Dynamic-Robust Photometric-Semantic Reconstruction for Open-Vocabulary 3D Scene Understanding

Boyu Cai[1,2], Li Yang[2†*], Yan Xu[3†], Wei Liu[2], Nian Liu[2], Sikui Zhang[2], Yan Wang[4‡], Chunfeng Yuan[2], Weiming Hu[1,2]

[1] School of Information Science and Technology, ShanghaiTech University
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information, Institute of Automation, Chinese Academy of Sciences (CASIA)
[3] Electronic Engineering Department, The Chinese University of Hong Kong
[4] Deepeleph Intelligent Technology
caiby2024@shanghaitech.edu.cn, li.yang@nlpr.ia.ac.cn

Abstract. The integration of novel view synthesis (NVS) and open-vocabulary segmentation (OVS) has recently yielded powerful feed-forward 3D foundation models. However, their inherent reliance on static-scene assumptions leads to severe misalignment of spatial features in unconstrained dynamic environments. To bridge this critical gap, we propose SPAR, a novel joint semantic-geometric encoding architecture that explicitly isolates transient dynamic noise prior to latent space aggregation. Furthermore, we introduce a dynamic-region-aware end-to-end training paradigm that structurally couples motion estimation with multi-view visual and semantic learning. This unified approach enables the network to inherently resolve motion conflicts and distill multi-view consistent, temporally stable scene representations from dynamic inputs. Extensive experiments on the challenging D-RE10K benchmark demonstrate that SPAR achieves state-of-the-art performance. Our end-to-end approach achieves exceptional novel view synthesis quality, yielding a PSNR of 22.15 dB and 23.33 dB given only 3 and 4 input views respectively. Despite being trained in a self-supervised manner, our model achieves an mIoU of 88.5% for motion mask prediction. Furthermore, our analysis reveals a strong inter-task synergy between photometric scene reconstruction and semantic understanding, where semantic synthesis learning consistently enhances photometric fidelity in novel view rendering. Code will be available at https://github.com/dmucby/SPAR.

Keywords: Dynamic Photometric-Semantic Reconstruction · 3D Scene Understanding

---

† Equal contribution.
* Corresponding author.
‡ Joint work with Zhejiang Deepeleph Intelligent Technology Co., Ltd.

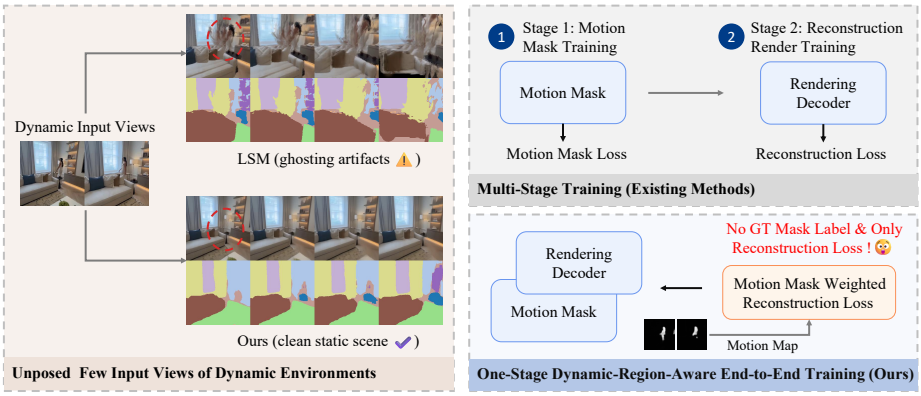

Fig. 1: Comparison under dynamic, unposed settings. Baselines such as LSM [9] struggle with moving objects, producing ghosting artifacts and inconsistent semantics (left top). By masking dynamic regions prior to latent encoding, our SPAR generates clean, ghost-free photometric and semantic renderings in a single forward pass (left bottom).

1 Introduction

Recent advances in novel view synthesis (NVS) [16, 27] and open-vocabulary segmentation (OVS) [20, 29] have pushed 3D perception beyond photorealistic rendering toward unified semantic scene understanding. Instead of optimizing a scene-specific language NeRF [17] or language Gaussian [28], feed-forward methods such as LSM [9], SIU3R [39] infer geometry, appearance, and semantic representations directly from sparse and unposed images. By combining large reconstruction models [37] and open-vocabulary segmentation models [6, 20], these methods jointly model geometry and semantics in 3D fields and achieve strong results on benchmarks such as ScanNet [8].

However, this paradigm relies on a strict assumption: all input views are expected to depict the same static 3D scene. In real-world videos, this assumption is often violated by moving humans, animals, and other transient foreground objects. Because a moving object may occupy different 3D locations across frames, its corresponding image regions cannot be reconciled within a single static 3D scene representation. When these regions are aggregated with static content, they introduce conflicting geometry and semantics into the scene representation, leading to ghosting artifacts in RGB rendering and inconsistent semantic predictions (see Figure 1, left).

To address this limitation, we propose **SPAR**, a **S**emantic-**P**hotometric-**A**ware **R**econstruction framework that jointly performs novel view synthesis and semantic understanding from a few unposed observations of dynamic environments. SPAR adopts an encoder-decoder architecture that jointly embeds RGB observations and open-vocabulary semantic features into a latent scene representation. A shared rendering decoder then queries this representation to synthesize target-view RGB images and semantic feature maps. To prevent moving objects from contaminating the latent scene representation, SPAR introduces a Cross-

View Dynamic Region Predictor (CV-DRP), which estimates motion masks from cross-view cues and filters dynamic foreground regions before scene-level encoding.

Furthermore, to avoid costly dynamic-objects labeling and maintain scalability, we introduce a dynamic-region-aware end-to-end training paradigm. This paradigm couples dynamic region estimation with multi-view photometric and semantic modeling within a unified optimization framework (See Figure 1, right). By using the predicted motion masks to spatially gate reconstruction errors, our approach focuses supervision on multi-view consistent static regions while suppressing gradients from transient foregrounds. As a result, photometric view synthesis and semantic modeling mutually reinforce each another, enabling the network to resolve motion conflicts and recover a clear static 3D scene representation in a single forward pass.

