ENTITY3R: AN END-TO-END ENTITY-AWARE GEOMETRIC FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

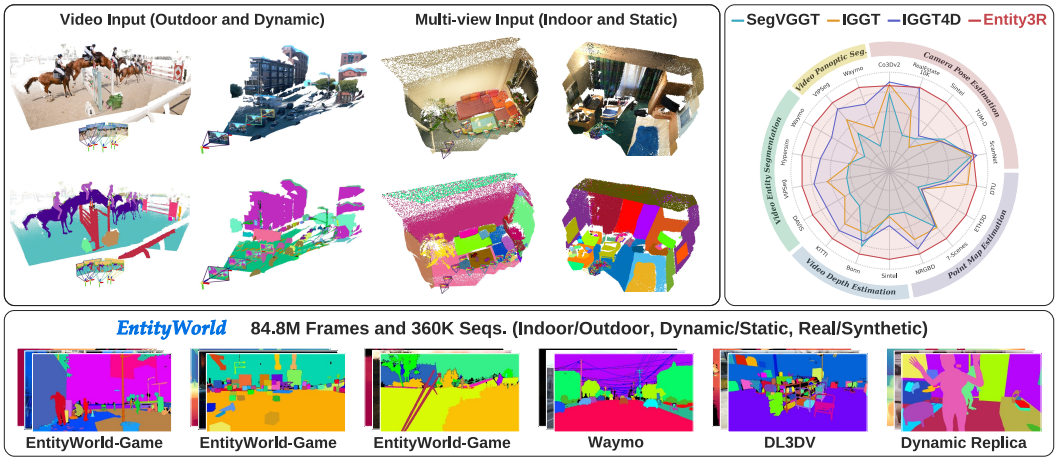


Figure 1: We present **Entity3R**, a unified end-to-end framework which can jointly perceive class-agnostic 3D entities in the open world and reconstruct scene geometry from several scene types. To address the data scarcity, we construct **EntityWorld**, an open-domain training corpus of nearly 84.8M frames and 360K sequences with geometry and entity masklet annotations, assembled from public geometric data resources and our collected game sequences. Entity3R achieves state-of-the-art performance on entity segmentation and geometric reconstruction benchmarks.

ABSTRACT

Class-agnostic 3D entity understanding aims to recover open-world entity identities with basic scene geometry from multi-view images or videos. Recent state-of-the-art methods built on instance-grounded geometric foundation models rely on hyperparameter-sensitive clustering to decode objects, which makes entity mask inference unreliable and costly. They also inherit the low native resolution of geometry transformers, limiting fine-grained entity perception. We present **Entity3R**, an end-to-end, entity-aware 3D/4D geometric foundation model. Entity3R extracts high-resolution geometry-instance representations with a dual-branch backbone, and decodes 3D/4D-consistent entity masks through learnable global object queries that interact with frame-level instance features. This design enables accurate and efficient 3D entity segmentation without post-hoc clustering. We further introduce a training-free, chunk-wise decoding strategy that allows Entity3R to decode more entities than its query budget and to process longer input sequences. To train Entity3R, we build **EntityWorld**, a large-scale corpus for joint geometry and instance learning, containing 360K sequences and 84.8M frames drawn from existing geometric datasets and newly collected photorealistic game data. Entity masklets are generated at scale by an automatic annotation pipeline. Trained on EntityWorld, Entity3R achieves state-of-the-art performance on both entity segmentation and geometric reconstruction benchmarks.

1 INTRODUCTION

Spatial intelligence underpins embodied robotics and autonomous driving that must perceive their surroundings in order to navigate, manipulate objects, and interact physically. One of its core ca-

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.

pabilities is 3D entity understanding, which requires recovering the basic geometry of a scene and perceiving the entities that occupy it beyond the predefined categories. This paper focuses on the task of class-agnostic 3D entity segmentation from the most widely used visual input (i.e., 2D images).

Earlier methods (Bhalgat et al., 2023; Zhu et al., 2025) mainly adopted an optimization-based multi-view fusion pipeline to lift 2D class-agnostic instance segmentation predictions to 3D scenes represented by reconstructed NeRF (Mildenhall et al., 2020) or 3DGS (Kerbl et al., 2023). Although promising results have been achieved, these methods require time-consuming optimization (e.g., more than 10 minutes), and cross-view inconsistencies in 2D instance segmentation predictions may further propagate to the 3D lifting results. More recently, the feedforward paradigm has revolutionized the field by designing class-agnostic instance-grounded geometric foundation models (Li et al., 2026; Zou et al., 2026). Specifically, these models directly predict instance embeddings alongside reconstructed geometry from multi-view or video inputs. Although improvements have been achieved by these methods, they are still limited by two limitations. First, these methods require hyperparameter-sensitive clustering post-processing for object decoding, leading to unreliable and costly entity mask inference. Second, they inherit the native resolution of current geometry transformers (518 pixels in width), which hinders the fine-grained segmentation for complicated scenes.

In this paper, we present **Entity3R**, a new unified end-to-end framework for effective class-agnostic entity understanding with geometric reconstruction, given multi-view images of static scenes or video frames of dynamic scenes. We explicitly impose entity-level perception in Entity3R by introducing global object queries coupled with frame-level instance maps, in addition to the basic geometric properties (i.e., camera poses and depth maps). In particular, we propose to learn fine-grained geometry-instance representation from high-resolution inputs with an efficient dual-branch design to preserve fine-grained entity information. The object queries are then decoded into cross-view entity masks by attending to the resulted high-resolution instance features, which forms the global scene context. During training, we match the mask predictions to ground-truth entities with the Hungarian algorithm (Kuhn, 1955), and each matched query is trained to predict the corresponding entity masklet. In this way, the queries and instance embeddings can be optimized via entity mask supervision, while the geometric properties are jointly learned with standard supervision following common practice (Wang et al., 2025a). During inference, Entity3R decodes the queries into entity predictions *in an end-to-end manner, improving accuracy over existing baselines while enabling superior efficiency* (e.g., $\mathbf{20}\times$ faster than the streaming clustering of IGGT4D (Zou et al., 2026)). Moreover, we introduce a chunk-wise object decoding strategy to further boost inference scalability in the number of frames without any retraining, allowing Entity3R to *cover more entities than a fixed set of queries can represent* in complex scenes and to *support longer input sequences* (e.g., from 80 to 180 frames on a 24-GB consumer GPU).

In addition to the architectural innovation, we further contribute **EntityWorld**, a new dataset for unified 3D entity perception and geometric reconstruction. To the best of our knowledge, *EntityWorld is the largest dataset for joint geometry and instance learning*, covering both static and dynamic scenes, as well as both indoor and outdoor environments. Besides existing public geometry datasets, EntityWorld includes a newly collected corpus of photorealistic game data with accurate geometry annotations, bringing the total to nearly 84.8M frames and 360K sequences. To obtain 3D/4D-consistent entity masks, we design an automatic annotation pipeline. Concretely, we first obtain initial entity mask proposals on the first frame with CropFormer (Qi et al., 2023a) and propagating them temporally with SAM2 (Ravi et al., 2025). An optical flow model (Teed & Deng, 2020) then associates re-detections in uncovered regions with lost SAM2 tracks or newly appearing objects, and flags large propagation jumps at which we split the sequence into new clips. The whole pipeline runs offline and automatically yields mask annotations at scale.

Our contributions are summarized as follows:

- We present **Entity3R**, a unified end-to-end framework for class-agnostic entity understanding with geometric reconstruction from multi-view images or video frames. It learns fine-grained geometry-instance representations from high-resolution inputs via an efficient dual-branch design, and jointly predicts cameras and depths while decoding 3D/4D-consistent entity masks by letting global object queries interact with instance features.
- We introduce a chunk-wise object decoding strategy that extends inference scalability without retraining, allowing Entity3R to cover more entities than a fixed query set can handle in one pass and to support longer input sequences.

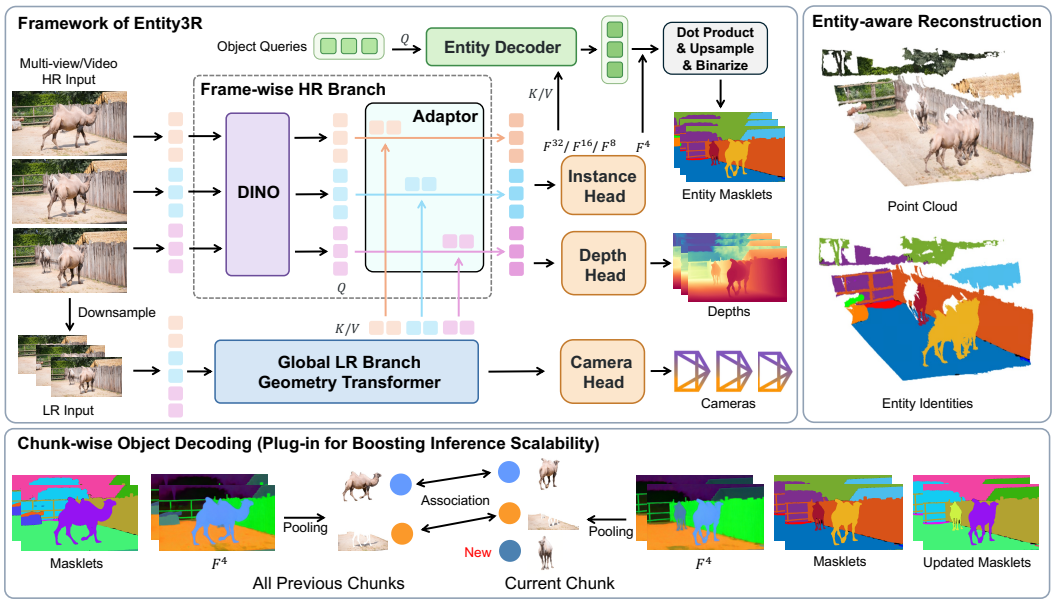


Figure 2: **Overview of Entity3R.** We achieve 3D class-agnostic entity understanding with geometric reconstruction in an end-to-end manner. Fine-grained geometry-instance representation is first extracted from high-resolution inputs using a dual-branch backbone. Then, the representation is fed into dense prediction heads for high-resolution depth and instance maps. Learnable object queries are updated in the entity decoder by alternately attending to the cross-view consistent instance features, and are decoded into class-agnostic entity masklets. We also propose a chunk-wise object decoding strategy as a memory-efficient plug-in for boosting inference scalability.

