FEATURE PUZZLES: SINGLE-IMAGE SUPERVISION FOR MULTI-VIEW 3D-AWARE SEGMENTATION

Anonymous authors
Paper under double-blind review

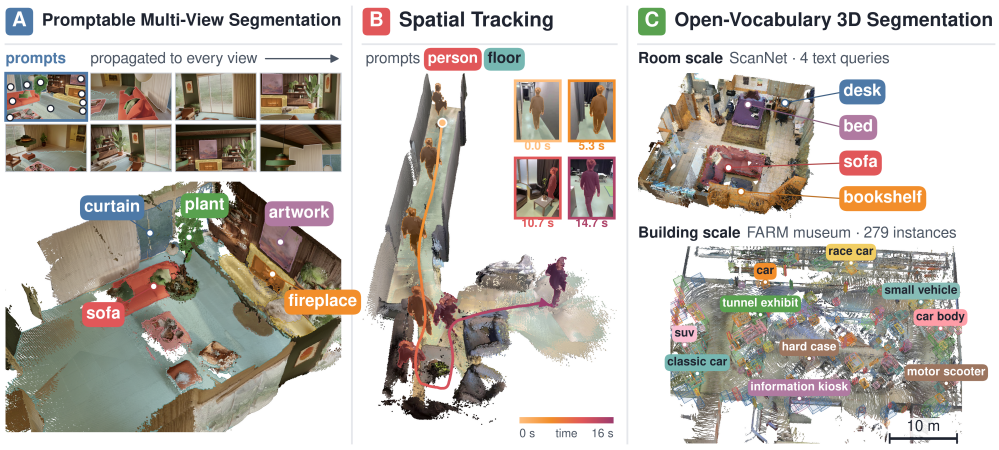

Figure 1: **Feature Puzzles** learns multi-view-consistent segmentation from *single images*, extending feed-forward reconstruction to (A) promptable segmentation that propagates to all views in one pass, (B) spatial tracking, and (C) open-vocabulary 3D segmentation.

ABSTRACT

Humans can easily recognize the same objects across different viewpoints of a 3D scene, whereas teaching models such multi-view-consistent perception typically requires videos or calibrated multi-view captures, which are considerably more expensive and less diverse than large-scale monocular data. In this paper, we introduce **Feature Puzzles**, a novel framework for learning multi-view-consistent segmentation using only single-image semantic segmentation supervision. Specifically, we propose a simple yet effective pipeline that constructs multi-view-consistent segmentation pseudo-labels through image cropping and geometric transformations. We further introduce a lightweight, plug-and-play network that leverages the multi-view-consistent image features learned by existing pre-trained 3D foundation models and the generated pseudo-labels to perform multi-view-consistent segmentation. Importantly, our flexible framework supports various types of prompts provided from one or multiple views and can be readily extended to open-vocabulary 3D segmentation. Under this single-image supervision, our feed-forward model achieves the best mIoU on NVOS, remains competitive on SPIn-NeRF and ScanNet++ against baselines using per-scene optimization or in-domain multi-view training, and demonstrates zero-shot transfer to open-vocabulary 3D segmentation. Visual comparisons are available on our anonymous project page: https://featurepuzzlessubmission.github.io/.

1 INTRODUCTION

Spatial intelligence requires not only recovering geometric structure but also recognizing and reasoning about scene content consistently across viewpoints (Yang et al., 2025). Multi-view-consistent perception, particularly multi-view-consistent segmentation, is essential for building coherent representations of the physical world and provides a foundation for downstream tasks such as robotic

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.

interaction (Qu et al., 2025; Zhang et al., 2026), embodied navigation (Zhang et al., 2024; Ma et al., 2026), and open-vocabulary 3D understanding (Peng et al., 2023).