We evaluate SPAR on the challenging D-RE10K benchmark [4], which contains unconstrained indoor sequences with complex motions of transient objects. Given only a few input views, SPAR achieves state-of-the-art performance, reaching a PSNR of 22.15 dB and 23.33 dB with 3 and 4 input views, respectively. Notably, our dynamic region predictor achieves an 88.5% mIoU under 3 input views, surpassing leading self-supervised methods by more than 36 percentage points. Furthermore, ablation studies reveal a strong inter-task synergy: rather than compromising rendering capacity, the semantic branch serves as a structural regularizer that improves the photometric fidelity of the reconstructed scene.

In summary, our main contributions are as follows:

- We propose SPAR, the first unified feed-forward framework for joint novel view synthesis and open-vocabulary semantic understanding from a few unposed views of dynamic environments.
- We design a novel Cross-View Dynamic Region Predictor with a dynamic-region-aware optimization paradigm. By using the predicted motion mask to spatially weight both photometric and semantic reconstruction losses, our method enables end-to-end joint training without requiring ground-truth motion mask labels.
- Extensive experiments on the D-RE10K benchmark demonstrate that SPAR establishes new state-of-the-art performance in both dynamic NVS and motion mask prediction. These results validate the mutual benefits of jointly learning photometric reconstruction and semantic scene understanding.

2 Related Work

2.1 Generalizable Novel View Synthesis

Instead of optimizing a scene-specific representation [16, 24, 27, 42], generalizable novel view synthesis (NVS) targets fast inference by training across diverse scenes such that novel views (or an intermediate 3D representation) can be predicted in a single forward pass. Early generalizable methods infer volumetric features from sparse inputs and exploit strong geometric inductive biases, e.g., epipolar constraints and plane-sweep cost volumes, as in PixelNeRF [48], MVSNeRF [3], and

IBRNet [36]. Subsequent works improve robustness in sparse-view settings and extend feed-forward reconstruction to efficient 3D Gaussian representations, including pixelSplat [2] and MVSplat [5], while pose-free Gaussian inference further reduces reliance on calibrated cameras at test time [21, 47]. More recently, large reconstruction models (LRMs) scale model capacity and data to learn generic 3D priors [11, 40, 49, 55]. In parallel, token-space renderers largely remove handcrafted 3D inductive biases, exemplified by LVSM [14]. RayZer [13] continues this transformer-based direction and introduces self-supervised, pose-free NVS without 3D supervision, yet it still assumes static imagery. WildRayZer [4] explicitly lifts this static-scene assumption by learning transient-region masks and gating both tokens and losses, enabling feed-forward NVS under dynamic, in-the-wild inputs.

2.2 Scene Understanding

Substantial progress has been made in 2D and 3D scene understanding. In 2D, vision-language pretraining such as CLIP [29, 43] enables open-vocabulary recognition, and LSeg [20] aligns dense predictions with language features, while object-level detection and grounding have also been extensively studied [44, 45]. Query-based transformers, e.g., MaskFormer [7] and its universal extensions such as Mask2Former [6], provide a unified formulation for semantic/instance/panoptic segmentation, and recent advances extend these paradigms to video settings [30, 51, 52]. However, 2D understanding is inherently viewpoint-dependent and does not directly enforce cross-view spatial consistency. For 3D understanding, recent approaches operate on pre-scanned point clouds (often clustered into superpoints) and employ mask/proposal-style formulations [25, 34, 41, 46], but they typically require costly 3D acquisition and preprocessing, limiting scalability in practical deployments.

2.3 Simultaneous Understanding and Novel View Synthesis

A growing line of work seeks to couple NVS with scene understanding by embedding semantic or language features into 3D representations. Representative approaches inject foundation-model features into NeRFs [17, 18] or 3D Gaussians [28, 53, 56], usually by aligning 3D fields with 2D features through differentiable rasterization. These pipelines often rely on dense captures and per-scene optimization, which limits efficiency at scale. LSM [9] mitigates this drawback by performing 2D-to-3D feature alignment within a large feed-forward reconstruction model. Nevertheless, the feature-alignment paradigm inherits two fundamental limitations: (i) instance-level understanding is constrained by the capabilities of the underlying 2D teacher, and (ii) semantic fidelity can degrade due to aggressive feature compression required by rasterization and memory constraints [39]. SIU3R [39] proposes an alignment-free alternative based on explicit pixel-aligned 2D-to-3D lifting and a unified query decoder, enabling native multi-task 3D understanding alongside reconstruction. Orthogonally, our method highlights that simultaneous modeling must also address dynamics: without explicitly disentangling transient content, multi-view consistency breaks and cor-

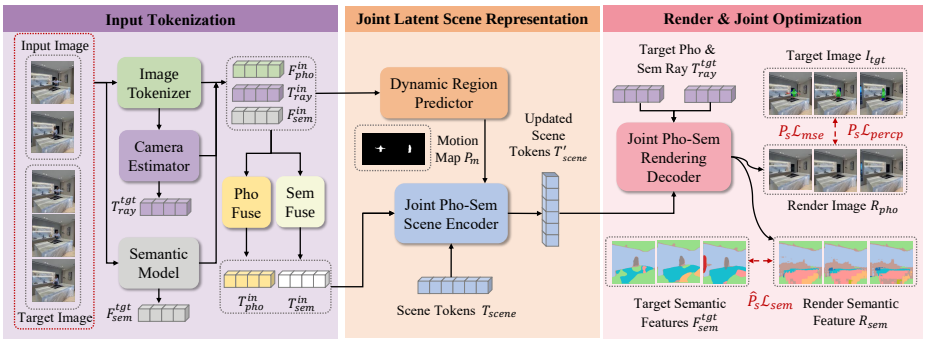

Fig. 2: Overview of SPAR. Our unified encoder-decoder jointly performs novel view synthesis and open-vocabulary scene understanding from unposed dynamic views. The pipeline first estimates camera poses to build pose-conditioned photometric and semantic tokens. A dynamic region predictor masks moving foregrounds to isolate static geometry. A scene encoder aggregates filtered tokens into a compact latent scene representation, which a dual-head decoder uses to render the target image and semantic features.

rupts both reconstruction and downstream understanding. Overall, recent evidence suggests that scalable simultaneous understanding and NVS benefit from (a) a joint photometric-semantic render and (b) explicit mechanisms to handle in-the-wild dynamics.