- We construct EntityWorld, an open-domain training corpus of nearly 84.8M frames and 360K sequences with geometry and entity masklet annotations, spanning existing geometric datasets and our newly collected game data. Our Entity3R achieves state-of-the-art performance on entity segmentation and geometric reconstruction benchmarks.

2 Related Work

Geometric Foundation Models. Feed-forward geometric foundation models (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026b; Lin et al., 2026; Wang et al., 2026a) reconstruct cameras and dense geometry from unposed multi-view or video inputs in a single forward pass, which is much more efficient than traditional SfM pipelines such as COLMAP (Schönberger & Frahm, 2016). DUSt3R (Wang et al., 2024) regresses dense point maps from image pairs. Extending the reconstruction beyond two views then requires aligning those pairwise reconstructions after inference, which is costly and can be unstable. VGGT (Wang et al., 2025a) removes this pairwise fusion and jointly predicts cameras, depths, and point maps for many views in one pass, with accuracy competitive with optimization-based pipelines. Subsequent geometry transformers follow this design: π^3 (Wang et al., 2026b) predicts permutation-equivariant cameras and point maps without a designated reference view, DA3 (Lin et al., 2026) predicts a singular depth-ray target with a plain transformer, and VGGT-Ω (Wang et al., 2026a) utilizes scales the training corpus with efficient register attention. Another line of work accelerates these transformers for long sequences (Shen et al., 2026; Wang & Xu, 2026; Jin et al., 2026) and enables streaming reconstruction (Wang et al., 2025b; Zhuo et al., 2026; Yuan et al., 2026; Chen et al., 2026). All above models remain focused on geometric reconstruction and do not assign object identity to the predicted point clouds, whereas Entity3R predicts open-world entity identities attached to the recovered 3D/4D geometry.

3D Scene Understanding. Integrating semantics into 3D reconstruction enables high-level reasoning and understanding of a scene. Optimization-based methods (Qin et al., 2024; Zhou et al., 2024; Zhai et al., 2026) inject vision-language features into 3D Gaussian Splatting (Kerbl et al., 2023). Feed-forward successors (Fan et al., 2024; Sun et al., 2026; Zhu et al., 2026) predict semantic 3D Gaussians from unposed multi-view images in a single pass, without per-scene optimization. Most

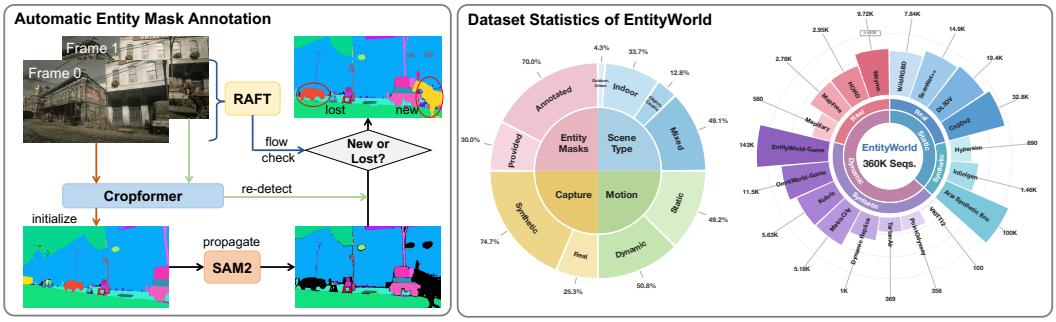


Figure 3: **Left:** Our automatic entity mask annotation pipeline for scalable entity masklet generation. **Right:** The dataset statistics of EntityWorld, which spanning diverse scene types.

of these works follow a two-stage design. They first extract per-view semantic features with 2D VLMs (Radford et al., 2021; Li et al., 2022) and then fuse these features across views in 3D, which can accumulate cross-view errors. Instance-grounded geometry transformers instead reconstruct geometry and object identity in one stage, learning directly from cross-view consistent instance masks. Among them, SegVGGT (Qu et al., 2026) trains 3D reconstruction and instance segmentation with semantic categories on ScanNet (Dai et al., 2017), and therefore does not transfer beyond this closed-set indoor setting. IGGT (Li et al., 2026) learns class-agnostic instance features through contrastive learning and obtains cross-view masklets by costly global clustering. Clustered masks provide instance identity rather than semantic features, and open-vocabulary understanding is executed by querying VLMs or MLLMs on these masks. Concurrent with our work, IGGT4D (Zou et al., 2026) replaces global clustering with a streaming procedure and extends training to dynamic indoor and driving data, yet object decoding still depends on scene-dependent clustering hyperparameters and can over- or under-segment. In contrast, Entity3R trains object queries (Carion et al., 2020) with Hungarian matching (Kuhn, 1955) and decodes entity masklets in an end-to-end and class-agnostic manner, using object queries without clustering.

3 Methodology

Entity3R is a class-agnostic 3D entity segmentation model that jointly reconstructs 3D/4D scene geometry and predicts open-world entity identities from unposed multi-view or video inputs. The overall architecture is illustrated in Fig. 2. In the following, Sec. 3.1 provides the preliminaries. Sec. 3.2 describes our fine-grained geometry-instance representation. Sec. 3.3 presents the end-to-end entity prediction. Sec. 3.4 gives the model training details.

3.1 Preliminaries

Visual Geometry Grounded Transformer. Given N unposed RGB images $\{I_i\}_{i=1}^N$, where each $I_i \in \mathbb{R}^{3\times H\times W}$ comes from a multi-view image set or a video stream, VGGT (Wang et al., 2025a) predicts cameras, depth maps and point maps in a single feed-forward pass: $f_\theta\big(\{I_i\}_{i=1}^N\big) = \big\{(\mathbf{g}_i, D_i, P_i)\big\}_{i=1}^N$. Here $D_i \in \mathbb{R}^{H\times W}$ is the depth map of image I_i, $P_i \in \mathbb{R}^{3\times H\times W}$ is the global point map, and $\mathbf{g}_i = (\mathbf{q}_i, \mathbf{t}_i, \mathbf{f}_i) \in \mathbb{R}^9$ is the concatenation of the rotation quaternion $\mathbf{q}_i \in \mathbb{R}^4$, the translation $\mathbf{t}_i \in \mathbb{R}^3$, and the field of view $\mathbf{f}_i \in \mathbb{R}^2$ parameterizing the camera of I_i. The geometry transformer f_θ consists of a DINO (Oquab et al., 2023) tokenizer and alternating-attention (Wang et al., 2025a) layers to aggregate global scene structure across views.

Class-agnostic Instance-grounded Geometry Transformer. IGGT series (Li et al., 2026; Zou et al., 2026) attaches an instance head to the geometry transformer to predict instance embeddings $\{\mathbf{F}_i \in \mathbb{R}^{8\times H\times W}\}_{i=1}^N$ besides geometry. This instance representation is learned by contrastive learning and can yield class-agnostic entity masklets by clustering. However, the clustering-based post-processing is time-consuming and hyperparameter-sensitive. Moreover, the instance embeddings are obtained from coarse geometry-instance feature tokens constrained by the native input resolution of about 518 pixels in width, which hinders fine-grained entity perception.

In this work, we seek an end-to-end framework for effective class-agnostic entity understanding without clustering, which directly operates on high-resolution inputs $\{I_i^{hr}\}_{i=1}^N$ for fine-grained

Table 1: **Class-agnostic video entity segmentation and class-aware video panoptic segmentation.** †We run offline inference with full global attention on IGGT4D (Zou et al., 2026) for fair comparison. ‡We additionally train an LR version of Entity3R, with HR branch removed. Runtime is measured with a 100-frame sequence on a single NVIDIA H20 GPU.