Recent advances in feed-forward 3D reconstruction (Wang et al., 2025a; 2026b; Lin et al., 2025; Wang & Agapito, 2025; Wang et al., 2025b; Lan et al., 2026) have substantially improved the efficiency of recovering 3D geometry while enabling the processing of long sequences. Building on these advances, IGGT (Li et al., 2026) and IGGT4D (Zou et al., 2026) extend feed-forward reconstruction models with instance-level understanding of static and dynamic scenes, while Embodied-Splat (Lee et al., 2026) enables online 3D reconstruction and semantic understanding. Despite their effectiveness, these approaches typically rely on multi-view-specific resources, such as posed RGB-D scans, curated cross-view instance datasets, or multi-view distillation and self-supervision signals, which are considerably more difficult and costly to acquire than single-image semantic supervision.

This reveals a key supervision gap: pretrained multi-view backbones already provide strong geometric priors, but consistent cross-view semantics still depend on scarce multi-view annotations. In contrast, single-image segmentation data is abundant but lacks cross-view correspondence. We therefore ask whether single-image semantic supervision alone can supply this missing semantic layer without real multi-view annotations. We emphasize the scope of this question. We do not claim that 3D or multi-view consistency itself can be learned from monocular images alone; the geometric prior comes from the pretrained backbone, and our contribution is a supervision framework that adapts this representation to multi-view-consistent semantics.

To address this gap, we propose **Feature Puzzles**, a training framework that turns large-scale single-image segmentation data into synthetic multi-view supervision. It decomposes each image into overlapping crops and treats them as synthetic views, avoiding the need for multiple observations of the same scene. For these shared-parent crop views, correspondence is exact in the parent-image coordinate system, enabling randomly sampled 2D masks and dense visual features to be transformed consistently across views without additional multi-view semantic annotation.

Building on this supervision, we introduce a lightweight adaptation scheme for feed-forward 3D reconstruction models. We fuse intermediate multi-scale features and adapt only late backbone blocks with low-rank adapters, which can be disabled at inference to recover the original depth and camera outputs. This equips existing backbones with multi-view-consistent perception without redesigning the reconstruction pipeline. We instantiate Feature Puzzles for two complementary tasks: *promptable multi-view segmentation*, which accepts prompts from one or more views and predicts masks jointly across all views in a single forward pass, and *open-vocabulary 3D segmentation*, which distills DINOv3 features to learn dense, multi-view-consistent representations for 3D recognition. As shown in Figure 1, Feature Puzzles extends one reconstruction backbone to promptable segmentation, spatial tracking, and open-vocabulary 3D understanding with minimal architectural changes.

We organize our evaluation around a single hypothesis: *shared-parent single-image supervision with exact synthetic-view correspondence can teach multi-view-consistent semantics to a pretrained 3D geometry representation*. Each experiment tests a consequence of this hypothesis. Promptable segmentation tests whether object identity transfers across genuine camera views. Instance spatial tracking tests whether this correspondence survives spatial and temporal displacement. Driving scenes test whether it transfers beyond indoor distributions. Open-vocabulary 3D segmentation tests whether the same principle extends from masks to dense semantic feature fields. For task-specific training, promptable segmentation uses only single-image SA-1B mask supervision, while open-vocabulary 3D segmentation uses single-image DINOv3 feature distillation. Neither task uses task-specific multi-view or video annotations. Under this supervision, our model achieves the best mIoU on NVOS in a single forward pass, remains competitive on SPIn-NeRF and ScanNet++, and demonstrates transfer to out-of-distribution driving scenes. Ablations further support the contributions of multi-layer features, backbone adaptation, and cross-view processing.

2 Related Work

Feed-forward 3D reconstruction. Reconstruction has moved from per-scene optimization (Schönberger & Frahm, 2016; Schönberger et al., 2016) to feed-forward prediction: VGGT (Wang et al., 2025a), π^3 (Wang et al., 2026b), MapAnything (Keetha et al., 2026), and Depth Anything 3 (Lin et al., 2025) regress cameras, depth, and point maps in a single pass, while Spann3R (Wang & Agapito, 2025), CUT3R (Wang et al., 2025b), LingBot-Map (Chen et al., 2026),

and STream3R (Lan et al., 2026) extend this to streaming or causal reconstruction. Closest to our data construction, Puzzles (Ma et al., 2025a) synthesizes multi-view training data from a single image or clip, but only for geometric pointmap regression, without transferring semantic masks or features or defining a cross-view perception objective. These reconstruction models primarily focus on geometric structure rather than task-specific semantics. We build on a pretrained multi-view backbone and ask how single-image supervision can add a consistent semantic layer.