3 Method

3.1 Overview

We propose **SPAR**, a unified framework that jointly performs novel view synthesis and semantic understanding from unposed multi-view observations of dynamic environments. As illustrated in Figure 2, SPAR follows a latent scene representation-based encoder-decoder architecture. Given multi-view images, the network first estimates per-image camera poses and constructs pose-conditioned photometric and semantic tokens. To handle dynamic content, a dynamic region predictor identifies the moving foreground regions and masks the associated photometric and semantic tokens to mitigate the impact of motion-induced inconsistencies. The scene encoder then aggregates the masked tokens into a set of scene tokens, forming a compact scene representation that jointly encodes photometric and semantic cues. Given the scene tokens and the target-view pose, the decoder predicts the target view via two parallel heads, producing both the rendered photometric image and semantic feature maps. The semantic features are further aligned with text embeddings to produce open-vocabulary semantic segmentation for the novel views.

Moreover, we design an end-to-end training paradigm that couples the optimization of dynamic region estimation with multi-view visual and semantic

learning, enabling SPAR to learn multi-view consistent and temporally stable scene representations from dynamic, multi-view inputs.

3.2 Model Architecture

Input Tokenization. Given multi-view images $I_i \in \mathbb{R}^{H\times W\times 3}$, we follow RayZer [13] to tokenize the images together with their estimated pose Plücker rays into patch-level tokens, which are fused to form the photometric tokens $T_{pho} \in \mathbb{R}^{N_{pho}\times D}$. Meanwhile, we extract the semantic features of the input images using a pre-trained 2D semantic segmentation model (e.g., LSeg [20]). The corresponding pose Plücker rays are also fused with these features to obtain the semantic tokens $T_{sem} \in \mathbb{R}^{N_{sem}\times D}$.

To learn stable scene representations from input views in dynamic environments, we suppress dynamic foreground regions in both the photometric and semantic tokens. Specifically, we employ a dynamic region predictor to estimate the masks of dynamic regions for the input images. These masks are then applied to the photometric and semantic tokens to suppress tokens corresponding to dynamic regions, producing the masked photometric tokens $\tilde{T}_{pho}$ and masked semantic tokens $\tilde{T}_{sem}$.

Joint Pho-Sem Scene Encoder. The scene encoder aggregates multi-view information from the masked photometric tokens $\tilde{T}^{in}_{pho}$ (in denotes the input views) and semantic tokens $\tilde{T}^{in}_{sem}$, and produces a compact set of scene tokens that capture the global scene representation.

To this end, we initialize a set of learnable global scene tokens $T_{scene} \in \mathbb{R}^{S\times D}$, and concatenate them with the masked photometric tokens and semantic tokens to form $T_{all} = \text{Concat}(T_{scene}, \tilde{T}^{in}_{pho}, \tilde{T}^{in}_{sem}) \in \mathbb{R}^{(S+N_{pho}+N_{sem})\times D}$. T_{all} is then fed into the scene encoder, which consists of L transformer layers with self-attention. This design enables deep cross-modal interaction among the scene, photometric, and semantic tokens, producing a latent representation T'_{all}. Finally, we select the first S tokens from T'_{all} to obtain $T'_{scene} \in \mathbb{R}^{S\times D}$, which serves as the compact scene representation encoding both the photometric and semantic information of the environment.

Joint Pho-Sem Rendering Decoder. Finally, we design a rendering decoder to jointly decode the photometric and semantic representations of the target views from the encoded scene representation.

We first obtain the pose of the target views using the camera pose estimator and compute their corresponding Plücker rays. Two separate linear layers are then applied to the target-view Plücker rays to generate their photometric and semantic query tokens $T^{tgt}_{ray\_pho} \in \mathbb{R}^{M\times D}$ and $T^{tgt}_{ray\_sem} \in \mathbb{R}^{M\times D}$, respectively ($tgt$ denotes the target views). We concatenate these target-view query tokens with the scene tokens T'_{scene} produced by the encoder to form $T_{joint} = \text{Concat}(T^{tgt}_{ray\_pho}, T^{tgt}_{ray\_sem}, T'_{scene}) \in \mathbb{R}^{(2M+S)\times D}$. The combined tokens are fed into the rendering decoder, which performs information interaction via self-attention and produces updated tokens T'_{joint}. We then select the

first $2M$ tokens from T'_{joint} and split them into two target-view feature sets $\{T_{out\_pho}, T_{out\_sem}\} \in \mathbb{R}^{M\times D}$. A dual-head MLP renderer is applied to separately predict the target-view image and semantic features from T_{out_pho} and T_{out_sem}. Specifically, the image rendering head focuses on reconstructing high-frequency appearance details, where MLP_{pho} maps T_{out_pho} to RGB values, while the semantic rendering head focuses on semantic alignment, where MLP_{sem} maps T_{out_sem} to continuous high-dimensional semantic embeddings rather than fixed classification logits:

$$R_{pho} = \text{MLP}_{pho}(T_{out\_pho}) \in \mathbb{R}^{M\times 3}, R_{sem} = \text{MLP}_{sem}(T_{out\_sem}) \in \mathbb{R}^{M\times C} \quad (1)$$

Cross-View Dynamic Region Predictor. Before feeding the patch-level photometric and semantic tokens into the encoder, we first need to identify dynamic object regions from the input views to reduce their interference with scene modeling. To this end, we design a *Cross-View Dynamic Region Predictor* that detects moving foreground regions across the multi-view inputs.

