S4VY: SEGMENT ANYTHING IN FEED-FORWARD 4D VISUAL GEOMETRY

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of S4VY.** *Left:* Given a set of RGB observations, S4VY predicts a consistent 4D mask for every instance in the scene. *Right:* Any instance can be grounded via either a natural-language prompt or an image-level point/box prompt.

ABSTRACT

Accurate instance segmentation in dynamic scenes is important for downstream applications such as robotics and autonomous driving. Existing Segment Anything models operate primarily on 2D image or video masks and preserve identity through sequential memory, while promptable 4D instance segmentation built upon feed-forward visual geometry remains underexplored. We introduce S4VY, a Segment Anything model built on feed-forward 4D visual geometry. From a set of RGB observations, S4VY transforms shared visual-geometric features into an exhaustive set of class-agnostic 4D instance masks through a space-time query decoder, with each persistent object query binding one entity across all observations. This representation supports prompt-independent segmentation as well as point- and box-conditioned selection, without requiring a seed mask or temporal ordering. We further develop an agentic harness for natural-language grounding in the large observation space of a 4D scene. Active tree search identifies relevant frames without scanning every fixed window; a dual-stream grounder combines fine-grained VLM visual priors with geometry-consistent instance features through complementary bounding-box prediction and object-query matching; and an independent critic selects the final 4D instance mask from their predictions. Extensive experiments demonstrate state-of-the-art 4D instance segmentation and strong language-guided grounding performance under a unified evaluation spanning static and dynamic scenes.

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

Instance-level segmentation of dynamic scenes is fundamental to perception and planning in embodied systems such as humanoid robots and autonomous vehicles. The SAM family (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) supports promptable segmentation across images and videos. Yet these models remain centered on predicting sets of 2D masks and propagating temporal memory to track instances among frames, without explicitly modeling the spatial geometry shared across observations. This becomes a critical bottleneck in challenging dynamic scenes, where multiple visually similar objects move simultaneously and must be re-identified after extended occlusion or absence from view. Real-world observations may also be sparse or unordered, further challenging sequential identity propagation. These settings call for object identities established through shared 3D/4D geometry, rather than inherited solely along an ordered sequence of 2D frames.

Segmentation on explicit 3D representations offers a direct route to spatial scene understanding (Jiang et al., 2020; Schult et al., 2023; Kolodiazhnyi et al., 2024; Choy et al., 2026), but typically assumes the inputs to be point clouds or RGB-D observations obtained from separate reconstruction pipelines or specialized sensors. This dependence potentially accumulates errors and limits data scalability. Recent feed-forward visual geometry models (Wang et al., 2024a; 2025a; 2026c;a) recover scene geometry directly from RGB observations in a single forward pass. Their consistent geometry representations have enabled methods that couple feed-forward reconstruction with semantic, panoptic, or instance segmentation (Zust et al., 2025; Li et al., 2025a; Qu et al., 2026; Zou et al., 2026). Within this visual-geometry paradigm, most of them target static scenes. IGGT4D (Zou et al., 2026) extends instance segmentation to dynamic 4D scenes, but associates local masks with an additional feature clustering, while flexible prompt interfaces remain underexplored.

To address these limitations, we present S4VY (pronounced "savvy"), named for its goal of segmenting anything in 4D visual geometry. As shown in Fig. 1, S4VY is a Segment Anything model that unifies prompt-free, exhaustive 4D instance segmentation with prompt-conditioned selection of any discovered instance in a shared 4D visual-geometric representation. Our segmentation model operates on geometry features jointly contextualized across all input observations by a feed-forward visual geometry backbone (Wang et al., 2026a), regardless of their order. A Space-Time Query Decoder uses global learnable object queries to predict instance masks, with each active query maintaining one instance identity across the entire observation set. A Prompt Encoder maps point or box prompts from one or multiple observations to a promptable object query,enabling any object in the observed scene to be selected.

Beyond point and box prompts, natural-language grounding provides an important interface for users and downstream systems to specify target instances in 4D scenes. A referent may appear in only a limited subset of a large observation set, making target discovery and consistent 4D association challenging. Vision-Language Models (VLMs) have advanced reasoning segmentation and visual grounding (Lai et al., 2024; Rasheed et al., 2024; Li et al., 2025c; Liu et al., 2025), while MVGGT (Wu et al., 2026a) extends referring segmentation to sparse multiview RGB inputs. These methods assume static scenes or fixed sparse views, leaving target discovery and consistent grounding over large dynamic observation sets unresolved. We develop an agentic grounding harness specifically designed for feed-forward 4D visual geometry. Its dual-stream Grounder combines the geometry-consistent feature space of the segmentation model with the fine-grained visual priors of the VLM, producing complementary object-query and bounding-box predictions. Guided by the Grounder's target-presence estimates, Active Tree Search recursively locates relevant observations across the frame space without evaluating every fixed window. A comparative Critic jointly evaluates representative overlays of the two mask candidates and selects the final 4D instance.

Our contributions are:

- **Feed-forward Segment Anything in 4D.** We introduce a unified model that transforms shared visual-geometric representations into exhaustive class-agnostic 4D instances with persistent identities, while allowing any discovered instance to be selected by point or box prompts.
- **Agentic language grounding in 4D.** We develop a grounding harness for large 4D observation spaces that combines VLM visual priors with geometry-consistent instance features through dual-stream prediction, active tree search, and comparative critic selection.
- **Evaluation across static and dynamic scenes.** Extensive experiments establish state-of-the-art 4D instance segmentation and strong language grounding performance under unified evaluation.

2 Related Work

Feed-forward multi-view geometry and understanding. With the scaling of vision transformers and training data, multi-view geometry has rapidly shifted from optimization pipelines to feed-forward reconstruction. DUSt3R (Wang et al., 2024a) initiated the modern line through pointmap regression, after which stronger correspondence, whole-set aggregation, and greater scale steadily widened its reach (Leroy et al., 2024; Wang & Agapito, 2025; Cabon et al., 2025; Yang et al., 2025). VGGT (Wang et al., 2025a) then brought cameras, depth, pointmaps, and tracks into one transformer. Geometry became a shared multi-view feature space. π^3 (Wang et al., 2026c) further removes reference-view dependence through permutation-equivariant visual geometry prediction, while VGGT-Ω (Wang et al., 2026a) extended this shared geometric representation across static and dynamic scenes. These representations have quickly become substrates for scene understanding, with semantic and panoptic prediction increasingly coupled to feed-forward reconstruction (Fan et al., 2024; Sun et al., 2026; Zust et al., 2025). At the instance level, IGGT (Li et al., 2025a) groups geometry-aware features, whereas SegVGGT (Qu et al., 2026) and FAST3DIS (Li et al., 2026a) decode object queries; IGGT4D (Zou et al., 2026) is among the few methods to extend joint reconstruction and instance understanding explicitly to 4D. This line differs from methods that segment an existing point cloud or space–time scan (Choy et al., 2019; Jiang et al., 2020; Aygun et al., 2021; Vu et al., 2022; Kreuzberg et al., 2023; Schult et al., 2023; Kolodiazhnyi et al., 2024; Choy et al., 2026; Steiner et al., 2026), as well as approaches that lift 2D masks into a preconstructed neural or Gaussian scene (Siddiqui et al., 2023; Bhalgat et al., 2023; Ye et al., 2024; Li et al., 2026d; Wu et al., 2026b). Within this paradigm, most methods remain limited to static multi-view scenes, while the closest dynamic extension relies on clustering-based identity association. Moreover, existing methods do not unify exhaustive 4D instance discovery with point, box, and language interfaces. S4VY addresses these limitations within a shared feed-forward 4D representation.

Segment Anything models. SAM (Kirillov et al., 2023) established promptable image segmentation, while subsequent models extended its interface to text and controllable mask granularity and improved segmentation quality, efficiency, and robustness (Zou et al., 2023; Li et al., 2024a; Ke et al., 2023; Xiong et al., 2024; Chen et al., 2023; Zhang et al., 2026). Video extensions either attach tracking or temporal memory to image masks (Yang et al., 2023a; Cheng et al., 2023; Rajič et al., 2025; Ravi et al., 2025), with SAM 3 (Carion et al., 2026) further supporting text and exemplar concepts, or represent instances with queries shared across a clip (Cheng et al., 2021; Wu et al., 2022b); both remain defined over 2D image or video representations. SAM-inspired 3D and 4D methods consume explicit point clouds, RGB-D observations, or camera–LiDAR streams (Cen et al., 2023; Yin et al., 2024; Zhou et al., 2024; Xu et al., 2025; Liu et al., 2023), or lift 2D masks into preconstructed neural or Gaussian scenes (Ye et al., 2024; Ji et al., 2024). SAM-V (Gong et al., 2026) integrates feed-forward geometry into SAM but targets prompted multi-view segmentation of static scenes. S4VY instead unifies exhaustive 4D instance discovery with point- and box-conditioned selection over potentially unordered RGB observations.

Language-guided segmentation and 4D scene grounding. Grounding DINO (Liu et al., 2024) and language-driven segmentation models such as CRIS (Wang et al., 2022) and Language as Queries (Wu et al., 2022a) align free-form text with image regions, masks, and video tracks. Related 3D and 4D work grounds text or open-vocabulary features in scene representations (Prabhudesai et al., 2020; Chen et al., 2020; Achlioptas et al., 2020; Zhao et al., 2021; Wu et al., 2023; Peng et al., 2023; Takmaz et al., 2023; Fiebelman et al., 2025; Li et al., 2025d; Zhou et al., 2025; Wu et al., 2025). With stronger VLMs, recent work has expanded grounding from direct semantic matching toward free-form descriptions, compositional relations, and implicit intent through reasoning segmentation, geometry-aware 3D localization, and language-guided 4D segmentation (Lai et al., 2024; Xia et al., 2024; Rasheed et al., 2024; Ren et al., 2024; Yan et al., 2024; Zhu et al., 2024; Chen et al., 2024; Yang et al., 2024; Xu et al., 2024; Li et al., 2025c; Liu et al., 2025; Wu et al., 2026a; Hu et al., 2026; Li et al., 2025b; Yuan et al., 2026b; Tang et al., 2026; Li et al., 2026b). Agentic long-video methods iteratively retrieve query-relevant observations or refine visual evidence through hierarchical and active observation (Wang et al., 2024b; 2025b; Li et al., 2026c; Wang et al., 2026d). Most existing scene-grounding methods operate on static scenes or fixed sparse views, while agentic video systems search an ordered timeline for video-level reasoning. S4VY instead grounds persistent instances across large 4D observation sets, incorporating Active Tree Search with VLM visual priors, geometry-consistent features, and Critic selection.

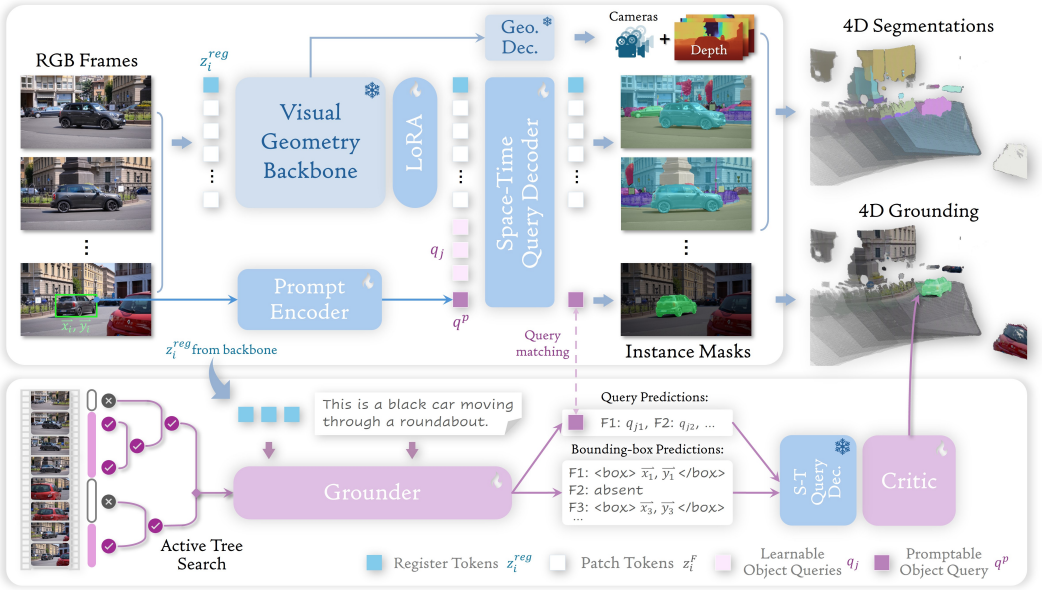


Figure 2: **Architecture of S4VY.** *Top:* The backbone encodes RGB observations into register and patch tokens. The Space-Time Query Decoder applies learnable object queries to the patch tokens to produce exhaustive instance masks across observations, while the Prompt Encoder maps image-level prompts to a promptable query. *Bottom:* For a language expression, Active Tree Search selects observation windows for the dual-stream Grounder, which predicts object-query matches and bounding boxes. The two predictions are decoded into mask candidates and compared by the Critic.

3 Segment Anything in 4D

As illustrated in Fig. 2, S4VY builds a query-based 4D segmenter on feed-forward visual geometry, supporting exhaustive 4D instance segmentation as well as point- and box-prompted segmentation.

3.1 Preliminaries and Problem Formulation

Feed-forward visual geometry. Given T RGB observations $\mathcal{I} = \{I_i\}_{i=1}^{T}$, where $I_i \in \mathbb{R}^{3\times H\times W}$, our feed-forward visual geometry backbone (Wang et al., 2026a) represents observation i with patch tokens $\mathbf{z}_i^F$, a camera token $\mathbf{z}_i^{\text{cam}}$, and register tokens $\mathbf{z}_i^{\text{reg}}$, and jointly contextualizes them through alternating within-frame and cross-frame attention. The segmenter decodes the contextualized patch tokens, while the register tokens are retained for the language-grounding interface in Section 4.

Exhaustive 4D instance segmentation. Let $\mathcal{S} \subset \mathbb{R}^3 \times \{t_i\}_{i=1}^{T}$ denote the spatiotemporal scene observed by $\mathcal{I}$, where t_i is the acquisition time of I_i. The input observations may be temporally sparse or presented out of chronological order. The task is to recover

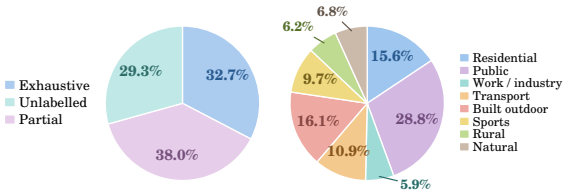
$$\mathcal{I} \mapsto \mathcal{M} = \{\mathbf{M}_j\}_{j=1}^{N}, \qquad \mathbf{M}_j : \mathcal{S} \to \{0,1\}, \tag{1}$$

where $\mathbf{M}_j$ denotes the 4D mask of instance j; exhaustive segmentation recovers the complete set $\mathcal{M}$ of all N instances in the observed scene. Following the SAM paradigm (Chen et al., 2023; Ravi et al., 2025), a point or box prompt can instead specify any individual $\mathbf{M}_j$.

3.2 Network Architecture

Space-Time Query Decoder. Our Space-Time Query Decoder exploits the rich cross-view and temporal correspondences encoded by feed-forward visual geometry to recover persistent 4D instances. We formulate this process as end-to-end set prediction following Mask2Former (Cheng et al., 2022), avoiding the post-clustering identity association used by the IGGT series (Li et al., 2025a; Zou et al., 2026). A set of global learnable object queries $\{\mathbf{q}_j\}$ jointly attends to the LoRA-adapted (Hu et al., 2021) contextualized patch tokens $\{\mathbf{z}_i^F\}_{i=1}^{T}$. The Space-Time Query Decoder f_{ST} predicts the complete 4D mask of each query as

$$\hat{\mathbf{M}}_j = f_{\text{ST}}\big(\mathbf{q}_j, \{\mathbf{z}_i^F\}\big)\,. \tag{2}$$



(a) Annotation coverage (b) Scene types

Supervision	Type		Origin		Scale	
	Sta.	Dyn.	Real	Syn.	Scenes	Frames
Exhaustive multi-view	✓	–	✓	✓	7.50k	1.50M
Exhaustive video	–	✓	✓	–	3.35k	83.1k
Partial video	–	✓	✓	✓	12.63k	958.6k
Unlabelled multi-view	✓	–	✓	–	9.74k	3.51M

(c) Data types and scale

Figure 3: **Training corpus.** (a & b) Annotation coverage and dominant scene types among the curated data. (c) Data composition and scale by supervision type.

During training, bipartite matching assigns each ground-truth 4D instance to a unique learnable object query, with unmatched predictions assigned to the no-object class.

