GROUND3R: STREAMING 3D REFERRING EXPRESSION SEGMENTATION WITH RECURRENT MULTIMODAL MEMORY

Anonymous authors
Paper under double-blind review

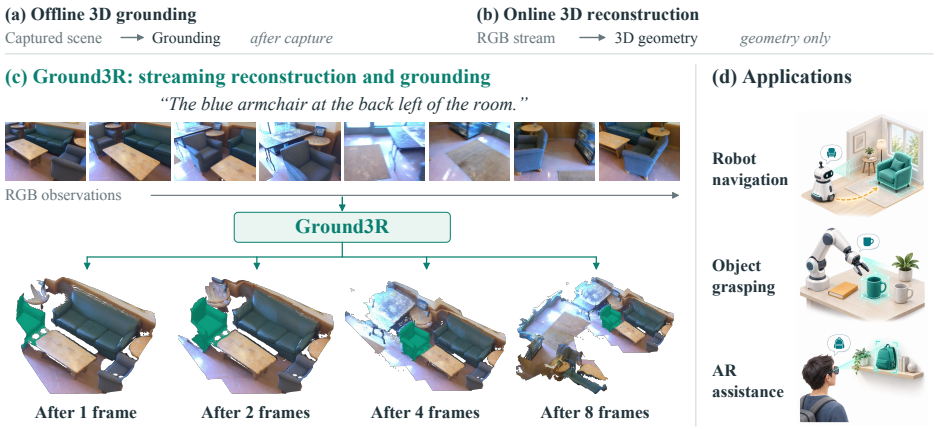


Figure 1: **Ground3R grounds language in 3D as RGB observations arrive.** (a) Offline grounding; (b) online reconstruction; (c) our streaming grounding, with the referred blue armchair in teal on the reconstruction accumulated after 1, 2, 4, and 8 frames; (d) illustrative applications.

ABSTRACT

3D referring expression segmentation (3D-RES) aims to segment objects described by natural language in 3D scenes. Existing methods typically rely on preconstructed point clouds or offline collections of RGB images, limiting their applicability to embodied agents that explore unfamiliar scenes through streaming observations. In such scenarios, agents must reconstruct the scene and identify the referred object incrementally as new images arrive, without access to future observations. To address this setting, we introduce streaming 3D referring expression segmentation (Stream3D-RES), which jointly reconstructs the scene and segments the referred object from incoming RGB images using only current and past observations. This task presents two key challenges: (1) identifying the target may require contextual information observed in earlier views, and (2) distinguishing the referent from similar objects becomes difficult as the target leaves and re-enters the field of view. To tackle these challenges, we propose Ground3R, an online framework that combines recurrent multimodal memory with spatially localized object queries. Its Multimodal Memory Branch maintains a language-conditioned scene state across frames, allowing current predictions to leverage historical context. The Spatial Query Localization and Refinement module generates language-independent object proposals and uses the expression and accumulated scene context to identify the referent and refine its mask. Query features are transferred across views through 3D centroid matching, while explicit target-visibility prediction enables empty masks when the target is absent. We further introduce StreamRefer, a benchmark covering four target-visibility patterns and three observation lengths. Extensive experiments demonstrate that Ground3R achieves state-of-the-art performance among online models, substantially outperforming causal 3D baselines and 2D-lifting alternatives across all visibility patterns.

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.

1 INTRODUCTION

3D referring expression segmentation (3D-RES) aims to segment objects described by natural language in 3D scenes (Huang et al., 2021; Wu et al., 2024c). As a fundamental capability for embodied agents, it enables language-guided object localization for downstream tasks such as navigation and manipulation (Huang et al., 2023; Shridhar et al., 2021). Existing 3D-RES methods typically assume preconstructed point clouds (Huang et al., 2021; Wu et al., 2024c) or offline collections of RGB images (Wu et al., 2026). However, embodied agents operating in unfamiliar environments observe scenes through a stream of partial RGB views, where complete scene geometry is not available in advance. In such scenarios, agents must reconstruct the scene and identify the referred object incrementally as new observations arrive, without access to future views. To address this setting, we introduce *streaming 3D referring expression segmentation* (Stream3D-RES), which jointly reconstructs the scene and segments the referred object from incoming RGB images using only current and past observations (Figure 1).

Existing methods cannot directly address the requirements of Stream3D-RES. On one hand, conventional 3D-RES methods rely on preconstructed scene representations or jointly process collected RGB views, making them unsuitable for incremental grounding as new observations arrive. While streaming reconstruction models such as CUT3R (Wang et al., 2025b) maintain a persistent state for online 3D reconstruction, they focus on geometry rather than language-guided segmentation. On the other hand, online 2D referring video models (Wu et al., 2023a; Zhang et al., 2026) can process incoming frames sequentially, but directly lifting their predictions into 3D yields limited segmentation accuracy in this setting. These limitations motivate an online framework that jointly reconstructs scene geometry and grounds the referring expression using context accumulated from preceding views.

Stream3D-RES poses two primary challenges. First, identifying the referred object may require contextual information observed in earlier views. For example, grounding "the chair beside the window" may depend on a previous observation of the window, even when the current view contains only the chair. As the camera moves, the objects and spatial relations needed to identify the target may no longer be visible. A streaming model must therefore retain language-relevant scene context across views and make it available for subsequent grounding. Second, reliable target grounding becomes difficult under changing visibility. Objects sharing the target's category or appearance can produce similar semantic responses, while the target itself may enter, leave, and re-enter the field of view. The model must distinguish the referent from similar candidates using accumulated scene context, while avoiding false predictions when the target is absent.

To address these challenges, we propose **Ground3R**, an online framework built on a pretrained recurrent reconstruction backbone (Wang et al., 2025b; Chen et al., 2026c). Ground3R combines recurrent multimodal memory with spatially localized object queries to ground each incoming observation using accumulated scene context. Its *Multimodal Memory Branch* (MMB) conditions scene features on the referring expression and maintains a language-conditioned scene state across frames. By incorporating language information into recurrent state updates, MMB accumulates scene context relevant to the referring expression rather than relying solely on geometry-oriented memory. Each incoming view updates this persistent state and retrieves historical context, providing multimodal features for current target segmentation.

Our *Spatial Query Localization and Refinement* (SQLR) module uses the accumulated context to identify and segment the target among spatially localized object candidates. Spatial queries first generate language-independent object proposals, establishing candidate regions before language-guided target selection. Each proposal then uses its spatial support to gather MMB features and interact with the referring expression, allowing the model to distinguish the referent from similar objects and refine its mask. To connect object observations across views, query features are transferred from the preceding view through 3D centroid matching. An explicit target-visibility prediction further determines whether to output the selected mask or an empty mask when the target is absent. Together with the reconstructed point clouds, these predictions yield streaming 3D target segmentation.

To evaluate this setting, we introduce *StreamRefer*, a benchmark for causal grounding across short, medium, and long observation lengths (8, 16, and 32 frames). StreamRefer explicitly categorizes sequences into four temporal visibility patterns: Persistent, Emerging, Disappearing, and Intermittent. These patterns evaluate grounding under different target-visibility trajectories, including initial

absence, disappearance, and reappearance. Extensive experiments show that Ground3R achieves state-of-the-art performance among online models, outperforming causal multi-view baselines and 2D-lifting alternatives across all four visibility patterns. Ablation studies further demonstrate the complementary contributions of multimodal memory and spatial query localization and refinement.

Our contributions are:

- We formulate Stream3D-RES for jointly reconstructing scenes and segmenting language-described objects from streaming RGB observations.
- We propose Ground3R with a Multimodal Memory Branch that maintains language-conditioned scene context in a persistent recurrent state, allowing current grounding to draw on historical observations.
- We design SQLR to combine language-independent object proposals, 3D centroid query transfer, and language-guided refinement with explicit target-visibility prediction.
- We introduce StreamRefer to evaluate grounding across four visibility patterns and three observation lengths, and demonstrate leading performance among online models.

2 Related Work

3D Referring Expression Segmentation. 3D visual grounding localizes objects from natural-language descriptions. Methods range from matching language to object candidates (Chen et al., 2020; Achlioptas et al., 2020) to language-guided localization and cross-modal alignment (Luo et al., 2022; Jain et al., 2022; Wu et al., 2023b; Park et al., 2026). 3D referring expression segmentation (3D-RES) extends object localization to point-level masks. Approaches use language-guided proposal selection (Huang et al., 2021) or end-to-end point-cloud–language fusion (Wu et al., 2024c; He & Ding, 2024; Wu et al., 2024a), with extensions to image-enhanced features (Chen et al., 2025). Generalized referring segmentation considers multiple or absent targets in images and 3D scenes (Liu et al., 2023; Wu et al., 2024b). Despite these advances, point-cloud-based methods depend on preconstructed geometry, limiting their use with incoming RGB observations. MVGGT (Wu et al., 2026) reconstructs geometry and segments targets directly from sparse RGB views, but jointly processes the collected images. In contrast, Ground3R reconstructs and grounds scenes incrementally, retaining language-relevant context as target visibility changes.