Specifically, for an input view $I_i \in \mathbb{R}^{H\times W\times 3}$, we first concatenate its semantic features $F^{(i)}_{sem}$, photometric tokens $F^{(i)}_{pho}$ and Plücker ray tokens $T^{(i)}_{ray}$, where $\{F^{(i)}_{sem}, F^{(i)}_{pho}, T^{(i)}_{ray}\} \in \mathbb{R}^{N\times D}$, along the channel dimension to form a joint feature representation $\mathcal{F}^{(i)} \in \mathbb{R}^{N\times 3D}$. Since a single-view image lacks sufficient references to determine whether a region belongs to a moving object, we leverage multi-view observations to capture cross-view inconsistencies caused by dynamic objects. Concretely, the joint features of paired source and reference views $(\mathcal{F}^{(i)}, \mathcal{F}^{(j)})$ are fed into multiple stacked Transformer self-attention layers to model cross-view relationships. The resulting features are then passed through an MLP followed by an upsampling layer to produce a dynamic region probability map $P^{(i)}_m \in \mathbb{R}^{H\times W}$ with the same spatial resolution as the input image. The dynamic mask $M^{(i)}$ is obtained by thresholding the predicted probability map:

$$M^{(i)} = \mathbb{I}(P^{(i)}_m > \tau). \quad (2)$$

At inference time, we optionally refine the predicted masks using SAM2 [30] to improve mask quality. The refined mask is then used to more precisely filter patch tokens of the input views before scene encoding.

3.3 Dynamic-Region-Aware Optimization

To learn stable photometric and semantic representations from dynamic and unposed multi-view inputs, we introduce a *dynamic-region-aware optimization* framework. The model jointly performs target-view rendering and dynamic region estimation, where the estimated dynamic regions are used to selectively weight supervision signals. This design allows the model to focus learning on geometrically consistent static regions while reducing the influence of cross-view inconsistent foreground motion.

Photometric and Semantic Supervision. During training, the model first encodes the scene from the given input view images and then predicts the rendered RGB image R_{pho} and semantic features R_{sem} for each target view. The rendered outputs are supervised by the ground-truth target-view image I_{tgt} and the corresponding semantic feature map F_{sem}^{tgt} extracted by a pre-trained semantic segmentation model. The photometric reconstruction loss is defined as:

$$\mathcal{L}_{pho} = \|R_{pho} - I_{tgt}\|_2^2 + \lambda \cdot \text{Percep}(R_{pho}, I_{tgt}), \tag{3}$$

where $\text{Percep}(\cdot)$ denotes the perceptual loss [15,23] and λ controls its loss weight. For semantic supervision, we adopt a cosine similarity loss to align the rendered semantic features with the pre-computed semantic embeddings:

$$\mathcal{L}_{sem} = 1 - \frac{R_{sem} \cdot F_{sem}^{tgt}}{\|R_{sem}\|_2 \, \|F_{sem}^{tgt}\|_2}, \tag{4}$$

Dynamic-Region-Aware Rendering Loss. Moving foreground objects introduce cross-view inconsistencies that can degrade scene representation learning. To mitigate this issue, we employ the dynamic region predictor to estimate a dynamic region probability map P_m for each target view. The corresponding static region probability map is defined as $P_s = 1 - P_m$. This static region probability map is used as a spatial weighting mask for the rendering losses, allowing the model to prioritize supervision from static regions. The dynamic-region-aware photometric loss is therefore defined as:

$$\mathcal{L}_{pho}^{dyn} = \frac{1}{\sum P_s} \sum P_s \odot \left(\|R_{pho} - I_{tgt}\|_2^2 + \lambda \cdot \text{Percep}(R_{pho}, I_{tgt}) \right). \tag{5}$$

For semantic feature supervision, the static region probability map P_s is downsampled to match the resolution of the semantic features, producing $\hat{P}_s$. The semantic loss is then defined as:

$$\mathcal{L}_{sem}^{dyn} = \frac{1}{\sum \hat{P}_s} \sum \hat{P}_s \odot \left(1 - \frac{R_{sem} \cdot F_{sem}^{tgt}}{\|R_{sem}\|_2 \, \|F_{sem}^{tgt}\|_2} \right). \tag{6}$$

To prevent the network from collapsing to a trivial solution that labels the entire image as dynamic, we introduce a regularization term that penalizes excessive dynamic predictions:

$$\mathcal{L}_{reg} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \text{BCE}(P_m^{(i,j)}, 0). \tag{7}$$

We also introduce a copy-paste data augmentation strategy. Specifically, objects are cropped from real images and randomly pasted onto static training scenes to simulate moving foreground objects. For such augmented samples, the original (unaugmented) target image I_{tgt} and its semantic features F_{sem}^{tgt} are used to compute the photometric reconstruction loss $\mathcal{L}_{pho}^{cp}$ and semantic loss $\mathcal{L}_{sem}^{cp}$ following Eq. (3) and Eq. (4). The pasted object mask is used as ground-truth supervision

for the predicted dynamic region probability map via a binary cross-entropy loss $\mathcal{L}_{mask}$. The overall optimization loss of the unified framework is defined as:

$$\mathcal{L}_{total} = \lambda_{pho}\mathcal{L}_{pho}^{\{dyn,cp\}} + \lambda_{sem}\mathcal{L}_{sem}^{\{dyn,cp\}} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{mask}\mathcal{L}_{mask}. \quad (8)$$

Our training strategy jointly optimizes multi-view rendering and dynamic region prediction. Rendering errors provide adaptive learning signals for identifying cross-view dynamic regions, while the predicted dynamic regions restrict supervision to geometrically consistent static areas. This mutually beneficial interaction enables SPAR to learn robust novel view synthesis and semantic understanding under challenging conditions of dynamic scenes and unknown camera poses.

4 Experiments

Implementation Details. In our joint framework, SPAR is implemented with a Transformer-based architecture. The joint scene encoder, camera estimator, CV-DRP, and joint decoder contain 8, 12, 4, and 12 Transformer layers, respectively. The scene token capacity is set to $S = 768$. We train SPAR on 8× A800 GPUs with a cosine learning rate schedule. During pre-training, we randomly mask out 10% of the input photometric tokens in a portion of the training data to simulate the discarding of dynamic regions in end-to-end training stages. The end-to-end training stage runs for 20K iterations with a learning rate of 2×10^{-4}. The loss weights are set to $\lambda_{\text{percp}} = 0.2$, $\lambda_{\text{sem}} = 0.1$, $\lambda_{\text{mask}} = 1$, and $\lambda_{\text{reg}} = 0.01$. All models are trained at 256×256 resolution with a 16×16 patch size. Additional details are provided in the supplementary material.

