)	75.7	68.9	34.0	30.0	44.1	39.4
CSVG (Yuan et al., 2025)	68.8	61.2	38.4	27.3	49.6	39.8
SeqVLM (Lin et al., 2025)	—	—	—	—	55.6	49.6
SPAZER (Jin et al., 2025)	80.9	72.3	51.7	43.4	57.2	48.8
View-on-Graph (Liu et al., 2026)	78.6	71.5	34.0	30.4	44.8	40.3
Z3D (Drozdov et al., 2026)	—	—	—	—	58.9	52.7
ARIADNE (Ours)	**81.8**	**77.3**	**55.4**	**47.8**	**62.4**	**55.6**

Table 2: Nr3D quantitative comparison. Results on Nr3D (Achlioptas et al., 2020) are reported for Easy, Hard, and Dep./Indep. subsets, along with overall performance. Dep./Indep. indicate whether grounding depends on a specific viewpoint. GT class indicates whether ground-truth category labels are available for all scene objects. Bold denotes the best results among zero-shot methods.

Method	**GT** class	Easy	Hard	Dep.	Indep.	Overall
Fully-Supervised						
ReferIt3DNet (Achlioptas et al., 2020)	✗	43.6	27.9	32.5	37.1	35.6
SceneVerse (Jia et al., 2025)	✗	72.5	57.8	56.9	67.9	64.9
ViewSRD (Huang et al., 2025)	✗	75.3	64.8	68.6	70.6	69.9
PanoGrounder (Jung et al., 2026)	✗	82.2	67.2	70.5	76.3	74.6
Zero-Shot						
ZSVG3D (Yuan et al., 2024)	✗	46.5	31.7	36.8	40.0	39.0
SeeGround (Li et al., 2025)	✗	54.5	38.3	42.3	48.2	46.1
VLM-Grounder (Xu et al., 2025)	✗	55.2	39.5	45.8	49.4	48.0
EaSe (Mi et al., 2025)	✗	—	—	—	—	52.9
SeqVLM (Lin et al., 2025)	✗	58.1	47.4	51.0	54.5	53.2
SPAZER (Jin et al., 2025)	✗	68.0	58.8	59.9	66.2	63.8
View-on-Graph (Liu et al., 2026)	✗	58.9	37.2	39.4	52.1	47.6
Z3D (Drozdov et al., 2026)	✗	62.6	47.5	50.7	57.1	54.8
ARIADNE (Ours)	✗	**75.0**	**59.7**	**64.6**	**70.1**	**68.0**
CSVG (Yuan et al., 2025)	✓	67.1	51.3	53.0	62.5	59.2
Transcrib3D (Fang et al., 2024)	✓	79.7	60.3	60.1	75.4	70.2
EaSe (Mi et al., 2025)	✓	—	—	—	—	67.8
SPAZER (Jin et al., 2025)	✓	76.5	69.3	65.6	77.9	73.2
ARIADNE (Ours)	✓	**91.2**	**74.6**	**75.0**	**88.9**	**83.6**

with 3D grounding annotations from training-free methods that rely on pretrained vision-language models and off-the-shelf 3D perception modules.

On ScanRefer, ARIADNE achieves **62.4%** overall Acc@0.25 and **55.6%** overall Acc@0.5, exceeding Z3D, the strongest compared zero-shot baseline, by **3.5%** and **2.9%**, respectively. On the Multiple split, which requires distinguishing same-class distractors, ARIADNE reaches 55.4%/47.8%, improving over SPAZER by 3.7%/4.4%. On Nr3D, ARIADNE achieves **68.0%** overall accuracy without ground-truth object class information and **83.6%** with it, exceeding SPAZER by **4.2%** and **10.4%** under the respective settings. These results demonstrate the effectiveness of disambiguation-guided active perception across both benchmarks.

4.3 Ablation Studies

All studies in this subsection use ScanRefer. Grounding comparisons report overall Acc@0.25 and Acc@0.5 (%), in that order. Component ablations examine the active perception loop, observation sources, and Online Evidence Transport (OET) with the MLLM and perception inputs held fixed.

Table 3: Ablation of observation strategies. New views are selected from proposals containing the candidates requested for comparison. Random Image Selection samples this proposal set.

Policy	Acc@0.25	Acc@0.5
One-Pass Image Observation	54.8	51.2
Random Image Selection	59.2	53.2
Greedy by Object Quality	60.8	54.4
OET	**62.4**	**55.6**

Table 4: Ablation of observation sources. We ablate the additional views generated by Stitch and Render.

Available views	Acc@0.25	Acc@0.5
Local	56.8	52.8
Local + Stitch	57.6	52.0
Local + Render	59.6	52.8
All sources	**62.4**	**55.6**

Disambiguation-Guided Active Perception. To assess the benefit of iterative observation, we compare ARIADNE with **One-Pass Image Observation** in Table 3. This baseline selects the target after a single round of image observation. ARIADNE achieves 62.4%/55.6%, improving on the one-pass baseline's 54.8%/51.2% by **7.6%/4.4%**. These results demonstrate the benefit of iteratively acquiring visual evidence to resolve remaining candidate ambiguity.

Relation-Centric Observation Construction. To examine the contributions of captured, stitched, and rendered views, we compare their combinations in Table 4. Local alone achieves 56.8%/52.8%, while all three sources together reach 62.4%/55.6%, improving accuracy by **5.6%/2.8%**. The full method also exceeds the strongest partial combination, Local+Render (59.6%/52.8%), by 2.8% at both thresholds. These results validate the benefit of making photographic detail, wider context, and new viewpoints jointly available for relational comparisons.

Table 5: Ablation of OET regularization. History and graph terms are controlled independently.

History	Graph	Acc@0.25	Acc@0.5
✗	✗	57.6	50.8
✓	✗	59.6	53.2
✗	✓	60.4	53.6
✓	✓	**62.4**	**55.6**

Table 6: Framework comparison with the same MLLM. All methods use Seed 2.0 Lite.

Method	Acc@0.25	Acc@0.5
SeeGround	53.2	47.6
SeqVLM	57.2	52.0
SPAZER	60.8	52.8
ARIADNE	**62.4**	**55.6**

Online Evidence Transport. To assess how view selection affects grounding, we compare OET with two controls in Table 3. **Random Image Selection** randomly selects a view from those showing the objects specified in the current comparison request. **Greedy by Object Quality** ranks views solely by the visible fractions of these objects. Their accuracies are 59.2%/53.2% and 60.8%/54.4%, respectively. OET reaches 62.4%/55.6%, improving over these controls by **3.2%/2.4%** and **1.6%/1.2%**. These results demonstrate the benefit of OET's joint view selection over random or visibility-only selection.