Referring Video Object Segmentation. Referring video object segmentation (RVOS) predicts target masks throughout a video from a natural-language expression. Transformer-based methods jointly model video and language (Botach et al., 2022; Wu et al., 2022), while recent work adapts pretrained vision-language, grounding, and segmentation models (Pan et al., 2025; Liang et al., 2025; Cuttano et al., 2025). Online approaches propagate object queries or recycle multimodal language-model embeddings to carry target information into subsequent frames (Wu et al., 2023a; Zhang et al., 2026). Lifting their image-space masks onto reconstructed geometry gives our 2D-Lift baselines, whereas Ground3R jointly reconstructs and segments each observation using language-conditioned memory.

Feed-Forward and Streaming 3D Reconstruction. Feed-forward models recover geometry through pointmap prediction and learned correspondence (Wang et al., 2024; Leroy et al., 2024), with multi-view architectures enabling joint reconstruction (Yang et al., 2025; Wang et al., 2025a; 2026). However, global alignment or joint processing requires recomputation as observations accumulate. Streaming approaches use spatial memory and recurrent states (Wang & Agapito, 2025; Wang et al., 2025b; Cabon et al., 2025) or cached causal attention (Zhuo et al., 2026; Lan et al., 2026). Long-sequence reconstruction further benefits from adaptive updates (Chen et al., 2026b), windowed geometric context (Li et al., 2026; Chen et al., 2026a), and test-time-trained scene representations (Jin et al., 2026). Beyond geometry, online perception extends to RGB-D instance segmentation (Xu et al., 2025) and joint human–scene or instance-aware reconstruction (Chen et al., 2026c; Zou et al., 2026). Notably, IGGT4D predicts temporally consistent instance features for downstream open-vocabulary segmentation and grounding, but builds its streaming representation independently of the referring expression. In contrast, Ground3R conditions recurrent memory and mask refinement on language.

3 STREAM3D-RES TASK AND STREAMREFER BENCHMARK

3.1 TASK FORMULATION

Given an RGB image stream $\mathcal{I} = (I_1, \ldots, I_T)$ and a referring expression e, Stream3D-RES learns a causal function

$$f : \big(\mathcal{I}_{\le t}, e\big) \mapsto \big(S_t, M_t\big), \qquad t = 1, \ldots, T, \tag{1}$$

where $\mathcal{I}_{\le t} = (I_1, \ldots, I_t)$ contains only the images observed up to time t. As I_t arrives, $S_t \in \mathbb{R}^{K_t \times 3}$ gives its reconstructed 3D point cloud in a shared scene coordinate frame, and $M_t \in \{0,1\}^{K_t}$ is the corresponding 3D binary mask identifying the referred object. Accumulating these outputs yields the scene point cloud and its target mask. Compared with 3D-RES (Huang et al., 2021; Wu et al., 2024c) and offline multi-view grounding (Wu et al., 2026), the model must reconstruct and segment from the observed prefix alone, retain earlier context to distinguish the target from similar objects as it leaves and re-enters the view, and output empty masks during its absence.

3.2 THE STREAMREFER BENCHMARK

We construct StreamRefer from ScanRefer expressions (Chen et al., 2020) and ScanNet RGB sequences and instance annotations (Dai et al., 2017). Observation episodes of 8, 16, or 32 chronological RGB frames cover short, medium, and long observation windows. Episodes are grouped by the target's visibility trajectory into Persistent (22.2%), Emerging (27.4%), Disappearing (32.9%), and Intermittent (17.6%). The benchmark allocates 20%, 60%, and 20% of expressions to 8-, 16-, and 32-frame episodes overall, and reports each temporal category separately at each length. Episodes provide consecutive target-visible observations for identification and retain target-absent frames to evaluate predictions during absence. StreamRefer contains 7,835 expression–episode pairs across 138 scenes; sampling and evaluation details are given in Appendix A.2.

Metrics. Each reconstructed point corresponds to an image pixel; let A_{nt} and B_{nt} denote the predicted and GT target pixel sets for pair n in view t. We evaluate the aligned point masks by pooling intersections and unions over each episode (Global) or averaging per-view IoU (View):

$$\mathrm{mIoU}_{\mathrm{Global}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_t |A_{nt} \cap B_{nt}|}{\sum_t |A_{nt} \cup B_{nt}|}, \qquad \mathrm{mIoU}_{\mathrm{View}} = \frac{\sum_n \sum_t \mathrm{IoU}(A_{nt}, B_{nt})}{\sum_n T_n}, \tag{2}$$

where N is the number of expression–episode pairs and t runs over the $T_n \in \{8, 16, 32\}$ views of episode n.

Subsets. *Intermittent* contains episodes with multiple visible intervals. For a single interval, *Emerging* denotes visibility after an initial absence. When the target is visible from the first view, we distinguish *Persistent* (at least half the episode) from *Disappearing* (less than half). We further distinguish episodes with no same-class distractors, those with same-class distractors visible during target absence, and the remaining episodes with same-class distractors visible elsewhere in the window (Appendix A.4).

4 METHOD

Section 4.1 outlines Ground3R, Section 4.2 and Section 4.3 introduce the Multimodal Memory Branch (MMB) and Spatial Query Localization and Refinement (SQLR), and Section 4.4 describes the training objectives.

4.1 OVERVIEW

Given an RGB image stream $\mathcal{I} = (I_1, I_2, \ldots)$ and a referring expression e, Ground3R reconstructs a point cloud S_t and predicts its 3D segmentation mask M_t as each image I_t arrives, using only the current and preceding observations (Section 3.1). As illustrated in Figure 2, a frozen image encoder extracts visual features F_t^{vis}, while a frozen RoBERTa encoder (Liu et al., 2019) provides word features F^{lang}. The recurrent geometry backbone (Wang et al., 2025b) reconstructs S_t and supplies

decoder features $F^{\text{mm}}_{t,0}$ to MMB. MMB conditions these features on the referring expression and maintains a persistent recurrent state H_t, producing multimodal features that carry historical context into current segmentation (Section 4.2). SQLR initializes spatial queries from local image features, incorporates matched query features from the preceding view through 3D centroid matching, and decodes them into language-independent object proposals. The object queries then combine text features F^{lang} and MMB features to identify the target object, refine its mask, and predict target visibility (Section 4.3). The selected image mask is lifted onto S_t to obtain M_t, with an empty output when the target is predicted to be absent.

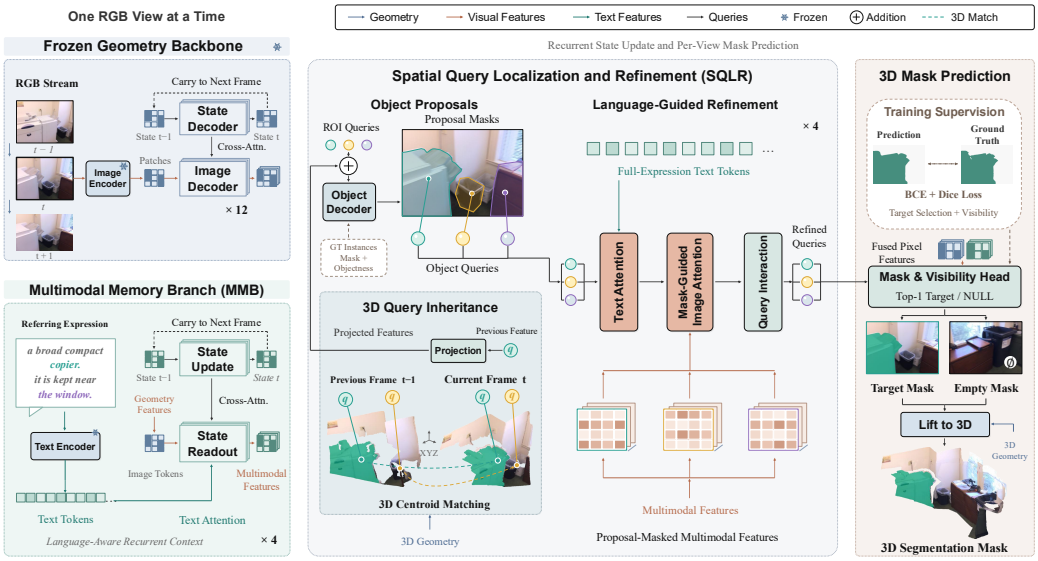


Figure 2: Overview of Ground3R. A frozen recurrent backbone reconstructs geometry from the RGB stream. The Multimodal Memory Branch keeps language-conditioned scene context in a recurrent state. SQLR initializes spatial queries with local image features and historical query features matched in 3D, decodes object proposals, and uses language and memory features for mask refinement and visibility prediction; the selected mask is lifted to 3D, or left empty when the target is absent.