Prompt encoder. To support image-level prompting, the prompt encoder f_{P} combines point or box prompts p from one or more observations by sampling local visual features from $\{\mathbf{z}_i^F\}_{i=1}^T$ to produce a promptable object query $\mathbf{q}^p$, which is appended to the learnable object queries $\{\mathbf{q}_j\}$ at the decoder input:

$$\mathbf{q}^p = f_{\mathrm{P}}\big(\{\mathbf{z}_i^F\}, p\big)\,, \qquad \hat{\mathbf{M}}^p = f_{\mathrm{ST}}\big(\mathbf{q}^p, \{\mathbf{q}_j\}, \{\mathbf{z}_i^F\}\big)\,. \tag{3}$$

Unidirectional self-attention lets $\mathbf{q}^p$ match the corresponding learnable object query while preventing potentially ambiguous prompt signals from affecting $\{\mathbf{q}_j\}$, preserving prompt-free instance discovery. Since prompts supplied by users or downstream localization tools may be spatially imprecise, we perturb point locations and box coordinates during training to improve tolerance to localization errors.

3.3 Data Curation and Model Optimization

Data curation. Learning to segment any instance in 4D requires broad supervision across scene domains and motion patterns. We consolidate 19 existing multi-view and video datasets into four complementary groups: exhaustive multi-view data from the released splits of InsScene-15K (Li et al., 2025a) including Infinigen (Raistrick et al., 2023), ScanNet++ (Yeshwanth et al., 2023), and RealEstate10K (Zhou et al., 2018); exhaustive video annotations from VIPSeg (Miao et al., 2022), Cityscapes-VPS (Kim et al., 2020), KITTI-STEP (Weber et al., 2021), and JRDB-PanoTrack (Le et al., 2024); partial video masklets from SA-V (Ravi et al., 2025), YouTube-VOS (Xu et al., 2018), MeViS (Ding et al., 2023a), MOSE (Ding et al., 2023b), UVO-Dense (Wang et al., 2021), DynamicReplica (Karaev et al., 2023), SAIL-VOS (Hu et al., 2019), LVOS (Hong et al., 2023), DAVIS (Perazzi et al., 2016; Pont-Tuset et al., 2017), and VOST (Tokmakov et al., 2023); and unlabelled multi-view observations from DL3DV (Ling et al., 2024) and Matterport3D (Chang et al., 2017).

As shown in Fig. 3(a), exhaustively annotated sources supervise the discovery of every instance, while partial and unlabelled sources are densified with temporally consistent pseudo-masklets generated by a teacher trained on the exhaustive subset. The scene-type and data distributions in Fig. 3(b,c) cover static and dynamic content as well as real and synthetic imagery. In total, this curation yields around $6M$ frames from $33k$ scenes.

Optimization. After Hungarian matching (Cheng et al., 2022; Carion et al., 2020) between predicted queries and ground-truth instances, we optimize the fully supervised set-prediction objective

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{obj}} + \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{Dice}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{obj}}$ is the object/no-object classification loss, $\mathcal{L}_{\mathrm{BCE}}$ is the binary cross-entropy mask loss, and $\mathcal{L}_{\mathrm{Dice}}$ is the soft Dice loss (Milletari et al., 2016) over the matched masks.

4 Ground Anything in 4D

Given a natural-language prompt x, 4D grounding aims to select the corresponding 4D mask $\mathbf{M}^*$ of the referred object throughout the observed scene masklets.

4.1 Grounding Components

Grounder. Vision-Language Models (VLMs) provide strong spatial priors for grounding referred objects in individual images, yet their native RGB interface does not exploit the cross-view consistency encoded by feed-forward visual geometry. The backbone register tokens form a compact,

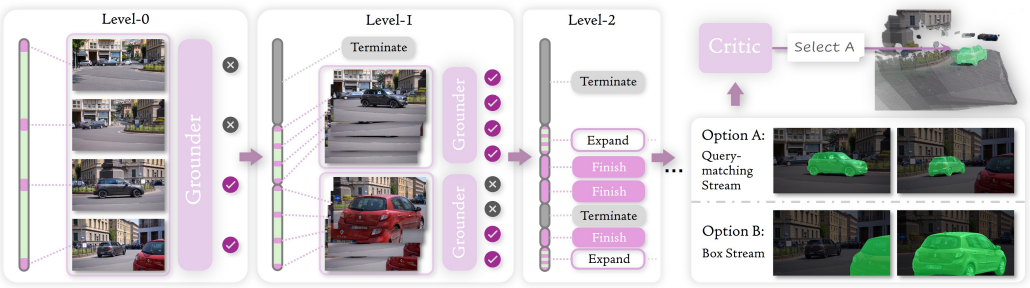


Figure 4: **Agentic grounding harness.** Active tree search locates relevant observations, the dual-stream grounder produces two 4D mask candidates, and the critic selects the final result. The illustrated example uses $N_w = 4$ observations per grounder window.

language-conditioned scene representation (Wang et al., 2026a). Supplying all patch tokens from multiple observations would introduce prohibitive context redundancy; we therefore inject the per-frame camera and register tokens as compact geometric context. We further supervise the grounder with a scene-geometry-guided chain of thought (CoT) (Wei et al., 2022) distilled from visual and metric scene cues, including object size, height, and inter-object distance, assisting scene reasoning and query and box predictions.

Our grounder retains the input flexibility of the base segmenter, accommodating conventional temporal sequences as well as sparse, non-consecutive observations. Each training window is formed from multiple observations sampled across $\mathcal{I}$ rather than restricted to consecutive frames. Let N_{w} denote the window size and $s_1, \ldots, s_{N_{\mathrm{w}}} \in \{1, \ldots, T\}$ the sampled observation indices. The dual-stream prediction is summarized as

$$\{(\hat{\mathbf{q}}_{s_k}, \hat{p}_{s_k})\}_{k=1}^{N_{\mathrm{w}}} = f_{\mathrm{G}}\left(x, \left\{I_{s_k}, \mathbf{z}_{s_k}^{\mathrm{cam}}, \mathbf{z}_{s_k}^{\mathrm{reg}}\right\}_{k=1}^{N_{\mathrm{w}}}\right), \tag{5}$$

where $\hat{\mathbf{q}}_{s_k}$ denotes the query prediction and $\hat{p}_{s_k}$ the predicted box prompt; a null query or absent prompt indicates that the referent is not present. In the query-matching stream, the query predictions are matched to a persistent object query; in the box stream, the predicted box prompts are jointly encoded by f_{P} into $\mathbf{q}^p$. The two streams yield complementary 4D mask candidates through f_{ST}.

Critic. Given the 4D mask candidates $\hat{\mathbf{M}}^q$ and $\hat{\mathbf{M}}^p$ produced by the query-matching stream and box stream, respectively, we render each candidate over the observed frames. An independent VLM-based critic f_{C} predicts their pairwise preference and selects the candidate that better matches the expression x:

$$s_{\mathrm{C}} = f_{\mathrm{C}}\left(x, \hat{\mathbf{M}}^q, \hat{\mathbf{M}}^p\right), \qquad \hat{\mathbf{M}}^* = \begin{cases} \hat{\mathbf{M}}^q, & s_{\mathrm{C}} > 0, \\ \hat{\mathbf{M}}^p, & \text{otherwise}, \end{cases} \tag{6}$$

where s_{C} is the pairwise preference logit, with a positive value favoring $\hat{\mathbf{M}}^q$. The critic is fine-tuned on candidate pairs, with the candidate that better agrees with the ground-truth mask providing the selection target.

Optimization. The grounder is optimized with cross-entropy for query matching and an autoregressive language-modeling loss for its generated outputs, while the critic uses binary cross-entropy to select the candidate with greater ground-truth overlap:

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{query}} + \mathcal{L}_{\mathrm{LM}}, \qquad \mathcal{L}_{\mathrm{C}} = \mathrm{BCE}(s_{\mathrm{C}}, y_{\mathrm{C}}), \tag{7}$$

where $y_{\mathrm{C}} = 1$ when $\hat{\mathbf{M}}^q$ has greater ground-truth overlap and 0 otherwise.

4.2 Agentic Grounding Harness

Grounding in a large 4D scene requires locating the observations that contain the referent and associating them with a consistent instance. Processing large observation sets with fixed sliding windows repeatedly evaluates irrelevant content. Agentic video methods mitigate this cost by exploiting temporal continuity to retrieve relevant evidence along an ordered timeline (Wang et al., 2024b; 2025b). However, our segmenter also accepts sparse or non-consecutive observations, for which temporal prior information is weakened. We therefore introduce Active Tree Search (ATS), driven by predicted target presence, to explore the observation set adaptively.

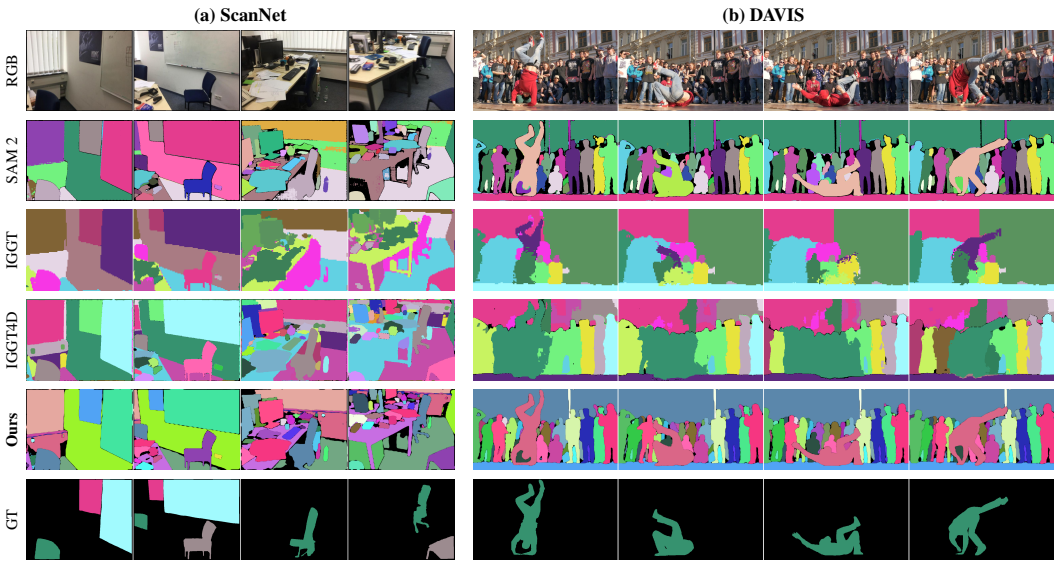


Figure 5: **Qualitative comparison of class-agnostic segmentation.** Within each row, masks of the same color share an instance identity; ground-truth masks cover only annotated instances. S4VY remains robust in both static 3D and dynamic 4D scenes.

As illustrated in Fig. 4, ATS forms each grounder window by uniformly sampling N_{w} observations from the current search range for joint evaluation. It begins with a global window spanning the full observation set. At each level, the grounder predicts target presence for every sampled observation; the neighboring sample positions around each positive prediction delimit the subranges examined at the next level. Each selected subrange is recursively resampled as a new grounder window, allowing the grounder to determine the next search step from its current predictions. The recursion stops once each grounder window contains only adjacent frames. Predictions from all examined windows are aggregated by voting to form two mask candidates. Representative observations are then selected, and the candidate masks are rendered as overlays and presented jointly with the expression x to the critic for comparative evaluation, determining the final mask $\hat{\mathbf{M}}^{*}$.

5 Experiments

5.1 Experimental setup

Benchmarks and metrics. We evaluate prompt-free instance segmentation on the ScanNet validation split from (Li et al., 2025a), DAVIS (Pont-Tuset et al., 2017), LVOS (Hong et al., 2023), and VIPSeg (Miao et al., 2022): T-mIoU, T-SR, and $\mathcal{J}\&\mathcal{F}$ on ScanNet; 4D-IoU, 4D-SR, and $\mathcal{J}\&\mathcal{F}$ on DAVIS and LVOS; and STQ (Weber et al., 2021) and mAP on VIPSeg. For language-guided grounding, our unified 3D/4D evaluation comprises ScanRefer (Chen et al., 2020), our annotated ScanNet++ split (Yeshwanth et al., 2023), Ref-DAVIS (Khoreva et al., 2018), and MeViS (Ding et al., 2023a), using 3D mask mIoU and box/mask accuracy at IoU 0.25 for static scenes and J&F for dynamic scenes.

Baselines. For segmentation, we compare lifted 3D segmentation (Yang et al., 2023b), image/video systems (Cheng et al., 2023; Wu et al., 2024; Li et al., 2024b; Ravi et al., 2025), and feed-forward visual-geometry methods (Zust et al., 2025; Li et al., 2025a; Qu et al., 2026; Zou et al., 2026; Gong et al., 2026). Grounding baselines comprise VLM and 3D proposal pipelines (Xu et al., 2024; Drozdov et al., 2026; Ravi et al., 2025), end-to-end image/video referring models (Carion et al., 2026; Yuan et al., 2026a), and multi-view geometry-language alignment (Wu et al., 2026a).

Implementation details. The segmenter initializes from the VGGT-Ω-1B checkpoint at 512-pixel resolution. We adapt it with rank-256 LoRA (Hu et al., 2021) adapters and train the segmenter for 50,000 steps on our curated training datasets. The grounder initializes from Qwen3-VL-4B (Bai et al., 2025) and is fully fine-tuned for 4 epochs on approximately 80k expressions from ScanRefer, our ScanNet++ annotations, Ref-DAVIS, MeViS, etc. The critic initializes from Qwen3-VL-2B and is trained separately on paired mask candidates from the corresponding training splits.

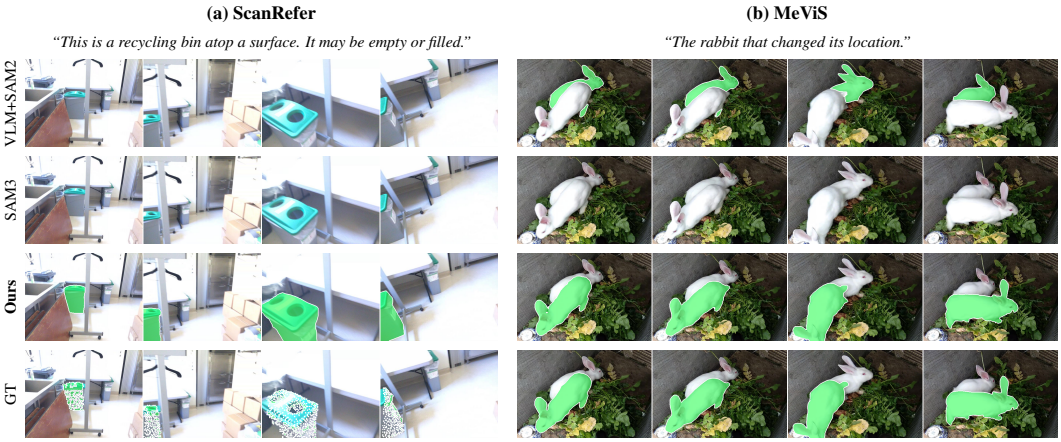


Figure 6: **Qualitative comparison of language-guided grounding.** ScanRefer ground-truth masks are projected from the annotated 3D point cloud.

Table 1: **Class-agnostic instance segmentation.** All methods are evaluated without target prompts or first-frame ground truth. Higher is better for segmentation metrics; latency is measured over 100 observations on an A100-80GB (s/clip).

Method	ScanNet			DAVIS			LVOS			VIPSeg		Latency
	T-mIoU	T-SR	$\mathcal{J}\&\mathcal{F}$	4D-IoU	4D-SR	$\mathcal{J}\&\mathcal{F}$	4D-IoU	4D-SR	$\mathcal{J}\&\mathcal{F}$	STQ	mAP	@100
SAM3D (Yang et al., 2023b)	0.686	0.296	0.574	–	–	–	–	–	–	–	–	–
DEVA (Cheng et al., 2023)	0.180	0.029	0.219	0.531	0.584	0.531	0.373	0.312	0.412	0.583	0.047	32.3
GLEE (Wu et al., 2024)	0.127	0.038	0.178	0.710	0.870	0.754	0.329	0.225	0.367	0.207	0.214	36.8
OMG-Seg (Li et al., 2024b)	0.189	0.067	0.218	0.632	0.789	0.691	0.167	0.102	0.191	0.554	0.205	13.7
SAM 2 (Ravi et al., 2025)	0.635	0.438	0.588	0.703	0.823	0.750	0.450	0.460	0.444	0.700	0.249	175
PanSt3R (Zust et al., 2025)	0.747	0.708	0.719	0.484	0.506	0.513	0.297	0.120	0.315	0.698	0.197	54.0
IGGT (Li et al., 2025a)	0.517	0.408	0.534	0.348	0.302	0.407	0.139	0.040	0.156	0.622	0.089	272
SegVGGT (Qu et al., 2026)	0.561	0.471	0.494	0.178	0.022	0.259	0.069	0.000	0.069	0.340	0.069	–
IGGT4D (Zou et al., 2026)	0.781	0.708	0.726	0.543	0.667	0.604	0.154	0.047	0.192	0.729	0.190	18.5
SAM-V (Gong et al., 2026)	0.729	0.692	0.685	0.307	0.145	0.336	0.197	0.070	0.174	0.344	0.083	–
S4VY (Ours)	**0.803**	**0.792**	**0.781**	**0.715**	**0.914**	**0.803**	**0.533**	**0.573**	**0.602**	**0.795**	**0.352**	**10.1**

5.2 Qualitative comparisons

Figure 5 shows that S4VY produces more complete, identity-consistent masks in both static and dynamic scenes. Even with automatic mask generation and incremental object discovery, SAM 2 misses newly visible objects or confuses their identities across non-contiguous indoor views; feed-forward geometry baselines handle such views better but struggle with dynamic outdoor objects. Figure 6 further demonstrates more robust grounding than VLM-assisted segmentation baselines for references requiring spatial reasoning.