Feed-forward semantic 3D scene understanding. Per-scene methods lift 2D vision–language features into a NeRF or 3DGS field (Kerr et al., 2023; Zhou et al., 2024; Qin et al., 2024; Shi et al., 2024; Wu et al., 2024; Cheng et al., 2025; Kim et al., 2025; Li et al., 2025b; Ma et al., 2024) and must re-optimize for every scene. Feed-forward alternatives predict semantics together with geometry in one pass. LSM (Fan et al., 2024), Uni3R (Sun et al., 2026), SAB3R (Chen et al., 2025), and Ov3R (Gong et al., 2026) distill CLIP- or DINO-style features into a reconstruction network, while EmbodiedSplat (Lee et al., 2026) builds a semantic 3DGS online from RGB(-D) streams. For instance understanding, PanSt3R (Zust et al., 2025) adds a panoptic head to frozen MUSt3R and DINOv2 features, EPS3D (Zhu et al., 2026) distills LSeg and SAM into a VGGT-based model, IGGT (Li et al., 2026) and IGGT4D (Zou et al., 2026) learn cross-view instance embeddings, and SceneDINO (Jevtić et al., 2025) predicts a 3D DINO feature field from a single image. Most rely on multi-view scene data—including posed scans, curated cross-view instance datasets, or video—during semantic training. In contrast, Feature Puzzles uses no task-specific real multi-view or video supervision, while inheriting its geometric prior from a pretrained multi-view backbone. Its promptable branch uses single-image SA-1B masks, and its open-vocabulary branch a frozen DINOv3 teacher; cross-view supervision is created by shared-parent pull-back. Optional in-domain finetuning is reported separately from zero-shot evaluation.

Promptable and cross-view segmentation. SAM (Kirillov et al., 2023a) and SAM2 (Ravi et al., 2025) provide promptable segmentation for images and video but do not explicitly model 3D cross-view consistency, so masks can drift across views of a static scene. NVOS (Ren et al., 2022), SPIn-NeRF (Mirzaei et al., 2023), SA3D (Cen et al., 2023), SAGA (Cen et al., 2025a;b), DCHM (Ma et al., 2025b) and OmniSeg3D (Ying et al., 2024) restore consistency by distilling SAM masks into a per-scene field, at a cost of minutes to hours per scene. MASA (Li et al., 2024) learns 2D instance association from augmented static images, while MV-SAM (Jeong et al., 2026) trains on SA-1B but relies on a frozen π^3 pointmap predictor for cross-view consistency at inference. The closest work is VGGT-Segmentor (Gao et al., 2026), which also builds on VGGT and offers a single-image self-supervised variant trained on SA-1B. It formulates cross-view segmentation as *pairwise* source–target matching guided by explicit VGGT point tracks. Because targets are processed independently, larger unordered view sets require additional pair selection, association, or tracking. Feature Puzzles instead trains and reasons jointly over shared-parent N-view sequences, predicting all views in one forward pass without pairwise matching, point tracking, or predefined view ordering.

3 Method

3.1 Overview

We propose Feature Puzzles, which turns a single image into overlapping synthetic views with exactly aligned supervision (Section 3.2). A pretrained VGGT-Omega backbone encodes these views (Section 3.3) and feeds two lightweight heads for promptable multi-view segmentation (Section 3.4) and open-vocabulary 3D features (Section 3.5).

3.2 Single-Image Multi-Crop Puzzles

Figure 2 shows how Feature Puzzles converts one annotated image into a pseudo-view sample. We call the original image the *parent image*, containing RGB, instance masks, dense teacher features, and semantic labels. It is cropped into overlapping *sub-images*, each treated as a synthetic view. Since all sub-images share the same parent, overlapping regions inherit exactly aligned supervision by construction, enabling multi-view-consistent training without real multi-view or video data.