4.2 Multimodal Memory Branch

Language-conditioned memory. MMB comprises $L = 4$ sequential interaction stages, each containing paired state-update and state-readout blocks following CUT3R (Wang et al., 2025b). Let $F^{\text{mm}}_{t,\ell}$ and H^{ℓ}_t denote the multimodal image features and state tokens after stage ℓ, initialized by the geometry decoder features $F^{\text{mm}}_{t,0}$ and the preceding state $H^0_t = H_{t-1}$. At each stage, image features are combined with word features through $\text{Attn}(A, B)$, which reads features from B for queries A. A learned scale α_ℓ controls the language contribution:

$$\widetilde{F}^{\text{mm}}_{t,\ell} = F^{\text{mm}}_{t,\ell-1} + \alpha_\ell \,\text{Attn}_\ell(F^{\text{mm}}_{t,\ell-1}, F^{\text{lang}}). \tag{3}$$

The state-update block $\mathcal{W}_\ell$ integrates the language-conditioned image features into the state, while the state-readout block $\mathcal{R}_\ell$ retrieves historical context for the current image:

$$H^{\ell}_t = \mathcal{W}_\ell(H^{\ell-1}_t, \widetilde{F}^{\text{mm}}_{t,\ell}), \qquad F^{\text{mm}}_{t,\ell} = \mathcal{R}_\ell(\widetilde{F}^{\text{mm}}_{t,\ell}, H^{\ell-1}_t). \tag{4}$$

Temporal accumulation and readout. Within each stage, state update and readout both use the state entering that stage. The final candidate state $\widehat{H}_t = \text{LN}(H^L_t)$ is carried forward as $H_t = (1 - \gamma_t)H_{t-1} + \gamma_t \widehat{H}_t$, where γ_t is an attention-derived update weight (Chen et al., 2026b). We initialize memory from the geometry backbone's initial state. A feature decoder combines features at different depths into the dense multimodal map used for mask refinement:

$$F^{\text{mm,pix}}_t = \mathcal{D}_{\text{mm}}(F^{\text{vis}}_t, F^{\text{mm}}_{t,L-2}, F^{\text{mm}}_{t,L-1}, F^{\text{mm}}_{t,L}). \tag{5}$$

4.3 Spatial Query Localization and Refinement

SQLR grounds the expression through spatially localized object candidates, using 3D correspondence to connect their observations across views.

Spatial query initialization and inheritance. We pool encoder features F_t^{vis} over a 10×10 image grid to initialize $Q = 100$ query features $q_{t,i}^{\text{ROI}}$, with grid anchors providing positional embeddings. Each anchor is assigned a 3D centroid $c_{t,i} = \text{median}_{u \in \mathcal{N}_i} P_t(u)$, the coordinate-wise median of valid reconstructed points in its local neighborhood $\mathcal{N}_i$, where $P_t(u)$ is the 3D point at pixel u in the predicted pointmap in the shared world coordinate frame. We find the nearest valid previous anchor $j_t(i) = \arg\min_{j \in \mathcal{J}_{t-1}} \|c_{t,i} - c_{t-1,j}\|_2$ and incorporate its decoded object query feature before object decoding:

$$q_{t,i}^0 = q_{t,i}^{\text{ROI}} + \delta_{t,i}\, \psi_{\text{3D}}\left(\text{sg}(q_{t-1,j_t(i)}^{\text{obj}})\right). \tag{6}$$

Here, sg stops gradients, ψ_{3D} is a learned projection, and $\delta_{t,i}$ accepts valid matches within distance ρ, with zero for unmatched queries and the first view.

Object proposal decoding. Four decoder layers transform the initialized queries into object features $q_{t,i}^{\text{obj}}$, with intermediate masks constraining their visual support. A text-free feature decoder supplies pixel features F_t^{obj}. Each decoded query predicts proposal logits

$$a_{t,i}(u) = \phi_{\text{obj}}(q_{t,i}^{\text{obj}})^\top F_t^{\text{obj}}(u) + b_{\text{obj}}, \tag{7}$$

where u indexes pixels and ϕ_{obj} maps queries to mask embeddings. The proposal probability is $m_{t,i}^{\text{obj}} = \sigma(a_{t,i})$, and an objectness head predicts $o_{t,i} = \phi_{\text{score}}(q_{t,i}^{\text{obj}})$. These candidates are supervised by their assigned GT instances independently of the expression.

Language-guided query refinement. The object queries localize candidate objects without using the referring expression. We refine them through multimodal cross-attention to identify the target among these candidates. A linear projection ψ_{in} initializes the refinement queries $R_t^0 = [\psi_{\text{in}}(\text{sg}(Q_t^{\text{obj}})),\, r_\emptyset]$, where a learned NULL query $r_\emptyset$ represents target absence. Cross-attention to text features gives language-conditioned queries $U = \text{LN}(R + \text{Attn}(R, F^{\text{lang}}))$. These queries then aggregate MMB features under proposal-mask guidance:

$$\text{MaskedAttn}(U, F^{\text{mm}}) = \text{softmax}\left(\frac{UW_q(F^{\text{mm}}W_k)^\top}{\sqrt{d}} + \mathcal{M}\right) F^{\text{mm}}W_v, \tag{8}$$

where F^{mm} denotes projected MMB features from the corresponding stage (time and stage indices omitted), W are learned projections, and d is the head dimension. The bias $\mathcal{M}_i(u) = \log \bar{m}_{t,i}(u)$ favors each candidate's proposal region, while $\mathcal{M}_\emptyset(u) = 0$ allows NULL to attend to the full image. Here, $\bar{m}_{t,i}$ denotes the soft proposal mask aligned with the MMB features. Self-attention then models dependencies among candidate queries and NULL. Four refinement layers yield $r_{t,i}$ and $r_{t,\emptyset}$ for mask and visibility prediction.

Mask and visibility prediction. We project and concatenate detached object features F_t^{obj} with $F_t^{\text{mm,pix}}$, then fuse them into F_t^{fuse} to decode refined masks:

$$z_{t,i}(u) = \phi_{\text{ref}}(r_{t,i})^\top F_t^{\text{fuse}}(u) + b_{\text{ref}}, \qquad m_{t,i}^{\text{ref}} = \sigma(z_{t,i}). \tag{9}$$

A relevance head scores each candidate as $s_{t,i}$ and NULL as $s_{t,\emptyset}$, giving target visibility

$$p_t = \sigma\left(\log\left[\frac{1}{Q}\sum_{i=1}^{Q}\exp(s_{t,i})\right] - s_{t,\emptyset}\right). \tag{10}$$

At inference, we select $i^* = \arg\max_i s_{t,i}$ and threshold z_{t,i^*} at zero. The resulting mask is lifted onto S_t when $p_t \geq 0.5$; otherwise, M_t is empty.

4.4 Training Objective

Query-to-instance assignment. Let $\{G_{t,k}\}_{k=1}^{J_t}$ be the binary masks of visible GT objects. Spatial assignment between the query grid and these masks defines $\pi_t : \{1, \dots, Q\} \to \{1, \dots, J_t\} \cup \{\emptyset\}$. Each assigned query is associated with one GT instance, while several queries may represent the same object. For the assigned set $\mathcal{O}_t = \{i : \pi_t(i) \neq \emptyset\}$, query i receives the mask label $g_{t,i} = G_{t,\pi_t(i)}$. With the mask loss $\ell_{\text{mask}} = \ell_{\text{BCE}} + \ell_{\text{Dice}}$ (Appendix A.1), every assigned proposal learns its own instance mask, and the refined masks of target-assigned queries receive additional supervision:

$$\mathcal{L}_{\text{prop}} = \frac{1}{|\mathcal{O}|} \sum_{(t,i)\in\mathcal{O}} \ell_{\text{mask}}(m_{t,i}^{\text{obj}}, g_{t,i}), \qquad \mathcal{L}_{\text{ref}} = \frac{1}{|\mathcal{T}|} \sum_{(t,i)\in\mathcal{T}} \ell_{\text{mask}}(m_{t,i}^{\text{ref}}, g_{t,i}), \tag{11}$$

where $\mathcal{O}$ collects all assigned query–view pairs in the training batch and $\mathcal{T}$ contains those assigned to the target object. Objectness BCE supervises assigned and unassigned queries, with reduced weight for the latter. Together with intermediate-layer supervision, it gives $\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{prop}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}$.

Grounding and visibility. For a target-visible view, let $\mathcal{T}_t$ contain the queries assigned to the target object. We maximize their combined relevance with a multi-positive grounding objective:

$$\mathcal{L}_{\text{grd}} = -\log \frac{\sum_{i\in\mathcal{T}_t} \exp(s_{t,i})}{\sum_{i=1}^{Q} \exp(s_{t,i})}. \tag{12}$$

We average this loss over target-visible views. Visibility loss $\mathcal{L}_{\text{vis}}$ applies BCE to p_t and the target-visibility label, balancing target-visible and target-absent views. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{obj}} + \lambda_{\text{grd}}\mathcal{L}_{\text{grd}} + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}}, \tag{13}$$

with grounding and refined-mask losses applied to target-visible views.