5.3 Quantitative comparisons

Class-agnostic segmentation. Table 1 compares prompt-free instance segmentation across static multi-view scenes and dynamic videos. S4VY leads every benchmark: its 0.803 T-mIoU and 0.792 T-SR on ScanNet reflect accurate and reliable cross-view instance recovery, while 0.803 and 0.602 J&F on DAVIS and LVOS show that the complete model maintains identity-consistent segmentation in dynamic scenes without temporal memory propagation. The corresponding gains in STQ and mAP further demonstrate strong performance under dense panoptic evaluation.

Language-guided grounding. Table 2 compares S4VY with VLM–segmenter pipelines, end-to-end language-guided segmenters, and Z3D's explicit 3D-proposal grounding. S4VY consistently outperforms the first two families across all four benchmarks, despite VLM+SAM 2 using a much larger Qwen3-VL-30B-A3B (Bai et al., 2025). Z3D approaches our performance only when supplied with benchmark GT point clouds, and still trails our 3D mIoU by 3.0 points on ScanRefer and 2.8 points on ScanNet++. Replacing GT geometry with VGGT-Ω reconstructions reduces its 3D mIoU from 0.431 to 0.114 and from 0.266 to 0.157, respectively, exposing the sensitivity of direct

Table 2: **Language-guided 3D/4D grounding.** Static benchmarks report 3D mIoU and box/mask accuracy at IoU 0.25; dynamic benchmarks report J&F. Z3D is evaluated with either benchmark GT point clouds or point clouds reconstructed by VGGT-Ω.

Method	VLM Size	ScanRefer			ScanNet++			Ref-DAVIS	MeViS
		3D mIoU	Box@.25	Mask@.25	3D mIoU	Box@.25	Mask@.25	J&F	J&F
VLM-Grounder (Xu et al., 2024)	30.5B	0.105	26.0	17.0	0.063	12.4	9.9	0.183	0.141
VLM+SAM 2 (Ravi et al., 2025)	30.5B	0.308	34.9	42.6	0.170	15.1	26.2	0.709	0.483
Sa2VA-8B (Yuan et al., 2026a)	8B	0.124	14.9	17.9	0.040	3.4	6.4	0.732	0.542
SAM 3 (Carion et al., 2026)	N/A	0.247	29.7	33.3	0.172	19.5	26.3	0.567	0.369
MVGGT (Wu et al., 2026a)	N/A	0.325	35.3	50.1	0.090	7.8	14.8	0.005	0.003
Z3D (GT) (Drozdov et al., 2026)	30.5B	0.431	52.4	58.1	0.266	**33.7**	40.2	–	–
Z3D (recon.) (Drozdov et al., 2026)	30.5B	0.114	26.1	16.1	0.157	31.1	28.6	0.147	0.180
S4VY (Ours)	4B+2B	**0.461**	**56.2**	**59.4**	**0.293**	30.4	**44.9**	**0.755**	**0.573**

Table 3: **Ablation studies.** (a) Grounder inputs and (b) output streams study on ScanRefer; (c) ATS versus sliding-window traversal. Presence is the mean percentage of observations containing the target, while Sliding and ATS report the mean number of grounder windows. ΔQ and ΔB report how the query- and box-stream scores change when sliding-window traversal is replaced by ATS, in percentage points; positive values favor ATS.

(a) Grounder inputs

Input	3D mIoU	Mask@.5	Box@.5
Single w/o $\mathbf{z}^{\text{reg}}$	0.206	22.9	20.7
Single	0.370	42.0	37.7
Multi	0.398	45.2	40.3
Multi + CoT	**0.407**	**46.0**	**41.1**

(b) Grounder outputs

Output	3D mIoU	Box @.25	Mask @.25
Query	0.406	49.3	51.8
Box	0.434	53.1	55.0
+ Critic	**0.461**	**56.2**	**59.4**

(c) Active Tree Search

Benchmark	Presence	Sliding	ATS	ΔQ	ΔB
ScanNet++	19.1	23.0	8.2	−1.9	−1.7
ScanRefer	27.9	28.5	9.7	−1.7	+2.7
MeViS	88.1	15.5	11.4	+0.7	+0.1
Ref-DAVIS	93.7	17.6	15.8	+3.7	+0.3

3D grounding to imperfect point clouds. In contrast, S4VY aligns VLM priors directly with the consistent feature space of feed-forward visual geometry.

5.4 Ablation studies

Input order robustness. On the same sparse frame sets, S4VY obtains 0.807/0.800 $\mathcal{J}\&\mathcal{F}$ before/after shuffling on DAVIS and 0.645/0.656 on LVOS, while SAM 2 obtains 0.766/0.749 and 0.579/0.559, respectively. This stability is consistent with joint space-time queries that do not propagate identity along a fixed temporal order.

Grounding architecture. Panels (a) and (b) of Table 3 ablate the grounder's inputs and output streams. Register tokens yield the largest gain (+17.0 Box@.5), demonstrating the importance of exposing the backbone's compact, geometry-consistent scene representation to the Grounder. Jointly reasoning over multiple observations adds another 2.6 points over single-frame inference, showing that cross-view context helps resolve the referred instance, while geometry-guided CoT contributes a further 0.8 points. The Critic raises 3D mIoU from 0.434 for the stronger box stream to 0.461, indicating that query matching and box prediction provide complementary grounding evidence.

Active Tree Search. Panel (c) relates target presence to search cost. On ScanRefer and ScanNet++, where targets appear in only 27.9% and 19.1% of the observations, ATS has roughly 60% fewer windows on average. On the video benchmarks, its mean window count remains below that of sliding windows, while query-stream accuracy is preserved or improved.

6 Conclusion

We introduced S4VY, a feed-forward Segment Anything model for 4D scenes. By exploiting a shared visual-geometric representation of an RGB observation set, S4VY exhaustively segments class-agnostic instances with persistent identities and supports point- and box-based selection without temporal ordering or memory propagation. We further developed an agentic language-grounding harness that combines VLM visual priors with geometry-consistent instance features, actively searches large observation spaces, and resolves complementary grounding predictions. Experiments across static and dynamic scenes demonstrate state-of-the-art 4D instance segmentation and strong language-guided grounding performance under our unified evaluation. A limitation is that jointly processing multiple high-resolution observations constrains the output mask resolution.

USE OF LARGE LANGUAGE MODELS

LLMs were used to generate and clean a subset of data captions, and our grounding method builds on open-source MLLMs. LLM-based tools also assisted with selected coding tasks and manuscript polishing. The authors reviewed all AI-assisted content and take full responsibility for this work.

REPRODUCIBILITY STATEMENT

We will release model checkpoints, source code, data-processing scripts, and releasable annotations and curation metadata. Third-party datasets remain subject to their original licences.

ETHICS STATEMENT

We use public research datasets under their respective licences. Some contain people, but our method neither identifies individuals nor predicts sensitive attributes. As with other grounding and tracking systems, misuse in surveillance or safety-critical applications remains a potential risk.

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 *European conference on computer vision*, pp. 422–440. Springer, 2020.

Mehmet Aygun, Aljosa Osep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, and Laura Leal-Taixé. 4d panoptic lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5527–5537, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Yash Bhalgat, Iro Laina, Joao F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *arXiv preprint arXiv:2306.04633*, 2023.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1050–1060. IEEE, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Jiazhong Cen, Zanwei Zhou, Jiemin Fang, chen yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, XIAOPENG ZHANG, and Qi Tian. Segment anything in 3d with nerfs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 25971–25990. Curran Associates, Inc., 2023. doi: 10.52202/075280-1130. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/525d24400247f884c3419b0b7b1c4829-Paper-Conference.pdf.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3367–3375, October 2023.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024.

Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326. IEEE, 2023.

Chris Choy, Junha Lee, Chunghyun Park, Minsu Cho, and Jan Kautz. Spaceformer: Fast proposal-free open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2604.20395*, 2026.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3070–3079. IEEE, 2019.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28652–28668, June 2026.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703. IEEE, 2023a.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20167–20177. IEEE, 2023b.

Nikita Drozdov, Andrey Lemeshko, Nikita Gavrilov, Anton Konushin, Danila Rukhovich, and Maksim Kolodiazhnyi. Z3d: Zero-shot 3d visual grounding from images. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 147–154, 2026.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems*, 37:40212–40229, 2024.

Gal Fiebelman, Tamir Cohen, Ayellet Morgenstern, Peter Hedman, and Hadar Averbuch-Elor. 4-legs: 4d language embedded gaussian splatting. *Computer Graphics Forum*, 44(2):e70085, 2025. doi: 10.1111/cgf.70085.

Jiangshan Gong, Yuqun Wu, Qiqian Fu, Yao Xiao, Chuhang Zou, Shenlong Wang, and Derek Hoiem. SAM-V: Geometry-aware segment anything for multi-view instance segmentation. *arXiv preprint arXiv:2609.25490*, 2026.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013. doi: 10.1109/TPAMI.2012.213.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13434–13446. IEEE, 2023.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2VLM: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9546, 2026.

Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation–a synthetic dataset and baselines. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3100–3110. IEEE, 2019.

Shengxiang Ji, Guanjun Wu, Jiemin Fang, Jiazhong Cen, Taoran Yi, Wenyu Liu, Qi Tian, and Xinggang Wang. Segment any 4d gaussians. *arXiv preprint arXiv:2407.04504*, 2024.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4866–4875. IEEE, 2020.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13229–13239. IEEE, 2023.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29914–29934. Curran Associates, Inc., 2023. doi: 10.52202/075280-1303. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5f828e38160f31935cfe9f67503ad17c-Paper-Conference.pdf.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian conference on computer vision*, pp. 123–141. Springer, 2018.

Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9856–9865. IEEE, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20943–20953. IEEE, 2024.

Lars Kreuzberg, Idil Esen Zulfikar, Sabarinath Mahadevan, Francis Engelmann, and Bastian Leibe. 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino (eds.), *Computer Vision – ECCV 2022 Workshops*, pp. 537–553, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25056-9.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Duy Tho Le, Chenhui Gou, Stavya Datta, Hengcan Shi, Ian Reid, Jianfei Cai, and Hamid Rezatofighi. Jrdb-panotrack: An open-world panoptic segmentation and tracking robotic dataset in crowded human environments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22325–22334. IEEE, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Changyang Li, Xueqing Huang, Shin-Fang Chng, Huangying Zhan, Qingan Yan, and Yi Xu. Fast3dis: Feed-forward anchored scene transformer for 3d instance segmentation. *arXiv preprint arXiv:2603.25993*, 2026a.

Chaoyue Li, Boxue Yang, Shengyao Zhou, Haoyang Wu, Rui Qian, and Linfeng Zhang. 4dvlt: Dynamic scene understanding with worldline-centered vision-language tracking. *arXiv preprint arXiv:2606.22631*, 2026b.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, pp. 467–484. Springer, 2024a.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, et al. Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction. *arXiv preprint arXiv:2510.22706*, 2025a.

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, et al. Sam3-i: Segment anything with instructions. *arXiv preprint arXiv:2512.04585*, 2025b.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*, 2026c.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3707–3717. IEEE, 2025c.

Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22001–22011. IEEE, 2025d.

Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27948–27959. IEEE, 2024b.

Yun-Jin Li, Mariia Gladkova, Yan Xia, and Daniel Cremers. Trase: Tracking-free 4d segmentation and editing. In *2026 International Conference on 3D Vision (3DV)*, pp. 1670–1679. IEEE, 2026d.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36:37193–37229, 2023.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3718–3727, June 2025.

Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21001–21011. IEEE, 2022.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

OpenAI. GPT-5.6 System Card, July 2026. URL `https://deploymentsafety.openai.com/gpt-5-6`. Published July 9, 2026.

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

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Adam W Harley, and Katerina Fragkiadaki. Embodied language grounding with 3d visual feature representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2217–2226. IEEE, 2020.

Jinyuan Qu, Hongyang Li, and Lei Zhang. Segvggt: Joint 3d reconstruction and instance segmentation from multi-view images. *arXiv preprint arXiv:2603.19926*, 2026.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12630–12641. IEEE, 2023.

Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 9284–9293, 2025.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018. IEEE, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8216–8223. IEEE, 2023.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9043–9052. IEEE, 2023.

Emily Steiner, Jianhao Zheng, Henry Howard-Jenkins, Chris Xie, and Iro Armeni. Rescene4d: Temporally consistent semantic instance segmentation of evolving indoor 3d scenes. *arXiv preprint arXiv:2601.11508*, 2026.

Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, Xinggang Wang, et al. Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33280–33290, 2026.

Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

Ruilin Tang, Yang Zhou, Zhong Ye, Wenxi Liu, Yan Huang, and Shengfeng He. Panopticquery: Unified query-time reasoning for 4d scenes. *arXiv preprint arXiv:2604.05638*, 2026.

Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, et al. Vidi: Large multimodal models for video understanding and editing. *arXiv preprint arXiv:2504.15681*, 2025.

Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22836–22845. IEEE, 2023.

Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2698–2707. IEEE, 2022.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, et al. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. In *European Conference on Computer Vision*, pp. 336–357. Springer, 2026b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024a.

Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10756–10765. IEEE, 2021.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024b.

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 *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026c.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11676–11685. IEEE, 2022.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025b.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9088–9099, June 2026d.

Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

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. *arXiv preprint arXiv:2601.06874*, 2026a.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4964–4974. IEEE, 2022a.

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 553–569, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-19815-1.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3783–3795. IEEE, 2024.

Junyi Wu, Van Nguyen Nguyen, Benjamin Planche, Jiachen Tao, Changchang Sun, Zhongpai Gao, Zhenghao Zhao, Anwesa Choudhuri, Gengyu Zhang, Meng Zheng, et al. Consistent instance field for dynamic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3793, 2026b.

Xianfeng Wu, Yajing Bai, Minghan Li, Xianzu Wu, Xueqi Zhao, Zhongyuan Lai, Wenyu Liu, and Xinggang Wang. 4dlangvggt: 4d language-visual geometry grounded transformer. *arXiv preprint arXiv:2512.05060*, 2025.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19231–19242. IEEE, 2023.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869. IEEE, 2024.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16111–16121. IEEE, 2024.

Jianyun Xu, Song Wang, Ziqian Ni, Chunyong Hu, Sheng Yang, Jianke Zhu, and Qiang Li. Sam4d: Segment anything in camera and lidar streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28535–28545. IEEE, 2025.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *European conference on computer vision*, pp. 603–619. Springer, 2018.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115. Springer, 2024.

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 *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7694–7701. IEEE, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023a.

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023b.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European conference on computer vision*, pp. 162–179. Springer, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3292–3302. IEEE, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with mllm for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Yuqian Yuan, Wentong Li, Zhaocheng Li, Yutong Lin, Juncheng Li, Siliang Tang, Jun Xiao, Yueting Zhuang, and Wenqiao Zhang. Instructsam: Segment any instance with any instructions. *arXiv preprint arXiv:2605.26102*, 2026b.

Yixin Zhang, Nicholas Konz, Kevin Kramer, and Maciej A. Mazurowski. Quantifying the limits of segmentation foundation models: Modeling challenges in segmenting tree-like and low-contrast objects. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5205–5215, 2026. doi: 10.1109/WACV61042.2026.00505.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2908–2917. IEEE, 2021.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pp. 519–535. Springer, 2020.

Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan, Suya You, Zhangyang Wang, Leonidas Guibas, et al. Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature fields. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14179–14190. IEEE, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, Fanbo Xiang, and Hao Su. Point-sam: Promptable 3d segmentation model for point clouds. *arXiv preprint arXiv:2406.17741*, 2024.

Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pp. 151–168. Springer, 2024.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023.

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.

Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. Panst3r: Multi-view consistent panoptic segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5856–5886. IEEE, 2025.

SUPPLEMENTARY MATERIAL

This supplement provides additional details of the segmentation and grounding architectures, data curation, training, and evaluation protocols, followed by further quantitative and qualitative results. The accompanying supplementary video presents additional 4D segmentation and language-grounding examples across static, dynamic, and in-the-wild scenes.

A ADDITIONAL METHOD DETAILS

A.1 SPACE-TIME QUERY DECODER

As illustrated in Fig. 7(b), the Space-Time Query Decoder separates global instance reasoning from dense mask prediction. The contextualized patch tokens $\{\mathbf{z}_i^F\}_{i=1}^T$ follow two complementary paths. A projection g_{mem} maps them to a joint decoder memory

$$\mathbf{Z} = \text{Concat}_{i=1}^T \, g_{\text{mem}}\big(\mathbf{z}_i^F\big), \tag{8}$$

while a dense feature head fuses intermediate backbone features into per-observation feature maps $\{\mathbf{F}_i\}_{i=1}^T$. The former supports reasoning across the complete observation set; the latter preserves the spatial detail required for mask prediction. No chronological-order embedding is added to $\mathbf{Z}$, allowing the same decoder to process ordered, sparse, or shuffled observations.