Datasets. We utilize three datasets for training: the static RealEstate [54] dataset, the dynamic D-RE10K [4] dataset and the semantically rich ScanNet [8] dataset. For a dynamic environment, model evaluation is performed on D-RE10K, which contains 74 Internet-curated indoor sequences featuring moving humans, pets, and vehicles. Each frame is paired with human-verified motion masks, enabling evaluation specifically restricted to transient regions. For semantic segmentation, we follow LSM [9] to conduct evaluation on 40 unseen scenes from ScanNet.

Evaluation Protocol and Metrics. Following [4, 13], we use masked PSNR [10], SSIM [38], and LPIPS [50] to evaluate the quality of novel view synthesis. To assess the quality of dynamic region predictions, we employ mIoU and Recall for motion mask evaluation. The inputs for all evaluations are unposed multi-view images, specifically 2, 3, or 4 input views, with 6, 5, or 4 target views, respectively. Moreover, our model can accept flexible dynamic multi-view inputs at test time, including 2, 3, and 4 input views. For semantic segmentation, we report mIoU and Accuracy following LSM [9], using two input views and one synthesized target view.

Table 1: Performance comparison on D-RE10K under view settings ($n = 2, 3, 4$).

Method	D-RE10K								
	Views = 2			**Views = 3**			**Views = 4**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Optimization-Based Methods									
NeRF On-the-go [31]	15.90	0.518	0.582	18.45	0.624	0.446	19.52	0.620	0.443
3DGS [16]	13.49	0.442	0.605	14.92	0.514	0.531	16.28	0.552	0.490
T-3DGS [26]	15.90	0.518	0.582	18.45	0.624	0.446	18.70	0.613	0.454
Spotless-Splats [32]	16.45	0.548	0.468	17.77	0.600	0.394	18.05	0.608	0.390
WildGaussians [19]	16.12	0.512	0.624	17.76	0.577	0.588	18.11	0.597	0.574
Feed-forward Methods									
RayZer + Co-Seg [1]	16.76	0.547	0.516	17.53	0.565	0.580	18.67	0.636	0.481
RayZer + MegaSAM [22]	–	–	–	20.13	0.688	0.336	20.66	0.702	0.310
RayZer + SAV [12]	19.01	0.628	0.397	20.35	0.696	0.332	20.73	0.711	0.308
SRT [33]	14.62	0.308	0.647	14.76	0.310	0.639	14.81	0.329	0.632
LSM [9]	10.92	0.271	0.636	11.48	0.307	0.638	11.42	0.309	0.644
SIU3R [39]	13.01	0.372	0.577	13.22	0.362	0.564	13.25	0.363	0.553
WildRayZer [4]	**21.78**	**0.734**	**0.308**	<u>21.98</u>	**0.754**	<u>0.314</u>	<u>22.38</u>	**0.773**	<u>0.290</u>
WildRayZer† [4]	19.47	0.657	0.318	-	-	-	-	-	-
Ours	<u>19.97</u>	<u>0.627</u>	<u>0.339</u>	**22.15**	<u>0.702</u>	**0.283**	**23.33**	<u>0.739</u>	**0.263**

4.1 Main Results

Novel View Synthesis in Dynamic Environments. Table 1 reports the quantitative comparison on D-RE10K under different input-view settings. With 4 input views, SPAR achieves the best PSNR and LPIPS among all compared methods, reaching 23.33 dB and 0.263, respectively. The same advantage is observed for 3 input views, where SPAR improves PSNR from 21.98 to 22.15 dB and reduces LPIPS from 0.314 to 0.283 compared with WildRayZer. In the more challenging 2-view setting, SPAR is below the WildRayZer result reported in the original paper, but achieves a higher PSNR than our reproduction of WildRayZer using the authors' released code and checkpoints. These results show that SPAR delivers strong feed-forward reconstruction quality for dynamic novel view synthesis, especially when three or more input views are available.

Qualitative Results for Photometric-Semantic Novel view Synthesis. We evaluate our single-stage joint semantic and photometric rendering framework against the static-only LSM [9] and the two-stage dynamic rendering approach, WildRayZer [4] + LSeg [20]. As illustrated in Figure 3, LSM lacks explicit support for dynamic scenes. It fails in moving regions and produces severely blurred renderings with corrupted segmentation maps. The two-stage WildRayZer + LSeg pipeline performs better in dynamic rendering, but its semantic prediction is limited by the quality of the intermediate rendered image. As a result, rendering artifacts can directly degrade segmentation accuracy. For example, it misclassifies the window as a curtain in the left sequence and the sofa

† The result is obtained by running the released code and checkpoint.

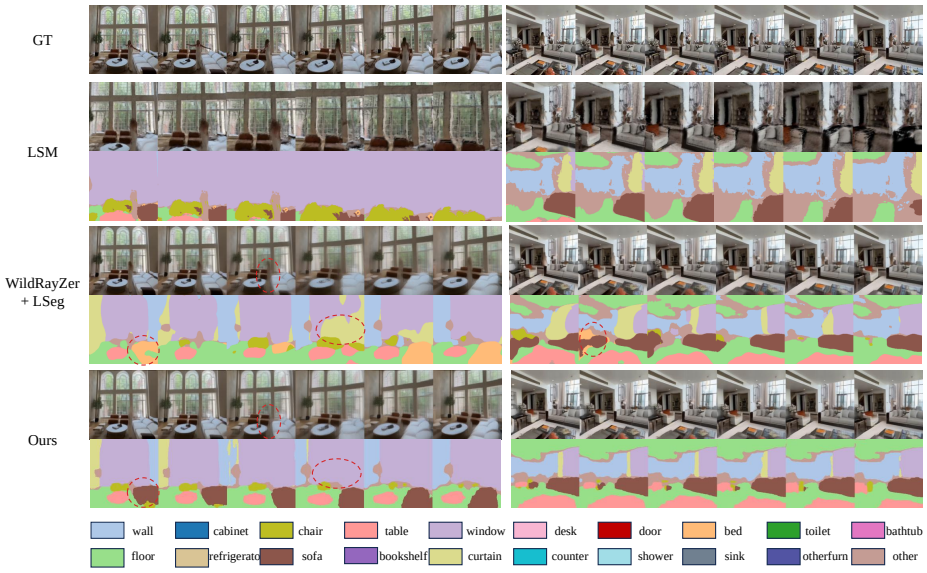

Fig. 3: Qualitative comparison of novel view synthesis and semantic segmentation on dynamic scenes.

as a bed in the right sequence. In contrast, our method jointly fuses multi-view photometric and semantic features in a unified single-stage rendering process. This design helps reduce the above errors and enables both high-fidelity novel view synthesis and structurally coherent semantic segmentation.