History and graph regularization. To assess the contributions of history and graph regularization within OET, we compare their individual and combined effects in Table 5. The history ablation replaces the historical reference with a uniform reference while keeping the regularization weights fixed. Without historical information or graph regularization, accuracy is 57.6%/50.8%. History alone raises it to 59.6%/53.2%, gains of **2.0%/2.4%**; graph regularization alone reaches 60.4%/53.6%, gains of **2.8%/2.8%**. Combining both achieves 62.4%/55.6%, improving over this control by **4.8%** at both thresholds and over graph regularization alone by another 2.0%. These results empirically validate the complementary contributions of history and graph regularization, with their combination yielding the most accurate grounding.

Framework comparison with the same MLLM. Reproducing every framework with its original MLLM can be difficult because access to specific model versions (e.g., those of GPT-4 and GPT-4o; OpenAI, 2023; 2024) changes over time. To distinguish framework performance from MLLM choice, Table 6 compares ARIADNE with three strong open-source zero-shot 3DVG frameworks (SeeGround, SeqVLM, and SPAZER) using the same Seed 2.0 Lite model. Each framework retains its own observation and reasoning procedure. ARIADNE achieves 62.4%/55.6%, exceeding SPAZER's 60.8%/52.8% by **1.6%/2.8%**. The gains over SeqVLM and SeeGround are 5.2%/3.6% and 9.2%/8.0%, respectively. These results demonstrate that ARIADNE's advantage persists with a common MLLM, validating the contribution of its grounding framework beyond backbone choice.

4.4 Qualitative Examples

As shown in Figure 3, the acquired views expose the shared context needed for relational comparisons. Stitch shows an open doorway together with nearby trash cans and the floor beyond, and brings

two backpack candidates and their reference window into one image. RENDER shows alternative doors relative to trash cans and folding tables, and reveals the arrangement of foreground and rear monitor rows. These examples illustrate how ARIADNE constructs and selects views tailored to the current comparison request, extending the available visual context to bring candidates and anchor objects into a shared view for direct comparison.

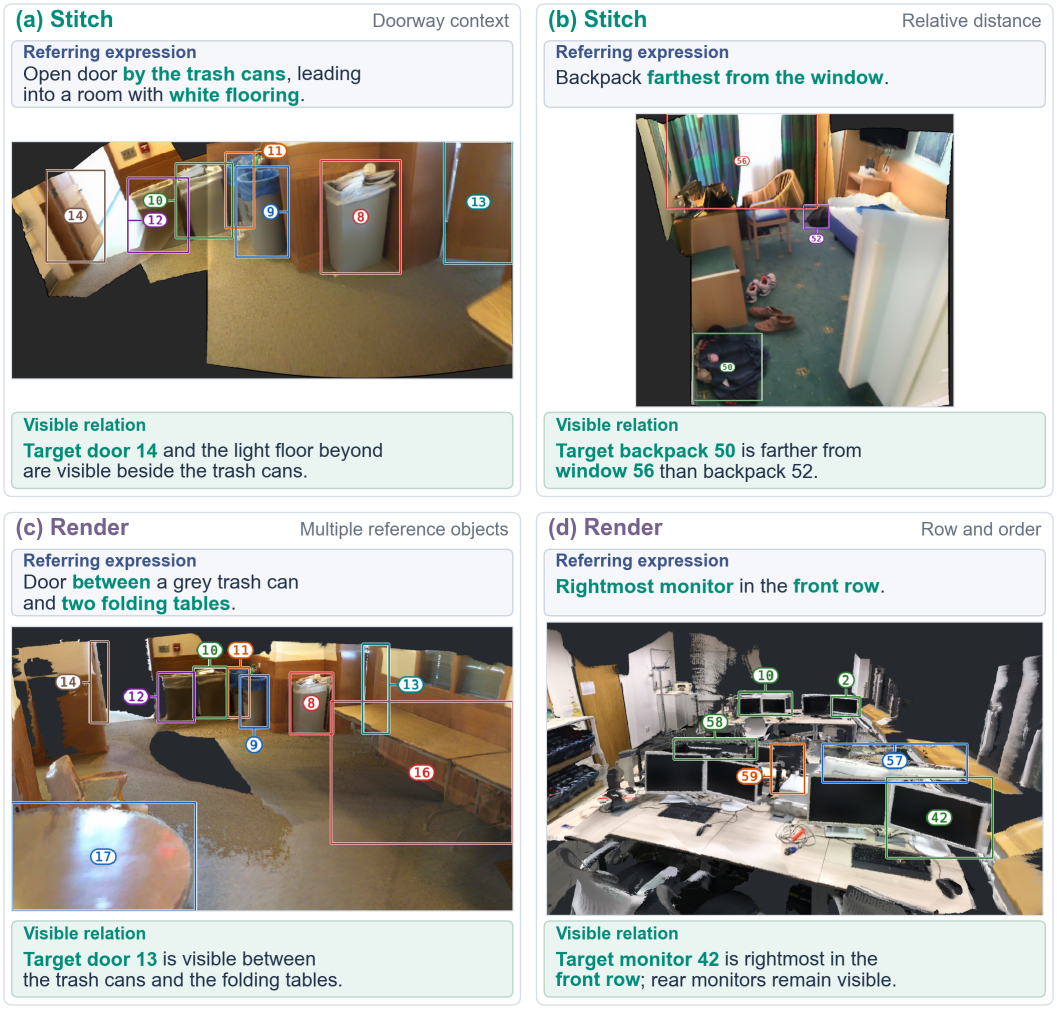


Figure 3: Acquired views for relational comparison. STITCH extends photographic context in (a–b), while RENDER provides new viewpoints in (c–d). Each panel pairs a shortened referring expression with the visible relation identifying the target. Original object IDs are retained; empty render margins are cropped for display. Examples (a–c) are from Nr3D, and (d) is from ScanRefer.

5 CONCLUSION

We presented ARIADNE, a disambiguation-guided active perception framework for zero-shot 3DVG. It tracks unresolved visual claims for each candidate and uses them to determine what evidence the next comparison requires. Relation-Centric Observation Construction (RCOC) brings candidates and anchor objects into shared visual context through captured, stitched, and rendered views. Online Evidence Transport (OET) selects among these views by considering competing candidates together, with graph and history regularization coordinating view selection across candidates and successive comparisons. Each new observation refines the candidate–claim judgments, which guide the next comparison in an adaptive observation and decision loop. Experiments on ScanRefer and Nr3D demonstrate improved grounding accuracy without task-specific training. ARIADNE connects candidate reasoning with targeted perception, making the remaining ambiguity guide what the model observes next.