Starting from N_q learnable object queries $\mathbf{Q}^{(0)} = \{\mathbf{q}_j\}_{j=1}^{N_q}$, decoder layer ℓ performs a Mask2Former-style update (Cheng et al., 2022)

$$\mathbf{Q}^{(\ell)} = \mathcal{D}_\ell\Big(\mathbf{Q}^{(\ell-1)}, \mathbf{Z}; \mathbf{A}^{(\ell-1)}\Big), \tag{9}$$

where $\mathcal{D}_\ell$ comprises masked cross-attention to $\mathbf{Z}$, self-attention among the object queries, and a feed-forward network. The attention mask $\mathbf{A}^{(\ell-1)}$ is obtained from the preceding mask prediction and restricts each query to its current spatiotemporal support; a query with empty support can again attend to the full memory. Because one query state is shared by all observations, cross-attention gathers different appearances of an instance into a single persistent representation rather than constructing a separate query trajectory for each frame.

Purely prompt-driven decoding, as in SAM-style models (Kirillov et al., 2023), treats each prompt independently and may favor locally coherent, low-level regions, leading to over-segmentation. Our learnable object queries instead exchange information through self-attention and are jointly optimized under exhaustive set supervision, establishing a global instance partition before any prompt is introduced.

When a promptable query $\mathbf{q}^p$ from Section A.2 is present, it is concatenated once with the learnable queries. For the ordering $\overline{\mathbf{Q}}^{(0)} = [\mathbf{q}^p; \mathbf{Q}^{(0)}]$ shown in Fig. 7(b), the decoder update becomes

$$\overline{\mathbf{Q}}^{(\ell)} = \mathcal{D}_\ell\Big(\overline{\mathbf{Q}}^{(\ell-1)}, \mathbf{Z}; \mathbf{A}^{(\ell-1)}, \mathbf{U}\Big), \qquad \mathbf{U} = \begin{bmatrix} 1 & \mathbf{1}_{1\times N_q} \\ [-\infty]_{N_q\times 1} & \mathbf{1}_{N_q\times N_q} \end{bmatrix}, \tag{10}$$

where 1 denotes an allowed connection and $[-\infty]_{N_q\times 1}$ denotes a block whose entries are all masked. Rows denote receiving queries and columns denote source queries. The lower-left block therefore prevents the learnable queries from reading $\mathbf{q}^p$, while the fully connected first row allows $\mathbf{q}^p$ to read every learnable query. This unidirectional mask is distinct from the spatiotemporal cross-attention mask $\mathbf{A}^{(\ell-1)}$ in Eq. (9). It lets the promptable query resolve its target against the globally discovered instance set, while preventing local prompt evidence from altering the convergence or predictions of the learnable queries.

For a refined query $\mathbf{q}_j$, a mask head produces an embedding $\mathbf{e}_j = f_{\text{mask}}(\mathbf{q}_j)$ and evaluates it against every dense feature map:

$$\hat{\mathbf{M}}_{j,i} = \mathbf{e}_j^\top \mathbf{F}_i, \qquad \hat{o}_j = f_{\text{obj}}(\mathbf{q}_j). \tag{11}$$

The common embedding $\mathbf{e}_j$ makes the query index the instance identity throughout the observation set, while $\hat{o}_j$ distinguishes object queries from no-object predictions. Hungarian matching assigns each ground-truth 4D mask to one query over the full observation set; unmatched queries are supervised as no-object. The decoder therefore predicts persistent masks end to end, without clustering dense embeddings or associating independently decoded instances.

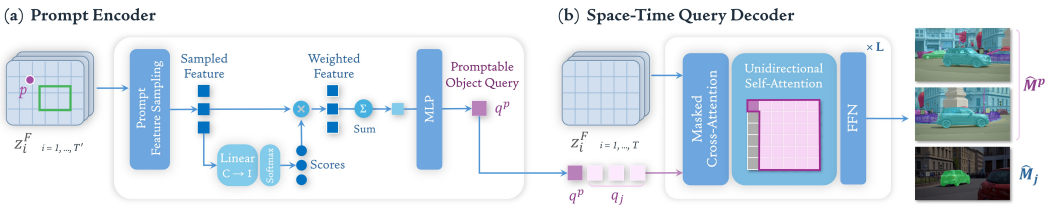


Figure 7: *(a)* **Prompt Encoder:** Point or box prompts sample dense features derived from the contextualized patch tokens and are aggregated into a promptable object query; T' is the number of observations containing prompts. *(b)* **Space-Time Query Decoder:** The promptable and learnable object queries decode persistent instance masks over the complete observation set. The attention mask blocks learnable queries from reading the prompt query while retaining the reverse direction.

A.2 Prompt Encoder

The prompt encoder in Fig. 7(a) converts image-level points and boxes into a promptable object query. Although the contextualized patch tokens are the shared backbone representation, prompt sampling operates on the spatially dense maps $\mathbf{F}_i$ derived from them. Let $\mathcal{P} = \{p_r\}_{r=1}^{R}$ collect prompts placed in one or more observations, and let τ_r identify the prompt type. A point samples a compact neighborhood around its location, whereas a box samples a regular grid over its enclosed region. A learned sampler then summarizes each set of samples and adds its type embedding:

$$\mathbf{v}_r = \mathcal{S}_{\tau_r}(\mathbf{F}_{i_r}, p_r) + \mathbf{t}_{\tau_r}. \tag{12}$$

This preserves local evidence for points while allowing boxes to aggregate evidence across their spatial extent.

Prompts from different observations need not be equally informative: sampled regions may contain different amounts of target evidence because of occlusion or background inclusion. Moreover, prompts supplied by users or downstream localization tools may contain localization errors or be inconsistent across observations. We therefore predict a relevance score for each sampled feature, enabling the network to filter unreliable evidence before forming the promptable query:

$$\alpha_r = \frac{\exp(a(\mathbf{v}_r))}{\sum_{s=1}^{R} \exp(a(\mathbf{v}_s))}, \qquad \mathbf{q}^p = f_{\mathrm{P}}\left(\sum_{r=1}^{R} \alpha_r \mathbf{v}_r\right). \tag{13}$$

A positional embedding derived from the mean prompt coordinates is supplied independently of the content query. Consequently, evidence from multiple observations can specify one target without collapsing its appearance and location into the same encoding. In Fig. 7(a), T' denotes the number of observations containing prompts.

The resulting $\mathbf{q}^p$ enters the decoder through the unidirectional self-attention in Eq. (10) and uses the same joint memory and dense mask readout as the learnable queries. It is supervised directly with the indicated target mask, rather than being assigned to a learnable query through an additional matching objective, and decodes the target's complete 4D mask even when prompts are supplied in only a subset of observations.

A.3 Training Data Details

Table 4 details the 19 sources summarized in Fig. 3(c). The reported counts refer to the subsets sampled for training rather than the full source releases. In particular, we sample 5,127 trajectories from RealEstate10K, 4,635 clips from SA-V, and 9,649 scenes from DL3DV. The resulting corpus contains 33,220 scenes, sub-scenes, or video clips and 6,052,440 RGB observations spanning indoor, outdoor, and mixed environments as well as real and synthetic imagery.

The exhaustive subset provides the set-level supervision required to discover every annotated instance. Its static component combines geometry-consistent synthetic views from Infinigen (Raistrick et al., 2023), real indoor captures from ScanNet++ (Yeshwanth et al., 2023), and large-scale camera trajectories from RealEstate10K (Zhou et al., 2018). Its dynamic component draws on the diverse panoptic videos of VIPSeg (Miao et al., 2022), urban driving sequences from Cityscapes-VPS (Kim

Table 4: **Training-data composition.** Counts report the sampled scenes or video clips and RGB observations used for training, rather than the full sizes of the source datasets.

Dataset	Supervision	Motion	Scene	Origin	Scenes/clips	Frames
Infinigen (Raistrick et al., 2023)	Exhaustive	Static	Indoor	Synthetic	1,466	146,034
ScanNet++ (Yeshwanth et al., 2023)	Exhaustive	Static	Indoor	Real	907	760,226
RealEstate10K (Zhou et al., 2018)	Exhaustive	Static	Mixed	Real	5,127	593,212
VIPSeg (Miao et al., 2022)	Exhaustive	Dynamic	Mixed	Real	2,806	66,767
Cityscapes-VPS (Kim et al., 2020)	Exhaustive	Dynamic	Outdoor	Real	394	1,970
KITTI-STEP (Weber et al., 2021)	Exhaustive	Dynamic	Outdoor	Real	12	5,027
JRDB-PanoTrack (Le et al., 2024)	Exhaustive	Dynamic	Mixed	Real	135	9,365
SA-V (Ravi et al., 2025)	Partial	Dynamic	Mixed	Real	4,635	362,391
YouTube-VOS (Xu et al., 2018)	Partial	Dynamic	Mixed	Real	3,471	94,588
MeViS (Ding et al., 2023a)	Partial	Dynamic	Mixed	Real	1,662	112,170
MOSE (Ding et al., 2023b)	Partial	Dynamic	Mixed	Real	1,496	92,715
UVO-Dense (Wang et al., 2021)	Partial	Dynamic	Mixed	Real	759	68,310
DynamicReplica (Karaev et al., 2023)	Partial	Dynamic	Indoor	Synthetic	236	70,800
SAIL-VOS (Hu et al., 2019)	Partial	Dynamic	Mixed	Synthetic	163	86,248
LVOS (Hong et al., 2023)	Partial	Dynamic	Mixed	Real	120	65,237
DAVIS (Perazzi et al., 2016)	Partial	Dynamic	Mixed	Real	60	4,209
VOST (Tokmakov et al., 2023)	Partial	Dynamic	Mixed	Real	32	1,920
DL3DV (Ling et al., 2024)	Pseudo	Static	Mixed	Real	9,649	3,316,851
Matterport3D (Chang et al., 2017)	Pseudo	Static	Indoor	Real	90	194,400
Total					**33,220**	**6,052,440**

et al., 2020) and KITTI-STEP (Weber et al., 2021), and the indoor-outdoor pedestrian sequences of JRDB-PanoTrack (Le et al., 2024). Together, these sources expose the decoder to both cross-view appearance changes in static geometry and persistent identities under scene motion.

The partially annotated video subset broadens the motion and appearance distribution beyond panoptic benchmarks. SA-V (Ravi et al., 2025), YouTube-VOS (Xu et al., 2018), and UVO-Dense (Wang et al., 2021) contribute diverse object masklets; MeViS (Ding et al., 2023a), MOSE (Ding et al., 2023b), LVOS (Hong et al., 2023), DAVIS (Perazzi et al., 2016; Pont-Tuset et al., 2017), and VOST (Tokmakov et al., 2023) emphasize motion-dependent references, dense distractors, long-term visibility changes, and object transformations. DynamicReplica (Karaev et al., 2023) and SAIL-VOS (Hu et al., 2019) complement these real videos with controlled synthetic motion. Their original masklets supervise annotated targets, while teacher predictions provide dense instance coverage outside the labelled subset.

Finally, DL3DV (Ling et al., 2024) and Matterport3D (Chang et al., 2017) provide 3.51M unlabelled multi-view observations from diverse real scenes. These sources are trained solely with teacher-generated pseudo-masklets and account for most of the corpus scale. Representative RGB observations and their corresponding annotations or pseudo-mask overlays are shown in Fig. 8.

A.4 Agentic Grounding in 4D

Grounder details. For a grounder window, each observation is represented by its RGB image and the corresponding camera and register tokens from the frozen segmentation backbone. The geometry tokens are projected into the VLM embedding space and condition a frame-specific grounding state. In the query stream, this state is compared with all persistent object-query embeddings, augmented with a learned null state for target absence. In the box stream, we retain Qwen's native text-generation interface to exploit its pretrained spatial priors, autoregressively predicting a normalized box or an explicit absence token for each observation. The resulting boxes are passed jointly to the prompt encoder, rather than being used to restrict the observations decoded by the segmenter.

An observation may be visited by several overlapping search windows. Let $\mathcal{V}$ denote the set of visited windows and $a_{w,i,j}$ the matching score assigned to persistent object query $\mathbf{q}_j$ for observation i in window w. The query stream selects

$$j^q = \arg\max_j \sum_{w \in \mathcal{V}} \sum_{i \in w} a_{w,i,j}, \tag{14}$$



Figure 8: **Representative training observations.** Four RGB observations and corresponding masks from twelve training sources not used in our evaluation; DL3DV uses pseudo-masks.

which aggregates query evidence across all visited windows without introducing an additional temporal association step. Repeated box predictions for the same observation are consolidated by presence voting and coordinate-wise median aggregation. The selected persistent query and the aggregated box prompts are then decoded by the Space-Time Query Decoder into $\hat{\mathbf{M}}^q$ and $\hat{\mathbf{M}}^p$, respectively.

Critic details. The critic operates on the final decoded masks. For each candidate, we select representative observations according to its visible mask area and render the original image beside a colored mask overlay. The two rendered candidates and the referring expression are presented to a VLM followed by a scalar preference head. Because a pairwise VLM judge may favor one presentation position irrespective of candidate quality, candidate order is randomized during training. At inference, we additionally evaluate both candidate orders and antisymmetrize their logits:

$$s_{\mathrm{C}} = h_{\mathrm{C}}\Big(x, \mathcal{R}(\hat{\mathbf{M}}^q), \mathcal{R}(\hat{\mathbf{M}}^p)\Big) - h_{\mathrm{C}}\Big(x, \mathcal{R}(\hat{\mathbf{M}}^p), \mathcal{R}(\hat{\mathbf{M}}^q)\Big), \tag{15}$$

where $\mathcal{R}$ denotes representative-view rendering and a positive score selects the query candidate. This comparison allows the critic to judge correspondence with the expression as well as the spatial and temporal consistency visible in the decoded masks.

Active Tree Search. ATS is defined over the supplied observation order; adjacency below therefore refers to neighboring input observations, which may follow a static scan trajectory or a dynamic sequence. The complete traversal is given in Algorithm 1.

Algorithm 1: Active Tree Search

Input: observations $\mathcal{I} = \{I_i\}_{i=1}^{T}$, expression x, window size N_{w}, and maximum depth D
Output: grounded 4D mask $\hat{\mathbf{M}}^*$

$\mathcal{E} \leftarrow \textsc{Search}([1, T], 0)$.
$(\hat{\mathbf{M}}^q, \hat{\mathbf{M}}^p) \leftarrow$ aggregate and decode the two prediction streams in $\mathcal{E}$.
$\hat{\mathbf{M}}^* \leftarrow f_{\mathrm{C}}(x, \hat{\mathbf{M}}^q, \hat{\mathbf{M}}^p)$; **return** $\hat{\mathbf{M}}^*$.
function $\textsc{Search}([a, b], d)$
    Uniformly sample N_{w} indices S from $[a, b]$; use every index if $b - a + 1 \leq N_{\mathrm{w}}$.
    $\{(\hat{\mathbf{q}}_i, \hat{p}_i)\}_{i \in S} \leftarrow f_{\mathrm{G}}(x, \{I_i\}_{i \in S})$; $\mathcal{E} \leftarrow \{(\hat{\mathbf{q}}_i, \hat{p}_i)\}_{i \in S}$.
    if the indices in S are adjacent, or $d = D$, **then return** $\mathcal{E}$.
    $P \leftarrow \{i \in S \mid \hat{p}_i \neq \varnothing\}$; $\mathcal{C} \leftarrow \textsc{Expand}([a, b], S, P, d)$.
    for each $C \in \mathcal{C}$ **do**
        $\mathcal{E} \leftarrow \mathcal{E} \cup \textsc{Search}(C, d + 1)$.
    end for
    return $\mathcal{E}$.
function $\textsc{Expand}([a, b], S, P, d)$
    if $P = \varnothing$ **then**
        if $d = 0$ **then return** $\textsc{UniformPartition}([a, b])$.
        return $\varnothing$.
    end if
    Group consecutive indices in P and form one child range per group, bounded by its nearest sampled neighbors in S.
    return the resulting child ranges $\mathcal{C}$.

The box stream therefore controls the recursive search, while predictions from every visited window contribute to the two candidates evaluated by the critic.

Grounding data construction. *Grounder data.* We combine the human referring annotations of ScanRefer (Chen et al., 2020), MeViS (Ding et al., 2023a), and Ref-DAVIS (Khoreva et al., 2018) with descriptions generated for ScanNet++ (Yeshwanth et al., 2023) and LVOS (Hong et al., 2023), which do not provide referring text. For ScanNet++, we use ground-truth 3D instances and their semantic labels as class anchors. Qwen3-VL-32B (Bai et al., 2025) receives the most visible RGB observation and up to two spatially separated context observations to generate up to five descriptions per instance. Quality control filters refusals, regenerates descriptions that omit the class anchor, checks hallucinations, contradictions, and neighboring-object errors with a VLM, and removes exact duplicates. For LVOS, we use GPT-5.6 (OpenAI, 2026) to generate referring-expression annotations for its video sequences. Each description is associated with its annotated instance or track and matched to a persistent object query by mask overlap over the available observations. For ScanNet++ and the video sources, we discard training pairs whose predicted masks have 2D mask IoU below 0.4 with the annotations, preventing incomplete or fragmented predictions from introducing incorrect grounding correspondences. After target filtering and source balancing, the training loader contains 79,856 records from 3,065 scenes or videos: 30,352 from ScanRefer, 31,764 from ScanNet++, 16,888 from MeViS, 435 from LVOS, and 417 from Ref-DAVIS.