Motion Mask Quality Comparison. As shown in Table 2, SPAR achieves state-of-the-art performance in dynamic region estimation. We report two variants of our method. *w/o Refine* directly uses the raw dynamic masks predicted by our Dynamic Region Predictor. *w/ Refine* further applies SAM2 only at test time to refine these predicted masks. The two variants use the same trained model, and SAM2 is not involved in training. Our method consistently outperforms both supervised and self-supervised baselines. Under the $n = 3$ setting, Ours w/ Refine reaches 88.5% mIoU, outperforming the leading self-supervised method WildRayZer [4] by a large margin. Moreover, compared with fully supervised motion-segmentation methods [12], our self-supervised framework still achieves better mIoU under the $n = 8$ setting. Notably, our Dynamic Region Predictor is trained without any ground-truth motion mask labels. Under reconstruction-driven optimization, regions that are hard to explain by static multi-view consistency tend to be identified as dynamic, allowing the model to reduce reconstruction errors caused by motion interference.

Motion Mask Qualitative Results. Figure 4 visualizes our Dynamic Region Predictor across diverse indoor multi-view sequences. The top row displays the

Table 2: Dynamic Region quality. Comparison of supervised and self-supervised motion segmentation methods under view settings ($n = 2, 3, 8$). w/ Refine denotes the use of SAM2 to refine the original motion masks.

Method	$n = 2$		$n = 3$		$n = 8$	
	mIoU ↑	Recall ↑	mIoU ↑	Recall ↑	mIoU ↑	Recall ↑
Supervised Methods						
MegaSAM [22]	-	-	12.7	37.1	35.4	60.7
Segment Any Motion [12]	31.9	47.2	41.2	57.1	50.9	70.8
Ours w/ Refine	**87.8**	<u>84.7</u>	**88.5**	<u>86.2</u>	**88.3**	85.6
Self-supervised Methods						
Co-segmentation [1]	9.6	45.0	13.7	53.2	16.3	45.5
WildRayZer [4]	53.9	85.1	52.1	84.3	54.2	**87.7**
Ours w/o Refine	<u>75.9</u>	**85.7**	<u>75.7</u>	**86.4**	<u>76.6</u>	<u>86.6</u>

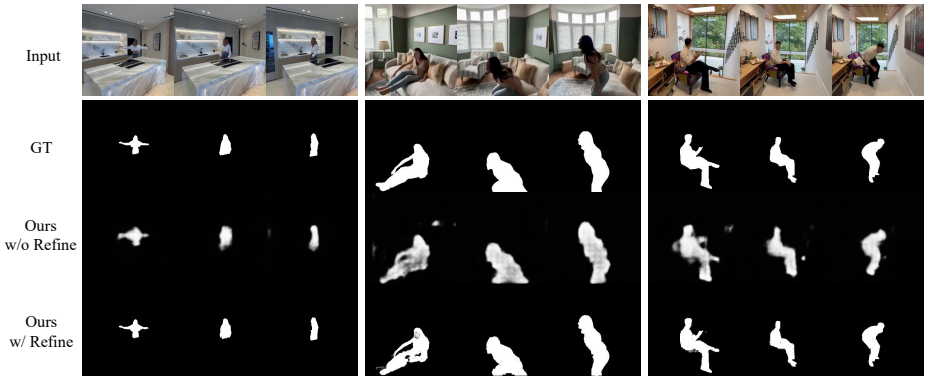

Fig. 4: Qualitative results of our Dynamic Region Predictor (DRP) on diverse multi-view sequences.

input frames, followed by the ground-truth (GT) masks. The third row shows our predictions, denoted as Ours *w/o Refine*. These predictions generate accurate probabilistic heatmaps that reliably localize dynamic human subjects. They remain effective even during complex non-rigid motions, such as sitting and turning. The bottom row presents the refined results, denoted as Ours *w/ Refine*. After refinement, these soft spatial priors are converted into precise binary masks that tightly align with the ground truth. This demonstrates the capability of our module as an efficient and robust motion predictor.

Qualitative Results for Novel View Synthesis and Dynamic Region Prediction on Unbounded Scenes. We directly evaluate our model on videos from SpatialVID [35] without any training or fine-tuning on this dataset. SpatialVID contains large-scale, open-world, primarily outdoor dynamic scenes with pedestrians, cyclists, vehicles, occlusions, viewpoint changes, and diverse scene

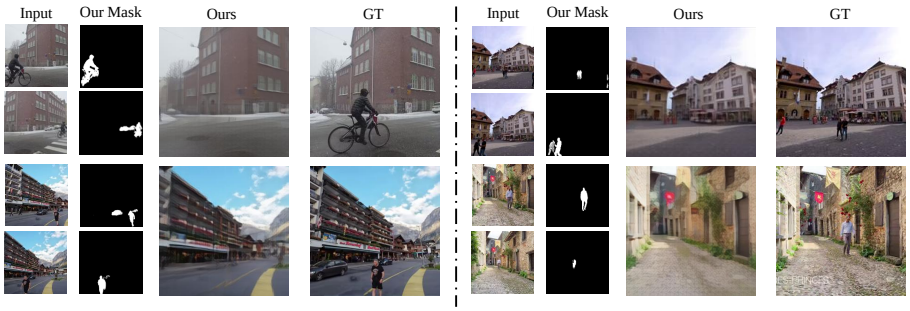

Fig. 5: Qualitative Results of novel view synthesis and dynamic region prediction on unbounded scenes.

layouts. As shown in Figure 5, these scenes pose substantial challenges for novel view synthesis due to low visibility, moving foreground objects, occlusions, and unbounded outdoor layouts. Despite these challenges, our method predicts coherent dynamic-region masks across input views and synthesizes plausible novel views. The predicted masks localize moving objects such as cyclists and pedestrians, while the rendered images preserve the overall scene structure. These results suggest that SPAR can generalize to unseen outdoor videos in a zero-shot manner.