Method	Video Entity Seg.							Video Panoptic Seg.				Runtime
	DAVIS	VIPSeg		Hypersim		Waymo		VIPSeg		Waymo		
	$\mathcal{J}\&\mathcal{F}\uparrow$	VEQ↑	$\text{STQ}_{\text{EN}}\uparrow$	VEQ↑	$\text{STQ}_{\text{EN}}\uparrow$	VEQ↑	$\text{STQ}_{\text{EN}}\uparrow$	VPQ↑	STQ↑	VPQ↑	STQ↑	
EntitySAM (Ye et al., 2025)	71.4	61.4	43.3	10.6	11.0	24.1	11.1	29.3	33.2	13.8	6.9	86s
SegVGGT (Qu et al., 2026)	31.4	26.3	29.4	11.8	16.1	8.3	11.2	13.9	20.5	3.8	6.2	21s
IGGT (Li et al., 2026)	37.0	44.5	41.2	11.2	17.4	11.9	10.5	21.3	28.7	8.2	5.2	OOM
IGGT4D (Zou et al., 2026)†	59.9	58.1	48.1	24.7	28.0	27.1	17.5	27.8	35.2	15.2	10.7	36s
Ours (w/o HR branch)‡	66.5	63.7	49.5	22.8	22.7	27.9	17.7	28.7	36.4	16.0	11.9	**9s**
Ours	**78.9**	**71.2**	**53.6**	**34.3**	**31.6**	**40.3**	**22.9**	**31.7**	**40.0**	**20.8**	**16.0**	19s

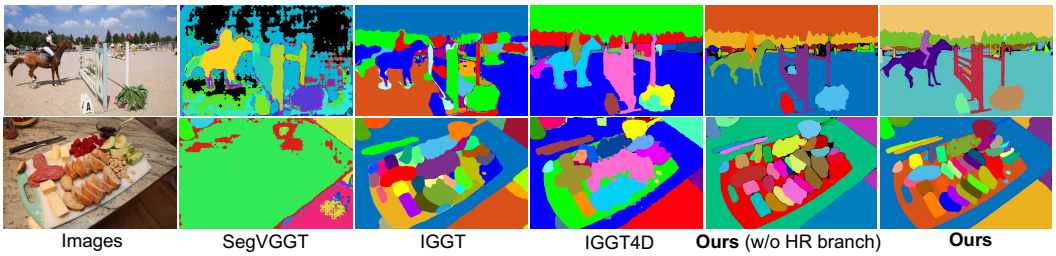


Figure 4: **Qualitative comparison of entity segmentation.** Entity3R produces higher-quality entity masks than others, and our high-resolution (HR) design yields more accurate object boundaries.

geometry-instance representation, formulated as:

$$f_\theta\big(\{I_i^{hr}\}_{i=1}^N\big) = \big\{(\mathbf{g}_i, D_i, M_i^{\text{ID}})\big\}_{i=1}^N, \tag{1}$$

Here, $M_i^{\text{ID}} \in \{0, 1, \ldots, O\}^{H\times W}$ is a dense object-ID map: pixels that share the same ID $j \geq 1$ belong to the same entity across views, and 0 denotes no-object. Unprojecting D_i and M_i^{ID} with $\mathbf{g}_i$ yields a global point cloud with each 3D point assigned an entity identity.

3.2 Fine-grained Geometry-Instance Representation

Directly processing high-resolution images with a geometry transformer is prohibitively expensive, as global attention constitutes the computational bottleneck. Inspired by DAGE (Ngo et al., 2026), we use a dual-branch backbone with high and low resolutions. The low-resolution (LR) branch follows the same procedure as geometry transformer to extract LR feature tokens $\{T_i^{lr}\}_{i=1}^N$ and camera tokens $\{T_i^{cam}\}_{i=1}^N$ from downsampled LR frames $\{I_i^{lr}\}_{i=1}^N$. The camera tokens are then fed to a camera head to predict camera parameters $\{\mathbf{g}_i\}_{i=1}^N$.

The High-resolution (HR) branch is responsible for extracting fine-grained semantic features and preserving object boundaries by encoding each HR frame independently. Since the DINO series (Oquab et al., 2023; Siméoni et al., 2025) provides strong dense representations for downstream depth estimation and semantic segmentation tasks, we directly leverage its pretrained priors and adopt it as the HR backbone to extract per-frame HR feature tokens $\{T_i^{hr}\}_{i=1}^N$. In this way, we keep the HR tokens out of the global attention layers, which reduces the computational cost significantly.

To fuse the HR tokens with the LR tokens such that the final representation is both fine-grained and globally consistent, we utilize a *frame-wise adaptor* in which the HR tokens serve as queries and the LR tokens as keys/values, allowing the HR tokens to retain fine-grained semantics while absorbing global scene information from the LR tokens. The adaptor is a lightweight network following DAGE with 3 transformer blocks, each structured as a cross-attention layer (CAttn) followed by a self-attention layer (SAttn) and a feed-forward MLP network (FFN), and formulated as:

$$T_i^{fuse} = \text{FFN}(\text{SAttn}(\text{CAttn}(Q = T_i^{hr}, K = T_i^{lr}, V = T_i^{lr}))), \tag{2}$$

Table 2: **Camera pose estimation.** We report pose AUC under angular error thresholds of 5/15 degrees for RealEstate10K (Zhou et al., 2018) and Co3Dv2 (Reizenstein et al., 2021), and distance errors of rotation/translation on Sintel (Butler et al., 2012), TUM-dynamics (Sturm et al., 2012), and ScanNet (Dai et al., 2017). All methods have seen Co3Dv2 and ScanNet in training.

Method	RealEstate10K		Co3Dv2		Sintel			TUM-dynamics			ScanNet		
	AUC/5 ↑	AUC/15 ↑	AUC/5 ↑	AUC/15 ↑	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓
DA3 (Lin et al., 2026)	61.02	79.47	**72.81**	**86.70**	0.117	0.043	0.336	0.012	0.009	0.312	0.033	**0.013**	0.636
π^3 (Wang et al., 2026b)	63.03	80.10	68.46	84.27	0.073	0.038	0.288	0.014	0.009	0.307	**0.030**	**0.013**	**0.345**
VGGT (Wang et al., 2025a)	30.71	60.01	67.83	83.95	0.171	0.062	0.471	0.012	0.010	0.309	0.035	0.015	0.376
VGGT-Ω (Wang et al., 2026a)	58.28	78.54	60.12	80.44	0.040	0.027	**0.239**	**0.007**	**0.006**	**0.279**	0.033	0.014	0.386
DAGE (Ngo et al., 2026)	52.49	74.67	62.90	81.62	0.141	0.054	0.416	0.014	0.011	0.323	0.031	**0.013**	0.386
SegVGGT (Qu et al., 2026)	19.80	50.78	57.64	79.40	0.109	0.053	0.442	0.012	0.009	0.314	0.035	0.015	0.382
IGGT (Li et al., 2026)	41.11	68.26	67.71	83.94	0.167	0.058	0.469	0.013	0.011	0.313	0.034	0.015	0.379
IGGT4D (Zou et al., 2026)	64.37	80.95	70.14	85.31	0.096	0.045	0.328	0.014	0.011	0.321	0.032	**0.013**	0.387
Ours	**65.27**	**81.64**	66.18	83.30	**0.038**	**0.024**	0.249	0.008	**0.006**	0.282	0.033	0.015	0.388

where $\{T_i^{fuse}\}_{i=1}^N$ serve as the *fine-grained geometry-instance representation* for the subsequent dense prediction heads. The depth head can predict HR depth maps $\{D_i\}_{i=1}^N$ from $\{T_i^{fuse}\}_{i=1}^N$ following the standard design in geometry transformers.

3.3 End-to-end Entity Prediction

End-to-end Entity Prediction. We introduce O learnable object queries $\{\mathbf{o}_j \in \mathbb{R}^C\}_{j=1}^O$ to detect entities in the scene, avoiding the hand-crafted and time-consuming clustering used by prior instance-grounded models (Li et al., 2026; Zou et al., 2026). To this end, we use Mask2Former-style (Cheng et al., 2022) pixel decoder, based on multi-scale deformable attention (Zhu et al., 2021), as our instance head to map the fused fine-grained tokens $\{T_i^{fuse}\}_{i=1}^N$ into multi-scale instance-specific features $\{F_i^s\}_{i=1}^N$, where $s \in \{32, 16, 8, 4\}$ and $F_i^s \in \mathbb{R}^{C \times \frac{H}{s} \times \frac{W}{s}}$. The instance features inherit the global consistency from the fine-grained geometry-instance representation. We therefore concatenate them across frames into a global scene context $F^s = \mathrm{Concat}(F_1^s, \ldots, F_N^s) \in \mathbb{R}^{C \times (N \times \frac{H}{s} \times \frac{W}{s})}$ for *full-context querying*. Following Mask2Former, our entity decoder updates the object queries using six transformer blocks that alternately attend to the intermediate features F^{32}, F^{16}, and F^8 over two rounds. Each block is formulated as:

$$\mathbf{o}_j = \mathrm{FFN}(\mathrm{SAttn}(\mathrm{CAttn}(Q = \mathbf{o}_j, K = F^s, V = F^s))). \tag{3}$$

where s takes values in $\{32, 16, 8, 32, 16, 8\}$ in order. After each block, the updated queries (including the initial queries) are dotted with F^4 to produce entity mask probabilities, $\mathrm{sigmoid}(\mathbf{o}_j^\top F^4)$, which provide deep supervision and are binarized into hard attention masks for the next block. Each query is also fed to an MLP binary classifier that predicts an entityness value $p_j \in \{0, 1\}$, indicating whether it corresponds to an object or to no-object. The dot product of the fully updated queries with F^4 yields the final results, which are split, reshaped, and upsampled into per-frame HR predictions and then resolved into dense object-ID maps $\{M_i^{\mathrm{ID}}\}_{i=1}^N$ under a non-overlapping constraint.

