OPENU3R: TOWARDS ZERO-SHOT GENERALIZABLE FEED-FORWARD OPEN-VOCABULARY UNDERSTANDING AND 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Jointly performing 3D Gaussian reconstruction and scene understanding in real-time is essential for embodied agents operating in unseen environments, yet existing feed-forward methods are restricted to closed vocabularies or struggle to generalize across datasets. Thus, we present OpenU3R, an open-vocabulary feed-forward framework that leverages complementary multi-view cues to support high-quality 3D Gaussian reconstruction together with semantic and instance-level scene understanding. Instead of using a fixed closed-set classifier, OpenU3R projects object queries into a shared vision-language embedding space and matches them against text embeddings, thereby endowing our method with open-vocabulary semantic capability. We further propose confidence-aware multi-view CLIP distillation, which pools frozen CLIP features within object masks across views and injects them into the object queries, improving cross-dataset transfer. On ScanNet, OpenU3R outperforms the best-performing baseline by 7.86 percentage points in mIoU, 5.43 percentage points in PQ, and 3.36 percentage points in mAP while maintaining competitive novel-view synthesis quality. Under zero-shot evaluation on RealEstate10K, OpenU3R improves over the same baseline by 29.00 percentage points in mIoU, 32.93 percentage points in PQ, and 11.95 percentage points in mAP, demonstrating strong cross-dataset generalization.

.

1 INTRODUCTION

The rapid development of embodied intelligence (Zhang et al., 2026) has increased the demand for agents to perceive, reconstruct, and understand unseen environments in real-time. Beyond recovering scene geometry and appearance, embodied agents also require semantic- and instance-level understanding to support downstream navigation and interaction (Qu et al., 2025). Recent advances in neural radiance fields (Mildenhall et al., 2021) and 3D Gaussian Splatting (Kerbl et al., 2023) have substantially improved 3D reconstruction quality and rendering efficiency, while large-scale visual representation learning has enabled increasingly powerful scene understanding. These advances motivate unified feed-forward frameworks that can jointly perform 3D reconstruction and scene understanding from sparse multi-view observations.

Early attempts to bridge reconstruction and understanding typically distill 2D semantic features from pretrained vision-language models, such as CLIP (Radford et al., 2021) and LSeg (Li et al., 2022), into NeRF or 3D Gaussian representations (Kobayashi et al., 2022; Kerr et al., 2023; Zhou et al., 2024; Qin et al., 2024). While these approaches enable language-aware 3D representations, they generally require costly per-scene optimization, limiting their applicability to real-time embodied intelligence systems. LSM (Fan et al., 2024) brings this paradigm into a feed-forward setting by predicting 3D Gaussians together with compressed vision-language features. However, feature compression may discard fine-grained semantic information, while features extracted from off-the-shelf 2D vision-language models provide limited instance-level reasoning (Xu et al., 2025). Uni3R (Sun et al., 2026) further improves global multi-view consistency through a Cross-View Transformer and supports open-vocabulary semantic segmentation with dense semantic feature fields, but it does not explicitly model object identities or instance-level relationships. SIU3R (Xu et al., 2025) moves be-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

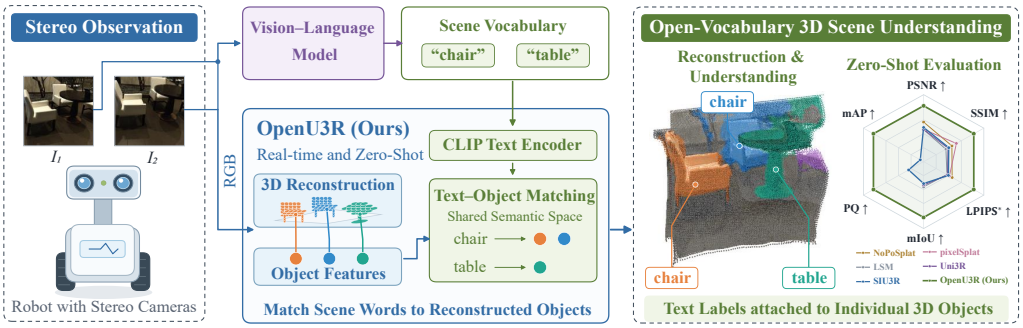

Figure 1: **Overview of OpenU3R.** Given a set of multi-view images, OpenU3R jointly performs real-time 3D Gaussian reconstruction and open-vocabulary scene understanding. A vision-language model identifies the scene vocabulary, while the feed-forward framework matches object features to text embeddings in a shared semantic space, associating open-vocabulary category labels with the corresponding reconstructed 3D objects. Colors in the reconstructed scene denote distinct object instances. The radar plot reports zero-shot transfer performance on RealEstate10K across reconstruction and scene-understanding metrics. LPIPS* is reversely normalized such that a larger radius consistently indicates better performance across all metrics.

yond feature alignment by introducing pixel-aligned representations, unified object queries, and bidirectional interactions between reconstruction and understanding, substantially improving semantic, instance, and panoptic prediction. However, its category prediction for these understanding tasks still relies on predefined training categories, and its generalization to broader visual domains remains limited. Overall, jointly achieving object-level open-vocabulary understanding and robust cross-dataset generalization within a real-time feed-forward reconstruction framework remains insufficiently explored.

To address these limitations, we propose OpenU3R, a real-time feed-forward framework for *zero-shot generalizable 3D Gaussian reconstruction and open-vocabulary scene understanding*. As illustrated in Figure 1, given sparse multi-view images captured by a robot or other vision-equipped device, OpenU3R jointly reconstructs the 3D scene and parses its object-level structure. A frozen vision-language model provides semantic cues for the observed scene, while OpenU3R associates reconstructed objects with textual concepts by matching object representations and text embeddings in a shared semantic space. Unlike closed-set classifiers, this design removes the dependence on a fixed vocabulary. Under zero-shot evaluation, OpenU3R consistently outperforms existing methods across scene-understanding metrics, demonstrating stronger cross-dataset generalization. Our main contributions are summarized as follows:

- We propose **OpenU3R**, a real-time feed-forward framework that jointly performs 3D Gaussian reconstruction and open-vocabulary semantic, instance, and panoptic scene understanding, while exhibiting strong cross-dataset zero-shot generalization.
- We introduce **Complementary Feature Fusion**, which integrates appearance semantics with multi-view geometric cues to better capture object identity and cross-view consistency, producing more robust object representations and substantially improving scene-understanding performance.
- We develop **Open-Vocabulary Query Recognition**, which projects object queries and text embeddings into a shared vision-language semantic space for language-guided category recognition, thereby removing the dependence on a fixed vocabulary.
- We further propose **Confidence-aware Multi-View Distillation**, which aggregates frozen CLIP features within object masks across views and selectively distills reliable semantic knowledge into object queries, strengthening object-level semantic representations and improving cross-dataset zero-shot transfer.
- We construct **RealEstate10K-483**, providing multi-view-consistent semantic- and instance-level annotations as a challenging zero-shot evaluation benchmark for cross-dataset open-vocabulary scene understanding.