Geometry-guided CoT supervision is generated for the known referent in a separate stage. For video datasets without 3D/4D ground truth, the visual geometry backbone (Wang et al., 2026a) reconstructs depth and camera geometry directly from the RGB observations. Back-projecting the ground-truth masks through this geometry yields an object-centric 4D point track, whose per-observation centroids and extents describe the referent's location, motion, and spatial relations to nearby instances. Static-scene geometry further provides metric object extent, height above the floor, and metric inter-object distances. Since monocular video reconstruction has neither an absolute scale nor a gravity-aligned frame, its facts instead use target visibility, size-normalized motion, relative object size and distance, and variation in camera distance. Qwen3-VL-235B receives these facts, the referring expression, up to four uniformly sampled RGB observations, and the known target,

and produces a short rationale that must combine an appearance cue with at least one geometric or spatiotemporal cue without revealing the target index.

Critic data. We construct critic training data from the grounder's full dual-stream predictions on the training splits. For each expression, the query and box outputs are decoded into two candidate masks and rendered as a comparison pair. Candidate quality is measured against the annotated referent using 3D box IoU for static scenes and $\mathcal{J}\&\mathcal{F}$ for videos; the higher-scoring candidate provides the preferred label. We discard pairs in which neither candidate reaches 0.5 and subsample near-ties, leaving 4,005 comparisons; this prevents jointly failed or ambiguous predictions from introducing noisy comparative supervision.

B Experimental Details

B.1 Segmenter and Prompt Encoder Training

Self-distillation with incomplete annotations. After Hungarian matching, exhaustive set prediction supervises every unmatched query as no-object. Applied directly to partial annotations, this objective would suppress unlabelled instances, while unlabelled sequences provide no positive targets. We therefore train a teacher on the exhaustive sources and use it to densify the remaining data. Original masklets are augmented with non-overlapping teacher predictions for partially annotated samples, whereas unlabelled samples are supervised entirely by the teacher. The shared teacher–student architecture thus transfers exhaustive supervision to the broader corpus through self-distillation.

Teacher predictions above 0.7 object probability are retained, regions between 0.3 and 0.7 are ignored, and the no-object loss is preserved elsewhere. Human annotations take precedence; overlapping teacher masks are discarded, and each retained pseudo-masklet is weighted by half its teacher confidence before Hungarian matching. This three-band supervision densifies partial labels without assigning uncertain predictions to either foreground or background.

The segmenter is optimized for $50{,}000$ steps from the same 1B visual geometry checkpoint using 8–16 randomly sampled observations at 512-pixel resolution. AdamW uses a global batch size of 64, weight decay 0.05, gradient clipping at 1.0, and 5% linear warm-up followed by cosine decay. Peak learning rates are 2.83×10^{-5} for rank-256 LoRA adapters and 5.66×10^{-4} for the Space-Time Query Decoder and mask heads. The number for learnable object queries is set as 400. Object classification, BCE, and Dice losses have weights 2, 5, and 5, with a no-object weight of 0.1. The exhaustive to incomplete sampling ratio increases linearly from $1{:}10$ to $2{:}1$. Since incomplete samples rely on teacher pseudo-masklets, this curriculum uses their diversity to accelerate early convergence, then progressively favors exhaustive annotations to reduce the influence of noisy labels. Training uses 64 NVIDIA A100 GPUs.

Prompt encoder training. The unidirectional attention in Eq. (10) isolates the learnable object queries from the promptable query, supporting either separate or joint training. Our reported model freezes the backbone, LoRA adapters, Space-Time Query Decoder, and mask heads, and optimizes only the prompt encoder for $15{,}000$ steps with AdamW on eight GPUs. The per-GPU batch size is one, the peak learning rate is 3×10^{-4}, and weight decay, warm-up, and gradient clipping are 0.05, 5%, and 1.0, respectively. The reported grounding pipeline uses the checkpoint at step $13{,}500$.

Prompts are sampled online from annotated or pseudo-labelled masklets, with at most four target instances per sample. We use box prompts with probability 0.4 and point prompts otherwise. A point prompt contains a positive point inside the target mask; for each point-prompted observation, an additional corrective point is sampled outside it with probability 0.3. Negative clicks thus supplement rather than replace positive evidence, teaching the encoder to exclude surrounding regions and nearby instances. Box prompts are not paired with negative boxes; their coordinates are instead jittered to model localization error. Prompts span a random number of target-visible observations. The promptable query is directly supervised by the corresponding 4D mask using BCE and Dice losses weighted by 5, with intermediate-layer supervision weighted by 0.4.

Alternatively, LoRA, decoder, mask-head, and prompt-encoder parameters are jointly optimized for $50{,}000$ steps with the segmenter settings above. We interleave exhaustive and prompt-conditioned

steps at 1:1 for the first $20{,}000$ steps, 2:1 for the next $20{,}000$, and 5:1 for the final $10{,}000$. Since the prompt-conditioned pool includes partially annotated and pseudo-labelled masklets, this curriculum accelerates prompt learning early and progressively reduces exposure to label noise toward convergence. For exhaustively annotated prompt batches, the learnable-query set loss is retained with weight 0.5. With probability 0.05, up to 20% of prompted observations are corrupted: a point is moved to a random location, while a box is replaced by one from another instance or by a random box. This augmentation directly trains the relevance weights in Eq. (13) to suppress inconsistent evidence across observations. The separate strategy applies an analogous noisy-to-clean curriculum during segmenter self-distillation; its subsequent frozen-segmenter prompt stage contains only prompt-conditioned updates. We systematically evaluate separate and joint training across all benchmarks and observe only marginal differences. All results reported in this paper therefore use the separately trained segmenter and Prompt Encoder checkpoints.

B.2 Grounder and Critic Training

Grounder. We use a Grounder window size of $N_{\mathrm{w}} = 8$ during both training and inference. During training, the observations are sampled from a span of at most 32: 80% of the windows are centred on a target-visible observation and balance target-present and target-absent observations, while the remainder are sampled without a presence constraint to include negative windows. We fully fine-tune Qwen3-VL-4B for 4 epochs, corresponding to approximately $5{,}000$ optimizer updates on eight GPUs. AdamW uses an effective batch size of 64, weight decay 0.01, and peak learning rates of 8×10^{-6} for the VLM and 10^{-4} for the newly initialized projection and prediction heads. We apply 300 warm-up steps followed by cosine decay, bfloat16 training, and gradient clipping at 1.0. The query-classification and autoregressive language-modeling losses in Eq. (7) are equally weighted.

Critic. The $4{,}005$ candidate pairs constructed in Section A.4 are shuffled during training, and their presentation order is randomized on every visit to remove branch-position bias. Near-tie examples constitute 30% of the training pairs, exposing the Critic to both clear and ambiguous comparisons. We fully fine-tune Qwen3-VL-2B with its scalar readout for three epochs on four GPUs. AdamW uses an effective batch size of 32, a learning rate of 5×10^{-6}, weight decay 0.01, bfloat16 precision, and gradient clipping at 1.0.

B.3 Baseline Configurations

Class-agnostic segmentation baselines. We compare lifted 3D segmentation, video segmentation, and feed-forward visual-geometry methods, using their official checkpoints on the same observations without semantic labels, target prompts, or first-frame ground truth. For SAM 2, the Automatic Mask Generator uses a 16×16 point grid to discover instances, whose masks initialize the video predictor. The predictor propagates their identities, while periodic automatic mask generation adds masks with less than 0.5 IoU to every active track as new instances. Re-discovery is performed every 2 observations for the short ScanNet clips and every 15 frames for the video benchmarks. When full-sequence inference exceeds GPU memory, we retain all observations through memory-bounded streaming-window inference. In particular, our IGGT implementation appends fixed scene-wide anchor views to each sliding window, concatenates the resulting dense instance features, and applies clustering over the complete sequence to maintain identities across windows. IGGT4D uses its released offline full-attention configuration. SAM-V uses the released checkpoint and automatic-mask settings; sequences longer than 32 observations are processed in overlapping 32-observation windows, with identities matched by mask IoU on the shared observation.

Language-grounding baselines. VLM-Grounder uses Qwen3-VL-30B-A3B-Thinking with YOLO-World and a single-view ensemble; for videos, it runs its 3D pipeline on the reconstructed geometry and reprojects the selected box to each observation. VLM+SAM 2 uses Qwen3-VL-30B-A3B-Instruct, and Sa2VA uses its released 8B model. SAM 3 and MVGGT use their official end-to-end checkpoints without a separately served VLM: SAM 3 receives category prompts on static scenes and full expressions on videos, while MVGGT receives eight uniformly sampled observations. Z3D (Drozdov et al., 2026) uses Qwen3-VL-30B-A3B-Thinking within a 3D-proposal pipeline combining MaskClustering, CLIP retrieval, VLM filtering, SAM 3 segmentation, and 3D proposal voting. To preserve the RGB-only input setting, its proposals are constructed from the

feed-forward geometry reconstructed for each scene rather than from ground-truth point clouds; on videos, the selected 3D proposal is projected back to the observations for evaluation. For static scenes, VLM+SAM 2 ranks the observations with CLIP and predicts a normalized box on the selected view. For the video results in Table 2, the VLM instead receives eight uniformly sampled observations jointly and predicts one box or absence per observation; these boxes initialize SAM 2 for bidirectional propagation over the complete sequence.

B.4 Benchmarks and Metrics

Benchmarks. For class-agnostic segmentation, we use the ScanNet setting from IGGT (Li et al., 2025a), DAVIS 2017, LVOS, and VIPSeg, comprising ten static multi-view clips, 30 short videos, 50 long-term videos, and 343 panoptic videos, respectively. Following (Li et al., 2025a; Zou et al., 2026), each ScanNet clip contains 8 sampled observations, while each LVOS video is uniformly sampled to 100 observations. For the separate robustness evaluation, each video produces three *sparse* subsets by sampling 16 observations without replacement from its first 32 observations, last 32 observations, and full temporal extent, respectively; the sampled observations retain their chronological order, and each subset contains annotated foreground. The *shuffle* setting randomly permutes exactly the same observations and their annotations. For the LVOS *hard* setting, farthest-point sampling over normalized RGB thumbnail descriptors selects 16 appearance-diverse observations across the full video; at least four contain annotated foreground when available, after which their order is shuffled. Supplementary evaluations further cover two zero-shot egocentric benchmarks: 48 HOI4D (Liu et al., 2022) clips and 42 EPIC-KITCHENS VISOR (Darkhalil et al., 2022) clips, as well as 9 KITTI-STEP (Weber et al., 2021) sequences for driving scenes. For language grounding, we sample 700 ScanRefer queries from 121 validation scenes and hold out 80 ScanNet++ scenes with 3,320 expressions. The video evaluation contains all 122 Ref-DAVIS expressions and 616 single-referent MeViS expressions from 47 videos.

Metrics. For class-agnostic segmentation, T-mIoU averages over ground-truth tracks the intersection-over-union accumulated across observations, while T-SR measures the fraction whose IoU exceeds 0.5 in every observation where they are present. We report the standard region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$ (Perazzi et al., 2016; Pont-Tuset et al., 2017) on DAVIS and LVOS, and STQ and class-agnostic thing mAP on VIPSeg. Because the video benchmarks provide only per-frame 2D masks, we reconstruct their 4D ground truth by lifting the annotated masks with the depth and cameras predicted by the visual geometry backbone (Wang et al., 2026a). Predicted masks are lifted through the same geometry. Within each clip, the point sets are voxelized in (x, y, z, t) with a separate time slice for each observed frame, and instances are matched one-to-one by voxel IoU. With N ground-truth instances in the clip, ι_j denoting the voxel IoU of the prediction matched to instance j, and $\iota_j = 0$ for an unmatched instance, we compute

$$\text{4D-IoU} = \frac{1}{N}\sum_{j=1}^{N} \iota_j, \qquad \text{4D-SR} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\iota_j > 0.5]. \tag{16}$$

Benchmark scores average these clip-level values across clips. Static grounding is evaluated by 3D mask mIoU and the proportions of predicted masks and axis-aligned boxes exceeding 0.25 IoU; Ref-DAVIS and MeViS use $\mathcal{J}\&\mathcal{F}$.

C Additional Results and Discussion

C.1 Additional Segmentation Results

We retain the metrics used in the main paper: $\mathcal{J}\&\mathcal{F}$ measures per-frame region and boundary quality, whereas T-mIoU and T-SR evaluate overlap and persistent recovery over the complete observation set.

Egocentric and driving scenes. To assess transfer toward downstream perception tasks, Table 5 evaluates the segmenter in egocentric and driving scenes. HOI4D and VISOR are evaluated strictly zero-shot, as neither dataset is included in our training corpus; for driving, we evaluate a held-out

validation split from KITTI-STEP. S4VY leads every reported metric on all three datasets, demonstrating robust transfer to both domains. All methods nevertheless obtain markedly lower T-SR than $\mathcal{J}\&\mathcal{F}$ and T-mIoU: many targets are small or visually similar, while rapid camera and object motion, occlusion, and exit from and re-entry into view make consistent full-clip recovery particularly challenging. Under this stricter criterion, S4VY still achieves the highest T-SR.

Table 5: **Egocentric and driving-scene segmentation.** Left: results on HOI4D and VISOR. Right: results on the KITTI-STEP evaluation split. IGGT4D is omitted on HOI4D because its training mixture contains part of the evaluation split.

Method	HOI4D			VISOR		
	$\mathcal{J}\&\mathcal{F}$	T-mIoU	T-SR	$\mathcal{J}\&\mathcal{F}$	T-mIoU	T-SR
DEVA (Cheng et al., 2023)	0.594	0.557	0.250	0.213	0.111	0.018
GLEE (Wu et al., 2024)	0.427	0.404	0.261	–	–	–
OMG-Seg (Li et al., 2024b)	0.272	0.248	0.192	0.020	0.016	0.000
SAM 2 (Ravi et al., 2025)	0.572	0.542	0.219	0.295	0.176	0.060
PanSt3R (Zust et al., 2025)	0.523	0.524	0.352	0.195	0.142	0.044
IGGT (Li et al., 2025a)	0.378	0.374	0.069	0.148	0.099	0.027
SegVGGT (Qu et al., 2026)	0.266	0.280	0.131	0.097	0.078	0.037
IGGT4D (Zou et al., 2026)	–	–	–	0.261	0.179	0.068
SAM-V (Gong et al., 2026)	0.465	0.462	0.122	0.199	0.157	0.035
S4VY (Ours)	**0.624**	**0.602**	**0.516**	**0.313**	**0.204**	**0.118**

Method	KITTI-STEP		
	$\mathcal{J}\&\mathcal{F}$	T-mIoU	T-SR
DEVA	0.181	0.171	0.037
OMG-Seg	0.164	0.161	0.088
SAM 2	0.274	0.266	0.075
IGGT	0.113	0.110	0.017
SegVGGT	0.065	0.071	0.002
IGGT4D	0.268	0.250	0.105
SAM-V	0.203	0.196	0.021
S4VY (Ours)	**0.501**	**0.519**	**0.191**

Complete ScanNet200 evaluation. The ScanNet setting from (Li et al., 2025a) used in the main paper contains 10 evaluation clips. To further determine the improvement at a larger scale, we additionally evaluate all 312 scenes in the ScanNet200 validation split. S4VY reaches 0.517 T-mIoU, 0.376 T-SR, and 0.454 $\mathcal{J}\&\mathcal{F}$ (Table 6), improving over IGGT4D by 2.9, 4.6, and 2.8 points, respectively. The consistent gains across all three metrics agree with the smaller setting and show that the advantage does not depend on the ten-clip subset.

Robustness to sparse and unordered inputs. Table 7 compares the sparse, shuffled, and hard protocols defined in Section B.4. When the same sparse observations are shuffled, every video baseline that propagates information along the input sequence declines: by 0.6–2.8 $\mathcal{J}\&\mathcal{F}$ points on DAVIS and 1.3–2.0 points on LVOS. In contrast, the feed-forward geometry methods are not showing a consistent downward trend after the shuffle; S4VY changes from 0.807 to 0.800 on DAVIS and from 0.645 to 0.656 on LVOS. This robustness is important for user casually captured dynamic scenes, whose observations may be sparse and nonchronological. LVOS-hard further selects and shuffles 16 appearance-diverse observations, under which the video baselines lose 9.8–16.1 points relative to their sparse results. Without temporal adjacency, frame-to-frame propagation must bridge large appearance and viewpoint changes without shared spatial geometry, making instance identities difficult to maintain. By jointly decoding identities in a consistent 4D geometry representation, S4VY retains 0.588 $\mathcal{J}\&\mathcal{F}$, 12.6 points above the strongest video baseline.

Table 6: **Complete ScanNet200 validation split.** RGB-only evaluation.