Chunk-wise Object Decoding to Boost Inference Scalability. Above full-context querying procedure lets every object query attend to the full scene context F^s over all N frames at once. This raises two issues. First, decoding over F^s makes memory grow with N. Second, as more frames bring more entities, a fixed query set gradually fails to cover them all. We therefore propose a memory-efficient *chunk-wise object decoding* strategy that is an optional plug-in to boost *inference scalability* without modifying or retraining the model, as illustrated in the bottom row of Fig. 2. Specifically, we partition the N frames into several non-overlapping chunks and predict entities chunk by chunk. Within each chunk, the queries attend only to that chunk's concatenated features. This bounds both the attended context memory and the number of entities per pass, while still producing cross-frame-consistent masklets inside each chunk. To keep entity IDs across chunks, we perform inter-chunk association by pooling each masklet over F^4 into a normalized feature center, and then matching centers in the current chunk to all previously established entity centers via cosine similarity. Matched centers are assigned to the corresponding entity IDs, while unmatched ones are treated as new entities and assigned new IDs. Note that the chunk-wise method applies only to object decoding, and all frames still enter the backbone (LR geometry transformer) jointly, which preserves the cross-frame consistency of the instance features F^4 and thereby ensures accurate inter-chunk association.

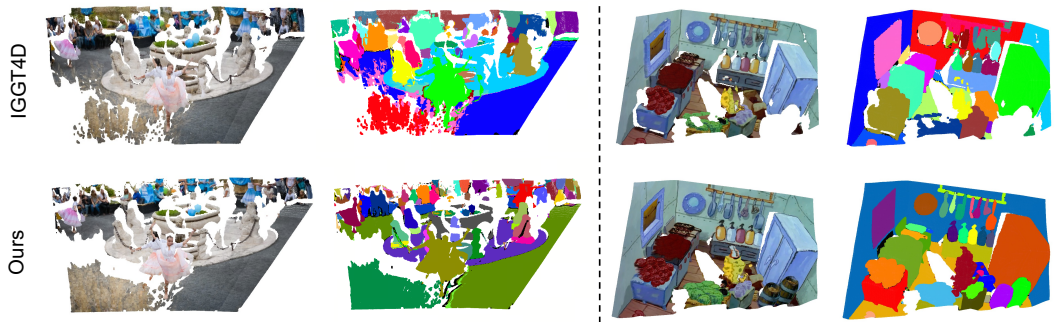


Figure 5: **Qualitative comparison of entity-colored point clouds.** For each scene we show the reconstructed point cloud and the corresponding entity identities. Compared with IGGT4D (Zou et al., 2026), Entity3R produces denser geometry with fewer holes and more accurate entity regions.

3.4 Model Training

We train Entity3R with the combination of geometry and entity supervisions: $\mathcal{L} = \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{ent}}$.

Geometry Loss. $\mathcal{L}_{\text{geo}}$ follows the standard design in geometry transformers (see in Sec. A.1).

Entity Loss. The entity loss comprises three terms: $\mathcal{L}_{\text{ent}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$. Following query-based image detection/segmentation (Carion et al., 2020; Cheng et al., 2021; 2022), we first match predicted masklets to ground-truth entity masklets with the Hungarian algorithm (Kuhn, 1955). For matched pairs, $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$ supervise the masklets with binary cross-entropy and Dice losses. $\mathcal{L}_{\text{cls}}$ is a binary classification loss on entityness as in (Qi et al., 2023b;a), where matched queries are trained as objects and unmatched queries as no-object. In this way, supervision is class-agnostic and uses no category vocabulary. We apply the entity loss with deep supervision (Cheng et al., 2022) to the object queries after they are updated by each block in entity decoder.

Implementation Details. Entity3R processes HR inputs at a width of 1024 and downsamples them to a width of 512 for LR branch. We initialize the LR geometry transformer, camera head and depth head from VGGT-Ω (Wang et al., 2026a), as its scalable training yields a stronger geometric backbone and its register attention improves training and inference efficiency. The HR branch uses a DINOv3 ViT-L/16 (Siméoni et al., 2025) encoder, followed by a 3-block frame-wise adaptor. The instance head is a Mask2Former-style pixel decoder based on multi-scale deformable attention (Zhu et al., 2021), with a ViT-Adapter (Chen et al., 2023) to extract multi-scale features. We prepare 200 learnable object queries to be updated by the 6-block entity decoder with masked attention. More details are provided in Sec. A.1.

4 EntityWorld Dataset

EntityWorld. Joint geometry reconstruction and entity perception need both geometric supervision and 3D/4D-consistent entity masklets at scale. We therefore contribute **EntityWorld**, a new dataset that contains nearly 84.8M frames and 360K sequences with cameras, depths, and class-agnostic entity masklets. EntityWorld is constructed by combining existing geometric datasets and our newly collected photorealistic game data. We follow OmniWorld (Zhou et al., 2026) to collect game sequences. Depth is accessed during rendering with ReShade (ReShade Contributors, 2024), and synchronized RGB frames are captured from the screen with OBS (OBS Contributors, 2024), yielding about 14.3M frames and 143K game sequences with accurate geometry labels. Game data can span diverse scene types including indoor, outdoor-urban, and outdoor-natural scenes, and thus complement the community sources that are dominated by indoor and outdoor-urban scenes. The full dataset composition of EntityWorld is summarized in Sec. A.2 and Tab. 7.

Automatic Entity Mask Annotation. Some of existing synthetic datasets already provide the entity mask GTs by the rendering engines and thus can be directly used. For the rest, we propose an automatic annotation pipeline for scalable entity masklet generation. Specifically, given RGB frames, we initialize entity mask proposals using CropFormer (Qi et al., 2023a) on the first frame. Compared to auto-segmentation in SAM series (Kirillov et al., 2023; Ravi et al., 2025) used by IGGT (Li

et al., 2026) labeling, CropFormer can produce higher-quality entity masks on an image without over-/under-segmentation (see the label comparison in Fig. 7). We then propagate the proposals temporally with SAM2 (Ravi et al., 2025). At each frame, we re-detect entities in the uncovered regions with CropFormer, while using an optical flow model (Teed & Deng, 2020) to determine whether a re-detected entity corresponds to a lost SAM2 track or a newly appearing object. We also use the optical flow model to detect large propagation jumps between adjacent frames, which may indicate an abrupt scene change. In such cases, we split the sequence at the jump and start a new clip from this frame, repeating the above procedure. We finally drop the sequences with less than 24 frames. The annotation pipeline and the dataset statistics are presented in Fig. 3.

5 EXPERIMENTS

This section compares Entity3R with the state-of-the-art methods on various benchmarks and explores the effectiveness of the key components in Entity3R.

5.1 VIDEO ENTITY SEGMENTATION

We evaluate Entity3R's basic entity perception ability on the video entity segmentation task, reporting the metrics VEQ and $\mathrm{STQ}^{\mathrm{EN}}$ proposed in EntitySAM (Ye et al., 2025), which are class-agnostic variants of the VPQ (Kim et al., 2020) and STQ (Weber et al., 2021) metrics for video panoptic segmentation. Considering that the ground-truth annotations are not exhaustive for all objects in the scene, we do not penalize false positive tracks directly in the metrics, as in Cheng et al. (2023). However, false positive tracks are indirectly penalized by requiring the output mask to be mutually exclusive. We adopt DAVIS (Pont-Tuset et al., 2017), VIPSeg (Miao et al., 2022), Hypersim (Roberts et al., 2021), and Waymo (Sun et al., 2020) for evaluation. Since DAVIS is a video object segmentation dataset, we report its widely used $\mathcal{J}\&\mathcal{F}$ (Perazzi et al., 2016) metrics instead. The comparison baseline models include EntitySAM, SegVGGT (Qu et al., 2026), IGGT (Li et al., 2026), and IGGT4D (Zou et al., 2026). The first is a video entity segmentation model based on SAM2 (Ravi et al., 2025) while the remaining are instance-grounded geometry transformers. To establish fair comparison, we additionally train an LR version of Entity3R, with HR branch removed. Results are shown in Tab. 1. Our LR model already surpasses the baselines on most metrics with the lowest runtime, while the stronger HR model comprehensively outperforms all methods at a still highly competitive inference speed. We present qualitative entity comparisons in Fig. 4 and Fig. 5, demonstrating Entity3R's superior entity masklet quality and object boundary accuracy. More qualitative results are available in the Sec. A.3.

5.2 DOWNSTREAM VIDEO PANOPTIC SEGMENTATION

Class-agnostic entity segmentation can be further applied to downstream open-vocabulary semantic understanding. Following IGGT (Li et al., 2026), we use the predicted entity masklets as prompts to an off-the-shelf VLM (Li et al., 2022). Vision-language features are average-pooled within each masklet and matched against text embeddings to assign a semantic category, yielding class-aware video panoptic predictions. We then evaluate on this video panoptic segmentation task, reporting VPQ/STQ metrics on VIPSeg and Waymo that contain category labels. As shown in Tab. 1, our method again leads on both datasets, indicating that our stronger entity perception also benefits class-aware semantic understanding.