2 RELATED WORK

Feed-Forward 3D Reconstruction. Classical Structure-from-Motion and Multi-View Stereo pipelines (Schonberger & Frahm, 2016; Schönberger et al., 2016), as well as 3D Gaussian Splatting (Kerbl et al., 2023), typically require scene-specific optimization. Feed-forward methods instead learn transferable geometric priors and directly infer 3D structure from images. DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) predict dense point maps and cross-view correspondences, while recent geometric foundation models such as VGGT (Wang et al., 2025), VGGT-Ω (Wang et al., 2026), and Depth Anything 3 (Lin et al., 2025) further improve camera, depth, and geometry estimation. These models provide strong geometric generalization and increasingly robust cross-dataset priors, but semantic and instance-level scene understanding is not their primary objective. Consequently, geometry must still be connected to an object-centric semantic representation when the goal is unified reconstruction and understanding.

Open-Vocabulary 3D Understanding. CLIP (Radford et al., 2021) enables open-vocabulary recognition through a shared image–text embedding space. OpenScene (Peng et al., 2023) transfers language-aligned features to 3D semantic recognition, and Open3DIS (Nguyen et al., 2024) extends this paradigm to 3D instances using multi-view 2D guidance. More recent feed-forward systems incorporate language-aware semantics directly into reconstruction: SpatialSplat (Sheng et al., 2025) learns efficient semantic Gaussian representations, Uni3R (Sun et al., 2026) distills pretrained semantic features into a unified 3D representation, and FF3R (Zhou et al., 2026) strengthens geometry–semantic interaction. Dense language-aware fields offer category flexibility, but they mainly transfer per-pixel or per-Gaussian semantics and do not explicitly organize recognition around persistent object queries. In contrast, OpenU3R aligns geometry-aware object queries with CLIP and performs runtime query–text matching, so the semantic vocabulary can be replaced without changing network parameters.

Unified Reconstruction and Scene Understanding. LSM (Fan et al., 2024) jointly predicts geometry, appearance, and semantics from sparse unposed views. SIU3R (Xu et al., 2025) introduces pixel-aligned 3D representations and shared object queries, while IGGT (Li et al., 2025) and InstanceSplat (Jiang et al., 2026) further emphasize instance-aware structure. PE3R (Hu et al., 2026) combines multi-view geometry with pretrained semantic priors, and Ov3R (Gong et al., 2026) and EmbodiedSplat (Lee et al., 2026) extend unified reconstruction and understanding to video or streaming observations. These systems improve the coupling between geometry and semantics, yet a unified representation does not by itself make object semantics transferable beyond the training taxonomy or guarantee robust zero-shot category recognition. OpenU3R therefore couples object-level geometry with confidence-aware multi-view CLIP distillation and decouples category prediction from a fixed classifier, targeting zero-shot transfer while retaining feed-forward reconstruction.

3 METHOD

3.1 OVERVIEW

Given V sparse context images $\mathcal{I} = \{I_v\}_{v=1}^{V}$, OpenU3R predicts pixel-aligned 3D Gaussians together with object-level semantic and instance predictions that can be rendered to arbitrary novel views (Fig. 2). The reconstruction branch (Sec. 3.2) uses frozen VGGT-Ω (Wang et al., 2026) geometry features to estimate dense geometry and provide multi-view geometric features. The semantic branch (Sec. 3.3) combines those geometry tokens with CroCo appearance tokens, decodes shared object queries with Mask2Former (Cheng et al., 2022), and projects the queries into the CLIP embedding space. At training time, multi-view CLIP distillation further regularizes this projected representation; at inference, the same representation is matched with scene-dependent text embeddings. Query masks and category scores are finally transferred to their pixel-aligned Gaussians, yielding a unified representation for reconstruction and scene understanding.

3.2 RECONSTRUCTION BRANCH

Geometry and Gaussian prediction. We adopt VGGT-Ω (Wang et al., 2026) as the geometry backbone. Its backbone is frozen and jointly processes all context views to extract intermediate geometry tokens $\{F_v^{(l)}\}$. A trainable dense head predicts the per-pixel depth D_v and geometry

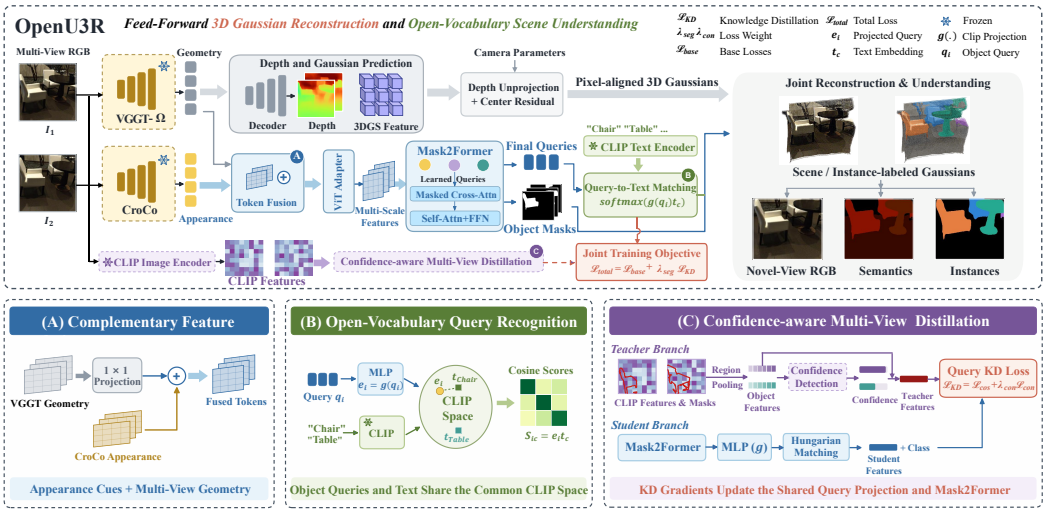

Figure 2: Overview of the OpenU3R pipeline. Given multi-view RGB images, OpenU3R jointly performs 3D Gaussian reconstruction and open-vocabulary scene understanding. The masks and scores predicted by the semantic branch associate instance identities and category labels with Gaussian primitives produced by the reconstruction branch, supporting novel-view RGB, semantic, and instance rendering. Our main contributions are as follows: **(A)** Fusing complementary VGGT-Ω geometry and CroCo appearance features produces richer representations that improve scene understanding performance. **(B)** Object queries are projected into the CLIP embedding space and matched with text embeddings for open-vocabulary recognition. **(C)** During training, confidence-aware multi-view CLIP distillation aligns student queries with reliable teacher prototypes, strengthening semantic representations and improving model generalization.

confidence C_v, while the known context-view camera intrinsics $\{K_v^{\mathrm{GT}}\}$ and poses $\{T_{v\to 0}^{\mathrm{GT}}\}$ are used for geometric unprojection. A Gaussian head predicts the center residual, opacity, scale, rotation, and spherical harmonic (SH) coefficients for each Gaussian. For a source pixel x, the initial Gaussian center is given by

$$\mu_v^0(x) = T_{v\to 0}^{\mathrm{GT}} \left(D_v(x) \left(K_v^{\mathrm{GT}}\right)^{-1} \tilde{x} \right), \tag{1}$$

where $\tilde{x}$ denotes the homogeneous image coordinate. Because the Gaussians remain pixel-aligned, each primitive maintains a direct correspondence with exactly one source pixel. This correspondence is central to our unified design: object masks and category evidence predicted in 2D can be directly transferred to the corresponding 3D Gaussian primitives, without requiring a separate 3D semantic decoder or an additional cross-representation association stage.