Performance Comparisons on the ScanNet Dataset. As shown in Table 3, our method achieves state-of-the-art performance for 3D-aware joint semantic segmentation and novel view synthesis on the semantic-rich ScanNet dataset. Our semantic branch predicts continuous LSeg/CLIP-aligned embeddings instead of fixed 20-class logits. Therefore, it can inherit the open-set recognition capability of LSeg/CLIP. For semantic segmentation, LSeg retains a marginal advantage in mIoU by directly segmenting the ground-truth target images, while our method achieves a highly competitive mIoU of 0.5271 and the highest accuracy of 0.8061 under joint rendering. For novel view synthesis, our formulation shows clear photometric advantages over existing baselines. We attribute this improvement to the way we integrate semantic information. Instead of compressing semantic features into individual 3D Gaussians as in LSM [9], our method embeds uncompressed semantics directly into global scene features, which helps reduce information loss.

4.2 Ablation Studies

Effectiveness of Dynamic-Region-Aware Optimization. As detailed in Table 4, the proposed dynamic-region-aware optimization improves robustness under severe dynamic interference. Random masking only provides limited regularization, while our targeted formulation brings a much larger gain. This shows that dynamic foreground estimation provides useful spatial guidance for separating dynamic interference and learning stable static scene representations.

Table 3: Quantitative comparison on the semantically rich ScanNet [8] dataset.

Method	Target View				
	mIoU ↑	Acc. ↑	PSNR ↑	SSIM ↑	LPIPS ↓
LSeg [20]	**0.5281**	0.7612	-	-	-
NeRF-DFF [18]	0.4037	0.6755	19.86	0.6650	0.3629
Feature-3DGS [53]	0.4223	0.7174	24.49	<u>0.8132</u>	0.2293
pixelSplat [2]	-	-	24.89	**0.8392**	<u>0.1641</u>
SRT [33]	-	-	17.01	0.4130	0.6350
LSM [9]	0.5078	<u>0.7686</u>	24.39	0.8072	0.2506
SIU3R [39]	-	-	<u>25.10</u>	0.8121	**0.1043**
Ours	<u>0.5271</u>	**0.8061**	**26.43**	0.8052	0.2407

Table 4: Ablation of Optimization Strategies. Quantitative comparison of novel view synthesis performance under severe dynamic interference.

Optimization Strategy	PSNR↑	SSIM↑	LPIPS↓
Unconstrained Baseline	15.66	0.489	0.532
Baseline + Random Masking	17.10 (+1.44)	0.536 (+0.047)	0.485 (-0.047)
+ **Dynamic-Region-Aware Opt. (Ours)**	**19.97** (+4.31)	**0.627** (+0.138)	**0.339** (-0.193)

Effectiveness of Semantic Branch. As shown in Table 5, adding the parallel semantic understanding branch consistently improves NVS performance on D-RE10K. Compared with the geometry-only baseline, the full unified framework achieves better results across all metrics. This indicates that joint semantic-geometric optimization benefits image rendering and does not compromise reconstruction fidelity.

Table 5: Effectiveness of Semantic Branch. Quantitative comparison demonstrating the performance gains after incorporating the semantic branch.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o semantic branch	19.77	0.615	0.356
w/ semantic branch (ours)	**19.97** (+0.20)	**0.627** (+0.012)	**0.339** (-0.017)

Ablation of Auxiliary Components. We provide additional ablations by separately removing pre-training, copy-paste augmentation (CPA), and SAM2 refinement. Removing any of these components leads to performance degradation, showing that they all contribute to the final performance. Among them, pre-training has the largest impact, indicating that a strong feed-forward reconstruction prior is important for dynamic view NVS. In contrast, removing CPA or test-time SAM2 refinement only causes a minor drop. This suggests that the improvement is not mainly driven by external mask augmentation or SAM2

post-processing. Instead, our dynamic-region-aware loss in Eq. (5)–(7) plays the primary role in guiding the model to handle dynamic interference during NVS.

Table 6: Ablation of Auxiliary Components. We ablate pre-training, copy-paste augmentation (CPA), and SAM2 refinement.

Method	PSNR↑	SSIM↑	LPIPS↓
Ours w/o pre-training	18.66	0.578	0.381
Ours w/o CPA	19.82	0.625	0.348
Ours w/o SAM2 refine	19.90	0.624	0.341
Ours	**19.97**	**0.627**	**0.339**

5 Conclusion

In this paper, we present SPAR, a unified feed-forward framework that alleviates the strict static-scene assumptions limiting many 3D foundation models. SPAR jointly performs novel view synthesis and open-vocabulary semantic understanding from a few unposed dynamic views, addressing spatial feature misalignment caused by transient moving objects. Our method introduces a joint semantic-geometric encoding architecture with a Cross-View Dynamic Region Predictor that suppresses dynamic regions before latent scene aggregation. Combined with a self-supervised dynamic-region-aware training scheme, SPAR encourages optimization to focus on multi-view consistent static regions. This training scheme relies only on reconstruction loss and does not require ground-truth motion mask labels.

Extensive evaluations on the challenging D-RE10K benchmark show that SPAR achieves state-of-the-art performance in both dynamic NVS and motion mask prediction under extreme few-view settings. Beyond these gains, we observe strong inter-task synergy: semantic learning acts as a structural regularizer that improves photometric reconstruction quality. By enabling robust and temporally stable 3D scene representations without explicit dynamic object annotations, SPAR provides an efficient and scalable foundation for embodied perception in complex real-world environments.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62403462, 62192782, 62532015, U22B2056, 62422317, 62202469), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (Grant No. 2023B03024), the Beijing Natural Science Foundation (Grant No. L223003, L243015), and the Beijing Major Science and Technology Project (Contract No. Z251100008425008). We thank the authors of LVSM [14], RayZer [13], WildRayZer [4] for clarifying details of their models.