5.3 GEOMETRY ESTIMATION

We assess Entity3R's geometry reconstruction ability following the same evaluation protocol of π^3 (Wang et al., 2026b) and ZipMap (Jin et al., 2026), covering camera pose estimation, video depth estimation, and point map estimation. For camera pose, we evaluate on RealEstate10K (Zhou et al., 2018), Co3Dv2 (Reizenstein et al., 2021), Sintel (Butler et al., 2012), TUM-dynamics (Sturm et al., 2012), and ScanNet (Dai et al., 2017) in Tab. 2. For video depth, we evaluate on Sintel (Butler et al., 2012), KITTI (Geiger et al., 2012), and Bonn (Palazzolo et al., 2019) in Tab. 4. For point map estimation, we evaluate on 7-Scenes (Shotton et al., 2013), NRGBD (Azinović et al., 2022), ETH3D (Schöps et al., 2017), and DTU (Jensen et al., 2014) in Tabs. 5–6. We can see that Entity3R

Table 3: **Ablation study results.** S denotes the chunk size for chunk-wise object decoding. Max #Frames is maximum number of frames that can be processed in an NVIDIA H20-96GB GPU. Runtime is measured with a 400-frame sequence on H20.

Method	**VIPSeg** (Shorter Seqs.)				**Waymo** (Longer Seqs.)				Max #Frames	Runtime @400F
	VEQ ↑	STQ_{EN} ↑	VPQ ↑	STQ ↑	VEQ ↑	STQ_{EN} ↑	VPQ ↑	STQ ↑		
(a) Inside-aggregation query interaction	64.8	48.2	27.5	35.1	33.6	18.4	16.2	11.8	–	–
(b) **Post-aggregation query interaction (Ours)**	**71.2**	**53.6**	**31.7**	**40.0**	**40.3**	**22.9**	**20.8**	**16.0**	–	–
(c) Chunk-wise object decoding (S=1)	**71.4**	**53.6**	31.7	**40.1**	**42.1**	**23.5**	**21.2**	**16.5**	**1200**	155s
(d) Chunk-wise object decoding (S=2)	**71.4**	**53.6**	31.7	**40.1**	41.9	**23.5**	20.9	16.4	**1200**	147s
(e) Chunk-wise object decoding (S=4)	**71.4**	53.5	31.7	**40.1**	41.7	23.4	21.0	**16.5**	**1200**	143s
(f) Chunk-wise object decoding (S=8)	71.3	53.5	**31.8**	40.0	41.5	23.3	20.9	16.3	**1200**	141s
(g) Chunk-wise object decoding (S=16)	71.3	53.5	31.7	40.0	41.2	23.1	20.7	16.3	**1200**	141s
(h) Full-context object decoding (w/o chunking)	71.2	**53.6**	31.7	40.0	40.3	22.9	20.8	16.0	400	**140s**

achieves performance comparable to state-of-the-art geometry models on most metrics across these benchmarks, while jointly providing entity perception.

5.4 Ablation Studies

Effect of Object Query Placement. We study two designs for how object queries interact with instance features: *inside-aggregation*, where object queries interact within the aggregation layers of the geometry transformer, following SegVGGT (Qu et al., 2026), an indoor segmentation method, and *post-aggregation*, our design where object queries are updated in a separate entity decoder after aggregation. We ablate this design choice in Tab. 3(a,b), showing that post-aggregation querying is clearly superior. Features inside the aggregation layers are still primarily geometric and not yet instance-specific, so early query interaction struggles to form reliable entity representations. In contrast, our entity decoder updates the queries with instance-specific features produced by the instance head, enabling more effective entity decoding.

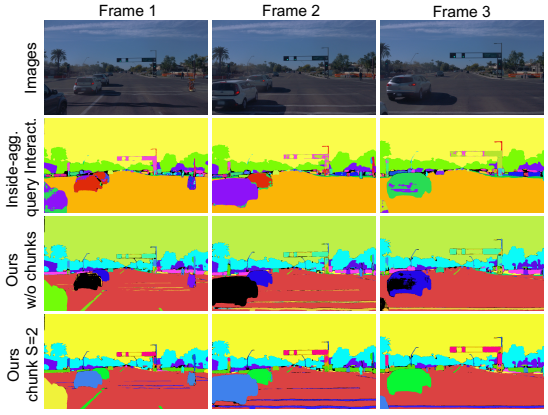


Figure 6: **Qualitative ablation results.** Our chunk-wise object decoding can handle longer sequences with more entities than a fixed set of queries can represent.

Effect of Chunk-wise Object Decoding. We explore the influence of chunk-wise object decoding at inference. As listed in Tab. 3(c–h), a small chunk size (S=1 or S=2) works best. On shorter sequences such as VIPSeg, chunking preserves accuracy without degradation. On longer sequences such as Waymo, splitting into chunks even improves performance by covering more entities under a fixed query budget. Chunk-wise decoding also supports more frames: full-context decoding OOMs beyond 400 frames, whereas chunk-wise alternatives can process up to 1200 frames in a memory-efficient manner. Moreover, the additional runtime cost introduced by inter-chunk association remains acceptable compared with full-context decoding. We also provide the qualitative results in Fig. 6.

6 Conclusion

We presented **Entity3R**, an end-to-end entity-aware geometric foundation model that jointly recovers class-agnostic entity identities with underlying scene geometry from multi-view images or videos. Entity3R learns fine-grained geometry-instance representations with a dual-branch backbone and decodes 3D/4D-consistent entity masks via learnable object queries, without clustering. A training-free chunk-wise strategy further improves scalability to support longer sequences. We further construct **EntityWorld**, a large-scale corpus of nearly 84.8M frames and 360K sequences for joint geometry and instance learning. Experiments show that Entity3R achieves state-of-the-art entity segmentation while remaining competitive on geometry reconstruction.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to generate synthetic datasets, or to design or provide feedback on research methodology or experiments. Helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, drafting parts of a research paper, summarizing or analysing existing literature, sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, formatting references, suggesting a structure for a research paper, and proposing a title or keywords for a research paper. We have carefully reviewed all AI-assisted outputs: AI-generated code for method implementations was verified and tested for correctness by the authors, translations were checked against the intended meaning, cleaned and reformatted data were inspected for correctness, qualitative analyses and result interpretations were validated, figures and images were inspected for correctness, drafted and edited manuscript text was revised as needed, literature summaries and identified references were cross-checked against the original sources, and suggested structure, title and keywords were finalized by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research methodology raises no ethical issues. No human subjects were involved.

REPRODUCIBILITY STATEMENT

Our Entity3R is built by integrating the publicly available codebases of VGGT-Ω (Wang et al., 2026a), pi3 (Wang et al., 2026b), DINOv3 (Siméoni et al., 2025), DAGE (Ngo et al., 2026), Mask2Former (Cheng et al., 2022), CropFormer (Qi et al., 2023a), SAM2 (Ravi et al., 2025), and RAFT (Teed & Deng, 2020). EntityWorld is constructed from publicly available geometric datasets together with our newly collected game sequences, for which we follow OmniWorld (Zhou et al., 2026) and use ReShade (ReShade Contributors, 2024) and OBS (OBS Contributors, 2024) for depth access and RGB capture. We include comprehensive data-preprocessing, data-annotation,training, inference, and implementation details in this paper and the appendix, which greatly facilitate reproducing our work. We will release our codes, models, and datasets if this work is accepted.

REFERENCES

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, 2022.

Dejan Azinović, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, 2022.

Yash Bhalgat, Iro Laina, João F. Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3D object instance segmentation by slow-fast contrastive fusion. In *NeurIPS*, 2023.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv:2001.10773*, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

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

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023.

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

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3D. In *NeurIPS*, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *CVPR*, 2022.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014.

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

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *CVPR*, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023.

Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

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

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, and Ziwei Liu. IGGT: Instance-grounded geometry transformer for semantic 3d reconstruction. In *ICLR*, 2026.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. MatrixCity: A large-scale city dataset for city-scale neural rendering and beyond. In *ICCV*, 2023.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, and et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024.

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 *CVPR*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Mapillary. Metropolis dataset, 2025. URL `https://www.mapillary.com/dataset/metropolis`.

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 *CVPR*, 2022.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Tuan Duc Ngo, Jiahui Huang, Seoung Wug Oh, Kevin Blackburn-Matzen, Evangelos Kalogerakis, Chuang Gan, and Joon-Young Lee. DAGE: Dual-stream architecture for efficient and fine-grained geometry estimation. In *CVPR*, 2026.

OBS Contributors. OBS studio, 2024. URL `https://obsproject.com/`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras combining statics and dynamics. In *IROS*, 2019.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *ICCV*, 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 *CVPR*, 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:1704.00675*, 2017.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. In *ICCV*, 2023a.

Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open-world entity segmentation. *TPAMI*, 45(7):8743–8756, 2023b.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *CVPR*, 2024.

Jinyuan Qu, Hongyang Li, and Lei Zhang. SegVGGT: Joint 3d reconstruction and instance segmentation from multi-view images. In *ECCV*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *CVPR*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021.