Bounded center refinement. Sparse-view depth can be slightly inconsistent across views, so we predict a bounded residual $\mathbf{r}(x) = (r_x, r_y, r_z)$ for each Gaussian. The image-plane offsets are limited to approximately ± 2 pixels and the depth correction to roughly $[-13\%, +15\%]$. The refined center is obtained by unprojecting the shifted pixel with the corrected depth. The bounded parameterization is deliberately conservative: it gives the renderer enough freedom to compensate for small cross-view depth errors but prevents the trainable head from drifting far from the geometry prior supplied by VGGT-Ω. The residual magnitude is further regularized by $\mathcal{L}_{\text{center-res}}$, with stronger penalties assigned to geometrically confident predictions.

Rendering and 3D semantic propagation. Novel RGB views are rendered with differentiable Gaussian splatting (Kerbl et al., 2023). Following the pixel-aligned lifting strategy of SIU3R (Xu et al., 2025), Gaussian g inherits the query–category evidence of its source pixel x_g, $A_{g,i,c} = p_{i,c}\sigma(m_i(x_g))$, where $p_{i,c}$ and m_i are the category probability and mask logit of query i. We render this evidence with the same alpha-compositing weights as RGB:

$$\widetilde{A}_{i,c}(x') = \sum_{g\in\mathcal{N}(x')} T_g(x')\,\alpha_g(x')\,A_{g,i,c}, \qquad T_g(x') = \prod_{j<g} \big(1-\alpha_j(x')\big). \tag{2}$$

At a target pixel, the strongest rendered category evidence determines the semantic label, while the corresponding query index provides the instance identity. This propagation uses exactly the same visibility ordering and alpha weights as RGB rendering, so the semantic prediction remains spatially aligned with the reconstructed scene. Consequently, one Gaussian representation supports novel-view appearance, semantic segmentation, and instance rendering without storing an additional dense semantic field.

3.3 SEMANTIC BRANCH

Complementary feature fusion. CroCo features emphasize appearance, whereas VGGT-Ω tokens encode multi-view geometry and correspondence. For view v, selected geometry tokens are injected into CroCo interaction levels through a residual projection,

$$\widetilde{\mathbf{H}}_v^{(c)} = \mathbf{H}_v^{(c)} + P_l\left(\mathbf{F}_v^{(l)}\right), \qquad c = \pi(l), \qquad \mathbf{W}_l\big|_{t=0} = \mathbf{0}, \tag{3}$$

where P_l is a bias-free 1×1 convolution with weight $\mathbf{W}_l$, and $\pi(\cdot)$ maps selected VGGT-Ω layers to CroCo interaction levels. The residual form allows appearance features to remain the default representation while injecting geometry only where useful. Zero initialization is important in practice: at the beginning of training the branch behaves exactly like the pretrained CroCo pathway, after which the projection gradually learns to contribute depth, correspondence, and cross-view structure.

Object-query decoding. The fused tokens are converted to multi-scale features with a ViT Adapter (Chen et al., 2022) and decoded by Mask2Former (Cheng et al., 2022). A shared set of N learnable queries is used across all context views, encouraging one query to represent the same object across viewpoints. Each query produces a feature $\mathbf{q}_i \in \mathbb{R}^{D_q}$ and mask logits $m_i(x)$. During training, standard Hungarian matching assigns queries to ground-truth instances using a weighted combination of classification, mask, and Dice costs; unmatched queries are supervised as no-object. Because the query set is shared across views, one matched query aggregates evidence for the same object from different observations and therefore serves as the natural unit for open-vocabulary recognition and multi-view distillation. We retain the original Mask2Former mask-prediction pathway and modify only the representation and category-recognition mechanism attached to each query.

Open-vocabulary query recognition. Let $\mathcal{V}_{\text{train}} = \{c_1, \ldots, c_K\}$ denote the training vocabulary. Each category name is encoded by the frozen CLIP text encoder (Radford et al., 2021) with a fixed prompt $P(\cdot)$:

$$\hat{\mathbf{t}}_k = \frac{E_{\text{text}}(P(c_k))}{\|E_{\text{text}}(P(c_k))\|_2}, \qquad \hat{\mathbf{T}}_{\mathcal{V}} = [\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_K]^\top. \tag{4}$$

Instead of learning category-specific classifier weights, we project each object query into the CLIP semantic space, such that object representations can be directly matched with text embeddings in the shared vision–language space. Specifically, for object query $\mathbf{q}_i$, we compute

$$\begin{aligned} \hat{\mathbf{e}}_i = \mathrm{Norm}(g(\mathrm{LN}(\mathbf{q}_i))), \qquad s_{i,k} = \tau\, \hat{\mathbf{e}}_i^\top \hat{\mathbf{t}}_k, \\ \mathbf{p}_i = \mathrm{softmax}([s_{i,1}, \ldots, s_{i,K}]), \end{aligned} \tag{5}$$

where $\tau = \exp(s)$ is a learnable logit scale. Here, $\hat{\mathbf{e}}_i$ and $\hat{\mathbf{t}}_k$ lie in the same CLIP semantic space, and their cosine similarity provides the category score for query $\mathbf{q}_i$. During training, the query representation is matched against the text embeddings of $\mathcal{V}_{\text{train}}$. Importantly, category identities are represented by text embeddings rather than stored in category-specific classifier parameters. At inference time, the same query-to-CLIP projection can therefore be matched against a different vocabulary without retraining or modifying the semantic decoder. Specifically, a frozen Qwen3-VL model (Bai et al., 2025) proposes a scene-dependent vocabulary $\mathcal{V}_{\text{scene}} = \mathcal{G}_\phi(\mathcal{I})$ from the context images. The corresponding category names are encoded by the same CLIP text encoder and replace the training text embeddings in Eq. (5). In this way, Qwen3-VL provides a compact set of candidate semantic concepts, while OpenU3R performs object-level recognition by matching reconstructed object queries with these textual concepts in the shared semantic space. The resulting query labels are subsequently propagated to the 3D Gaussians through Eq. (2).

Confidence-aware multi-view distillation. Training only with the seen taxonomy can still overfit the projected queries to training categories. We therefore construct an instance-level CLIP teacher from multiple views. For instance k visible in view v with mask $\mathbf{M}_{k,v}$, we first pool the frozen dense CLIP feature $\mathbf{F}_v^{\text{clip}}$ inside the mask:

$$\mathbf{u}_{k,v} = \text{Norm}\left(\frac{\sum_x \mathbf{M}_{k,v}(x)\mathbf{F}_v^{\text{clip}}(x)}{\sum_x \mathbf{M}_{k,v}(x) + \epsilon} \right). \tag{6}$$

Single-view features can be unreliable under occlusion, truncation, motion blur, or large viewpoint changes. For each visible set Ω_k, we score an observation by its average cosine agreement with the remaining views,

$$a_{k,v} = \frac{1}{|\Omega_k| - 1} \sum_{v' \in \Omega_k \setminus \{v\}} \mathbf{u}_{k,v}^\top \mathbf{u}_{k,v'}. \tag{7}$$

The most consistent observation is used as a medoid. Views whose similarity to this medoid falls below an adaptive threshold are discarded rather than allowed to corrupt the teacher target. The retained set Ω_k^{in} is aggregated into a teacher prototype

$$\mathbf{p}_k = \text{Norm}\left(\sum_{v \in \Omega_k^{\text{in}}} w_{k,v}\mathbf{u}_{k,v} \right), \qquad \sum_{v \in \Omega_k^{\text{in}}} w_{k,v} = 1, \tag{8}$$

where $w_{k,v}$ reflects semantic consistency and visibility. The aggregation favors clear, mutually consistent observations while retaining complementary viewpoints. Each prototype also receives a confidence $\gamma_k \in [0, 1]$ derived from the quality of its retained views, and instances with too few reliable observations are excluded from distillation.