5 Experiments

5.1 Implementation Details and Setup

We use the CUT3R geometry backbone (Wang et al., 2025b), freezing it and the image and RoBERTa encoders (Liu et al., 2019). The grounding components are trained at a learning rate of 2×10^{-4} on four NVIDIA RTX 4090 GPUs (details in Appendix A.1), and training and evaluation use the ScanRefer and ScanNet training/validation splits (Chen et al., 2020; Dai et al., 2017). We report Global mIoU, with View mIoU additionally in Table 2 and the appendix.

5.2 Main Results

Compared methods. We compare against (1) MVGGT (Wu et al., 2026), run offline on the full episode and as MVGGT-causal, which reprocesses the observed prefix $I_1, \dots, I_t$ at each step; and (2) 2D-Lift baselines, which lift per-frame masks from referring video segmentation models onto Pi3 (Wang et al., 2026) geometry for the offline 2D-Lift (ReferDINO) (Liang et al., 2025) and 2D-Lift (SSA) (Pan et al., 2025), and onto CUT3R (Wang et al., 2025b) geometry for the online 2D-Lift (OnlineRefer) (Wu et al., 2023a) and 2D-Lift (StreamingRVOS-1B) (Zhang et al., 2026).

Overall segmentation. Table 1 compares Global mIoU across the twelve category–length combinations and the full benchmark. Ground3R achieves the highest overall score among online methods at 41.5, exceeding MVGGT-causal by 11.2 points and 2D-Lift (StreamingRVOS-1B) and 2D-Lift (OnlineRefer) by 15.9 and 26.6 points, respectively. It also surpasses the offline baselines 2D-Lift (ReferDINO) and 2D-Lift (SSA), which obtain 26.5 and 20.0. Offline MVGGT reaches 47.4 with access to the full episode. Ground3R also outperforms offline MVGGT at all three observation lengths on Persistent and Intermittent episodes.

Table 1: StreamRefer results (Global mIoU, %; category shares in parentheses). Bold and underline mark the best and second-best online results. Inference-time protocol: Appendix A.6.

Method	Persistent (22.2%)			Emerging (27.4%)			Disappearing (32.9%)			Intermittent (17.6%)			Overall	Inference time↓ (ms)
	8	16	32	8	16	32	8	16	32	8	16	32		
Offline														
2D-Lift (ReferDINO)	34.5	34.3	32.5	24.7	24.6	23.6	18.5	20.2	21.9	28.4	34.3	29.6	26.5	–
2D-Lift (SSA)	24.9	24.0	22.2	18.8	17.4	20.9	17.9	17.8	17.1	16.4	22.8	22.8	20.0	–
MVGGT	41.4	45.5	43.2	35.1	46.2	53.1	59.4	55.3	53.0	32.2	41.7	34.5	47.4	–
Online														
2D-Lift (OnlineRefer)	19.3	20.2	17.7	13.9	13.4	13.3	10.2	12.4	13.3	11.6	16.8	19.3	14.9	<u>174.3</u>
2D-Lift (StreamingRVOS-1B)	<u>32.6</u>	<u>31.9</u>	<u>31.6</u>	23.4	24.0	23.6	20.3	21.2	<u>21.5</u>	21.9	31.2	26.5	25.6	622.7
MVGGT-causal	26.2	24.8	23.7	<u>32.6</u>	<u>39.0</u>	**39.1**	<u>31.4</u>	<u>25.6</u>	15.4	<u>32.6</u>	<u>37.6</u>	<u>30.7</u>	<u>30.3</u>	3112.3
Ground3R	**42.3**	**48.2**	**48.6**	**37.3**	**40.2**	<u>36.3</u>	**32.3**	**41.3**	**44.2**	**35.9**	**44.7**	**36.9**	**41.5**	**128.0**

Table 2: Core-component comparison on StreamRefer (mIoU, %).

Variant	Components		Overall		8 frames		16 frames		32 frames	
	SQLR	MMB	Global	View	Global	View	Global	View	Global	View
Baseline	—	—	24.53	25.79	23.15	25.48	24.65	25.82	25.54	25.83
w/o SQLR	—	✓	34.53	58.66	35.39	53.75	35.37	59.34	31.14	58.88
w/o MMB	✓	—	33.83	53.74	30.53	47.61	35.20	54.33	33.04	54.40
Ground3R (full)	✓	✓	**41.47**	**65.16**	**36.70**	**54.64**	**43.07**	**65.31**	**41.44**	**67.58**

Table 3: SQLR ablation (Global mIoU, %).

Object queries	Query refine.	3D inherit.	Overall	8 fr.	16 fr.	32 fr.
×	×	×	34.53	35.39	35.37	31.14
✓	×	✓	34.90	30.46	36.28	35.20
✓	✓	×	40.56	36.40	41.89	40.72
✓	✓	✓	**41.47**	**36.70**	**43.07**	**41.44**

Table 4: MMB ablation (Global mIoU, %).

Language cond.	Recurrent state	Overall	8 fr.	16 fr.	32 fr.
×	×	33.83	30.53	35.20	33.04
×	✓	34.92	31.29	36.52	33.76
✓	×	36.99	34.51	37.75	37.22
✓	✓	**41.47**	**36.70**	**43.07**	**41.44**

Temporal visibility patterns. Among online methods, Ground3R achieves the highest Global mIoU in all four temporal visibility categories and leads in eleven of the twelve category–length combinations. The largest gains over MVGGT-causal occur in the 32-frame Persistent and Disappearing subsets: Ground3R reaches 48.6 and 44.2, improving by 24.9 and 28.8 points, respectively. It also leads at all three lengths on Intermittent, reaching 44.7 vs. 37.6 for MVGGT-causal on 16-frame episodes. Appendices A.3–A.5 report detailed results, same-class distractor subsets, and Nr3D/Sr3D expression transfer; Appendix A.7 examines the visibility threshold.

Inference efficiency. After 31 observed frames, Ground3R has the lowest per-frame inference time at 128.0 ms, compared with 174.3 ms for 2D-Lift (OnlineRefer), 622.7 ms for 2D-Lift (StreamingRVOS-1B), and 3112.3 ms for MVGGT-causal. Ground3R carries recurrent states and object query features forward, whereas MVGGT-causal re-evaluates the full observed prefix for each incoming frame. Appendix A.6 gives the timing protocol and results over longer sequences.

5.3 Ablation Studies

Core components. Table 2 compares a dense CNN baseline (24.53) with MMB- and SQLR-based variants (34.53 and 33.83), while the full model reaches 41.47 Global mIoU.

Spatial query localization and refinement. Table 3 examines object queries, language-guided query refinement, and 3D query inheritance. Replacing SQLR with dense prediction reduces Global

mIoU by 6.94 points, while retaining the queries but removing refinement yields 34.90, a 6.57-point drop that supports language-guided refinement after proposal generation. Transferring historical query features through 3D centroid matching further improves Global mIoU from 40.56 to 41.47, with gains at all three lengths (0.30, 1.18, and 0.72 points).

Multimodal memory. Removing language conditioning or the recurrent state lowers Global mIoU from 41.47 to 34.92 and 36.99, respectively (Table 4). The full MMB outperforms both variants at every observation length, showing the value of conditioning scene features on language and retaining them across views.

Qualitative results. In Figure 3, Ground3R distinguishes the end table beside the couch in (a), where the baselines select nearby objects, and accumulates the armchair across all eight views in (b). It also follows spatial descriptions, selecting the cabinets above the sink in (c) and the left door beside another door in (e), and isolates the monitor in the cluttered scene in (d).

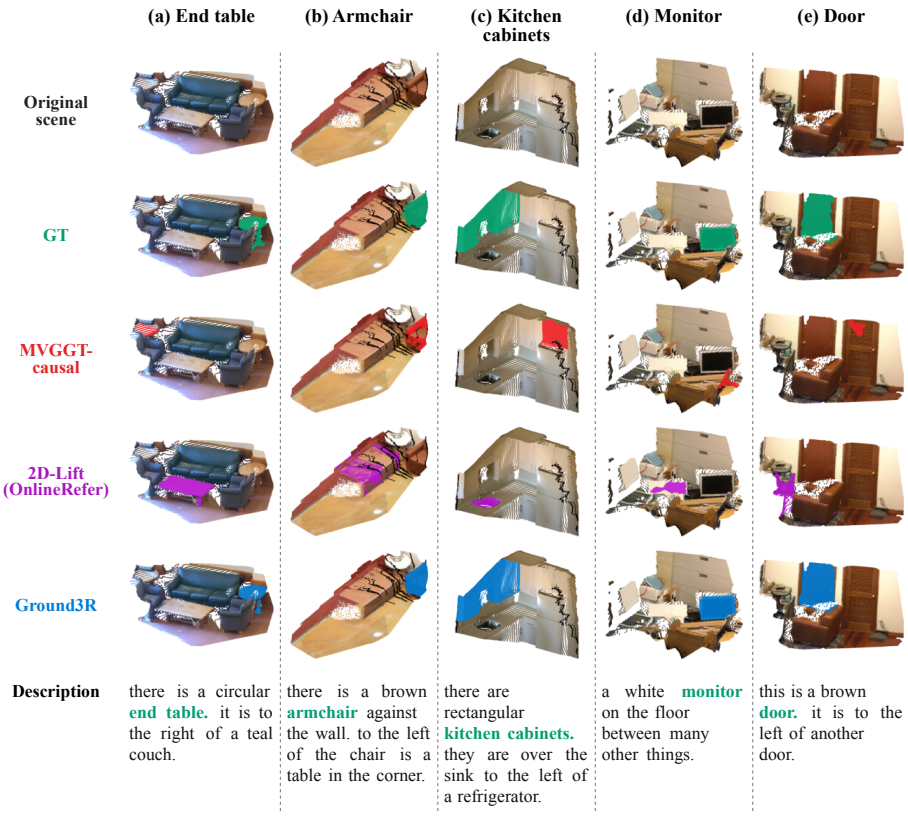


Figure 3: Qualitative comparison on StreamRefer. Panel (b) shows the accumulated reconstruction and predictions from all eight frames; the other panels show selected frames.