Method	T-mIoU	T-SR	$\mathcal{J}\&\mathcal{F}$
VGGT-Ω + Mask3D (Schult et al., 2023)	0.053	0.017	0.050
SAM3D (Yang et al., 2023b)	0.463	0.207	0.338
DEVA (Cheng et al., 2023)	0.088	0.029	0.086
GLEE (Wu et al., 2024)	0.042	0.024	0.053
OMG-Seg (Li et al., 2024b)	0.101	0.050	0.101
SAM 2 (Ravi et al., 2025)	0.327	0.153	0.288
PanSt3R (Zust et al., 2025)	0.461	0.318	0.403
IGGT (Li et al., 2025a)	0.339	0.210	0.298
SegVGGT (Qu et al., 2026)	0.425	0.319	0.362
IGGT4D (Zou et al., 2026)	0.488	0.330	0.426
SAM-V (Gong et al., 2026)	0.376	0.205	0.333
S4VY	**0.517**	**0.376**	**0.454**

Table 7: **Robustness to sparse and unordered inputs.** All entries are class-agnostic $\mathcal{J}\&\mathcal{F}$.

Method	DAVIS		LVOS		
	Sparse	Shuffle	Sparse	Shuffle	Hard
DEVA	0.532	0.526	0.433	0.416	0.307
GLEE	0.749	0.743	0.560	0.547	0.462
OMG-Seg	0.693	0.665	0.452	0.435	0.311
SAM 2	0.766	0.749	0.579	0.559	0.418
PanSt3R	0.579	0.581	0.447	0.464	0.359
IGGT	0.448	0.459	0.269	0.273	0.171
SegVGGT	0.328	0.332	0.111	0.106	0.078
IGGT4D	0.613	0.600	0.268	0.268	0.248
S4VY (Ours)	**0.807**	**0.800**	**0.645**	**0.656**	**0.588**

Prompted 4D segmentation. For a paired comparison, Table 8 supplies all methods with the same sampled prompts. SAM-V supports only point prompts and natively decodes one prompted observation at a time; for $K > 1$, we run these decodes independently and retain the output with the highest model-predicted IoU. With clean inputs, S4VY is consistently strongest with points, whereas SAM 2 is generally stronger with boxes once multiple observations are prompted. SAM-V improves only marginally with additional ScanNet prompts and degrades substantially on dynamic DAVIS scenes, reflecting the limits of independently selecting among single-prompt predictions. In practical use, prompts across observations may also be corrupted by user annotation errors or an upstream grounding tool. When half of eight prompts are irrelevant, S4VY exceeds SAM 2 by $41.2/25.1$ $\mathcal{J}\&\mathcal{F}$ points for box/point prompts on ScanNet and by $11.9/31.6$ points on DAVIS. The learned relevance scoring in our Prompt Encoder, together with training-time prompt corruption, suppresses inconsistent evidence; without explicit cross-observation prompt filtering, an erroneous prompt can instead redirect mask propagation to the wrong instance and cause severe degradation.

Table 8: **Point- and box-prompted segmentation** on ScanNet and DAVIS. K is the maximum number of evenly spaced target-present observations used for prompting; corruption replaces the specified fraction with target-irrelevant prompts. All entries report $\mathcal{J}\&\mathcal{F}$ for S4VY and SAM 2 (Ravi et al., 2025); SAM-V (Gong et al., 2026) supports point prompts only. For $K > 1$, SAM-V independently decodes each prompted observation and retains the output with the highest predicted IoU. $K = 8$ covers all visible ScanNet views.

K	Corruption	ScanNet					DAVIS				
		Box		Point			Box		Point		
		S4VY	SAM 2	S4VY	SAM 2	SAM-V	S4VY	SAM 2	S4VY	SAM 2	SAM-V
1	0%	0.806	0.753	0.793	0.649	0.567	0.790	0.809	0.793	0.771	0.134
2	0%	0.804	0.844	0.825	0.713	0.598	0.789	0.818	0.802	0.764	0.124
4	0%	0.800	0.877	0.825	0.760	0.589	0.779	0.827	0.805	0.760	0.145
8	0%	0.801	0.878	0.825	0.769	0.589	0.807	0.829	0.810	0.780	0.129
16	0%	–	–	–	–	–	0.796	0.828	0.805	0.775	0.127
8	25%	0.768	0.401	0.662	0.346	0.488	0.733	0.733	0.800	0.659	0.091
8	50%	0.665	0.253	0.507	0.256	–	0.651	0.532	0.761	0.445	–
16	25%	–	–	–	–	–	0.727	0.722	0.808	0.668	0.082

C.2 Grounding Tools and Efficiency

Grounding tools. Table 9 compares localization tools under the same Prompt Encoder and mask decoder. Qwen3-VL, GLM-4.1V, and GPT-5.6 are general-purpose MLLMs queried for boxes; Molmo 2 provides native multi-frame point localization and tracking, whereas LocateAnything is specialized for image grounding with parallel box decoding. Simply enlarging the visual context of a generic MLLM is insufficient: Qwen3-VL declines from 0.316 to 0.202 3D mask mIoU when its input changes from a single view to multiple views. LocateAnything is effective for single-image localization, but its independent per-view application here cannot enforce cross-view identity or reject false positives when the referent is absent. Molmo 2 is consequently the strongest external tool at 0.432 mIoU, benefiting from its native multi-frame pointing and tracking, but its coordinate predictions remain disconnected from the persistent 4D instance space. Our dual-stream Grounder instead retains the VLM's pretrained spatial prior through box prediction while aligning the query stream with the feed-forward visual-geometry feature space, reaching the best results.

Agentic frame localization. Table 10 compares VLM prediction and agentic search for locating the observations that contain a referent. This setting is most relevant to long inputs with sparse target presence: each scene contains an average of 80.3 views, while its target appears in only 28.6% of them. VideoAgent (Wang et al., 2024b), VideoTree (Wang et al., 2025b), and LensWalk (Li et al., 2026c) originally rely on proprietary LLMs; we preserve their published search harnesses while replacing these backends with Qwen3-VL-235B (Bai et al., 2025) for reproducibility. VideoTree is the closest structural baseline, but uses EVA-CLIP visual embeddings to cluster frames and an LLM to score the relevance of cluster captions. A missed relevance judgment can therefore prevent further exploration of an entire branch, and its hierarchy is constructed for temporally structured video. Our Active Tree Search instead expands the frame space according to the Grounder's target-

Table 9: **Grounding-tool comparison** on 160 queries randomly sampled from ScanRefer. Coordinates from each tool are decoded by the same promptable segmenter. Both configurations predict boxes. Model size refers to the grounding model only.

Method	Model Size	Configuration	3D mask mIoU	Mask@0.5	Box@0.5
Molmo 2 (Clark et al., 2026)	4B	Multi-frame point tracking	0.432	45.4	46.1
Molmo 2 (Clark et al., 2026)	4B	Multi-frame point localization	0.407	43.4	42.1
Qwen3-VL (Bai et al., 2025)	4B	Single-view box	0.316	32.2	32.9
Qwen3-VL (Bai et al., 2025)	4B	Multi-view box	0.202	21.1	20.4
GLM-4.1V (Hong et al., 2025)	9B	Per-view box	0.303	32.9	29.6
LocateAnything (Wang et al., 2026b)	3B	Per-view box	0.287	31.6	31.6
GPT-5.6 (OpenAI, 2026)	N/A	Per-view box	0.421	47.4	42.8
Ours	4B	Box stream	0.424	45.3	42.0
Ours	4B	Query stream	**0.468**	**50.0**	**50.7**

presence predictions, which incorporate both language and consistent visual geometry. It achieves 0.708 precision and 0.535 recall, yielding the highest F1 (0.573) and frame-set IoU (0.455) while selecting only 21.6% of the observations. In comparison, Vidi (Team et al., 2025) attains high precision (0.626) but selects only 7.0% of the observations and recalls merely 14.4% of the target-present frames; VideoTree reaches comparable recall (0.562) only by selecting 56.6% of the input, with precision falling to 0.283. These results demonstrate a substantially better balance between localization precision and coverage.

Table 10: **Agentic frame localization** on ScanRefer. Model size denotes the VLM used for frame selection; selected frames reports the returned fraction rather than a ranked metric.

Method	Model Size	Precision	Recall	F1	IoU	Selected frames (%)
VideoAgent (Wang et al., 2024b)	235B	0.305	0.108	0.144	0.082	9.5%
VideoTree (Wang et al., 2025b)	235B	0.283	0.562	0.341	0.221	56.6%
Vidi 1.5 (Team et al., 2025)	9B	0.626	0.144	0.192	0.121	7.0%
LensWalk (Li et al., 2026c)	235B	0.343	0.181	0.192	0.128	12.6%
Uniform top-8	N/A	0.291	0.144	0.173	0.099	13.8%
Uniform top-16	N/A	0.285	0.274	0.250	0.151	27.4%
Select all	N/A	0.286	1.000	0.416	0.286	100.0%
Ours	4B	0.708	0.535	**0.573**	**0.455**	21.6%

Segmentation efficiency. Table 11 reports end-to-end runtime in seconds per clip. By jointly processing the observation set through feed-forward visual geometry, S4VY avoids the repeated per-frame detection and propagation of video methods; the absolute efficiency gap therefore widens as the number of observations increases. Among visual-geometry approaches, direct set prediction also avoids the post-hoc clustering used by the IGGT pipeline. The resulting system processes an eight-observation ScanNet scene in 0.5 seconds and a 100-observation HOI4D clip in 10.2 seconds. As shown in Table 12, S4VY processes four independent 100-frame clips (400 frames per batch) at 42.11 frames/s on a single GB200, while single-clip inference reaches 40.63 frames/s.

Table 11: **Segmentation efficiency.** Average wall-clock seconds per clip on a sampled benchmark subset using an A100-80GB. Parentheses report the mean number of observations in the sampled clips. Lower is better.

Method	ScanNet (8)	VIPSeg (41)	DAVIS (71)	LVOS (100)	HOI4D (100)
DEVA (Cheng et al., 2023)	3.0	9.3	32.3	32.3	26.7
GLEE (Wu et al., 2024)	2.5	13.8	24.6	36.8	35.9
OMG-Seg (Li et al., 2024b)	0.9	5.4	10.1	13.7	14.6
SAM 2 (Ravi et al., 2025)	3.4	46	183	175	168
PanSt3R (Zust et al., 2025)	3.3	18.3	35.3	54.0	50.7
IGGT (Li et al., 2025a)	13.3	83	162	272	230
IGGT4D (Zou et al., 2026)	2.6	6.9	14.4	18.5	18.0
S4VY (Ours)	**0.5**	**3.0**	**6.6**	**10.1**	**10.2**

Table 12: **Hardware scaling.** Best-tested model-forward throughput at 384×688. Parentheses report batch size B.

Frames	A100	GB200	Ratio
1	25.85 (32)	63.15 (32)	2.44×
4	25.44 (32)	65.19 (32)	2.56×
8	24.37 (16)	64.40 (32)	2.64×
16	22.28 (8)	61.61 (16)	2.77×
32	18.98 (4)	56.69 (8)	2.99×
64	14.72 (2)	48.64 (4)	3.30×
100	11.64 (1)	42.11 (4)	3.62×

C.3 More Ablations

Segmenter. Table 13(a–b) reports class-agnostic instance AP on IGGT-Benchmark (Li et al., 2025a), computed over multi-view mask volumes: mAP averages AP across IoU thresholds from 0.50 to 0.95 in increments of 0.05, while AP_{50} and AP_{75} use fixed IoU thresholds of 0.50 and 0.75. Panel (a) shows that LoRA matches or improves full fine-tuning, with the largest gain at AP_{75}, indicating that effective segmentation adaptation does not require updating the entire backbone. Panel (b) further shows that removing frame positional embeddings preserves segmentation quality relative to learned embeddings, whereas sinusoidal embeddings consistently degrade it. These results support our decoder design without frame-order priors for unordered observation sets.

Grounding decomposition. Table 13(c–d) extends the output-stream analysis from ScanRefer to our annotated ScanNet++ split and both dynamic benchmarks. The box stream is stronger on ScanRefer, Ref-DAVIS, and MeViS, whereas neither stream dominates on ScanNet++: box prediction yields slightly higher 3D mask mIoU (0.271 versus 0.268), but query matching performs better at both accuracy thresholds. The Critic raises 3D mask mIoU from 0.434 to 0.461 on ScanRefer and from 0.271 to 0.293 on ScanNet++, while improving $\mathcal{J}\&\mathcal{F}$ from 0.744 to 0.755 on Ref-DAVIS and from 0.543 to 0.573 on MeViS. These branch-dependent results confirm that box prediction retains the VLM's spatial prior while query matching exploits the geometry-consistent instance space. By comparing their rendered candidates, the Critic selects their instance-specific strengths and consistently improves over either stream alone.

Table 13: **Additional ablations.** (a) Backbone adaptation and (b) frame positional encoding in the decoder on IGGT-Benchmark (Li et al., 2025a); (c & d) output-stream decomposition on static and dynamic grounding benchmarks.

(a) Backbone adaptation

Strategy	mAP	AP_{50}	AP_{75}
Full fine-tuning	0.200	0.286	0.233
LoRA	**0.210**	**0.287**	**0.260**

(b) Frame positional encoding

Encoding	mAP	AP_{50}	AP_{75}
Learned	0.208	0.288	**0.252**
Sinusoidal	0.186	0.262	0.220
None	**0.212**	**0.294**	0.251

(c) Static 3D grounding

Benchmark	Output	3D Mask		
		mIoU	@.25	@.5
ScanRefer	Query	0.406	51.8	45.0
	Box	0.434	55.0	47.6
	+Critic	**0.461**	**59.4**	**50.5**
ScanNet++	Query	0.268	41.7	30.9
	Box	0.271	41.0	29.8
	+Critic	**0.293**	**44.9**	**33.1**

(d) Dynamic 4D grounding

Benchmark	Output	Video Mask		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Ref-DAVIS	Query	0.626	0.730	0.678
	Box	0.695	0.793	0.744
	+Critic	**0.706**	**0.805**	**0.755**
MeViS	Query	0.470	0.579	0.524
	Box	0.494	0.591	0.543
	+Critic	**0.522**	**0.623**	**0.573**

C.4 Additional Qualitative Results

More comparisons with state-of-the-art methods. Figures 9 and 10 compares S4VY with existing class-agnostic segmenters on ScanNet multi-view scenes and the DAVIS, LVOS, and HOI4D video evaluation sets. The basketball sequence in Fig. 9(c) contains many similarly appearing players, frequent occlusion, and rapid motion. S4VY maintains their identities and follows the ball through a shot and the subsequent change of possession from the purple-masked player to the red-masked player. In Fig. 9(d), it likewise preserves the cyan-masked kangaroo through all shown observations and recovers the gray-masked cat in the first and fourth despite rapid motion and large deformation. These examples illustrate how spatially grounded 4D correspondence resolves ambiguities that appearance-only propagation cannot reliably distinguish.

Frame-space and 4D-space results. Figures 11 and 12 evaluates zero-shot transfer to two indoor datasets absent from our training corpus: Structured3D (Zheng et al., 2020), whose panoramas are decomposed into overlapping perspective patches, and ARKitScenes (Baruch et al., 2021). IGGT and IGGT4D exhibit visible geometric misalignment and incomplete scene structure in these ex-

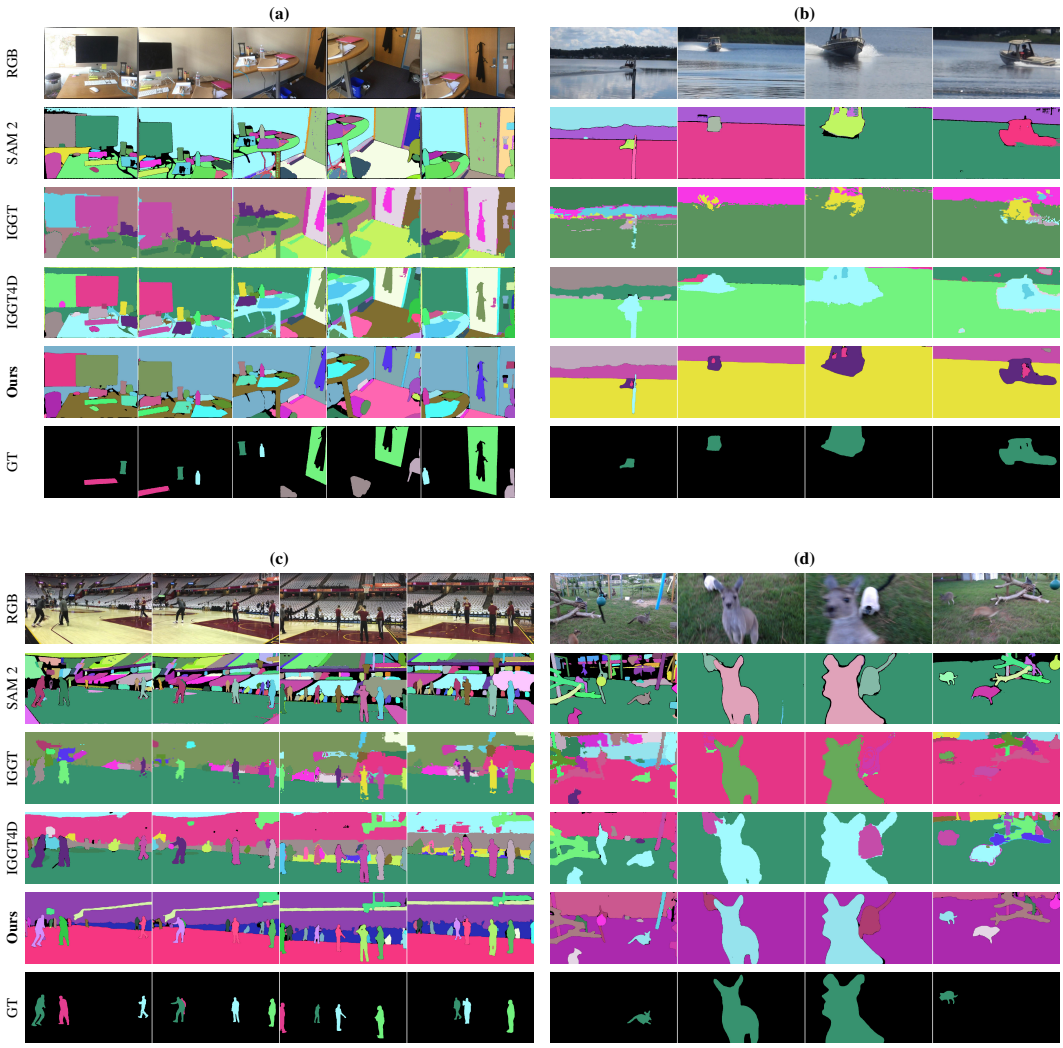


Figure 9: **Additional qualitative comparisons of class-agnostic segmentation.** A consistent mask color denotes the same instance within each sequence.

amples, whereas S4VY yields more spatially coherent predicted geometry and object identities. The dynamic and driving results in Figs. 13 and 14 further illustrate that these identities remain attached to reconstructed scene entities under camera and object motion. Together, these visualizations provide qualitative evidence that the predicted identities remain coherent beyond individual image planes.