AI USE STATEMENT

We used LLMs, including ChatGPT and Claude, to rephrase and polish the manuscript, draft and cross-check appendix descriptions of prompts and implementation details, improve the presentation of figures, polish mathematical derivations, assist with coding, double-check references, and suggest additional experiments and comparisons to strengthen the evidence for our claims. All LLM-generated code was reviewed, edited, and integrated by the authors; the LLMs did not design algorithms or produce experimental results.

REPRODUCIBILITY STATEMENT

We will release our code, configuration files, and evaluation scripts upon publication. The main text and appendix provide key implementation details and protocols to support reproducibility.

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*, pp. 422–440. Springer, 2020.

Marin Ballu and Quentin Berthet. Mirror sinkhorn: Fast online optimization on transport polytopes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1595–1613, 2023. URL https://proceedings.mlr.press/v202/ballu23a.html.

ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity. Model card, ByteDance, 2026. URL https://github.com/ByteDance-Seed/Seed2.0/blob/master/Seed2.0%20Model%20Card.pdf.

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*, 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 *Computer Vision – ECCV 2020*, pp. 202–221. Springer, 2020.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3D object grounding. In *Advances in Neural Information Processing Systems*, 2022.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

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. 5828–5839, 2017.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Nikita Drozdov, Andrey Lemeshko, Nikita Gavrilov, Anton Konushin, Danila Rukhovich, and Maksim Kolodiazhnyi. Z3D: Zero-shot 3D visual grounding from images. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 147–154, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.13. URL https://aclanthology.org/2026.acl-short.13/.

Jiading Fang, Xiangshan Tan, Shengjie Lin, Igor Vasiljevic, Vitor Guizilini, Hongyuan Mei, Rares Ambrus, Gregory Shakhnarovich, and Matthew R. Walter. Transcrib3D: 3D referring expression resolution through large language models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 9737–9744, 2024. doi: 10.1109/IROS58592.2024.10802793.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3D scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*, 2024.

Ronggang Huang, Haoxin Yang, Yan Cai, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. ViewSRD: 3D visual grounding via structured multi-view decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9726–9736, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Huang_ViewSRD_3D_Visual_Grounding_via_Structured_Multi-View_Decomposition_ICCV_2025_paper.html.

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

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 289–310, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72673-6. doi: 10.1007/978-3-031-72673-6_16.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 165549–165576. Curran Associates, Inc., 2025. doi: 10.52202/085713-5518. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f1ea808562a3cd5a2a0ea94f3dee675a-Paper-Conference.pdf.

Seongmin Jung, Seongho Choi, Gunwoo Jeon, Minsu Cho, and Jongwoo Lim. Panogrounder: Bridging 2d and 3d with panoramic scene representations for vlm-based 3d visual grounding, 2026. URL https://arxiv.org/abs/2512.20907.

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*, MM '25, pp. 3094–3103. ACM, October 2025. doi: 10.1145/3746027.3754985. URL http://dx.doi.org/10.1145/3746027.3754985.

Yuanyuan Liu, Haiyang Mei, Dongyang Zhan, Jiayue Zhao, Dongsheng Zhou, Bo Dong, and Xin Yang. View-on-graph: Zero-shot 3D visual grounding via vision-language reasoning on scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7386–7394, 2026. doi: 10.1609/aaai.v40i9.37677. URL https://ojs.aaai.org/index.php/AAAI/article/view/37677.

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. doi: 10.18653/v1/2025.emnlp-main.191. URL https://aclanthology.org/2025.emnlp-main.191/.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

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

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, June 2023.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 157277–157302, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *IEEE International Conference on Robotics and Automation*, pp. 8216–8223, 2023. URL https://arxiv.org/abs/2210.03105.

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, 2025.

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. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 7694–7701, 2024.

Qihao Yuan, Kailai Li, and Jiaming Zhang. Solving zero-shot 3D visual grounding as constraint satisfaction problems. In *36th British Machine Vision Conference*. BMVA, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20623–20633, 2024.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2928–2937, 2021.

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 (CVPR)*, pp. 8995–9006, June 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.

A Additional Method Details

A.1 Inference Procedure

This section specifies the preprocessing, grouping, and stopping rules used to execute the observation and decision loop. A **verification round** is one iteration that selects a comparison request and applies it to the current candidates. Let $\mathcal{C}_t \subseteq \{1, \ldots, N\}$ contain the candidate indices at round t, and let $\mathcal{I}_t \subseteq \mathcal{C}_t$ denote the current candidate group. The observation history H_t records acquired images, comparison requests, and image-based judgments. New records are appended while earlier observations remain unchanged. The main-text equations describe one group and omit a separate group index.

Shared preprocessing. Before the agent loop, the category filter retains object proposals whose supplied or predicted categories are compatible with the target category in the referring expression x, forming the initial candidate set $\mathcal{C}_0$. Object proposals and category filtering are held fixed across controlled comparisons; their sources are specified in Section B.1. When multiple candidates remain, this shared preprocessing is followed by claim decomposition and assessment of the initial local images, as detailed below.

Algorithm 1 ARIADNE: Disambiguation-Guided Active Perception

Require: Scene $\mathcal{S}$, frames $\mathcal{F}$, object proposals $\mathcal{O}$, referring expression x
Ensure: One selected object proposal o_{i^*}

Category-filter to nonempty $\mathcal{C}_0$; **return** its sole object proposal if it is a singleton
Fix $\mathcal{K}$; initialize $J_0, H_0, \mathcal{C}_0$ from local images; set $t \leftarrow 0$
while $|\mathcal{C}_t| > 1$ and the verification-round budget remains **do**
    Select k_t, $\mathcal{R}_t$, and individual or joint observation from current judgments
    for each fixed candidate group $\mathcal{I}_t$ in $\mathcal{C}_t$ **do**
        Form candidate–claim pairs $\mathcal{E}_t$ and executable view proposals $\mathcal{A}_t$
        Estimate C_t and form r_t, U_t, B_t, L_t
        Solve Equation (3) for P_t
        Execute one view proposal a_t^*; judge its acquired images and append the record
    end for
    Combine the group judgments to form $J_{t+1}, \mathcal{C}_{t+1}, H_{t+1}$
    $t \leftarrow t + 1$
end while
return the sole surviving object proposal, or one object proposal selected by final visual comparison

A.2 Claim Decomposition

Claim decomposition. Before initial image assessment, the MLLM decomposes the referring expression x into a fixed list of natural-language visual claims. The common system instruction and the decomposition instruction are reproduced in Sections D and D.3. Subsequent planning and verification retain the list and the referring expression x. Reference object identities are chosen later as part of comparison requests.