6 Conclusion

We introduce Stream3D-RES for jointly reconstructing scenes and segmenting language-described objects from streaming RGB observations, together with StreamRefer for evaluating grounding across target-visibility patterns and observation lengths. Ground3R combines a language-conditioned recurrent memory with spatial queries that are matched across views in 3D and refined with language, and predicts when the target is absent. It achieves state-of-the-art performance among online models on StreamRefer with the lowest per-frame inference time.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and assist with qualitative analysis and interpretation of experimental results. Additionally, we used these tools for drafting and revising the manuscript, literature search and citation checking, mathematical notation, LaTeX typesetting, table formatting, and figure preparation. The authors selected and revised the AI-assisted manuscript text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies language-guided 3D perception using existing ScanNet and ScanRefer data. StreamRefer organizes these observations and annotations into chronological evaluation episodes. Use and redistribution of the source data remain subject to the original datasets' access conditions. RGB-based scene reconstruction can reveal details of private spaces, making privacy and consent relevant considerations for deployment. Applications involving physical interaction should account for errors in target identification and visibility prediction.

REPRODUCIBILITY STATEMENT

Section 4 describes the model architecture, recurrent updates, query initialization, and training objectives. Section 3 and Appendix A.2 specify the StreamRefer construction, temporal categories, dataset statistics, and evaluation metrics. Training settings and baseline inference procedures are described in Section 5.1 and Appendix A.1. The main results and ablations report performance across observation lengths, with detailed length-specific comparisons in Appendix A.3.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *ECCV*, 2020.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pp. 4985–4995, 2022.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *CVPR*, pp. 1050–1060, 2025.

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

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a.

Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, Danni Yang, and Xiaoshuai Sun. IPDN: Image-enhanced prompt decoding network for 3D referring expression segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2132–2140, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *ICLR*, 2026b.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *ICLR*, 2026c.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAM-WISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *CVPR*, pp. 3395–3405, 2025.

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

Shuting He and Henghui Ding. RefMask3D: Language-guided transformer for 3D referring segmentation. In *ACM MM*, pp. 8316–8325, 2024.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *ICRA*, 2023.

Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3D instance segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1610–1618, 2021.

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

Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T. Barron, Noah Snavely, and Aleksander Hołyński. ZipMap: Linear-time stateful 3D reconstruction via test-time training. In *CVPR*, pp. 21748–21759, 2026.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *ICLR*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, 2024.

Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. WinT3R: Window-based streaming reconstruction with camera token pool. In *ICLR*, 2026.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. ReferDINO: Referring video object segmentation with visual grounding foundations. In *ICCV*, pp. 20009–20019, 2025.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, pp. 23592–23601, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3D-SPS: Single-stage 3D visual grounding via referred point progressive selection. In *CVPR*, pp. 16454–16463, 2022.

Feiyu Pan, Hao Fang, Fangkai Li, Yanyu Xu, Yawei Li, Luca Benini, and Xiankai Lu. Semantic and sequential alignment for referring video object segmentation. In *CVPR*, pp. 19067–19076, 2025.

GwangWook Park, Hyo-Jun Lee, Jong-Hyeon Baek, Hanul Kim, and Yeong Jun Koh. EG-3DVG: Expression and geometry aware grounding decoder for 3D visual grounding. In *CVPR*, pp. 2625–2634, 2026.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and where pathways for robotic manipulation. In *CoRL*, 2021.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *3DV*, pp. 78–89, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025a.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, 2026.

Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. RG-SAN: Rule-guided spatial awareness network for end-to-end 3D referring expression segmentation. In *NeurIPS*, volume 37, pp. 110972–110999, 2024a.

Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun, and Rongrong Ji. 3D-GRES: Generalized 3D referring expression segmentation. In *ACM MM*, pp. 7852–7861, 2024b.

Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3D-STMN: Dependency-driven superpoint-text matching network for end-to-end 3D referring expression segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5940–5948, 2024c.

Changli Wu, Haodong Wang, Jiayi Ji, Yutian Yao, Chunsai Du, Jihua Kang, Yanwei Fu, and Liujuan Cao. MVGGT: Multimodal visual geometry grounded transformer for multiview 3D referring expression segmentation. In *CVPR*, pp. 16800–16810, 2026.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. OnlineRefer: A simple online baseline for referring video object segmentation. In *ICCV*, pp. 2761–2770, 2023a.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pp. 4974–4984, 2022.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. EDA: Explicit text-decoupling and dense alignment for 3D visual grounding. In *CVPR*, pp. 19231–19242, 2023b.

Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. EmbodiedSAM: Online segment any 3D thing in real time. In *ICLR*, 2025.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *CVPR*, pp. 21924–21935, 2025.

Wenkang Zhang, Kaicheng Yang, Xiang An, Qiang Li, Ziyong Feng, Wankou Yang, and Jiankang Deng. Towards streaming referring video segmentation via large language model. In *CVPR*, pp. 24598–24607, 2026.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *ICLR*, 2026.

Zhengyu Zou, Hao Li, Kuixuan Jiao, Liu Liu, Tingyang Xiao, Xiaolin Zhou, Fangzhou Hong, Zhizhong Su, Dingwen Zhang, and Ziwei Liu. IGGT4D: Streaming 4D instance-grounded geometry transformer. *arXiv preprint arXiv:2607.19228*, 2026.

A Appendix

A.1 Implementation Details

Architecture and trainable components. We use the frozen CUT3R geometry backbone (Wang et al., 2025b) and the frozen RoBERTa-base language encoder (Liu et al., 2019), with a maximum text length of 64 tokens. The trainable grounding components comprise the Multimodal Memory Branch (MMB, $L = 4$ interaction stages), the Spatial Query Localization and Refinement (SQLR) decoder with $Q = 100$ queries and 4 refinement layers, mask prediction heads, and the relevance head for target-visibility prediction. For a predicted mask m and its assigned GT mask g on the output pixel grid Ω, the mask loss combines binary cross-entropy and Dice terms with unit weights, $\ell_{\text{mask}} = \ell_{\text{BCE}} + \ell_{\text{Dice}}$, where

$$\ell_{\text{BCE}}(m, g) = -\frac{1}{|\Omega|} \sum_{u \in \Omega} \big[g(u) \log m(u) + (1 - g(u)) \log(1 - m(u)) \big], \tag{14}$$

$$\ell_{\text{Dice}}(m, g) = 1 - \frac{2 \sum_u m(u) g(u) + \varepsilon}{\sum_u m(u) + \sum_u g(u) + \varepsilon}, \tag{15}$$

with a small stabilizer ε. Query-to-instance assignment uses area-normalized overlap between grid cells and instance masks, reserving queries for individual instances before distributing the remaining cells. Objectness classification and auxiliary supervision on intermediate Object Decoder layers are assigned loss weights of 0.1 and 0.5, respectively. We set $\lambda_{\text{grd}} = \lambda_{\text{ref}} = \lambda_{\text{vis}} = 1$ and down-weight unassigned queries by 0.1 in the objectness loss.

Training configuration. We train the grounding parameters using AdamW with an initial learning rate of 2×10^{-4}, weight decay 0.05, and momentum coefficients $(\beta_1, \beta_2) = (0.9, 0.95)$. The learning rate follows a linear warmup and decays to a minimum rate of 1×10^{-6}. We use bfloat16 mixed precision and an effective batch size of 32 through gradient accumulation. We train for 64 epochs. Training episodes consist of eight chronologically ordered views, balanced with four target-visible and four target-absent frames. Data augmentations include sequence-consistent color jitter (brightness, contrast, saturation) and multi-scale random cropping.