For the query $\sigma(k)$ matched to instance k, let $\hat{\mathbf{e}}_{\sigma(k)} = \mathbf{e}_{\sigma(k)}/\|\mathbf{e}_{\sigma(k)}\|_2$ and $\Gamma = \sum_k \gamma_k$. We first align each matched query with its multi-view teacher using

$$\mathcal{L}_{\text{KD-cos}} = \frac{1}{\Gamma} \sum_k \gamma_k \left(1 - \hat{\mathbf{e}}_{\sigma(k)}^\top \mathbf{p}_k \right), \tag{9}$$

and impose category-level structure with

$$\mathcal{L}_{\text{KD-con}} = -\frac{1}{\Gamma} \sum_k \gamma_k \log \frac{\sum_{j: y_j = y_k} \exp(\hat{\mathbf{e}}_{\sigma(k)}^\top \mathbf{p}_j / T_d)}{\sum_j \exp(\hat{\mathbf{e}}_{\sigma(k)}^\top \mathbf{p}_j / T_d)}. \tag{10}$$

We use $\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{KD-cos}} + \lambda_{\text{con}}\mathcal{L}_{\text{KD-con}}$. Teacher prototypes and confidences are detached, so gradients update only the semantic branch and projection g, not the frozen CLIP teacher. The cosine term transfers instance-level visual semantics, the contrastive term organizes same-category instances in the shared CLIP space, and confidence weighting prevents noisy teacher instances from dominating optimization. Because the same projection g is later compared with runtime text embeddings, this supervision acts directly on the representation used for unseen categories, while the original decoder query preserves the spatial and instance information required for mask prediction.

3.4 Training Objective

OpenU3R is optimized with reconstruction, object-level understanding, vision–language distillation, and geometric regularization. Following SIU3R (Xu et al., 2025), we use the same formulations for the RGB reconstruction loss $\mathcal{L}_{\text{RGB}}$, the perceptual loss $\mathcal{L}_{\text{LPIPS}}$, and the Mask2Former objective $\mathcal{L}_{\text{Mask2Former}}$. In our model, the classification term in $\mathcal{L}_{\text{Mask2Former}}$ is evaluated using the open-vocabulary query–text probabilities in Eq. (5). The semantic objective is

$$\mathcal{L}_{\text{semantic}} = \mathcal{L}_{\text{Mask2Former}} + s_{\text{distill}}(e)\mathcal{L}_{\text{distill}}, \tag{11}$$

where $s_{\text{distill}}(e)$ gradually activates CLIP query distillation during training.

For geometric regularization, we use instance-aware depth smoothness and a confidence-weighted Gaussian-center residual penalty:

$$\begin{aligned} \mathcal{L}_{\text{depth-smooth}} &= \sum_{d \in \{x,y\}} \text{mean}\Big(M_d \left| \nabla_d \hat{D} \right| \Big), \\ \mathcal{L}_{\text{center-res}} &= \frac{\sum_g w_g \frac{1}{3}\|r_g\|_1}{\sum_g w_g}, \qquad w_g = 0.25 + 0.75 c_g. \end{aligned} \tag{12}$$

Table 1: Quantitative comparison of 3D reconstruction and scene understanding on ScanNet (Dai et al., 2017). Best results are shown in **bold**.

Method	Novel-View Synthesis			Depth Estimation		Novel-View Understanding		
	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	mIoU ↑	PQ ↑	mAP ↑
pixelSplat	24.93	0.8065	0.2003	0.1812	0.4106	–	–	–
NoPoSplat	25.91	0.8147	0.1878	0.0944	0.2434	–	–	–
Uni3R	25.65	0.8042	0.1991	**0.0618**	**0.1796**	0.4879	–	–
LSM	21.88	0.7336	0.3035	0.0747	0.2190	0.2707	–	–
SIU3R	25.96	0.8220	**0.1841**	0.0742	0.2081	0.5920	0.6495	0.2714
Ours	**26.01**	**0.8330**	0.1849	0.0778	0.2122	**0.6706**	**0.7038**	**0.3050**

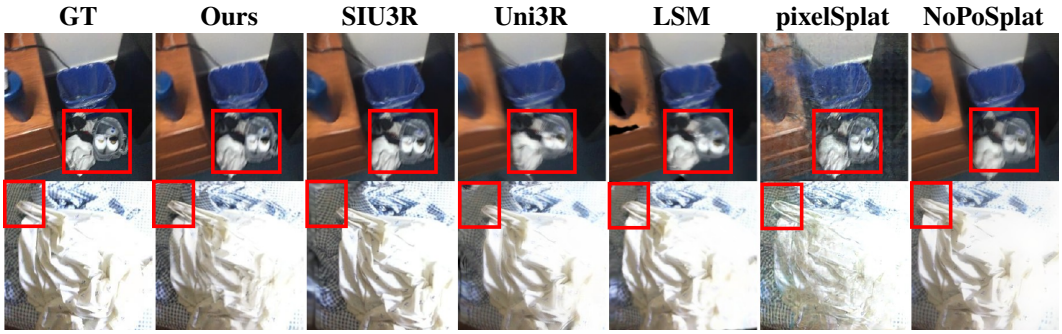

Figure 3: Qualitative comparison of reconstruction results on ScanNet.

Here, $\hat{D}$ is the rendered depth, and M_d retains neighboring pixel pairs assigned to the same predicted instance, so smoothing does not cross object boundaries. r_g denotes the normalized center residual of Gaussian g, while c_g is its geometry confidence; therefore, reliable VGGT-Ω estimates are constrained more strongly.

The overall objective is

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{RGB}}\mathcal{L}_{\text{RGB}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{semantic}}\mathcal{L}_{\text{semantic}} \\ + \lambda_{\text{smooth}}\mathcal{L}_{\text{depth-smooth}} + \lambda_{\text{center}}\mathcal{L}_{\text{center-res}}.\end{aligned} \tag{13}$$

Detailed loss weights, teacher filtering thresholds, and the distillation schedule are provided in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Datasets. Following SIU3R (Xu et al., 2025), we train and evaluate on the standard scene-disjoint ScanNet (Dai et al., 2017) splits. To further evaluate zero-shot transfer beyond the training distribution, we select 483 video clips from the RealEstate10K (Zhou et al., 2018) test split and design a multi-view-consistent automatic annotation pipeline to generate multi-view-consistent semantic and instance-level annotations, thereby constructing the RealEstate10K-483 benchmark. The resulting benchmark contains 2,896 sparse-view evaluation pairs (Appendix A.2). Its broader layouts, camera trajectories, and object concepts provide a complementary cross-dataset evaluation setting.