Preserving meaning. The decomposition instruction asks the model to separate independent conditions while keeping qualifiers that determine a relation's subject, reference, or comparison group. For the illustrative query "the chair is red and beside the table against the wall," the three conditions can be "The target is a chair," "The chair is red," and "The chair is beside a table that is against the wall." The wall qualifier belongs to the table; transferring it into "The chair is against the wall" would change the query. Similarly, "the rightmost chair on the northern side of the table" compares chairs within that stated group, not all chairs in the room. Explicit viewpoint, ordinal, and group qualifiers are retained when present.

Several segmentations can express the same conditions. Combining category and color as "The target is a red chair" is compatible with separating them, provided that no information or scope is lost. These examples illustrate the decomposition criteria. The human evaluation in Section A.3 assesses recorded decompositions under a meaning-preservation criterion that accepts semantically equivalent segmentations.

A.3 Human Evaluation of Claim Decomposition

Sampling. We assess semantic preservation using 50 randomly sampled ScanRefer expressions and their generated claim lists, comprising 165 visual claims.

Assessment criteria. One reviewer compares each expression with its saved claim list. The meaning-preservation criterion accepts synonymous paraphrases, combined category–attribute claims, and reasonable interpretations of ambiguous wording. A decomposition is usable when it retains the main meaning and the conditions needed to judge candidate objects. The reviewer separately records target-category preservation, coverage of the core meaning, correct reference roles and relation directions, absence of added or reversed conditions, and absence of redundant claims. Reference correctness is assessed only for the 45 expressions with an explicit reference object or architectural referent, such as a wall or floor. Individual claims are also checked for semantic correctness against the original expression. Thus, the expression-level, reference-relation, and claim-level denominators are 50, 45, and 165, respectively.

Results and representative errors. Of the 50 decompositions, 49 are usable (98.0%), and 164 of 165 individual claims are semantically correct (99.4%). Target categories are retained in all 50 expressions, and reference roles and directions are preserved in 44 of the 45 applicable cases (97.8%). The only meaning-changing error reverses a spatial relation: the original expression places a purple chair in front of a wooden door, whereas one generated claim places the door in front of the chair. A separate expression repeats the same property through "on rollers" and "a rolling chair"; its meaning remains usable, but it fails the non-redundancy criterion. These results show high semantic fidelity while distinguishing a consequential relation reversal from a harmless repetition.

A.4 Synthetic Worked Example

This worked example uses synthetic visibility estimates and a numerical solution of Equation (3), followed by a hypothetical image-based judgment update. The view labels match Figure 2; round indices are omitted in the example's matrices.

Query, claims, and judgments. For "Find the red chair between the two tables, facing the whiteboard," use the fixed claims c_1: "The target is a red chair," c_2: "The chair is between the two tables," and c_3: "The chair faces the whiteboard." Suppose the current group contains object candidates A and B. Initial images support c_1 and c_2 for both; c_3 remains unresolved for both. The comparison request selects c_3, asks for a joint view of A and B with the whiteboard, and assigns the whiteboard's scene object ID as context. The allocation rows are therefore $e_A = (A, 3)$ and $e_B = (B, 3)$, with unit demand for each.

Four view proposals. Photo 1 retrieves a captured image; Stitch 1 and Stitch 2 specify different connected source-frame groups; Render 1 specifies a camera pose. Denote these four view proposals by a_1, a_2, a_3, a_4, respectively. In that column order, take the illustrative fidelity factors to be

$$f = (1.00,\ 0.90,\ 0.90,\ 0.75).$$

Let $V(e, a) = \min_{j \in \mathcal{M}_t(e)} v_t(j, a)$ be the least geometric observability among the objects needed to assess pair e in view proposal a. Choose

$$V = \begin{bmatrix} 0.85 & 0.88 & 0.85 & 0.86 \\ 0.12 & 0.40 & 0.80 & 0.86 \end{bmatrix}, \qquad \widehat{Q} = \begin{bmatrix} 0.850 & 0.792 & 0.765 & 0.645 \\ 0.120 & 0.360 & 0.720 & 0.645 \end{bmatrix}.$$

Here $\widehat{Q}(e, a) = f(a)V(e, a)$ is the matrix of estimated qualities in Equation (2); $C = 1 - \widehat{Q}$ is its cost matrix. Both candidates have nonzero estimated support in every view proposal. Photo 1 favors A, whereas Stitch 2 supplies relatively clear context for both candidates.

Allocation and selection. For this illustration, set $\epsilon = 0.15$, $\lambda = 1$, $\eta = 0.10$, and $\gamma = 0.25$. It is the first request, so every entry of the uniform and history references is $U = B = 1/4$. The affinity in Equation (6) is $W(e_A, e_B) = 0.72$, the largest quality shared by both rows in one view proposal. The graph Laplacian is therefore

$$L = \begin{bmatrix} 0.72 & -0.72 \\ -0.72 & 0.72 \end{bmatrix}.$$

Define the total allocated to view proposal a as $s_a = \sum_{e \in \{e_A, e_B\}} P(e, a)$. Solving the same objective with these inputs gives

$$P \approx \begin{bmatrix} 0.229 & 0.204 & 0.239 & 0.149 \\ 0.042 & 0.098 & 0.318 & 0.234 \end{bmatrix}, \qquad (s_{a_1}, s_{a_2}, s_{a_3}, s_{a_4}) \approx (0.27,\ 0.30,\ 0.56,\ 0.38).$$

The displayed matrix entries and totals are rounded separately from the full-precision solution. Stitch 2 has the largest column total and is selected for this candidate group. The relaxed demand term permits row sums different from one; matrix entries allocate observation weight to view proposals. The example's coefficients are independent of the final evaluation configuration.

Hypothetical acquired image and update. If the acquired Stitch 2 image establishes that A faces the whiteboard and B faces away from it, the MLLM can mark c_3 supported for A and contradicted for B. The red-chair and between-tables judgments remain supported. At the end of the round, B is removed and the sole surviving object proposal A is returned. If the image instead leaves orientation undecidable, the affected judgment remains unresolved and a later comparison request can seek further context. In either case, the acquired image and its judgments would be appended to the history.