Baseline inference. Online baselines process video streams sequentially: 2D-Lift (OnlineRefer) (Wu et al., 2023a) uses the official ResNet-50 checkpoint and propagates its tracking state frame-by-frame; 2D-Lift (StreamingRVOS-1B) (Zhang et al., 2026) uses the released StreamingRVOS-1B checkpoint. Offline baselines evaluate entire video episodes: 2D-Lift (ReferDINO) (Liang et al., 2025) employs Swin-B and selects queries via clip-averaged class scores; 2D-Lift (SSA) (Pan et al., 2025) uses FC-CLIP with a ConvNeXt-Large backbone. For mask unprojection into 3D, offline lifting baselines use Pi3 (Wang et al., 2026) reconstruction, while online lifting baselines use CUT3R's recurrent geometric module at 512×384 resolution. MVGGT-causal and offline MVGGT (Wu et al., 2026) use the official released checkpoints with 518×336 resolution; MVGGT-causal re-evaluates each expanding observed prefix at every step, whereas offline MVGGT operates over the full sequence with bidirectional attention.

A.2 StreamRefer Details

A.2.1 Sampling and Temporal Categories

We construct chronological observation episodes with $T \in \{8, 16, 32\}$ frames for scene–target pairs and associate them with referring expressions. The overall allocation is 20%, 60%, and 20% of expression–episode pairs at 8, 16, and 32 frames, respectively. The category-specific counts are reported in Table 5. Target visibility is determined from ground-truth instance annotations, and expressions associated with a selected scene–target pair share its observation episode. Candidate windows provide consecutive target-visible observations and retain target-absent views when present.

Temporal categories follow the target's visible intervals. Episodes with multiple visible intervals form Intermittent. Among episodes with a single interval, those beginning with target absence form Emerging. If the target is visible from the first view, an interval lasting at least half the episode gives Persistent; a shorter interval gives Disappearing. These definitions scale with episode length.

A.2.2 Evaluation Protocol

We follow the metrics in Eq. 2. An empty union receives a score of one, both at the sequence level and for individual views. Temporal labels and visibility subsets are determined from the original ground-truth annotations, and IoU is computed on each method's evaluation mask grid.

A.2.3 Benchmark Statistics

StreamRefer contains 7,835 expression–episode pairs associated with 1,708 scene–target pairs across 138 ScanNet validation scenes. The 8-, 16-, and 32-frame partitions contain 1,567, 4,701, and 1,567 expressions, yielding 137,896 expression–view observations in total. Target-visible observations account for 48.05% of the benchmark and target-absent observations for 51.95%. Table 5 gives the temporal-category composition.

Table 5: Expression–episode counts by category and length.

Category	**8 fr.**	**16 fr.**	**32 fr.**	**Total**
Persistent	418	990	329	1,737
Emerging	366	1,323	455	2,144
Disappearing	540	1,551	485	2,576
Intermittent	243	837	298	1,378
Total	**1,567**	**4,701**	**1,567**	**7,835**

The distractor partition contains 3,556 expressions without same-class distractors, 3,727 with same-class distractors visible during target absence, and 552 in the remaining same-class-visible category.

A.3 Results by Observation Length

Tables 6–8 report the complete results across 8-, 16-, and 32-frame observation lengths, with category distributions indicated in headers. Bold and underline mark the best and second-best online scores. Across all observation lengths, Ground3R consistently achieves the highest overall Global mIoU among online methods, outperforming MVGGT-causal by 6.20, 11.73, and 14.52 points at 8, 16, and 32 frames, respectively. At 32 frames, the gains over MVGGT-causal reach 28.78 points on Disappearing and 24.88 points on Persistent episodes.

Table 6: Results on 8-frame StreamRefer episodes (mIoU, %).

Method	**Persistent** (26.7%)		**Emerging** (23.4%)		**Disappearing** (34.5%)		**Intermittent** (15.5%)		**Overall**	
	Global	View	Global	View	Global	View	Global	View	Global	View
Offline										
2D-Lift (ReferDINO)	34.47	35.03	24.66	25.63	18.53	19.36	28.42	28.47	25.75	26.42
2D-Lift (SSA)	24.92	27.70	18.82	24.57	17.94	26.29	16.41	20.86	19.77	25.42
MVGGT	41.43	56.54	35.07	54.50	59.40	80.65	32.20	45.29	44.71	62.63
Online										
2D-Lift (OnlineRefer)	19.30	20.96	13.94	14.45	10.19	13.89	11.62	13.70	13.72	15.88
2D-Lift (StreamingRVOS-1B)	<u>32.63</u>	34.46	23.42	27.73	20.29	27.35	21.88	24.33	24.56	28.87
MVGGT-causal	26.25	<u>49.26</u>	<u>32.62</u>	<u>54.42</u>	<u>31.39</u>	**72.93**	<u>32.61</u>	**47.28**	<u>30.50</u>	**58.31**
Ground3R	**42.32**	**55.83**	**37.33**	**55.33**	**32.26**	<u>58.74</u>	**35.92**	<u>42.43</u>	**36.70**	<u>54.64</u>

Table 7: Results on 16-frame StreamRefer episodes (mIoU, %).

Method	Persistent (21.1%)		Emerging (28.1%)		Disappearing (33.0%)		Intermittent (17.8%)		Overall	
	Global	View	Global	View	Global	View	Global	View	Global	View
Offline										
2D-Lift (ReferDINO)	34.31	31.34	24.58	24.51	20.25	20.07	34.35	32.47	26.94	25.90
2D-Lift (SSA)	24.01	26.22	17.44	28.76	17.85	29.31	22.75	26.96	19.90	28.09
MVGGT	45.54	59.62	46.24	68.13	55.28	78.06	41.66	54.35	48.26	67.16
Online										
2D-Lift (OnlineRefer)	20.17	18.99	13.35	15.21	12.42	14.22	16.82	17.86	15.10	16.15
2D-Lift (StreamingRVOS-1B)	31.85	32.38	24.01	30.90	21.23	28.43	31.21	32.71	26.03	30.72
MVGGT-causal	24.84	48.38	38.95	**65.82**	25.60	69.24	37.65	53.03	31.34	61.00
Ground3R	**48.18**	**63.75**	**40.25**	65.56	**41.34**	**71.98**	**44.68**	**54.39**	**43.07**	**65.31**

Table 8: Results on 32-frame StreamRefer episodes (mIoU, %).

Method	Persistent (21.0%)		Emerging (29.0%)		Disappearing (31.0%)		Intermittent (19.0%)		Overall	
	Global	View	Global	View	Global	View	Global	View	Global	View
Offline										
2D-Lift (ReferDINO)	32.48	30.59	23.63	23.86	21.86	23.56	29.64	27.77	26.08	25.92
2D-Lift (SSA)	22.22	28.71	20.88	33.85	17.10	29.06	22.77	27.10	20.35	30.01
MVGGT	43.22	59.03	53.08	74.69	53.01	77.71	34.47	50.24	47.45	67.69
Online										
2D-Lift (OnlineRefer)	17.68	18.20	13.27	16.19	13.30	16.50	19.29	17.86	15.35	17.02
2D-Lift (StreamingRVOS-1B)	31.60	32.11	23.56	30.45	21.45	28.14	26.48	28.16	25.15	29.65
MVGGT-causal	23.71	48.78	**39.08**	**70.22**	15.40	65.97	30.67	48.49	26.92	60.27
Ground3R	**48.59**	**65.74**	36.31	69.15	**44.18**	**78.17**	**36.95**	**49.96**	**41.44**	**67.58**

A.4 Grounding with Same-Class Distractors

Table 9 reports Global mIoU on the distractor subsets defined in Section 3.2. The "visible elsewhere" subset contains the remaining episodes with same-class distractors visible in the observation window, excluding those with distractors visible during target absence. Bold and underline mark the best and second-best online results. Ground3R leads the online methods on all three subsets. When same-class distractors are visible during target absence, it achieves 28.32% Global mIoU, exceeding 2D-Lift (StreamingRVOS-1B) by 9.14 points and MVGGT-causal by 11.06 points.

Table 9: Same-class distractor results (Global mIoU, %).

Method	No same-class distractor	Visible during target absence	Visible elsewhere in the window
Offline			
2D-Lift (ReferDINO)	33.80	18.70	32.55
2D-Lift (SSA)	23.89	15.47	25.06
MVGGT	62.73	33.59	41.66
Online			
2D-Lift (OnlineRefer)	16.78	12.50	18.59
2D-Lift (StreamingRVOS-1B)	31.81	19.17	28.37
MVGGT-causal	45.46	17.26	20.56
Ground3R	**55.53**	**28.32**	**39.69**

A.5 Expression Transfer to Nr3D and Sr3D

We evaluate expression transfer on subsets of Nr3D and Sr3D (Achlioptas et al., 2020), which provide human-written descriptions and synthetic spatial descriptions of ScanNet objects, respectively. The expressions are paired with chronological observation episodes and evaluated directly under the StreamRefer protocol without any fine-tuning. Table 10 compares Ground3R with MVGGT-causal and 2D-Lift (OnlineRefer) across observation lengths. Ground3R achieves 35.58% and 40.06% Global mIoU on Nr3D and Sr3D, outperforming MVGGT-causal by 13.78 and 15.05 points, respectively. On Sr3D's template-based spatial descriptions, Ground3R achieves 42.83% Global mIoU at 32 frames compared to 18.78% for MVGGT-causal. Ground3R leads at each observation length on both human-written and template-based expressions.