ReShade Contributors. ReShade, 2024. URL `https://reshade.me/`.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Fast visual geometry transformer. In *ICLR*, 2026.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew William Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv:2508.10104*, 2025.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, Xinggang Wang, and Eunbyung Park. Uni3R: Unified 3D reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In *CVPR*, 2026.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

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

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026a.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IROS*, 2020.

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

Zipeng Wang and Dan Xu. FlashVGGT: Efficient and scalable visual geometry transformers with compressed descriptor attention. In *CVPR*, 2026.

Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljoša Ošep, Laura Leal-Taixé, and Liang-Chieh Chen. STEP: Segmenting and tracking every pixel. In *NeurIPS*, 2021.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. RGBD objects in the wild: Scaling real-world 3D object learning from RGB-D videos. In *CVPR*, 2024.

Mingqiao Ye, Seoung Wug Oh, Lei Ke, and Joon-Young Lee. EntitySAM: Segment everything in video. In *CVPR*, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *ICCV*, 2023.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual geometry grounded transformer for endless streams. *arXiv:2601.02281*, 2026.

Hongjia Zhai, Qi Zhang, Xiaokun Pan, Xiyu Zhang, Yitong Dong, Huaqi Zhang, Dan Xu, and Guofeng Zhang. OnlinePG: Online open-vocabulary panoptic mapping with 3D gaussian splatting. In *CVPR*, 2026.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D gaussian splatting to enable distilled feature fields. In *CVPR*, 2024.

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

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. OmniWorld: A multi-domain and multi-modal dataset for 4d world modeling. In *ICLR*, 2026.

Runsong Zhu, Shi Qiu, Zhengzhe Liu, Ka-Hei Hui, Qianyi Wu, Pheng-Ann Heng, and Chi-Wing Fu. Rethinking end-to-end 2D to 3D scene segmentation in Gaussian splatting. In *CVPR*, 2025.

Runsong Zhu, Jiaxin Guo, Xiaoyang Guo, Zhengzhe Liu, Ka-Hei Hui, Wei Yin, Kai Chen, Wei Chen, Weiqiang Ren, Yunhui Liu, Pheng-Ann Heng, and Chi-Wing Fu. EPS3D: End-to-end feed-forward 3D panoptic segmentation. In *ICML*, 2026.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.

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

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

Table 4: **Video depth estimation.** We report results under scale-only alignment.

Method	Sintel		KITTI		Bonn	
	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
DA3 (Lin et al., 2026)	0.302	0.687	0.058	0.978	0.052	0.972
π^3 (Wang et al., 2026b)	0.229	0.672	**0.038**	**0.986**	0.051	0.975
VGGT (Wang et al., 2025a)	0.295	0.649	0.073	0.963	0.055	0.971
VGGT-Ω (Wang et al., 2026a)	**0.101**	**0.911**	0.056	0.979	<u>0.044</u>	0.972
DAGE (Ngo et al., 2026)	0.274	0.665	<u>0.042</u>	<u>0.983</u>	0.055	<u>0.978</u>
SegVGGT (Qu et al., 2026)	0.288	0.674	0.116	0.907	0.048	0.977
IGGT (Li et al., 2026)	0.272	0.663	0.066	0.971	0.062	0.970
IGGT4D (Zou et al., 2026)	0.212	0.723	0.056	0.968	0.053	0.972
Ours	<u>0.109</u>	<u>0.860</u>	0.047	0.980	**0.042**	**0.980**

Table 5: **Point map estimation on ETH3D (Schöps et al., 2017) and DTU (Jensen et al., 2014).** Following π^3 (Wang et al., 2026b) setting, we report the mean and median of accuracy (Acc.), completeness (Comp.), and normal-consistency (NC.) for keyframes selected every 5 images.

Method	ETH3D						DTU					
	Acc. ↓		Comp. ↓		NC. ↑		Acc. ↓		Comp. ↓		NC. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
DA3 (Lin et al., 2026)	0.267	0.204	0.327	0.213	0.839	0.919	1.409	0.716	1.581	0.687	0.676	0.764
π^3 (Wang et al., 2026b)	0.187	0.125	0.210	0.129	0.872	0.967	**1.153**	0.623	1.797	0.632	0.668	0.754
VGGT (Wang et al., 2025a)	0.271	0.173	0.305	0.180	0.841	0.942	1.310	0.763	1.932	1.018	0.666	0.751
VGGT-Ω (Wang et al., 2026a)	**0.061**	**0.038**	**0.056**	**0.036**	**0.906**	**0.984**	<u>1.207</u>	0.624	1.921	0.628	0.670	0.758
DAGE (Ngo et al., 2026)	0.207	0.140	0.213	0.131	0.858	0.968	1.305	<u>0.619</u>	1.687	<u>0.485</u>	0.645	0.714
SegVGGT (Qu et al., 2026)	0.299	0.210	0.306	0.185	0.839	0.949	2.334	1.365	1.976	0.831	0.685	0.774
IGGT (Li et al., 2026)	0.254	0.164	0.270	0.154	0.866	0.966	1.455	0.830	1.953	1.036	<u>0.687</u>	<u>0.780</u>
IGGT4D (Zou et al., 2026)	0.253	0.186	0.296	0.174	0.853	0.926	2.041	1.062	**1.544**	0.678	**0.691**	**0.784**
Ours	<u>0.076</u>	<u>0.041</u>	<u>0.070</u>	<u>0.037</u>	<u>0.893</u>	<u>0.981</u>	1.253	**0.600**	<u>1.561</u>	**0.470**	0.648	0.719

A Appendix

A.1 Model Training Details

Geometry Loss. The geometry loss consists of point, depth, and camera terms: $\mathcal{L}_{\text{geo}} = \lambda_{\text{point}}\mathcal{L}_{\text{point}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}}$. We apply π^3-style (Wang et al., 2026b) scale-invariant local geometry reconstruction loss and affine-invariant camera pose loss. For this purpose, we remove the designated reference camera view and supervise relative poses together with scale-aligned local point maps, making the geometry objective reference-free. Concretely, we unproject the predicted depths with the predicted cameras to obtain per-frame local point maps, estimate a scene-level scale that aligns them to the ground truth, and compute the point, depth and relative camera pose losses after scale alignment. $\mathcal{L}_{\text{point}}$ and $\mathcal{L}_{\text{cam}}$ are computed same as in π^3, while $\mathcal{L}_{\text{depth}}$ follows VGGT (Wang et al., 2025a) setting, containing aleatoric uncertainty and a gradient consistency term.

Implementation. We train Entity3R on EntityWorld dataset with AdamW (Loshchilov & Hutter, 2019) optimizer for 160K iterations, using 32 NVIDIA H20-96GB GPUs. The learning rate follows a linear warm-up over 5% of training and cosine decay for the remaining 95%. We set a peaking learning rate of 1×10^{-5} for the HR DINOv3 (Siméoni et al., 2025) encoder and the LR VGGT-Ω (Wang et al., 2026a) geometry transformer, and 1×10^{-4} for the remaining modules. Training proceeds in two phases: the first 80K iterations train with only the geometry loss $\mathcal{L}_{\text{geo}}$, and the remaining 80K iterations jointly optimize with $\mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{ent}}$. Following VGGT (Wang et al., 2025a), we randomly apply color jitter, Gaussian blur, and grayscale to the input frames as data augmentation. For training stability, we normalize the ground-truth cameras, depths, and local points using the global point cloud scale. During training, HR input images are resized to a width of 1024 pixels with a random aspect ratio sampled from $[0.33, 1.0]$, and the LR branch receives frames downsampled by $1/2$ to a width of 512. We sample 1-12 frames per sequence and use 12 images per GPU. To keep

Table 6: **Point map estimation on 7-Scenes (Shotton et al., 2013) and NRGBD (Azinović et al., 2022).** Metrics are the same as in Tab. 5. Keyframes are selected every 200 (sparse) and 40 (dense) frames for 7-Scenes, and every 500 (sparse) and 100 (dense) frames for NRGBD.