Feature puzzles generation. Let a parent image I have spatial size $H \times W$. As in Figure 2B, we sample N overlapping *sub-images* $\{x_n\}_{n=1}^{N}$, each a box in parent coordinates of side length $s \in [s_{\min}, s_{\max}]$ (fraction of the shorter side; default $[0.5, 0.9]$), resampled to square resolution R

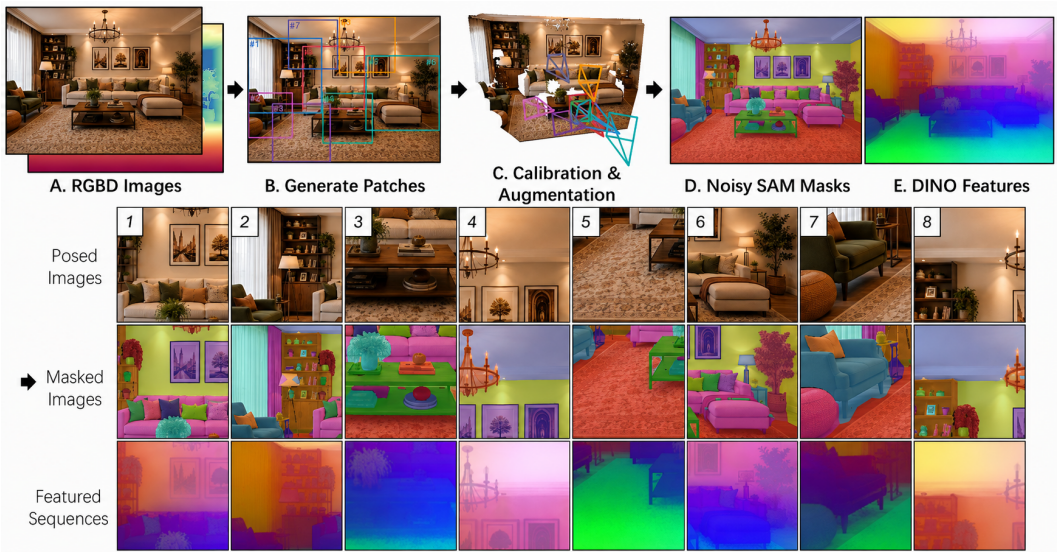

Figure 2: **Feature Puzzles pipeline.** From a single RGB(-D) parent image (**A**), we sample overlapping crops as a puzzle (**B**), optionally calibrate their cameras and apply 3D augmentation (**C**), and pull back SAM masks (**D**) and DINO features (**E**) into every crop (bottom).

so $x_n \in \mathbb{R}^{R\times R\times 3}$. We also write x_n for the box, so $x_i \cap x_j$ is the overlap of two sub-images. A candidate is accepted only if it overlaps an accepted sub-image with IoU $\geq o_{\min}$, so the N sub-images form a connected overlap graph—the condition for cross-view warping.

3D-augmented views. Affine crops contain no parallax or occlusion. Optionally, following Image-to-Clips (Ma et al., 2025a), we estimate parent depth with MoGe-2 (Wang et al., 2025c), calibrate each view's camera, and apply 6-DoF augmentation, pulling labels back through depth projection. This optional construction is distinct from the affine crop mapping defined below.

The sub-image-to-parent map. For the affine crop views, sub-image x_n is related to the parent I by a closed-form affine map π_n that we chose when sampling it, so $x_n = I \circ \pi_n$ maps sub-image coordinates to parent coordinates. Unlike a camera pose, π_n is never estimated: it is fixed by the sampled box, requires no depth and no camera model, and is invertible in closed form.