Table 10: Expression transfer on Nr3D and Sr3D (mIoU, %).

Method	**Global by observation length**			**Overall**	
	8	16	32	Global	View
Nr3D					
2D-Lift (OnlineRefer)	14.84	13.87	14.56	14.43	16.64
MVGGT-causal	24.04	22.70	18.64	21.80	55.19
Ground3R	**30.14**	**37.47**	**39.14**	**35.58**	**62.44**
Sr3D					
2D-Lift (OnlineRefer)	12.45	10.39	13.49	12.11	13.24
MVGGT-causal	31.27	24.96	18.78	25.01	56.99
Ground3R	**37.66**	**39.71**	**42.83**	**40.06**	**66.56**

A.6 Computational Efficiency

Per-frame inference time. For the inference time in Table 1, we process the first 31 frames and measure the time to process the newly arriving 32nd frame on a single NVIDIA RTX 3090. The timed interval starts with a decoded RGB image in CPU memory and includes image preprocessing, device transfers, inference, and returning masks and pointmaps to the CPU to extract the segmented 3D points. For 2D-Lift, video segmentation, CUT3R reconstruction, and mask lifting run within the same timed interval; MVGGT-causal re-evaluates the 32-frame prefix. We synchronize the GPU before and after timing, repeat each measurement three times with the preceding state rebuilt, and average the per-expression median times. A fixed evaluation subset includes both target-visible and target-absent final frames. Ground3R takes 128.0 ms, compared with 174.3 ms for 2D-Lift (OnlineRefer), 622.7 ms for 2D-Lift (StreamingRVOS-1B), and 3112.3 ms for MVGGT-causal. Offline methods use full-episode context and are omitted from this online inference-time comparison.

Scaling with sequence length. We compare throughput and peak GPU memory as the observation length increases from 8 to 512 frames within fixed temporal spans. Ground3R carries its recurrent state across frames and releases intermediate features after each update. Figure 4 shows approximately 10 FPS and 6.22–6.25 GiB of peak GPU memory throughout the evaluated lengths. MVGGT-causal recomputes the observed prefix at each step: its throughput decreases from 2.86 FPS at 8 frames to 0.28 FPS at 64 frames, and it runs out of GPU memory at 128 frames. Here throughput measures model processing of preprocessed images; Ground3R additionally includes host-to-device transfers.

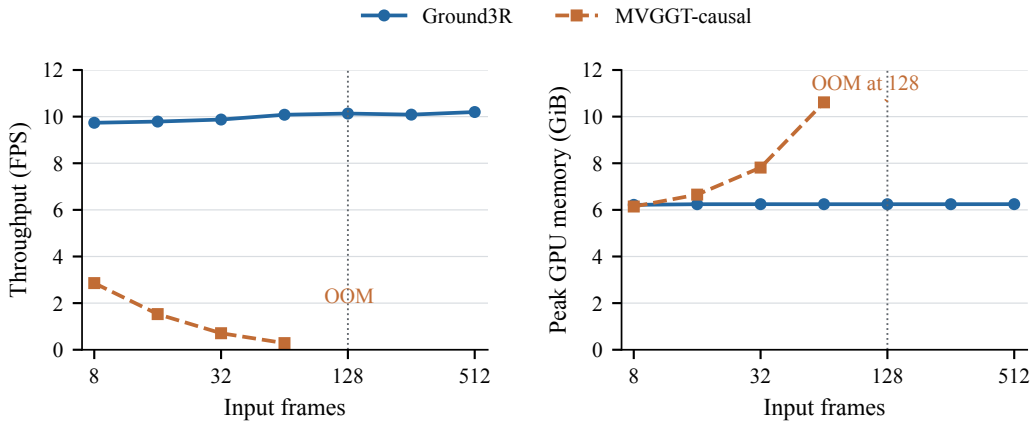


Figure 4: Throughput and peak GPU memory as observation length increases on an RTX 3090. OOM indicates an out-of-memory failure.

A.7 Visibility Threshold Sensitivity

We vary the threshold $\tau \in [0, 1]$ on the target-visibility probability p_t (Eq. 10); the main experiments use $\tau = 0.50$. During inference, a prediction is suppressed to an empty mask when its predicted visibility probability falls below τ. We report Global mIoU and Neg, the percentage of frames with an empty ground-truth mask on the evaluation grid that are correctly predicted as empty.

As shown in Figure 5, increasing τ progressively suppresses false-positive masks, raising Neg from 80.61% at $\tau = 0.30$ to 91.00% at $\tau = 0.70$. Across $\tau \in [0.30, 0.60]$, Global mIoU ranges from 40.65% to 41.98%. The default threshold $\tau = 0.50$ yields 86.54% Neg.

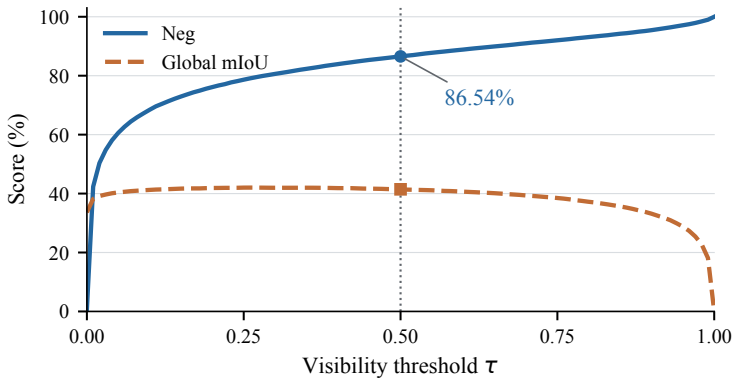


Figure 5: Global mIoU and Neg across visibility thresholds. Markers denote the default threshold $\tau = 0.50$.

A.8 Qualitative Results

Spatial references across views. Figures 6 and 7 show objects described relative to spatial references observed earlier in the sequence. In the cabinet and curtain examples, the trash can and cabinet are observed before the respective targets enter the view. In the stool and TV-stand examples, the door and chair leave the current view while the target remains visible. Ground3R segments the described objects as the observations accumulate. The point clouds contain predictions from the entire sequence; all methods' masks are displayed on the same Ground3R geometry to compare segmentation.