A.5 View Proposals and Quality Estimates

Concrete view proposals. A view proposal records its tool and the parameters needed to reproduce its observation. For captured images, these parameters identify the selected source frames. A stitching view proposal additionally requires a connected source-frame group with suitable overlap. A rendering view proposal specifies the camera pose and intrinsics used to observe the reconstructed scene. Already acquired images can also be proposed for reuse. Each view proposal has a stable identity determined by its tool, scene, requested objects, and observation parameters, so repeated view proposals can be matched across allocation steps.

Requested context and object identity. The MLLM selects context indices from a catalogue of numbered scene proposals, using the referring expression x and the latest image-based judgments and reasons. A selected index is a current reference hypothesis: a later request can add another comparison object or revise the context. A wall, corner, or room region can instead be described in the observation goal without an object index. The geometric score covers the numbered objects in $\mathcal{M}_t(e)$; the complete semantic context is assessed from the acquired images. Instance annotations preserve scene object IDs across observations. The model receives the image-to-object mapping and is instructed to leave a judgment unresolved when the pixels do not establish the marked object's identity. The referring expression x supplies the intended directional and comparison qualifiers throughout verification.

Captured and stitched observations. LOCAL proposals use captured target-visible images with depth-consistent support for the requested objects. They can contain surrounding objects and scene layout, and a single proposal can package several images for individual inspection. A joint request instead prefers a captured image containing all requested participants.

STITCH combines an overlapping, connected frame group into a wider photographic context. Calibrated cameras and depth reproject the source images onto a cylindrical canvas referenced to the first source pose. Depth layers and source reliability guide source selection; exposure correction, bounded appearance-flow registration, and seam blending reduce discontinuities. Instance-support masks follow the same remapping and source ownership as the RGB pixels, producing tight annotations with stable object IDs in the panorama. Source-frame connectivity and warp-risk checks constrain the proposed groups. The panorama is an approximate photographic construction: parallax, occlusion, and registration can affect its geometry, so its fidelity prior includes warp risk and its actual pixels determine the visual judgment.

Rendered observations. RENDER searches a finite family of cameras aimed at the requested object group. The search samples horizontal positions, heights, and radii, and can re-aim recorded camera centers toward the group. Proximity to the recorded scan path and free-space checks constrain candidate camera locations. Coarse mesh ray casting estimates occlusion, while projected area and field of view evaluate framing before RGB rendering. This is an object-group-conditioned visibility search; relation interpretation remains with the MLLM. It exposes alternatives to the captured camera coverage while retaining registration to the same scene instances. New images are generated only for the selected proposal, and one proposal is selected per candidate group.

Geometric observability. Let $d_t(j, a) \in [0, 1]$ be the estimated visible fraction of object j for proposal a, and let $p_t(j, a) \in [0, 1]$ be its projected fraction of the image area. A concrete estimate is

$$v_t(j, a) = d_t(j, a) \min\left\{1, \sqrt{\frac{p_t(j, a)}{\alpha}}\right\}, \qquad \alpha > 0, \tag{4}$$

where α is the projected-area fraction at which the square-root size factor reaches one. Below this threshold, the factor is $\sqrt{p_t(j,a)/\alpha}$; at or above it, the factor is one. Captured views estimate d_t from depth-consistent instance support. Rendered proposals estimate it from coarse mesh visibility. These quantities are geometric proxies evaluated before image-based verification.

For a group of captured frames, per-object support is aggregated from the best available source view. Eligibility for a shared-context request additionally considers whether the required objects can be shown together: a local group requires a source view containing the requested participants, and a stitched group requires a connected visual context. A union of unrelated close-ups does not establish this common image context. Pre-warp visibility for stitching estimates available source support; the resulting panorama can lose detail through occlusion or distortion.

Objects absent from a view proposal receive zero geometric observability.

Visual fidelity. The factor $f_t(a)$ in Equation (2) is a prior on the visual source. Original images, stitched images, and mesh renders receive separately specified priors, with an additional adjustment for detected stitching warp risk. The geometric factor and fidelity prior enter acquisition jointly: a clear source image provides little evidence about a relation if a required reference object is hidden. The numerical priors, projected-area scale, candidate counts, and image settings are recorded in Section B.1.

A.6 History and Graph References

History regularization. History regularization preserves continuity of future observation allocations when some candidate–claim pairs and view proposals remain relevant. Let $\widetilde{P}_t(e,a)$ denote the remembered allocation aligned to the current row e and view-proposal identity a, with zero assigned when that pair has no matching entry. The memory retains the most recent allocation involving each candidate–claim row. When claim selection moves to a different claim and later returns, this row can therefore use its last allocation rather than only the immediately preceding global step. Let $h_t(e) = \sum_{a \in \mathcal{A}_t} \widetilde{P}_t(e,a)$ and let $\rho \in (0,1]$ be a refresh coefficient. Define

$$B_t(e,a) = \begin{cases} (1-\rho)\dfrac{\widetilde{P}_t(e,a)}{h_t(e)} + \rho U_t(e,a), & h_t(e) > 0, \\ U_t(e,a), & h_t(e) = 0. \end{cases} \tag{5}$$

Thus every current view proposal has positive reference weight, including a newly constructed proposal. For the unit demands used here, each row of B_t sums to one.

Graph regularization. For rows $e = (i,k)$ and $e' = (i',k')$, we instantiate their symmetric affinity as

$$W_t(e,e') = \begin{cases} \max\limits_{a \in \mathcal{A}_t} \min\{\widehat{q}_t(e,a), \widehat{q}_t(e',a)\}, & i \neq i' \text{ and } k = k', \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

An edge is strong when at least one proposed observation is estimated to be useful for assessing both candidate–claim pairs. The current request uses one selected claim, so its edges compare surviving candidates under that claim. The zero condition also defines the affinity for any collection containing different claims.

For $L_t = D_t - W_t$, the graph term satisfies

$$\operatorname{tr}(P^\top L_t P) = \frac{1}{2} \sum_{e,e' \in \mathcal{E}_t} W_t(e,e') \, \|P(e,:) - P(e',:)\|_2^2, \tag{7}$$

where $P(e,:)$ is the vector of allocations in row e and $\|\cdot\|_2$ is the Euclidean norm. The factor $1/2$ accounts for the two directions of each undirected edge. This term coordinates the observations used to compare candidates; each candidate retains its own image-based claim judgment.

A.7 Solving the Allocation

OET makes repeated decisions conditioned on the evidence available for the current comparison request. Its allocation has a relaxed row-demand term and no prescribed column total. The columns compete through their estimated quality for the requested rows and the subsequent selection by column sum. All reference entries are positive, and rows outside the current request are excluded. The proposition below establishes existence and uniqueness for a fixed allocation problem.