Implementation details. OpenU3R is trained on ScanNet for 100 epochs with AdamW (Loshchilov & Hutter, 2019) on two RTX A6000 GPUs (batch size 2, base learning rate 1×10^{-4}) at 256×256 resolution. VGGT-Ω (Wang et al., 2026) is frozen except for its dense prediction head; remaining optimization details are in Appendix A.3.

4.2 Main Results

In-domain reconstruction and scene understanding on ScanNet. Table 1 shows that OpenU3R preserves competitive reconstruction while substantially improving object-level understanding.

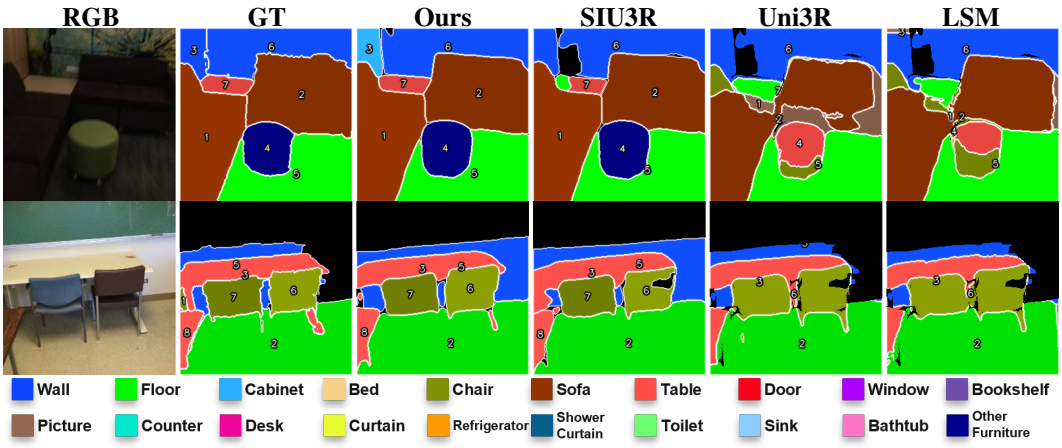

Figure 4: Qualitative comparison of segmentation results on ScanNet.

Table 2: Quantitative results on our RealEstate10K-483 benchmark under zero-shot cross-dataset evaluation. Best results are shown in **bold**.

Method	Novel-View Synthesis			Depth Estimation		Novel-View Understanding		
	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	mIoU ↑	PQ ↑	mAP ↑
NoPoSplat	13.37	0.3614	0.4899	0.2470	0.2844	–	–	–
pixelSplat	10.93	0.4181	0.5416	0.4140	0.3882	–	–	–
LSM	10.86	0.2844	0.6012	0.2897	0.3061	0.2160	–	–
Uni3R	10.86	0.3107	0.5867	0.3293	0.3290	0.2011	–	–
SIU3R	11.56	0.3206	0.5639	0.3043	0.3198	0.1878	0.1402	0.0085
Ours	**18.87**	**0.6436**	**0.2791**	**0.1763**	**0.2241**	**0.4778**	**0.4695**	**0.1280**

Relative to SIU3R, mIoU/PQ/mAP increase from 0.5920/0.6495/0.2714 to 0.6706/0.7038/0.3050 (+0.0786/+0.0543/+0.0336). PSNR/SSIM also improve from 25.96/0.8220 to 26.01/0.8330, while LPIPS remains comparable (0.1849 vs. 0.1841) and depth changes are small (AbsRel/RMSE 0.0778/0.2122 vs. 0.0742/0.2081). Qualitatively, Figs. 3 and 4 show that our method preserves fine reconstructed structures while producing cleaner and more complete semantic masks, consistent with the quantitative gains.

Zero-shot cross-dataset generalization on RealEstate10K-483. We evaluate the ScanNet-trained model on RealEstate10K-483 without fine-tuning, mapping annotations to the common ScanNet-20 vocabulary only for class-aligned metric computation. Table 2 shows strong transfer: OpenU3R reaches 18.87 dB PSNR, 0.6436 SSIM, and 0.2791 LPIPS; compared with the strongest reconstruction baselines, this is +5.50 dB PSNR, +0.2255 SSIM, and 0.2108 lower LPIPS. AbsRel/RMSE decrease to 0.1763/0.2241 from the best competing 0.2470/0.2844. For understanding, mIoU rises to 0.4778 from 0.2160 for LSM (+0.2618), while PQ/mAP improve over SIU3R from 0.1402/0.0085 to 0.4695/0.1280 (+0.3293/+0.1195). Figs. 5 and 6 further show that OpenU3R retains scene structure and object appearance more faithfully under domain shift and yields more spatially coherent masks with fewer incomplete predictions.

4.3 Ablation Study

Component analysis. Table 3 evaluates the three key components of OpenU3R. Removing **Complementary Feature Fusion** causes the most consistent degradation across the reported metrics: PSNR decreases from 26.01 to 26.00, LPIPS increases from 0.1849 to 0.1850, AbsRel/RMSE worsen from 0.0778/0.2122 to 0.0834/0.2250, and mIoU/PQ/mAP drop from 0.6706/0.7038/0.3050 to 0.5530/0.5760/0.2082, while SSIM remains unchanged at 0.8330. This demonstrates that sufficiently fusing CroCo appearance semantics with VGGT-Ω multi-view geometric cues benefits reconstruction, geometry estimation, and scene understanding as a whole. **Confidence-aware Multi-View Distillation** and **Open-Vocabulary Query Recognition** bring slight improvements in seman-

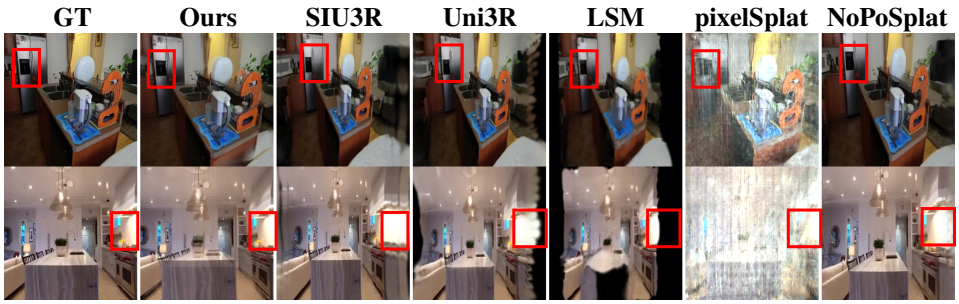

Figure 5: Qualitative comparison of reconstruction results on RealEstate10K-483.