More grounding comparisons. The results in Figs. 15 and 16 show that language references are resolved to the same persistent masks across observations, including expressions that require spatial reasoning rather than appearance matching alone. For example, Fig. 15(c) contains several visually similar laptops; by resolving the relation "across from an armchair" in the reconstructed scene, S4VY consistently selects the intended silver laptop across views.

In-the-wild evaluation. Figs. 17 and 18 extends the evaluation to unconstrained indoor, outdoor, and driving sequences. Despite substantial changes in viewpoint, scale, camera motion, and scene content, S4VY maintains coherent instance identities and 4D reconstructions, demonstrating robustness to capture conditions and domains beyond the curated benchmarks. For visualization, we follow He et al. (2013) and apply guided image filtering to refine predicted mask boundaries using the corresponding RGB observations.

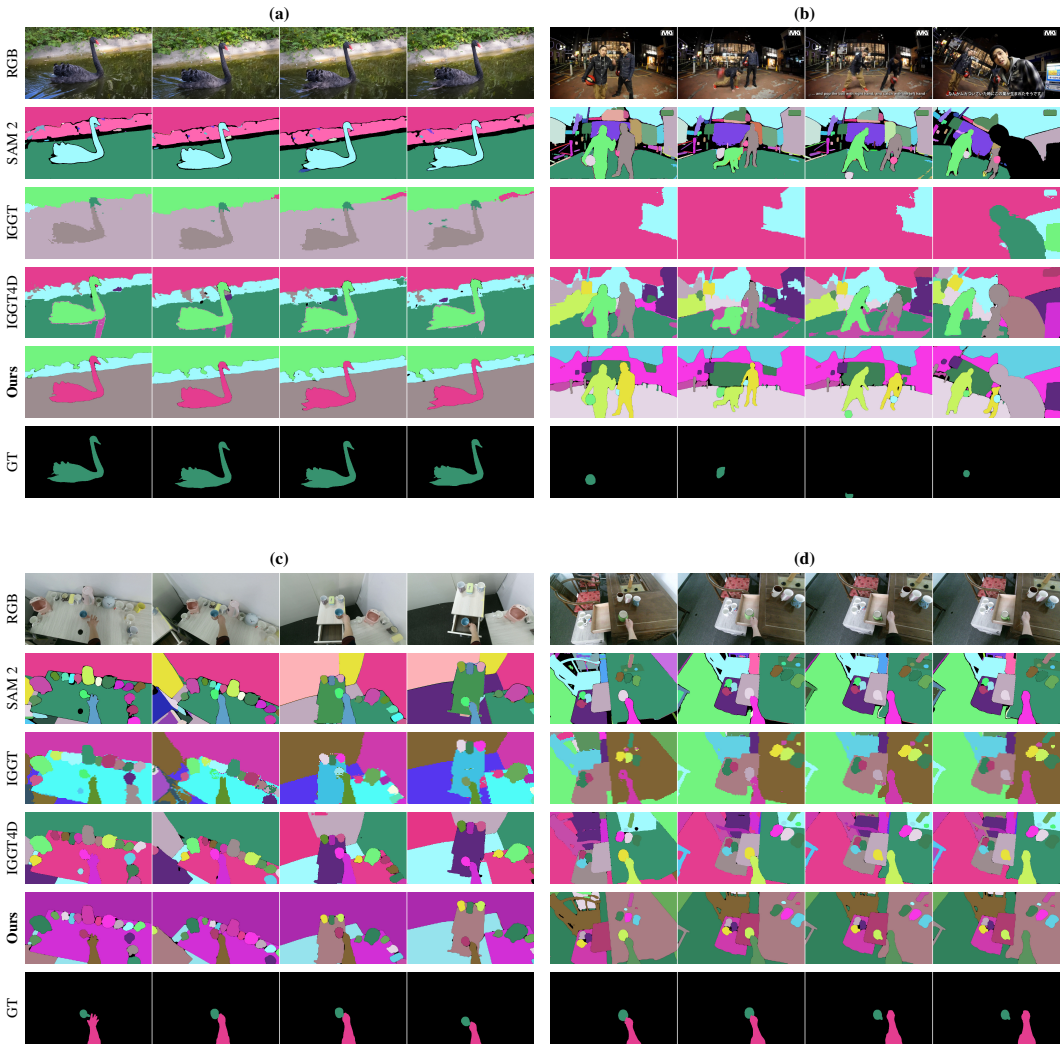


Figure 10: **Further qualitative comparisons of class-agnostic segmentation.** A consistent mask color denotes the same instance within each sequence.

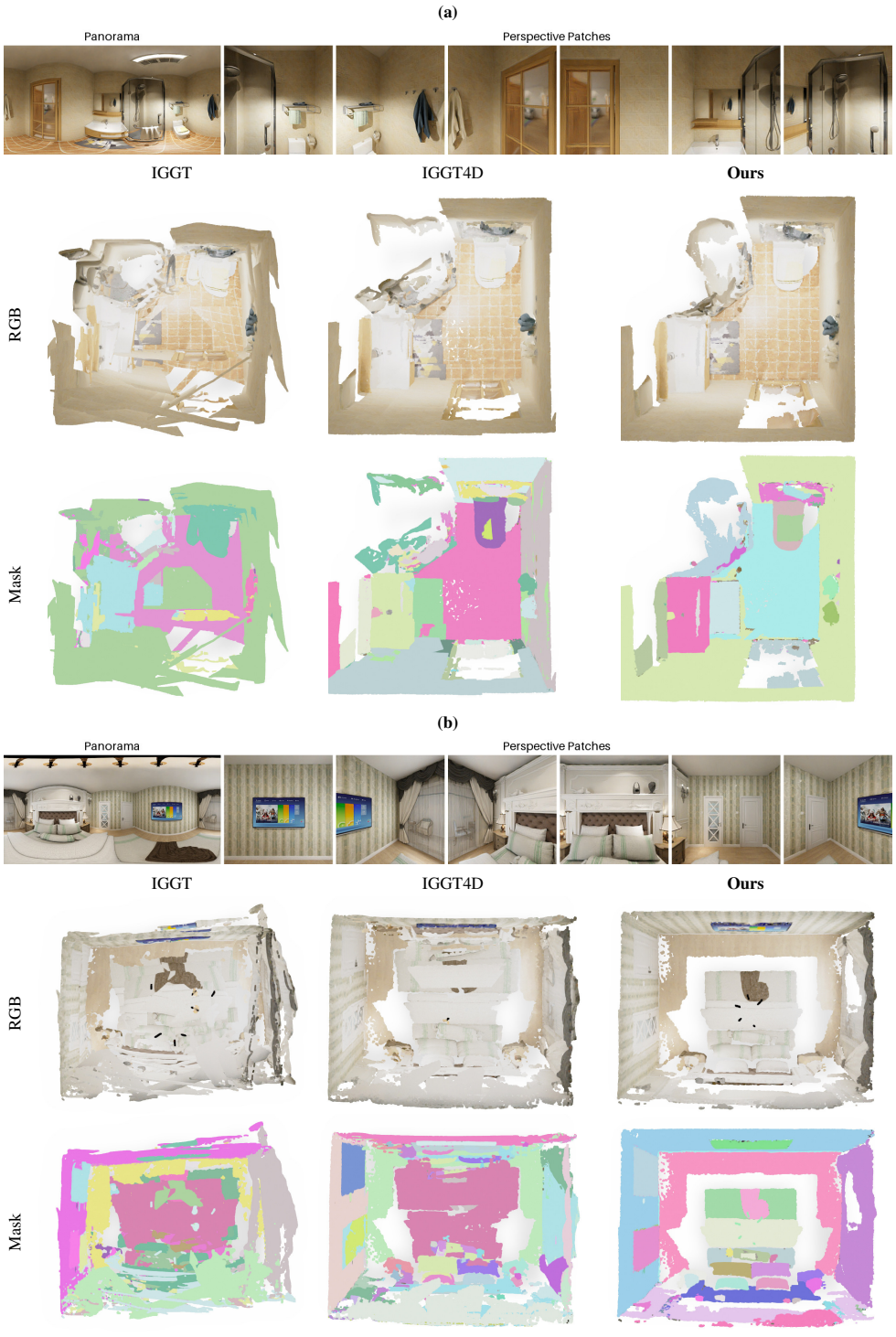


Figure 11: **Zero-shot segmentation of panoramic scenes.** Each Structured3D panorama (Zheng et al., 2020) is decomposed into overlapping perspective patches; six representative patches are shown above the aligned top-view RGB and instance-colored point clouds.

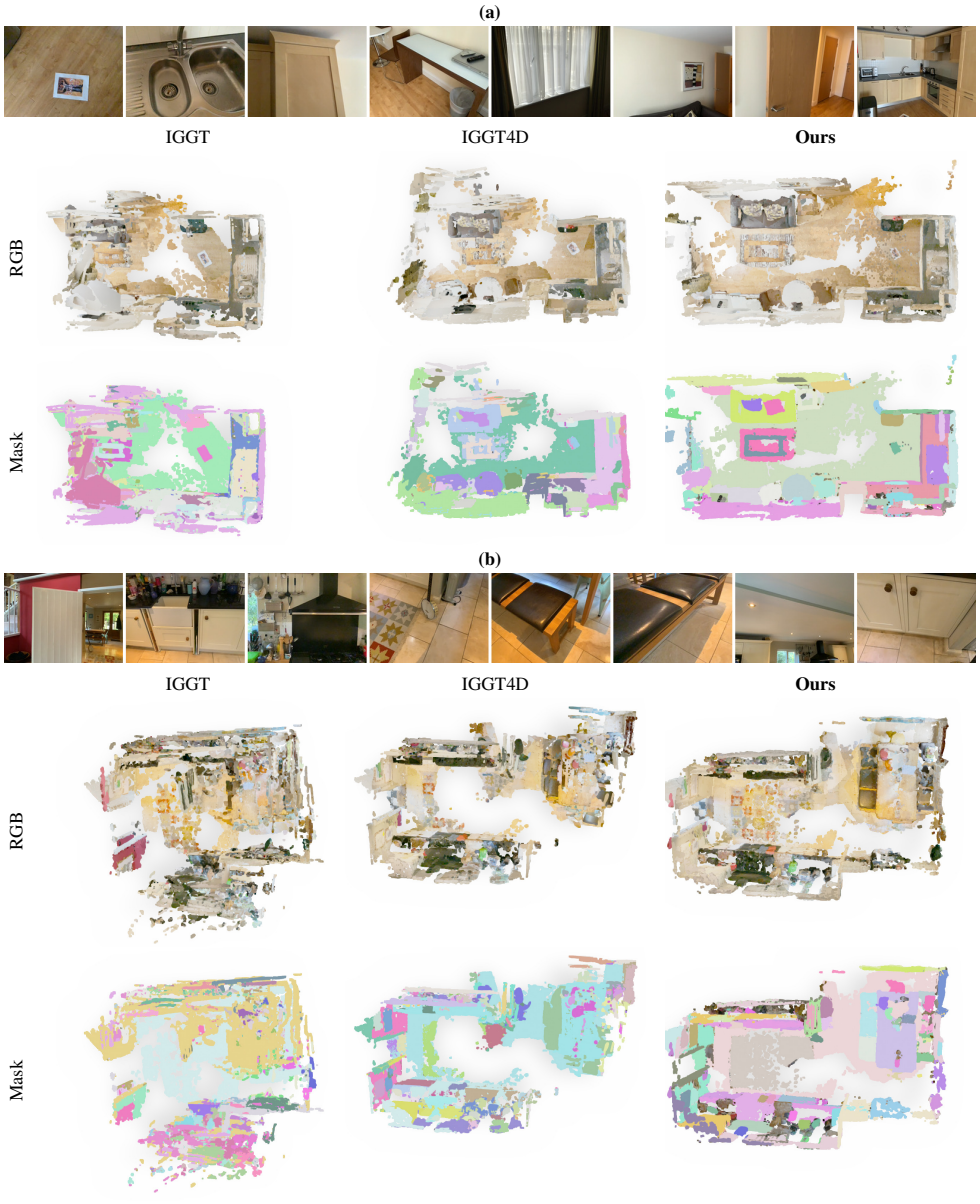


Figure 12: **Feed-forward 3D reconstruction and instance segmentation.** Eight RGB observations from held-out ARKitScenes are followed by aligned top-view RGB and instance-colored point clouds; all methods are evaluated zero-shot.



Figure 13: **Feed-forward 4D reconstruction and instance segmentation in dynamic scenes.** Selected RGB observations are followed by aligned RGB and instance-colored point-cloud views; earlier observations are rendered with lower opacity.

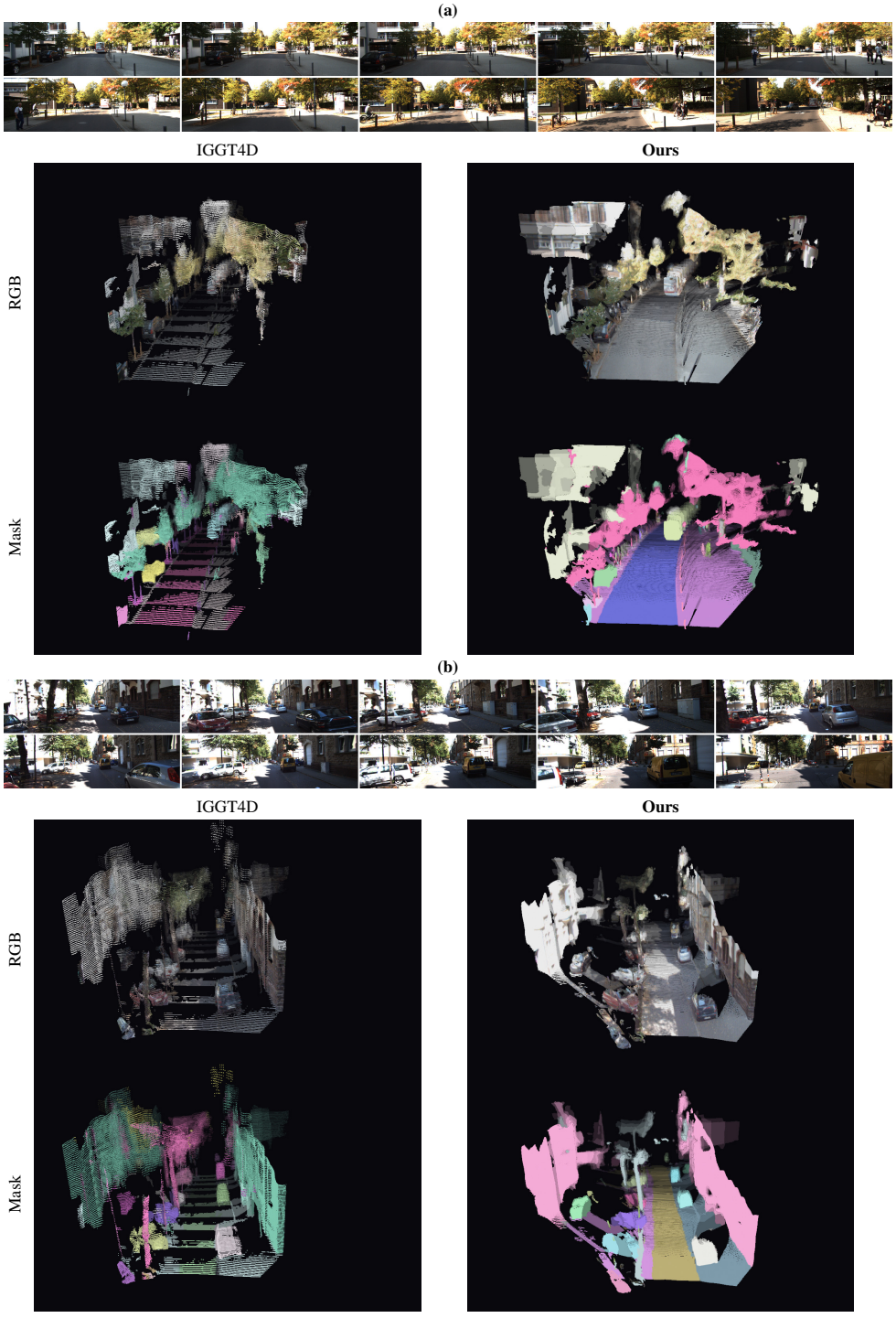


Figure 14: **Feed-forward 4D reconstruction and instance segmentation in driving scenes.** Selected RGB observations are followed by aligned RGB and instance-colored point-cloud views; earlier observations are rendered with lower opacity.