References

1. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 **2**(3), 4 (2021)
2. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19457–19467 (2024)
3. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14124–14133 (2021)
4. Chen, X., Zhou, W., Cheng, Z.: Wildrayzer: Self-supervised large view synthesis in dynamic environments. arXiv preprint arXiv:2601.10716 (2026)
5. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European Conference on Computer Vision. pp. 370–386. Springer (2024)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 17864–17875 (2021)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
9. Fan, Z., Zhang, J., Cong, W., Wang, P., Li, R., Wen, K., Zhou, S., Kadambi, A., Wang, Z., Xu, D., Ivanovic, B., Pavone, M., Wang, Y.: Large spatial model: End-to-end unposed images to semantic 3d (2024), `https://arxiv.org/abs/2410.18956`
10. Gonzalez, R.C.: Digital image processing. Pearson education india (2009)
11. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: The Twelfth International Conference on Learning Representations (2024)
12. Huang, N., Zheng, W., Xu, C., Keutzer, K., Zhang, S., Kanazawa, A., Wang, Q.: Segment any motion in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3406–3416 (2025)
13. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., et al.: Rayzer: A self-supervised large view synthesis model. arXiv preprint arXiv:2505.00702 (2025)
14. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=QQBPWtvtcn`
15. Jin, L., Tucker, R., Li, Z., Fouhey, D., Snavely, N., Holynski, A.: Stereo4d: Learning how things move in 3d from internet stereo videos. arXiv preprint arXiv:2412.09621 (2024)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
17. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19729–19739 (2023)

18. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation (2022), https://arxiv.org/abs/2205.15585
19. Kulhanek, J., Peng, S., Kukelova, Z., Pollefeys, M., Sattler, T.: Wildgaussians: 3d gaussian splatting in the wild. arXiv preprint arXiv:2407.08447 (2024)
20. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=RriDjddCLN
21. Li, W., Xu, Y., Yao, M., Liang, F., Sun, J., Wang, M., Zhang, G., Huang, L., Li, H.: Noperoomgs: Indoor 3d gaussian splatting optimization without camera pose input. Advances in Neural Information Processing Systems **38**, 49124–49144 (2026)
22. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10486–10496 (2025)
23. Li, Z., Xian, W., Davis, A., Snavely, N.: Crowdsampling the plenoptic function. In: European Conference on Computer Vision. pp. 178–196. Springer (2020)
24. Lu, F., Xu, Y., Chen, G., Li, H., Lin, K.Y., Jiang, C.: Urban radiance field representation with deformable neural mesh primitives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 465–476 (2023)
25. Lu, J., Deng, J., Wang, C., He, J., Zhang, T.: Query refinement transformer for 3d instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18516–18526 (2023)
26. Markin, A., Pryadilshchikov, V., Komarichev, A., Rakhimov, R., Wonka, P., Burnaev, E.: T-3dgs: Removing transient objects for 3d scene reconstruction. arXiv preprint arXiv:2412.00155 (2024)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
28. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20051–20060 (2024)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
31. Ren, W., Zhu, Z., Sun, B., Chen, J., Pollefeys, M., Peng, S.: Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8931–8940 (2024)
32. Sabour, S., Goli, L., Kopanas, G., Matthews, M., Lagun, D., Guibas, L., Jacobson, A., Fleet, D., Tagliasacchi, A.: Spotlesssplats: Ignoring distractors in 3d gaussian splatting. ACM Transactions on Graphics **44**(2), 1–11 (2025)
33. Sajjadi, M.S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., et al.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6229–6238 (2022)

34. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8216–8223. IEEE (2023)
35. Wang, J., Yuan, Y., Zheng, R., Lin, Y., Gao, J., Chen, L.Z., Bao, Y., Zhang, Y., Zeng, C., Zhou, Y., et al.: Spatialvid: A large-scale video dataset with spatial annotations. arXiv preprint arXiv:2509.09676 (2025)
36. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2021)
37. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
39. Wei, D., Zhao, L., Li, W., Huang, Z., Ji, S., Liu, P.: Siu3r: Simultaneous scene understanding and 3d reconstruction beyond feature alignment. Advances in Neural Information Processing Systems **38**, 110800–110830 (2026)
40. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. Advances in Neural Information Processing Systems **35**, 3502–3516 (2022)
41. Xu, W., Shi, C., Tu, S., Zhou, X., Liang, D., Bai, X.: A unified framework for 3d scene understanding. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
42. Xu, Y., Wang, Y., Yu, S.X.: Novel view synthesis from a few glimpses via test-time natural video completion. Advances in Neural Information Processing Systems **38**, 46541–46561 (2026)
43. Yang, L., Cai, B., Liu, W., Wang, Y., Yuan, C., Li, B., Hu, W.: Sra-det: Learning omni-grained open-vocabulary detection beyond category names. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27611–27620 (June 2026)
44. Yang, L., Xu, Y., Wang, S., Yuan, C., Zhang, Z., Li, B., Hu, W.: Pdnet: Toward better one-stage object detection with prediction decoupling. IEEE Transactions on Image Processing **31**, 5121–5133 (2022)
45. Yang, L., Xu, Y., Yuan, C., Liu, W., Li, B., Hu, W.: Improving visual grounding with visual-linguistic verification and iterative reasoning. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9499–9508 (2022)
46. Yang, L., Zhang, Z., Qi, Z., Xu, Y., Liu, W., Shan, Y., Li, B., Yang, W., Li, P., Wang, Y., et al.: Exploiting contextual objects and relations for 3d visual grounding. Advances in Neural Information Processing Systems **36**, 49542–49554 (2023)
47. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In: International Conference on Learning Representations (ICLR) (2024)
48. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4578–4587 (2021)

49. Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., Xu, Z.: Gs-lrm: Large reconstruction model for 3d gaussian splatting. In: European Conference on Computer Vision (ECCV) (2024)
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
51. Zhang, T., Tian, X., Wu, Y., Ji, S., Wang, X., Zhang, Y., Wan, P.: Dvis: Decoupled video instance segmentation framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1282–1291 (2023)
52. Zhang, T., Tian, X., Zhou, Y., Ji, S., Wang, X., Tao, X., Zhang, Y., Wan, P., Wang, Z., Wu, Y.: Dvis++: Improved decoupled framework for universal video segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
53. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21676–21685 (2024)
54. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
55. Ziwen, C., Tan, H., Zhang, K., Bi, S., Luan, F., Hong, Y., Fuxin, L., Xu, Z.: Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4349–4359 (2025)
56. Zuo, X., Samangouei, P., Zhou, Y., Di, Y., Li, M.: Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. International Journal of Computer Vision **133**(2), 611–627 (2025)