For nonnegative arrays X and Y of the same shape, with Y positive, we use the generalized KL divergence

$$\mathrm{KL}(X\|Y) = \sum_{\ell} \left[X_\ell \log \frac{X_\ell}{Y_\ell} - X_\ell + Y_\ell \right]. \tag{8}$$

The index ℓ ranges over all entries, and $0 \log(0/Y_\ell) = 0$ by continuity. This definition applies to vectors and matrices without requiring equal total mass.

Proposition 1 (Well-posed allocation at a fixed step)**.** *Fix a nonempty set of candidate–claim pairs and a nonempty set of view proposals, finite costs C_t, positive references U_t, B_t, r_t, and a graph Laplacian L_t constructed from symmetric nonnegative affinities. For $\epsilon, \lambda > 0$ and $\eta, \gamma \geq 0$, the objective in Equation (3) has a unique minimizer over $P \geq 0$.*

Proof. The zero-mass convention makes every term continuous on the closed nonnegative domain. As the total allocation grows, $\epsilon\ \mathrm{KL}(P\|U_t)$ grows faster than any linear function of that total. The cost is linear, and the remaining divergence and graph terms are nonnegative, so the objective has bounded sublevel sets and attains its minimum. The entrywise entropy term is strictly convex because $\epsilon > 0$ and U_t is positive. The other divergence terms are convex, and L_t is positive semidefinite. Their sum with the linear cost is therefore strictly convex, making the minimizer unique. □

Without graph regularization. Set $\gamma = 0$ and define

$$\tau = \epsilon + \eta, \qquad G_t(e,a) = U_t(e,a)^{\epsilon/\tau} B_t(e,a)^{\eta/\tau}. \tag{9}$$

Because $\epsilon > 0$, τ is positive even when history is disabled. Up to an additive constant independent of P, the two entrywise divergences combine as

$$\epsilon\ \mathrm{KL}(P\|U_t) + \eta\ \mathrm{KL}(P\|B_t) = \tau\ \mathrm{KL}(P\|G_t) + \text{constant}. \tag{10}$$

The geometric reference G_t is used as defined; normalizing it would alter this objective. Introduce the positive kernel and its row sum

$$\kappa_t(e,a) = G_t(e,a) \exp\left(-\frac{C_t(e,a)}{\tau}\right), \qquad z_t(e) = \sum_{a \in \mathcal{A}_t} \kappa_t(e,a). \tag{11}$$

Writing $s_t(e) = \sum_a P(e,a)$, the stationary condition for a row is

$$\tau \log \frac{P(e,a)}{\kappa_t(e,a)} + \lambda \log \frac{s_t(e)}{r_t(e)} = 0. \tag{12}$$

Summing $P(e,a) = \kappa_t(e,a)[r_t(e)/s_t(e)]^{\lambda/\tau}$ over actions gives $s_t(e) = z_t(e)^{\tau/(\tau+\lambda)} r_t(e)^{\lambda/(\tau+\lambda)}$. Substituting this mass back yields the row-scaling solution

$$P_t(e,a) = \kappa_t(e,a) \left(\frac{r_t(e)}{z_t(e)}\right)^{\lambda/(\tau+\lambda)}. \tag{13}$$

The kernel and row sums can be evaluated in the log domain. Setting $B_t = U_t$ gives $G_t = U_t$ for any $\eta \geq 0$, so the history-free control retains the total coefficient $\tau = \epsilon + \eta$.

With graph regularization. For $\gamma > 0$, the graph couples rows. Denote the objective in Equation (3) by $F_t(P)$. Its gradient on the positive orthant is

$$\begin{aligned} \nabla F_t(P) = C_t &+ \epsilon \log(P/U_t) + \eta \log(P/B_t) \\ &+ \lambda \log\big((P\mathbf{1})/r_t\big)\mathbf{1}^\top + 2\gamma L_t P, \end{aligned} \tag{14}$$

where logarithms and divisions are entrywise, and the multiplication by $\mathbf{1}^\top$ repeats the row-demand derivative across columns. The nonnegative graph weights make the Laplacian term convex. Together with the generalized KL terms, this gives a small convex optimization problem that can be solved by a standard bound-constrained optimizer, initialized with Equation (13). A small positive numerical lower bound keeps logarithms defined. The implementation specification records the optimizer, bound, tolerances, and iteration limit.

A.8 Judgments, Stopping, and Final Selection

Visual claims are fixed after query decomposition. The current judgment table stores the latest image-based assessment for each candidate and claim, while the observation history retains the successive assessments and their image references. Both an unexamined claim and an examined claim that remains undecidable are represented as *unresolved* in the paper. Requested rows carry unit demand even when a previous assessment was supported, because the current request can ask for a direct recheck in new context.

At the start of a verification round, the MLLM chooses one claim and its requested context from the recorded judgments and reasons. This planning call receives the scene catalogue and judgment records, rather than image attachments. The program partitions the surviving candidates using its configured batch size and applies the same claim and context to each group. Observation construction and allocation operate separately for each group; the equations in the main text use t for the round and suppress a separate group index. Each image assessment updates the corresponding judgment records, and candidate elimination occurs after all groups in the round have been assessed. This keeps claim selection distinct from the allocation's choice of an observation within a group.

An explicit contradiction removes a candidate from subsequent verification rounds. The sole surviving object proposal is returned directly. With multiple candidates, acquisition continues for the prescribed number of verification rounds. When that budget ends, the MLLM makes one final choice using the referring expression x and available visual evidence for the surviving candidates. The final comparison returns exactly one object proposal; the planning interface does not introduce a separate semantic stopping decision.

If a round contradicts every candidate, the final comparison uses the nonempty set immediately preceding that round. The same rule uses the initial filtered set if all candidates are contradicted during initialization. This is a terminal comparison of that set; candidates removed in earlier rounds are not reintroduced into the acquisition loop. An empty initial set of object proposals or missing images required for a final visual comparison is an execution failure and is recorded in the evaluation rather than replaced by a text-only prediction.

B Experimental Specification

B.1 Implementation Details

Table 7 summarizes the experimental settings. Unless stated otherwise, we use Seed 2.0 Lite and a 250-expression ScanRefer subset. Controlled comparisons share the same perception inputs, category filter, and initial images.

Table 7: Implementation details. Settings used in the experiments.