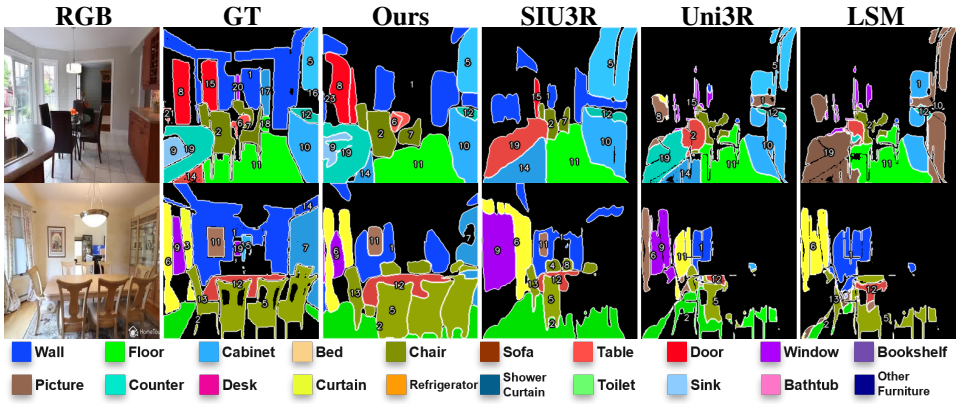

Figure 6: Qualitative comparison of segmentation results on RealEstate10K-483.

tic understanding while maintaining comparable reconstruction and overall in-domain performance. These two modules are designed primarily to enable open-vocabulary semantic understanding while preserving the underlying reconstruction and scene-understanding capability. Open-Vocabulary Query Recognition maps object features and text semantics into a shared CLIP embedding space, allowing the model to perform inference without relying on a fixed vocabulary, whereas Confidence-aware Multi-View Distillation strengthens transferable query semantics through reliable multi-view CLIP supervision. Overall, the ablations show that OpenU3R introduces open-vocabulary capability without materially compromising the original reconstruction and understanding performance.

Table 3: Ablation study of the main components of OpenU3R on ScanNet. Best results are shown in **bold**.

Method	Novel-View Synthesis			Depth Estimation		Novel-View Understanding		
	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	mIoU ↑	PQ ↑	mAP ↑
Ours w/o Confidence-aware Multi-View Distillation	**26.05**	0.8330	0.1854	0.0802	0.2177	0.6634	0.6989	**0.3364**
Ours w/o Open-Vocabulary Query Recognition	25.93	0.8300	0.1871	0.0825	0.2238	0.6626	0.7029	0.3324
Ours w/o Complementary Feature Fusion	26.00	0.8330	0.1850	0.0834	0.2250	0.5530	0.5760	0.2082
Ours	26.01	**0.8330**	**0.1849**	**0.0778**	**0.2122**	**0.6706**	**0.7038**	0.3050

5 Conclusion

We present OpenU3R, a feed-forward framework for joint 3D Gaussian reconstruction and open-vocabulary scene understanding. By integrating complementary appearance and multi-view geometric cues, OpenU3R builds object-centric representations for both reconstruction and semantic- and instance-level understanding. Open-vocabulary query recognition aligns object queries and text semantics in a shared vision-language space, removing the dependence on a fixed vocabulary, while Confidence-aware Multi-View Distillation strengthens transferable semantic representations through reliable multi-view CLIP supervision. Experiments across in-domain and zero-shot cross-dataset settings demonstrate strong reconstruction quality, open-vocabulary understanding, and cross-dataset generalization.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with translation, and assist with the interpretation of experimental results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, or clean and reformat datasets. Additionally, we used generative AI tools to assist with literature search and summarization, software code editing, figure preparation, paper drafting, language editing for readability, and suggestions on paper structure and presentation. We reviewed all AI-assisted work. AI-assisted code was verified and tested for correctness by the authors, literature-related information was checked against the original sources, and all experimental results and technical claims were independently reviewed and verified by the authors. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

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.

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.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

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.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

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.

Ziren Gong, Xiaohan Li, Fabio Tosi, Jiawei Han, Stefano Mattoccia, Jianfei Cai, and Matteo Poggi. Ov3r: Open-vocabulary semantic 3d reconstruction from rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34206–34216, 2026.

Jie Hu, Shizun Wang, and Xinchao Wang. Pe3r: Perception-efficient 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26877–26887, 2026.

Minchao Jiang, Xiaoxuan Ma, Shunyu Jia, Haoru Wang, Zhang Liang, and Wentao Zhu. Instance-splat: Instance-aware feed-forward 3d gaussian splatting for scene understanding. *arXiv preprint arXiv:2608.07144*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19672–19682. IEEE, 2023.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022.

Seungjun Lee, Zihan Wang, Yunsong Wang, and Gim Hee Lee. Embodiedsplat: Online feed-forward semantic 3dgs for open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23774–23784, 2026.

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.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

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*, 2025.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

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. *Communications of the ACM*, 65(1):99–106, 2021.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4018–4028. IEEE, 2024.

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.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060. IEEE, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European conference on computer vision*, pp. 125–141. Springer, 2022.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pp. 501–518. Springer, 2016.

Yu Sheng, Jiajun Deng, Xinran Zhang, Yu Zhang, Bei Hua, Yanyong Zhang, and Jianmin Ji. Spatialsplat: Efficient semantic 3d from sparse unposed images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26404–26414. IEEE, 2025.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

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.

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, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω, 2026. URL https://arxiv.org/abs/2605.15195.

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, 2024.

Qi Xu, Dongxu Wei, Lingzhe Zhao, Wenpu Li, Zhangchi Huang, Shunping Ji, and Peidong Liu. Siu3r: Simultaneous scene understanding and 3d reconstruction beyond feature alignment, 2025. URL https://arxiv.org/abs/2507.02705.

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.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, et al. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *The Fourteenth International Conference on Learning Representations*, 2026.

Chaoyi Zhou, Run Wang, Feng Luo, Mert D Pesé, Zhiwen Fan, Yiqi Zhong, and Siyu Huang. Ff3r: Feedforward feature 3d reconstruction from unconstrained views. *arXiv preprint arXiv:2604.09862*, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21676–21685. IEEE, 2024.

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.

A APPENDIX

In this Appendix, we provide complimentary demonstrations:

- automatic scene annotation pipeline in Appendix A.1;
- RealEstate10K-483 zero-shot evaluation benchmark construction in Appendix A.2;
- detailed training objectives in Appendix A.3;
- supplementary visualization results in Appendix A.4.

A.1 AUTOMATIC ANNOTATION PIPELINE

We design an end-to-end automatic scene annotation pipeline that takes only multi-view RGB images or videos as input and generates large-scale 3D scene understanding annotations. Given an RGB sequence, the pipeline automatically estimates depth and camera poses, and produces semantic, instance, and panoptic annotations while maintaining object-level consistency across views. It combines feed-forward geometry estimation, vision–language models, foundation segmentation models, and geometry-based instance association to obtain geometrically and semantically consistent annotations. The complete pipeline is highly efficient, requiring less than 5 minutes to process a 100-frame video and less than 2 minutes for a 50-frame video. A detailed breakdown of annotation process is provided in Algorithm 1.

Our pipeline employs three foundation models for geometry estimation, semantic understanding, and pixel-level segmentation. For geometric reconstruction, we use VGGT-Ω (Wang et al., 2026) to estimate dense depth maps and camera poses from RGB image clips, providing the geometric information required for subsequent 3D lifting. For semantic annotation, we use Qwen3-VL-8B (Bai et al., 2025) to jointly analyze multiple frames within each clip and generate a consistent object vocabulary shared across the sequence. Given these semantic prompts, SAM3 (Carion et al., 2026) produces pixel-level segmentation masks on the selected frames.

Maintaining consistent instance identities and semantic labels across multiple views remains challenging because independently predicted 2D masks do not provide globally consistent 3D instance information. To address this issue, we lift the predicted 2D semantic masks into 3D point clouds, yielding a set of 3D instance fragments. We first establish temporal correspondences between fragments using semantic consistency and mask overlap across neighboring frames. We then perform voxel-based 3D association, where the lifted fragments are organized in a voxel grid and merged with spatially neighboring fragments according to geometric proximity and point-level overlap. Through this spatiotemporal association process, fragmented multi-view observations are progressively consolidated into globally consistent semantic instances across the entire sequence.