Method	7-Scenes						NRGBD					
	Acc. ↓		Comp. ↓		NC. ↑		Acc. ↓		Comp. ↓		NC. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Sparse View												
DA3 (Lin et al., 2026)	0.041	0.030	0.060	0.044	0.764	0.877	0.027	0.014	0.028	0.013	**0.922**	**0.995**
π^3 (Wang et al., 2026b)	0.047	0.028	0.074	0.048	0.741	0.840	0.024	0.014	0.028	0.014	0.909	0.991
VGGT (Wang et al., 2025a)	0.044	0.024	0.056	0.033	0.733	0.846	0.050	0.027	0.067	0.038	0.881	0.979
VGGT-Ω (Wang et al., 2026a)	0.026	0.016	0.031	0.020	**0.773**	**0.891**	**0.021**	**0.012**	**0.018**	**0.011**	0.911	0.992
DAGE (Ngo et al., 2026)	0.042	0.023	0.058	0.036	0.675	0.747	0.035	0.017	0.037	0.015	0.882	0.991
SegVGGT (Qu et al., 2026)	0.048	0.031	0.063	0.039	0.741	0.850	0.046	0.022	0.051	0.024	0.878	0.988
IGGT (Li et al., 2026)	0.045	0.030	0.062	0.040	0.752	0.866	0.038	0.020	0.046	0.020	0.910	0.993
IGGT4D (Zou et al., 2026)	0.037	0.024	0.064	0.047	0.760	0.871	0.028	0.017	0.031	0.014	**0.922**	**0.995**
Ours	**0.021**	**0.012**	**0.028**	**0.016**	0.708	0.794	0.023	**0.012**	**0.018**	**0.011**	0.895	0.989
Dense View												
DA3 (Lin et al., 2026)	0.018	0.007	0.023	0.009	0.689	**0.795**	0.013	**0.006**	0.014	0.006	0.890	0.987
π^3 (Wang et al., 2026b)	0.016	0.007	0.022	0.011	0.689	0.792	0.013	0.007	0.014	0.006	0.874	0.981
VGGT (Wang et al., 2025a)	0.022	0.008	0.026	0.012	0.666	0.760	0.015	0.008	0.015	0.006	0.871	0.982
VGGT-Ω (Wang et al., 2026a)	0.016	0.007	**0.018**	0.007	**0.691**	**0.795**	**0.012**	**0.006**	0.008	0.005	0.860	0.979
DAGE (Ngo et al., 2026)	0.015	0.006	0.020	0.009	0.638	0.709	0.015	0.008	0.014	0.005	0.842	0.976
SegVGGT (Qu et al., 2026)	0.016	0.007	0.027	0.011	0.681	0.780	0.032	0.014	0.027	0.010	0.838	0.973
IGGT (Li et al., 2026)	0.016	0.007	0.024	0.009	0.687	0.791	0.016	0.010	0.014	0.006	0.890	0.990
IGGT4D (Zou et al., 2026)	0.017	0.007	0.025	0.011	0.688	0.793	0.013	0.008	0.013	0.005	**0.905**	**0.991**
Ours	**0.012**	**0.005**	**0.018**	**0.006**	0.645	0.717	**0.012**	**0.006**	**0.006**	**0.004**	0.848	0.973

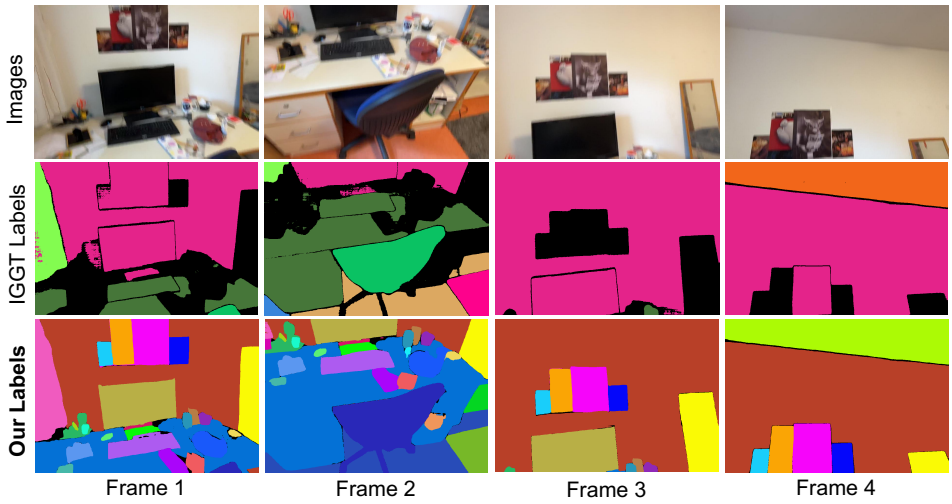


Figure 7: Comparison of entity masklet annotations. Compared with the IGGT (Li et al., 2026) labels, our annotation pipeline generates more complete and higher-quality entity masklets.

training efficient, we use bf16 mixed precision, gradient checkpointing, and gradient accumulation with 2 steps. For the loss weights, we set λ_{point}=1, λ_{depth}=1, λ_{cam}=5, λ_{cls}=1 and λ_{bce}=λ_{dice}=2.

A.2 EntityWorld Dataset Details

We first collect several publicly available geometric datasets in the community: Mapillary (Mapillary, 2025), Waymo (Sun et al., 2020), VirtualKITTI2 (Cabon et al., 2020), MatrixCity (Li et al., 2023), ScanNet++ (Yeshwanth et al., 2023), HOI4D (Liu et al., 2022), Hypersim (Roberts et al.,

2021), Aria Synthetic Environments (Pan et al., 2023), PointOdyssey (Zheng et al., 2023), Dynamic Replica (Karaev et al., 2023), Infinigen (Raistrick et al., 2024), WildRGBD (Xia et al., 2024), Co3Dv2 (Reizenstein et al., 2021), Kubric (Greff et al., 2022), MapFree (Arnold et al., 2022), DL3DV (Ling et al., 2024), TartanAir (Wang et al., 2020), and OmniWorld-Game (Zhou et al., 2026). We further add our newly collected EntityWorld-Game sequences. Tab. 7 lists every source used to build EntityWorld, together with its scene type, motion type, capture type, entity-mask origin, and the number of frames and sequences. In total, EntityWorld contains nearly 84.8M frames and 360K sequences, spanning diverse domains including indoor/ourdoor and dynamic/static scenes. We provide some data samples of EntityWorld in Fig. 9.

Table 7: **Dataset composition in EntityWorld.** *Provided* means the dataset provides entity mask labels; *Annotated* means we annotate the masks using our CropFormer+SAM2 pipeline.

Dataset	Scene Type	Dynamic	Real	Entity Masks	# Frames	# Seqs
Mapillary (Mapillary, 2025)	Outdoor-urban	Dynamic	Real	Annotated	34K	580
Waymo (Sun et al., 2020)	Outdoor-urban	Dynamic	Real	Annotated	703K	9.72K
VirtualKITTI2 (Cabon et al., 2020)	Outdoor-urban	Dynamic	Synthetic	Provided	42.5K	100
MatrixCity (Li et al., 2023)	Outdoor-urban	Dynamic	Synthetic	Annotated	516K	5.18K
ScanNet++ (Yeshwanth et al., 2023)	Indoor	Static	Real	Annotated	1.7M	14.9K
HOI4D (Liu et al., 2022)	Indoor	Dynamic	Real	Annotated	891K	2.95K
Hypersim (Roberts et al., 2021)	Indoor	Static	Synthetic	Provided	66.8K	690
Aria Synthetic Env. (Pan et al., 2023)	Indoor	Static	Synthetic	Provided	58.1M	100K
PointOdyssey (Zheng et al., 2023)	Indoor	Dynamic	Synthetic	Provided	106.5K	358
Dynamic Replica (Karaev et al., 2023)	Indoor	Dynamic	Synthetic	Annotated	300.6K	1K
Infinigen (Raistrick et al., 2024)	Indoor	Static	Synthetic	Provided	146K	1.46K
WildRGBD (Xia et al., 2024)	Object-centric	Static	Real	Annotated	1.41M	7.84K
Co3Dv2 (Reizenstein et al., 2021)	Object-centric	Static	Real	Annotated	2.65M	32.8K
Kubric (Greff et al., 2022)	Object-centric	Dynamic	Synthetic	Provided	135.1K	5.63K
MapFree (Arnold et al., 2022)	Mixed	Dynamic	Real	Annotated	504.4K	2.78K
DL3DV (Ling et al., 2024)	Mixed	Static	Real	Annotated	2.06M	19.4K
TartanAir (Wang et al., 2020)	Mixed	Dynamic	Synthetic	Provided	306.6K	369
OmniWorld-Game (Zhou et al., 2026)	Mixed	Dynamic	Synthetic	Annotated	813.5K	11.5K
EntityWorld-Game	Mixed	Dynamic	Synthetic	Annotated	14.3M	143K
EntityWorld	–	–	–	–	**84.8M**	**360K**

A.3 Additional Qualitative Results

Beyond the qualitative comparisons in the main paper, we provide more results in Fig. 8.

A.4 Runtime Analysis

Tab. 8 reports the per-component runtime on a 100-frame sequence with an NVIDIA H20-96GB GPU. First, our query-based object decoding costs less than 3 s across settings, whereas IGGT4D (Zou et al., 2026) spends 20.9s / 57.3s on clustering post-processing at the two resolutions. Second, feeding high-resolution inputs directly into the geometry transformer is prohibitively expensive: at a width of about $\sim$1024, IGGT4D and our LR variant (w/o HR branch) both run the geometry transformer at full resolution, taking 231.7s / 62.1 GB and 100.8s / 76.9 GB, respectively. In contrast, our full Entity3R model keeps the geometry transformer on low-resolution tokens and recovers fine-grained details with a frame-wise HR encoder that bypasses global attention, reducing the cost to 18.7s with only 27.3 GB peak memory—about $5\times$–$12\times$ faster and much more memory-efficient than pushing high-resolution inputs through the geometry transformer.