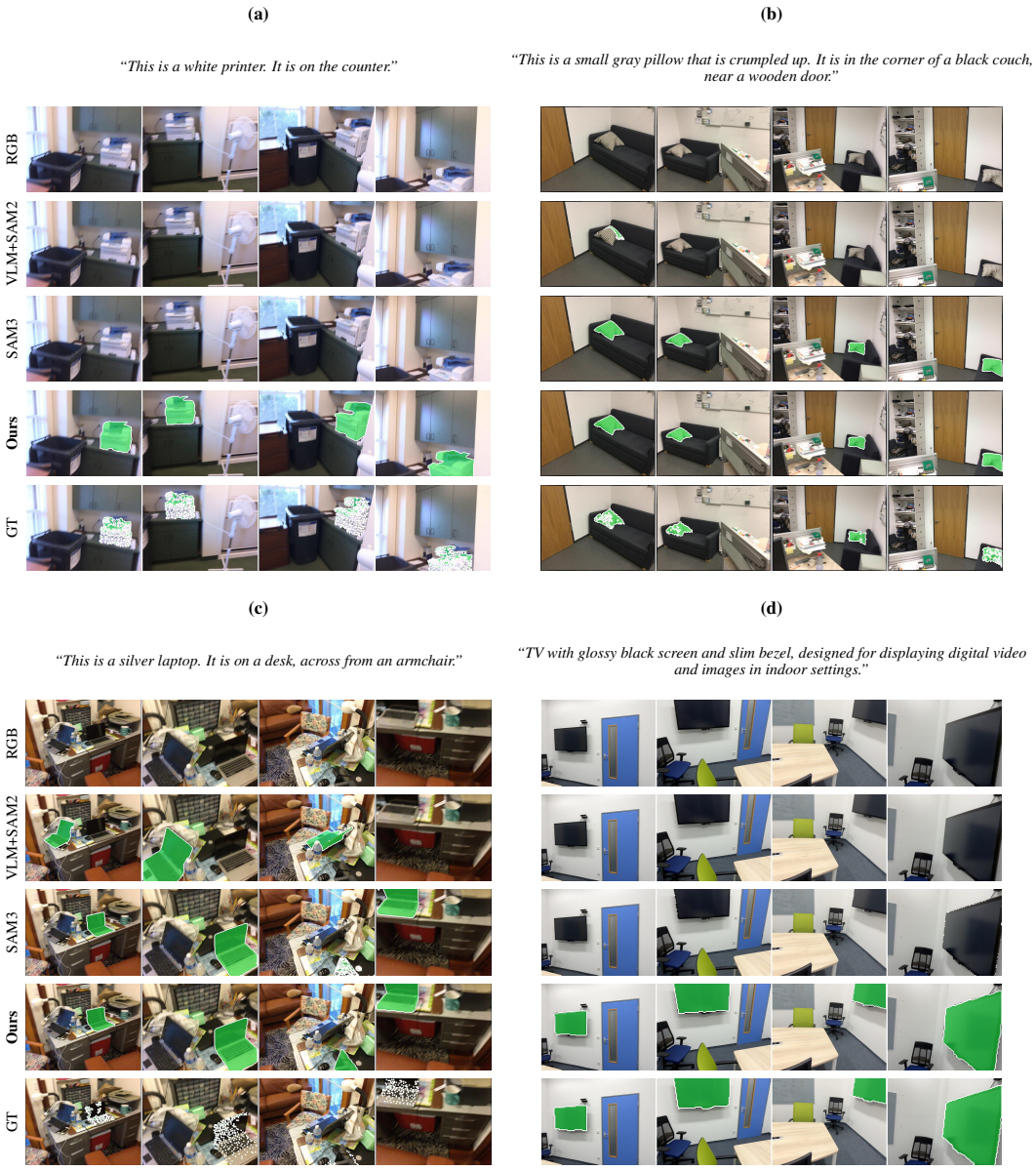


Figure 15: **Additional qualitative comparisons of language-guided grounding (I).** Ground-truth masks are projected from the annotated 3D point clouds.

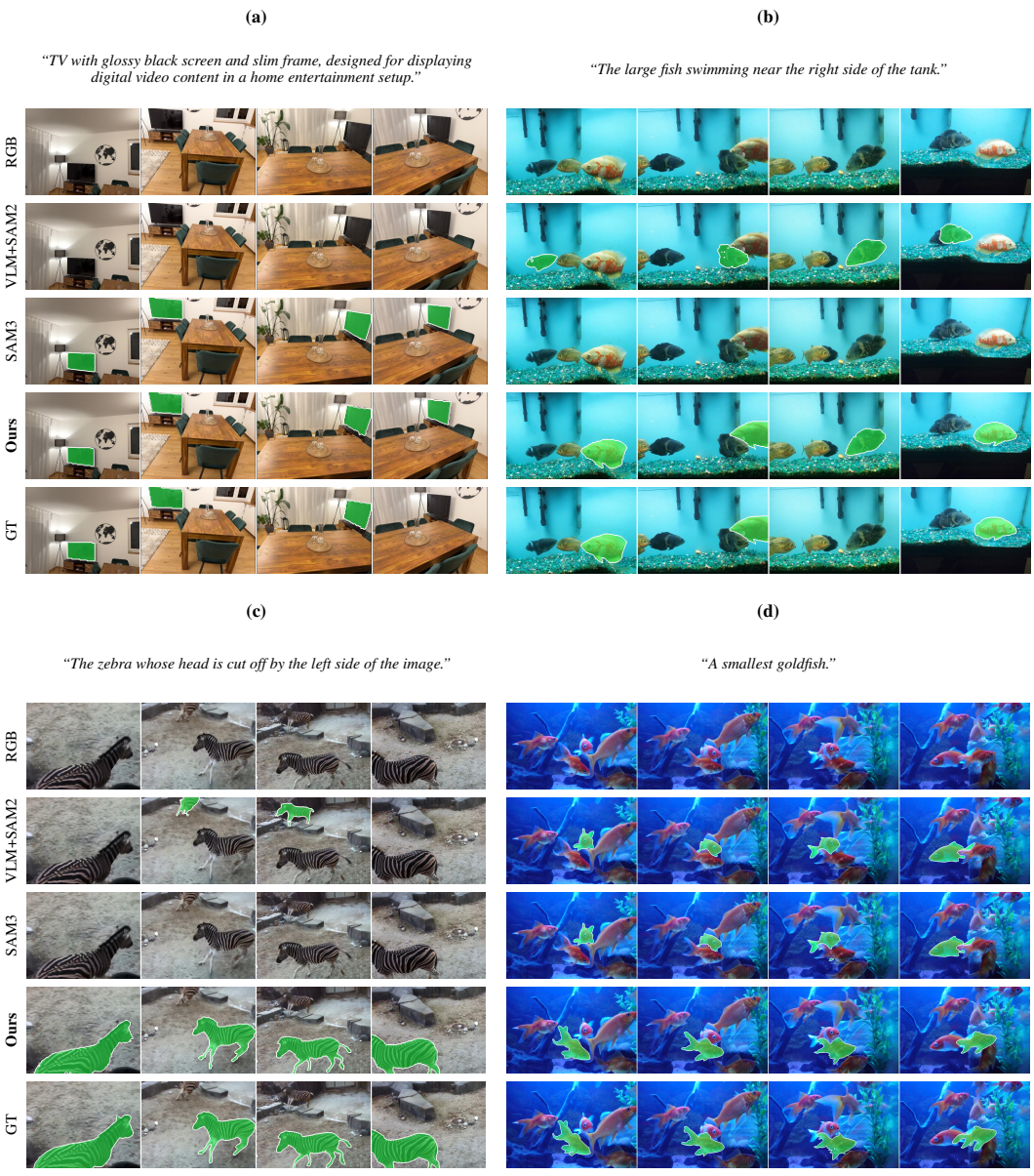


Figure 16: **Additional qualitative comparisons of language-guided grounding (II).**

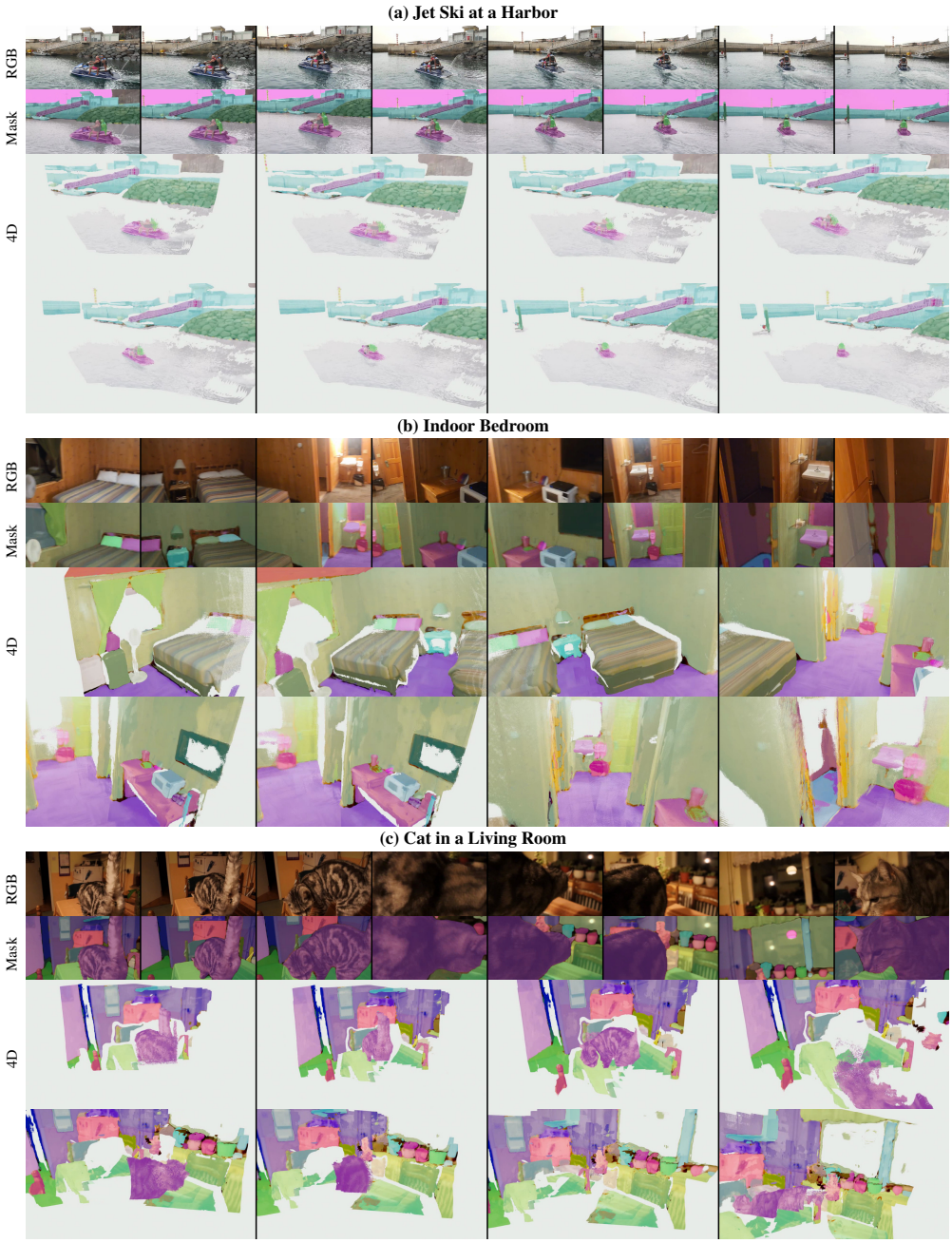


Figure 17: **In-the-wild 4D segmentation (I).** Uniformly sampled observations from in-the-wild data, with frame-wise masks and reconstructed 4D instances shown in the same temporal order.

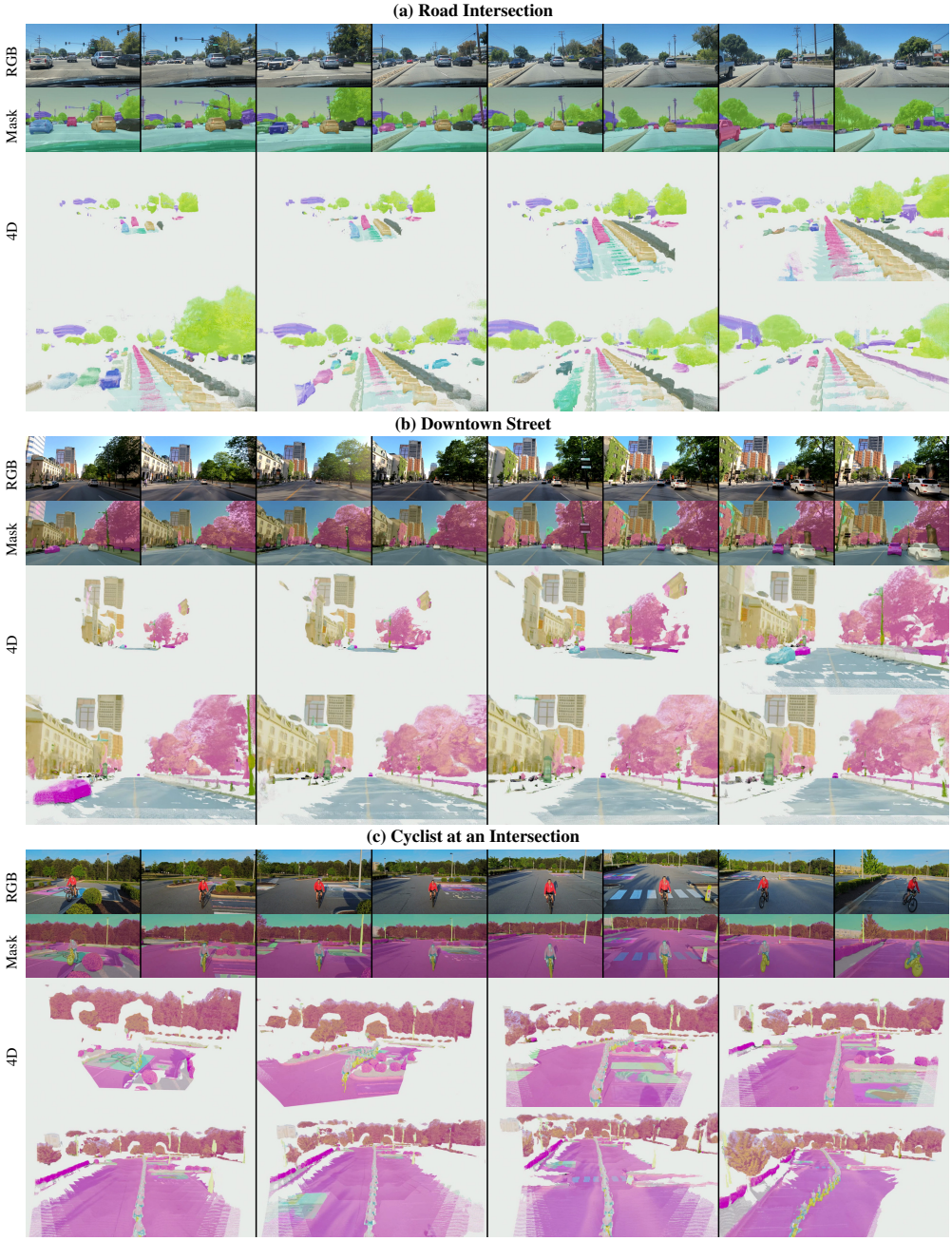


Figure 18: **In-the-wild 4D segmentation (II).** Uniformly sampled observations from in-the-wild data, with frame-wise masks and reconstructed 4D instances shown in the same temporal order.