A.2 REALESTATE10K-483 ZERO-SHOT EVALUATION BENCHMARK CONSTRUCTION

Existing 3D scene understanding datasets and benchmarks are primarily designed to support model training and in-domain evaluation, rather than to explicitly measure zero-shot generalization across different scene distributions. Replica (Straub et al., 2019) provides high-quality semantic and instance annotations, but contains only a small collection of photorealistic indoor scene reconstructions. ScanNet (Dai et al., 2017) provides large-scale RGB-D sequences together with camera poses, reconstructed geometry, and semantic annotations, while its standard semantic benchmark evaluates models over a predefined category taxonomy. ScanNet200 (Rozenberszki et al., 2022) extends this setting to 200 long-tailed semantic categories, but follows the original ScanNet train/validation/test scene splits and remains based on a predefined evaluation vocabulary.

ScanNet++ (Yeshwanth et al., 2023) further provides high-fidelity indoor reconstructions with rich and open-vocabulary semantic annotations, enabling more comprehensive semantic scene understanding evaluation. However, its benchmark is still constructed around held-out scenes from the same underlying dataset distribution. Similarly, InsScene-15K (Li et al., 2025) provides a large-scale collection of RGB images, camera poses, depth maps, and 3D-consistent instance masks without category-level semantic annotations. And it is primarily introduced as a training resource for learning geometry and instance-consistent representations rather than as a dedicated benchmark for cross-dataset zero-shot evaluation.

Algorithm 1 End-to-End Automatic Semantic Annotation Pipeline

Require: RGB clip $\mathcal{I} = \{I_i\}_{i=1}^{N}$
Ensure: Global semantic-instance annotations $\mathcal{A}$
  Initialize an empty set of 3D instance fragments $\mathcal{P}$
  ▷ **Phase 1: Geometry and 2D Semantic Annotation**
  $\{D_i, \mathbf{K}_i, \mathbf{T}_i\}_{i=1}^{N} \leftarrow \mathrm{VGGT}_{\Omega}(\mathcal{I})$
  $\mathcal{V} \leftarrow \mathrm{Qwen3VL}(\mathrm{SampleViews}(\mathcal{I}))$
  $\mathcal{V} \leftarrow \mathrm{NormalizeVocabulary}(\mathcal{V})$
  for $i = 1, 2, 3, \ldots$ **do**
    $\mathcal{M}_i \leftarrow \mathrm{SAM3}(I_i, \mathcal{V})$
  end for
  ▷ **Phase 2: 2D-to-3D Lifting and Temporal Association**
  for each segmented frame i and mask $m \in \mathcal{M}_i$ **do**
    $P_{i,m} \leftarrow \mathrm{BackProject}(m, D_i, \mathbf{K}_i, \mathbf{T}_i)$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{P_{i,m}\}$
  end for
  for each pair of adjacent segmented frames (i, j) **do**
    $\mathrm{MergeTemporal}(\mathcal{P}_i, \mathcal{P}_j)$ using semantic consistency and 2D IoU
  end for
  ▷ **Phase 3: Voxel-based 3D Spatial Merging**
  Assign each instance fragment to a $0.60\,\mathrm{m}$ voxel using its 3D centroid
  for each voxel and its 26 neighboring voxels **do**
    for each same-category candidate pair (P_a, P_b) **do**
      if $d(c_a, c_b) < 0.30\,\mathrm{m}$ **and** $\mathrm{Overlap}(P_a, P_b) \geq 0.10$ **then**
        $\mathrm{Merge}(P_a, P_b)$
      end if
    end for
  end for
  Remove short-lived or fragmented instances
  ▷ **Phase 4: Global ID Propagation and Conversion**
  Assign a unique global instance ID to each persistent 3D instance
  Replace local IDs with global IDs on SAM3-annotated frames
  Propagate global IDs to unannotated frames using geometry-guided bidirectional warping
  Clean small mask fragments
  $\mathcal{A} \leftarrow \mathrm{ConvertToSIU3R}(\mathcal{I}, \{D_i, \mathbf{K}_i, \mathbf{T}_i\}, \mathcal{P})$
  return $\mathcal{A}$

To provide a more challenging setting for evaluating zero-shot cross-dataset generalization in 3D scene understanding, we construct our benchmark from the test split of RealEstate10K (Zhou et al., 2018). We select 483 video clips and annotate each clip as a continuous sequence using the automatic annotation pipeline described in Appendix A.1. Rather than annotating frames independently, we process each clip jointly at the sequence level, which helps preserve semantic labels, instance identities, and geometric correspondences across temporally related viewpoints. These sequence-consistent annotations provide a reliable basis for constructing sparse-view evaluation samples.

Sparse-view evaluation construction. Building on the continuous sequence annotations, we convert each video clip into a collection of sparse-view evaluation samples. Specifically, we select two frames with a pairwise view-overlap IoU between 0.3 and 0.8 as the context views. This criterion excludes both near-duplicate viewpoints and view pairs with insufficient shared scene content, while preserving meaningful viewpoint variation. Frames temporally located between the two context views are then selected as target views. Each evaluation sample therefore requires the model to reconstruct the scene from only two sparse observations and generalize its predicted appearance, geometry, and semantic understanding to unseen intermediate viewpoints.

Benchmark statistics. Applying this sampling protocol to the 483 annotated clips yields 2,896 sparse-view evaluation samples. In total, RealEstate10K-483 contains 660 unique semantic concepts and 62,260 annotated object instances, corresponding to approximately 21.5 instances per evaluation

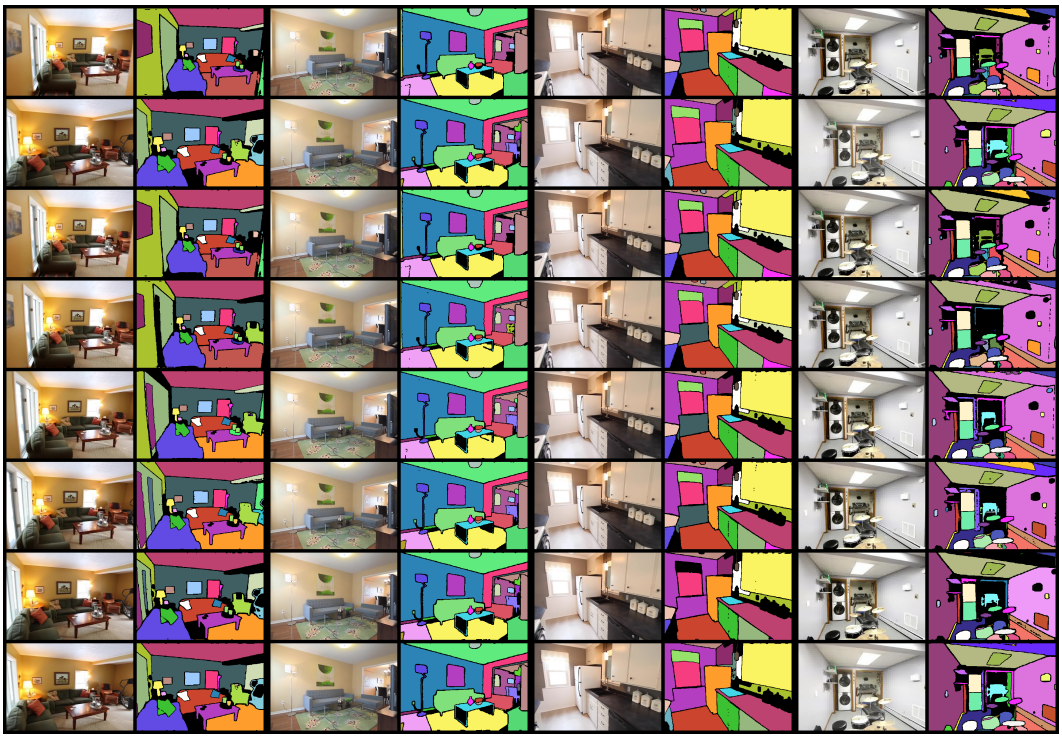

Figure 7: Visualization of our RealEstate10K-483 zero-shot evaluation benchmark annotated by our automatic annotation pipeline A.1 Each scene consists of 6 views in total, obtaining instance level consistency.