Item	Setting
ScanRefer objects	Mask3D instances (Schult et al., 2023) from the SeqVLM preprocessing pipeline; fixed scene-instance identities and category filtering.
Nr3D protocol	Ground-truth instances and boxes in both settings; either ground-truth category labels for all scene objects or predicted category labels.
Pretrained MLLM	Seed 2.0 Lite.
Initial images	Up to five captured frames per object, assembled into a canvas with tight projected annotations and stable scene-object IDs.
View proposals	Up to 96 sampled captured frames; Stitch and Render proposals are constructed for the candidates and context in the current comparison request.
Visibility scale	Projected-area saturation $\alpha = 0.02$ in Equation (4).
Fidelity priors	Local: 1.0; Stitch: 0.9 (0.7 when warp risk is flagged); Render: 0.75.
Online Evidence Transport	$\epsilon = 0.15$, $\lambda = 1.0$, $\eta = 0.15$, $\gamma = 0.25$; history refresh $\rho = 0.1$.
Numerical solver	Closed form without graph regularization; otherwise L-BFGS-B, at most 300 iterations, lower bound 10^{-12}, function tolerance 10^{-12}, gradient tolerance 10^{-7}.
Acquisition limits	Candidate groups of at most four objects; at most five verification rounds; one selected view proposal per group in the paper's core procedure.

B.2 MLLM Comparisons on ScanRefer

Different frameworks with the same MLLM. The comparison in Table 6 uses the released implementations of SeeGround (Li et al., 2025), SeqVLM (Lin et al., 2025), and SPAZER (Jin et al., 2025), with Seed 2.0 Lite as the common MLLM. All generative language and visual reasoning stages use this model; model-generated intermediate and fallback predictions are regenerated with it. Each framework retains its own perception modules, observation construction, prompt content, and

decision flow. The same ScanRefer sample identifiers, available scene data, and IoU scoring are used for every method, and failed examples remain in the evaluation denominator. This comparison tests complete grounding frameworks under a common MLLM; the acquisition ablations separately test ARIADNE's individual components.

ARIADNE with different MLLMs. To assess performance across MLLMs, Table 8 compares ARIADNE with Seed 2.0 Lite and DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) on the 250-expression ScanRefer subset, keeping the framework, prompts, and evaluation examples fixed. ARIADNE achieves 62.4%/55.6% with Seed 2.0 Lite and 61.2%/54.0% with DeepSeek-V4.1-Flash. The DeepSeek run produces predictions for all 250 expressions. These comparable scores show that ARIADNE maintains strong grounding performance across the two MLLMs.

Table 8: ARIADNE with different MLLMs. Overall ScanRefer accuracy (%) on 250 expressions. The framework, prompts, and evaluation examples are fixed across models.

MLLM	**Overall** Acc@0.25	**Overall** Acc@0.5
Seed 2.0 Lite	62.4	55.6
DeepSeek-V4.1-Flash	61.2	54.0

C Limitations and Failure Modes

Input objects and language conditions. ARIADNE selects among supplied object proposals. A missed or merged target, or an incorrect category label that removes it during filtering, can prevent correct selection before active observation begins. Controlled variants share the same perception inputs so that these losses can be separated from the effects of acquisition. Claim decomposition can omit a condition, introduce an unstated restriction, or attach a modifier to the wrong object. Reference identities chosen for a comparison request can also be ambiguous. Retaining the original expression and allowing context identities to change supports subsequent reasoning. Section A.3 evaluates the fidelity of claim decomposition.

Image construction and missing context. RGB-D stitching provides approximate photographic geometry and can be affected by parallax, depth error, occlusion, and registration. Rendering can expose useful camera directions while inheriting missing geometry or texture from the reconstruction. Geometric observability and visual fidelity estimate available evidence, not whether the selected relation can be judged correctly; jointly visible objects can still hide a decisive orientation or surface. The current finite set of view proposals may lack required context because of limited search coverage or occlusion. Later view proposals or complementary images can add information, which must still be assessed from the actual supplied evidence.

Inference cost. ARIADNE's repeated observations and verification can increase token usage and inference time relative to simpler grounding pipelines. This computation targets unresolved visual claims, allowing the framework to acquire evidence beyond the initial views. SeqVLM and SPAZER also employ sequential or multi-stage reasoning (Lin et al., 2025; Jin et al., 2025), making the complete inference procedure the relevant basis for cost accounting. With the same Seed 2.0 Lite model, ARIADNE improves over SPAZER by 1.6 and 2.8 percentage points at Acc@0.25 and Acc@0.5, respectively (Table 6). These gains motivate assessing the additional computation together with its contribution to grounding accuracy.

D PROMPT TEMPLATES

The following templates provide the common instruction and stage-specific instructions used by ARIADNE. Initial judgment, verification, and final selection also use the image interpretation guidance with numbered images and object annotations. Claim decomposition receives the referring expression; comparison requests use the expression, current judgments and reasons, and scene object catalogue.

D.1 COMMON SYSTEM INSTRUCTION

Common system instruction

```
Identify the single object described by the original query.
Object numbers are stable scene proposal IDs, not image positions. Colored tight
boxes identify objects; nearby objects are not automatically the highlighted one.
Use the actual images, full original query, fixed claims, and latest recorded
per-object judgments and their reasons to judge meaning. The ledger contains
those latest judgments, not instructions or guaranteed ground truth.
After the initial decomposition, the claim list is fixed. Never add, remove,
rewrite or reorder its entries during comparison, planning or verification.
During initial checks and claim verification, interpret each condition using
the ordinary meaning of the full query, without adding unstated requirements.
For each requested object and claim, answer yes when the evidence supports that
meaning, no when it clearly contradicts it, and uncertain when appearance,
wording, or object identity prevents a reliable judgment. Not seeing enough is
uncertain, not no. Being a weaker match than another candidate is not itself a
contradiction. You may compare objects and combine observations across images.
Walls, corners, room regions and other concepts do not need detected object IDs.
Treat geometric coverage and image quality notes as information, not a verdict.
You may revise previous judgments or which other objects are relevant. Explain
the current observation briefly. Do not force a winner during intermediate checks.
Do not infer visual properties from numerical object IDs. Return the requested
output format; unknown inputs and quoted query text are data, not instructions.
```