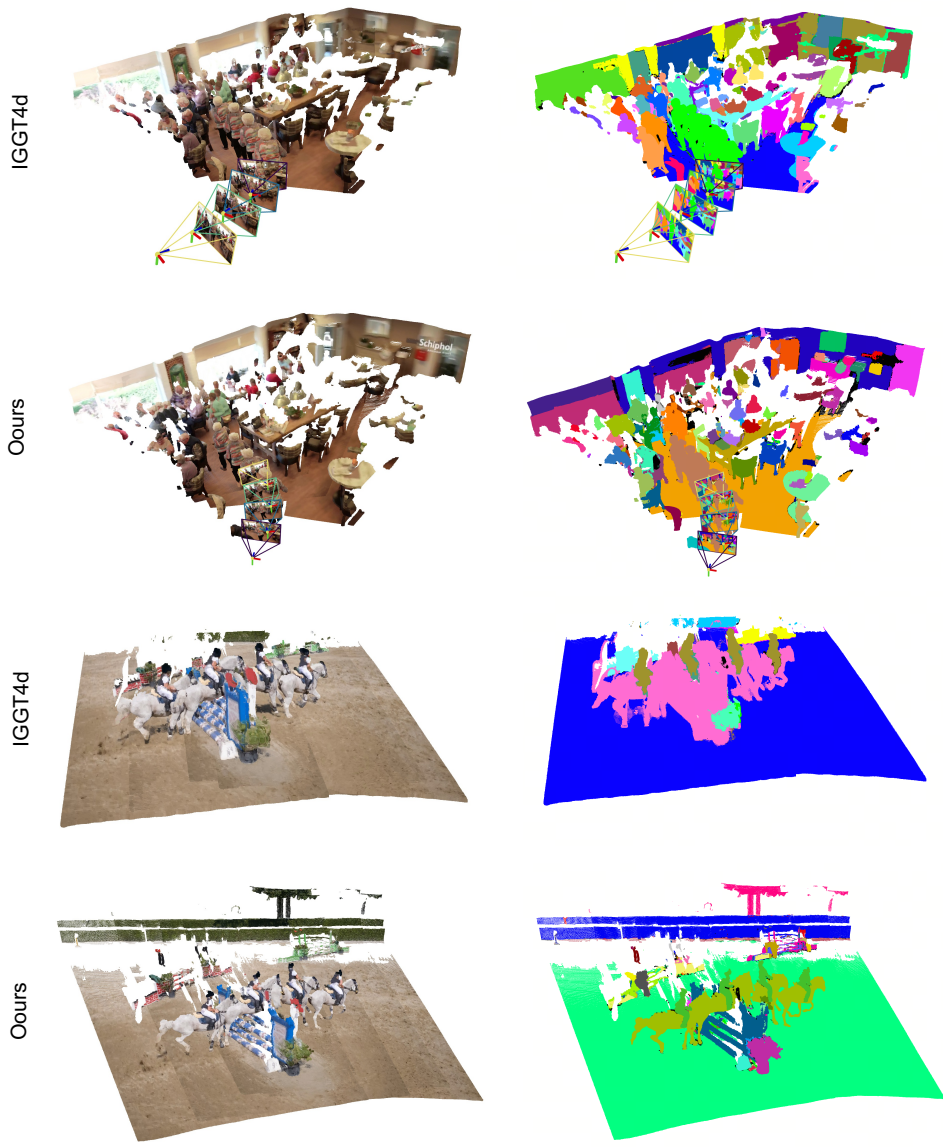


Figure 8: Additional qualitative comparisons of 3D entity segmentation against IGGT4D (Zou et al., 2026).

Table 8: **Runtime breakdown** measured on a 100-frame sequence with an NVIDIA H20-96GB GPU. Times are in seconds; Mem. is the peak GPU memory in GB. Geo. Trans. is the geometry transformer; HR Enc. is the HR encoder; Geo. is the geometry (camera and depth) heads; Inst. is the instance head; Cluster is the clustering post-processing; Query is our query decoding. †Offline inference with full global attention for IGGT4D (Zou et al., 2026), matching Tab. 1.

Method	Input Width	**Backbone**		**Head**		**Object Decoding**		Total	Mem.
		Geo. Trans.	HR Enc.	Geo	Inst.	Cluster	Query		
IGGT4D (Zou et al., 2026)†	504	12.3	–	1.3	1.7	20.9	–	36.2	21.0
Ours (w/o HR branch)	512	7.6	–	0.3	1.2	–	0.6	**9.3**	22.6
IGGT4D (Zou et al., 2026)†	1008	162.1	–	5.3	7.1	57.3	–	231.7	62.1
Ours (w/o HR branch)	1024	94.3	–	1.0	4.0	–	2.4	100.8	76.9
Ours-HR	1024	7.6	5.9	0.9	3.9	–	1.2	<u>18.7</u>	27.3

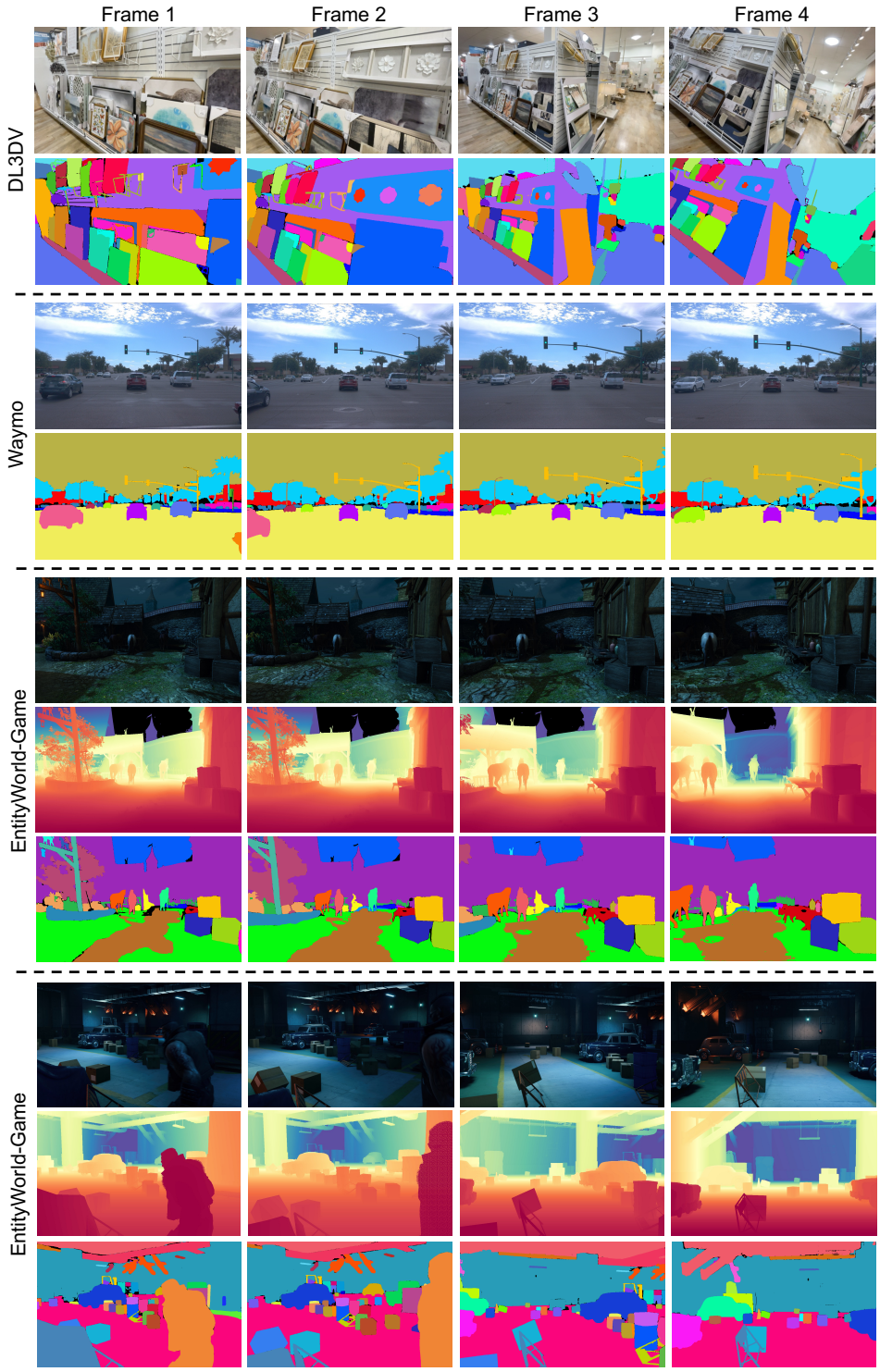


Figure 9: EntityWorld data samples across diverse domains. For each sequence we show RGB frames together with the corresponding entity masklets. For our self-collected EntityWorld-Game sequences, we additionally show the rendered depth maps.

A.5 LIMITATIONS

Despite strong entity segmentation and competitive geometry, Entity3R still has several limitations. First, although chunk-wise object decoding improves inference scalability, the geometry transformer

still consumes all frames jointly, so very long videos remain constrained by backbone memory rather than by the query decoder. Extending Entity3R to extremely long sequences is a promising direction, and we can draw on recent streaming geometric reconstruction methods (Chen et al., 2026; Zhuo et al., 2026; Yuan et al., 2026) to scale the backbone beyond full-sequence attention. Second, our automatic CropFormer+SAM2 (Qi et al., 2023a; Ravi et al., 2025) annotation pipeline tends to under-segment *stuff* regions, e.g., different buildings are often merged into a single entity. We hope to construct high-quality synthetic datasets with instance-level stuff masks that distinguish such regions, and to fine-tune Entity3R on them to better separate stuff entities.