sample on average. The resulting annotations cover a substantially broader semantic space than conventional small closed-set evaluation settings, while preserving the visual diversity and naturally occurring object compositions of real-world video scenes.

Evaluation characteristics. The resulting benchmark evaluates not only scene understanding in the observed context views, but also whether appearance, geometry, semantic labels, and instance identities can be consistently transferred from sparse observations to novel viewpoints. For novel-view synthesis, we report PSNR, SSIM, and LPIPS; for depth estimation, we use AbsRel and RMSE; and for scene understanding, we report mIoU, PQ, and mAP on both context and target views. Together with the controlled context-view overlap and the diverse scene layouts, appearances, occlusions, and camera trajectories present in RealEstate10K, RealEstate10K-483 provides a unified setting for evaluating sparse-view reconstruction, semantic and instance understanding, novel-view scene understanding, and zero-shot cross-dataset generalization.

A.3 Detailed Training Objectives

We provide here the implementation details of the training objectives summarized in Sec. 3.4, including the loss weights, Mask2Former settings, CLIP teacher filtering criteria, and the distillation schedule. Following SIU3R (Xu et al., 2025), we use the same formulations for $\mathcal{L}_{\text{RGB}}$, $\mathcal{L}_{\text{LPIPS}}$, and $\mathcal{L}_{\text{Mask2Former}}$ as described in the main text.

Loss weights. The objective in Eq. 13 is instantiated as

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RGB}} + 0.5\,\mathcal{L}_{\text{LPIPS}} + 0.05\,\mathcal{L}_{\text{semantic}} \\ + 0.05\,\mathcal{L}_{\text{depth-smooth}} + 0.005\,\mathcal{L}_{\text{center-res}}.\end{aligned} \tag{14}$$

The semantic objective is

$$\mathcal{L}_{\text{semantic}} = \mathcal{L}_{\text{Mask2Former}} + s_{\text{distill}}(e)\,(\mathcal{L}_{\text{KD-cos}} + 0.1\,\mathcal{L}_{\text{KD-con}})\,, \tag{15}$$

where the two distillation terms respectively encourage query–teacher feature alignment and category-aware semantic consistency.

Mask2Former settings. Object queries are associated with ground-truth instances using Hungarian matching. We use classification, mask, and Dice costs with weights 1, 5, and 5, respectively. After matching, the final and auxiliary decoder outputs are supervised using classification, mask, and Dice losses with weights 2, 5, and 5, respectively. The classification term is computed using the open-vocabulary query–text probabilities defined in Eq. 5. Unmatched queries are assigned to the no-object class with an EOS weight of 0.1. Mask supervision is computed over $12{,}544$ sampled points following the Mask2Former training protocol.

Confidence-aware CLIP query distillation. For each matched object query, we construct a multi-view CLIP teacher prototype from frozen visual features pooled within the corresponding ground-truth instance masks. Teacher observations are retained only when the confidence is at least 0.05, the visible mask area contains at least 64 pixels, the instance is observed in at least two views, and the cross-view feature similarity is at least 0.5. Teacher features and confidence values are detached during optimization.

The distillation objective combines a confidence-weighted cosine loss and a category-aware contrastive loss:

$$\mathcal{L}_{\text{KD-cos}} = \frac{\sum_i c_i \left(1 - \mathbf{s}_i^\top \mathbf{t}_i\right)}{\sum_i c_i}, \tag{16}$$

where $\mathbf{s}_i$ and $\mathbf{t}_i$ denote the normalized student query embedding and teacher prototype, respectively, and c_i is the corresponding teacher confidence.

The contrastive component is defined as

$$\mathcal{L}_{\text{KD-con}} = -\frac{1}{\sum_i c_i} \sum_i c_i \log \frac{\sum_{j:y_j=y_i} \exp\left(\mathbf{s}_i^\top \mathbf{t}_j / T_d\right)}{\sum_j \exp\left(\mathbf{s}_i^\top \mathbf{t}_j / T_d\right)}, \tag{17}$$

with temperature $T_d = 0.07$.

Geometric regularization details. The instance-aware depth smoothness and confidence-weighted Gaussian-center regularization follow the formulations in Sec. 3.4. For the center residual, the image-plane correction is bounded to approximately ± 2 pixels, while the depth residual permits approximately $+15\%/-13\%$ relative correction. These bounds allow local reconstruction refinement while preventing excessive deviation from the geometry initialization.

Distillation schedule. CLIP query distillation is disabled during the initial optimization stage and is progressively introduced according to

$$s_{\text{distill}}(e) = \begin{cases} 0, & e < 5, \\ \dfrac{e-5}{15}, & 5 \le e < 20, \\ 1, & e \ge 20. \end{cases} \tag{18}$$

Thus, the semantic branch first establishes task-specific segmentation structure before the vision–language distillation signal is gradually increased to its full strength.

A.4 More Visualization results

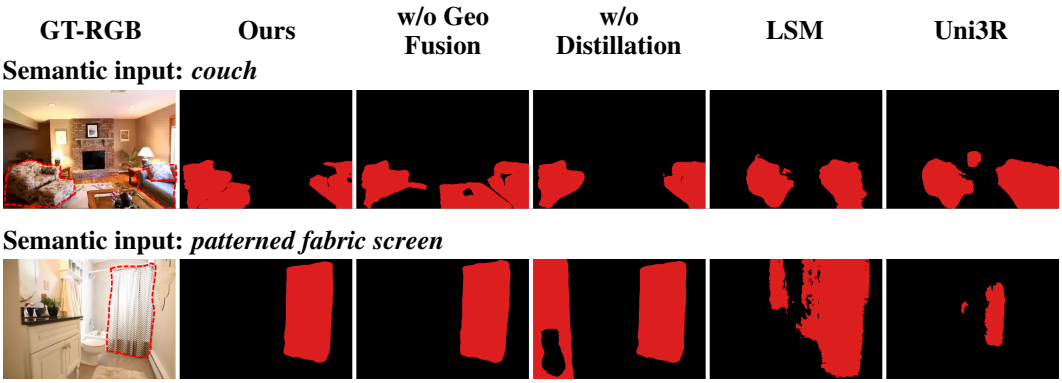

Figure 8: Qualitative comparison of open-vocabulary segmentation results under different semantic inputs. Red regions indicate the predicted target masks.