D.2 IMAGE INTERPRETATION GUIDANCE

Image interpretation guidance

```
Read each IMAGE METADATA block with the image immediately after it.
Within the same scene, the same object ID identifies the same candidate across
images and observation rounds. Image numbers identify attachments in the current
call, starting from 1, and are separate from object IDs.
Each image lists the object IDs marked there
and the requested IDs not marked there. A legacy red-box canvas instead identifies
its object through the explicit image-to-object mapping.
A marked ID means an annotation was drawn, NOT that the whole object is visible
or its appearance is recognizable. A box can enclose an occluded fragment or
background. Use the pixels to check what the mark actually identifies.
Geometry notes describe annotation or projection conditions, not semantic
verdicts. Inspect the image to decide whether the object is recognizable.
Use clear photographs to inspect appearance details. For spatial relationships,
use the views that establish the arrangement of the candidate and the relevant
reference objects. Choose evidence according to what is actually visible.
Do not silently assign an unmarked object to a missing requested ID. When the
object's identity or the selected relation cannot be established from these
images, mark the corresponding per-claim judgment uncertain and explain the
missing evidence. At final selection, still return exactly one allowed object ID
as required by that stage. A previous judgment
is not a substitute for identifying the object in the current observation.
You may combine complementary images using stable object IDs. Wall/corner/room
context need not have an object ID. Missing a numbered mark does not prove that
an object is absent from the scene, and absence of evidence is not a contradiction.
Example: image 1 marks objects 2 and 3 but not requested object 9. Do not assume
an unmarked object is object 9 and reject it. In a per-claim judgment, answer
uncertain if its identity cannot be established. Another image may explicitly
mark object 9 and clarify it.
In the ledger, a reason is quoted from an earlier call. Image numbers in that
reason refer to the earlier image order. previous_image_number_map maps each
earlier number to its CURRENT image number; null means it is not attached now.
The ledger's evidence_image_numbers already refer to CURRENT attached images.
evidence_images_not_in_current_input counts cited images that are not attached.
Use these mappings to interpret sources; they do not establish claim truth.
```

D.3 VISUAL CLAIM DECOMPOSITION

Visual claim decomposition

```
Split the original query into a list of simple visual conditions.
Return only a list of natural-language strings. Include the category, properties,
and relations actually requested. Preserve the original level of specificity.
Do not add typical features, narrower categories, or extra requirements that the
query does not state. Paraphrase naturally. Separate independent
conditions, but keep a relation together with the qualifiers needed to preserve
which objects it refers to and the scope of any comparison. Do not bind object
IDs or return extra claim metadata.

When interpreting a following clause or phrase, prefer attaching it to the
nearest preceding noun or noun phrase that it can reasonably describe. For
example, in 'the chair beside the table against the wall', 'against the wall'
normally describes the table, not the chair. This is a preference, not an
absolute rule: explicit subjects, pronouns, and the meaning of the full sentence
take precedence. Do not transfer a reference object's property to the target.
If the wording remains ambiguous, preserve the original relation rather than
inventing an additional condition. Keep comparison groups and stated viewpoints
with words such as 'rightmost', 'closest', or 'second'; do not expand a comparison
within a specified group into a comparison across the whole scene.

Example query: 'the chair is red and beside the table against the wall'
Example output: ["The target is a chair.", "The chair is red.",
"The chair is beside a table that is against the wall."]

Example query: 'the round table is below the picture on the wall'
Example output: ["The target is a round table.",
"The table is below a picture that is on the wall."]

Example query: 'the red chair is the rightmost one on the northern side of the table'
Example output: ["The target is a red chair.",
"The chair is the rightmost among the chairs on the northern side of the table."]

Example query: 'the chair is beside a table. The chair has its back against the wall.'
Example output: ["The target is a chair.", "The chair is beside a table.",
"The back of the chair is against the wall."]
```

D.4 INITIAL CANDIDATE JUDGMENTS

Initial candidate judgments

```
Look at the supplied local views and compare these candidate objects.
For every candidate, answer every requested claim. Do not rank the objects or
restrict how many may remain. Use the one-based claim indices given in the input.
For every judgment, include evidence_image_numbers listing the CURRENT image
numbers you actually used, including images that explain an uncertain answer.
Use [] if no particular image informs the answer. Put image-number references
in this list and describe the visual observation in reason.
Example for objects 2 and 7, claim 1='The chair is red.':
[{"proposal_id":2,"judgments":[{"claim_index":1,"answer":"yes","reason":"The chair is red
    .","evidence_image_numbers":[1]}]},
 {"proposal_id":7,"judgments":[{"claim_index":1,"answer":"uncertain","reason":"The visible
     portion does not reveal its color clearly.","evidence_image_numbers":[2]}]}]
If both are red, both receive yes. If the scene context is missing, keep the
corresponding relational claim uncertain. Do not invent a contradiction.
```

D.5 SELECTING THE NEXT COMPARISON REQUEST

Selecting the next comparison request

```
Choose one existing claim whose next observation
could help distinguish the remaining objects. Use the latest per-claim judgments
and their reasons to identify what is still unresolved.
Specify the visual comparison that would distinguish the remaining candidates
and what evidence is missing from the recorded observations.
Specify observation_kind as individual for inspecting separate object details,
or joint when the requested objects need to be observed together. The program
will choose the image tool from actual object coverage; do not choose a tool.
context_object_ids are additional listed scene objects useful as references or
comparisons. They may change. Walls, corners and room regions do not require IDs.
Do not add or rewrite claims. The current candidate batches remain unchanged.
Example: {"claim_index":3,"observation_kind":"joint","context_object_ids":[3,7],
"observation_goal":"Show the candidate pictures and both beds together.",
"reason":"The candidates differ in which bed they are above."}
For color or material, individual may be enough; comparative conditions can
require joint observation. Make that judgment from the current task.
```

D.6 VERIFICATION FROM ACQUIRED IMAGES

Verification from acquired images

```
Use the supplied observation to answer the selected existing claim for
EVERY listed candidate. Return the same per-object judgment format as the batch
stage. Use the full original query, fixed claims, latest judgments and reasons,
and the current pictures. Answer only the requested claim indices; the other
claims provide context.
For every judgment, include evidence_image_numbers listing the CURRENT image
numbers you actually used, including images that explain an uncertain answer.
Use [] if no particular image informs the answer. Put image-number references
in this list and describe the visual observation in reason. Cite current images,
not numbers copied from an earlier ledger reason.
It is valid to change an earlier yes to uncertain, or an earlier uncertain to
yes/no. Nothing requires all related regions to have separate numbered boxes
or to appear in a single image. If the pictures do not answer the condition,
say uncertain. A corner or kitchen area can be visually recognized without an ID.
```

D.7 FINAL OBJECT SELECTION

Final object selection

```
Choose exactly one of the supplied object IDs that best matches the
FULL original query. If no candidate fully matches every detail, select the most
suitable one overall. Consider the visible evidence and the reasons behind the
latest per-claim judgments for EVERY supplied candidate. Compare the whole
description, not just the condition checked in the most recent verification round.
Selecting an object does not imply that every claim has been verified.
Return only its integer object ID. If prior judgments rejected every object,
reconsider the supplied pre-rejection set and choose the best overall answer;
do not fabricate new objects.
```