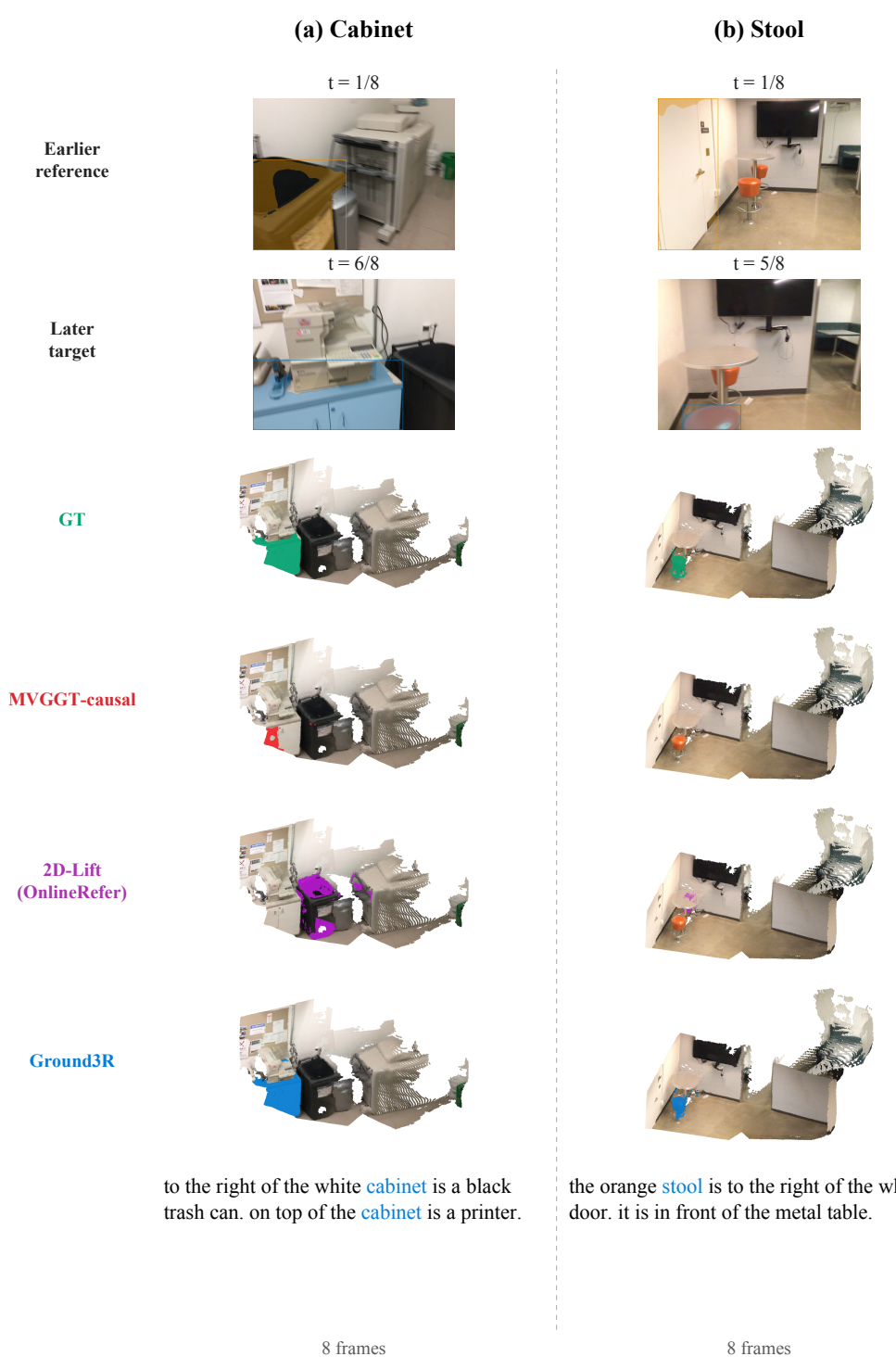


Figure 6: Grounding with spatial references across views. Amber and blue in the RGB images mark the reference and target, respectively. Point clouds accumulate all frames.

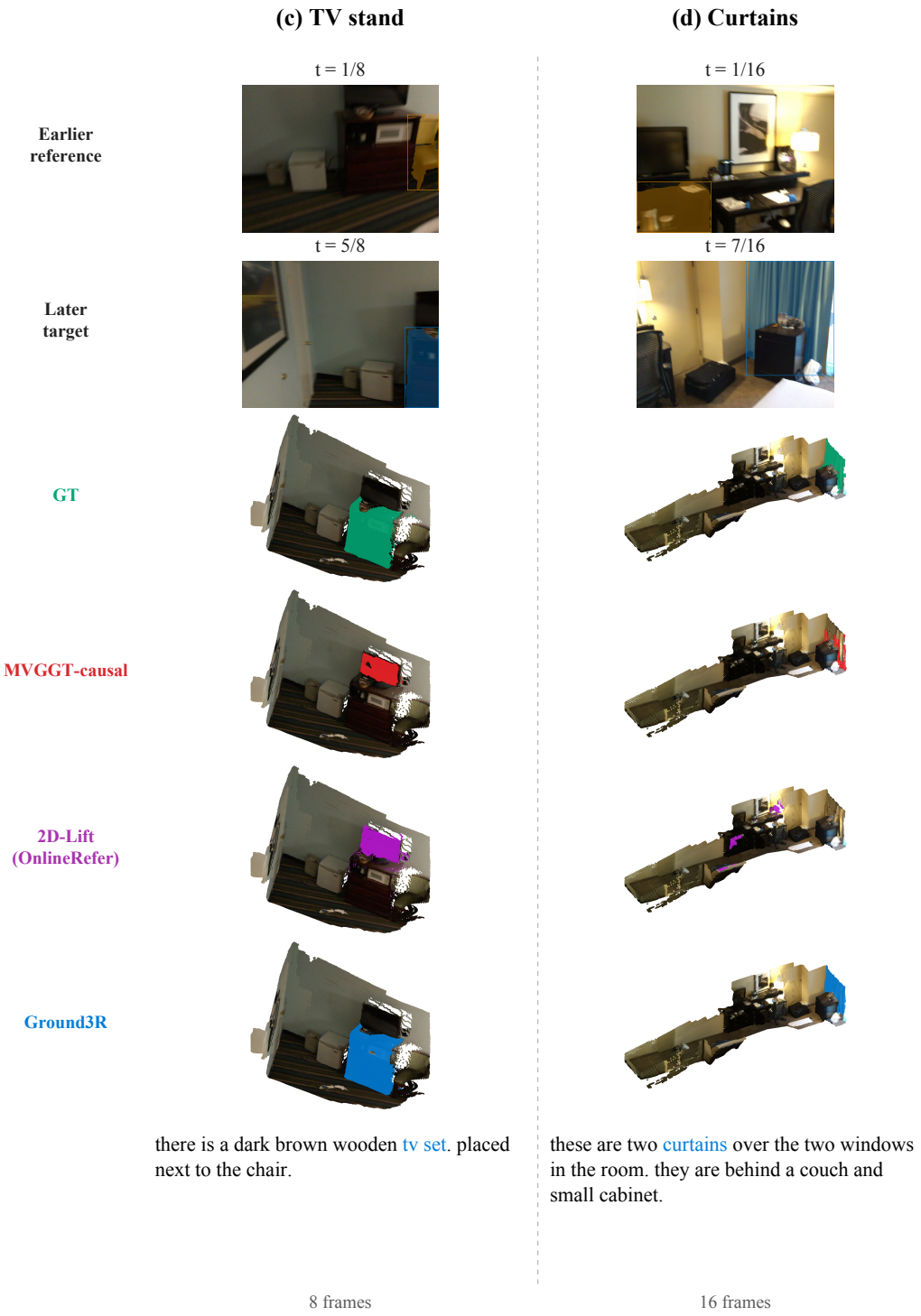
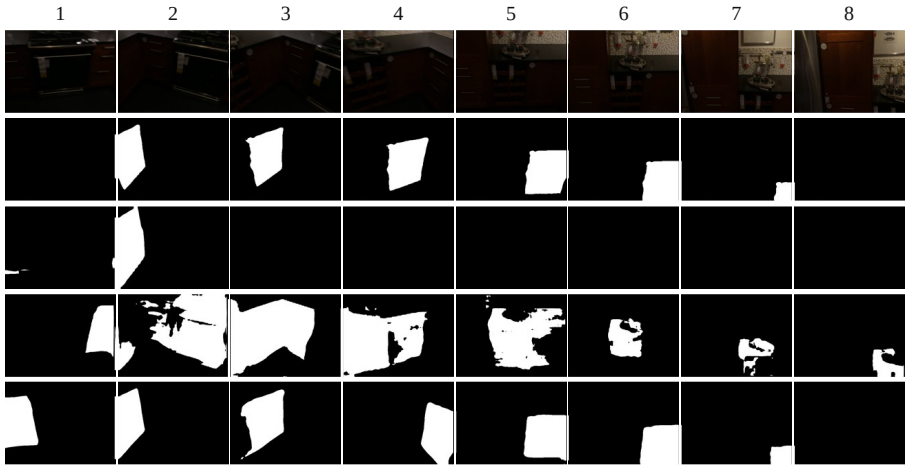


Figure 7: Grounding with spatial references across views (continued).

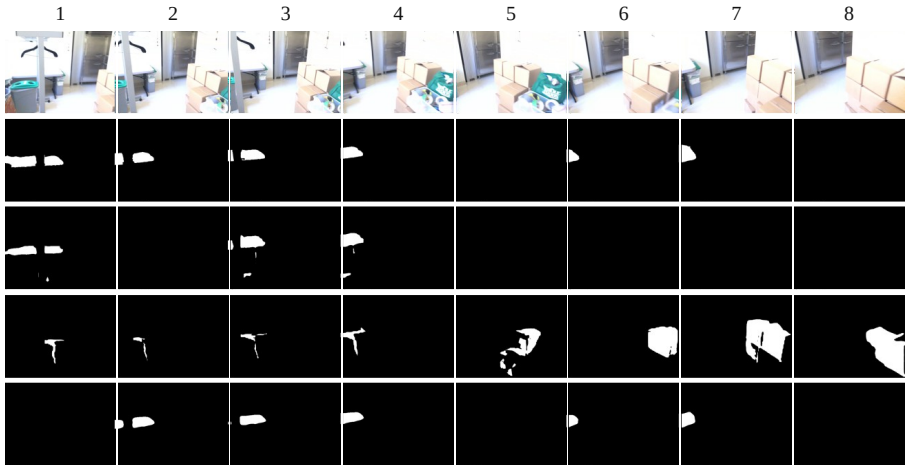
Lighting conditions. Figure 8 compares predictions for a cabinet in a dark scene and a table in a highlight-rich scene across eight chronological views.

(a) Dark scene0328_00 / object 15



Description. the kitchen is left of the stove. the **kitchen cabinet** is a dark brown box.

(b) Highlight-rich scene0412_00 / object 4



Description. the **table** is on the right side of the room. there are trash cans to the left of it.

Figure 8: Segmentation under (a) dark and (b) highlight-rich conditions. Highlighted words identify the target. 2D-Lift uses OnlineRefer.