Label transfer by pull-back. For the affine crop construction, the same map π_n that relates a sub-image to the parent RGB transfers labels just as well: rather than annotating each crop independently, we pull back a single parent-space label field ℓ—which may carry instance masks, semantic classes, or dense teacher features—by composition with π_n and resample to crop resolution,

$$\ell_n = \mathcal{R}_{\mathcal{Y}}\big[\,\ell \circ \pi_n\,\big]. \tag{1}$$

Here, $\mathcal{R}_{\mathcal{Y}}$ resamples nearest-neighbour for discrete labels (instance identity, semantic class) and bilinearly with re-normalization for dense features. The parent is the single source of truth and each view a restriction of it, so two views cannot disagree: for any point p in the overlap $x_i \cap x_j$,

$$\ell_i\big(\pi_i^{-1}(p)\big) \;=\; \ell_j\big(\pi_j^{-1}(p)\big) \;=\; \ell(p). \tag{2}$$

Here, $\pi_i^{-1}(p)$ is p in sub-image i's local coordinates, so both sides read the same parent label $\ell(p)$ through different crop coordinates—cross-view consistency is exact *by construction* for these shared-parent affine views.

What the parent field is. The field ℓ takes two forms. *For promptable segmentation*, ℓ is the set of SA-1B instance masks $\{\mathcal{A}_m\}$; pulling each mask $\mathcal{A}_m$ back to view n gives that view's binary target $\mathbf{G}_{n,m}$ for instance m. Since m indexes the parent field, not a view, it is a *global instance identifier*: channel m is the same object in all N views, so *no* Hungarian matching (Li et al., 2025a) or cross-view association (Ravi et al., 2025) is needed and the prompt–target pairing of Section 3.4 is one-to-one. *For open-vocabulary 3D segmentation*, ℓ is a frozen DINOv3 teacher (Siméoni et al., 2025) evaluated *once on the whole parent image* and pulled back by grid sampling. Evaluating before

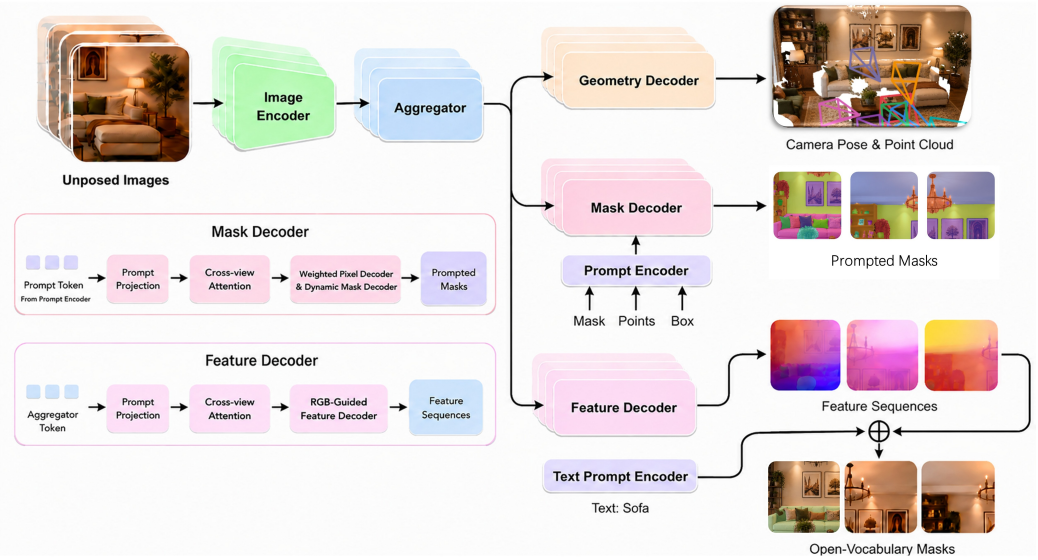

Figure 3: **Feature Puzzles network architecture.** A shared *Image Encoder* + *Aggregator* produce tokens feeding three decoders: *Geometry* (poses, 3D points), prompt-conditioned *Mask* (mask/point/box $\rightarrow$ masks), and *Feature* ($\oplus$ text-prompt $\rightarrow$ open-vocab masks). Mask and feature decoders share cross-view attention and a weighted pixel decoder.

cropping—not per-view—is what guarantees consistency: transformer features change under crop-and-resize, so per-view targets would disagree on the overlap; pull-back instead gives every view a target from the same parent field, so distillation and cross-view consistency share one optimum (Appendix A.1). Appendix A.2 gives the augmentation policy, prompt sampling, and annotation filters; Appendix A.3 contrasts this with the Image-to-Clips augmentation of Puzzles (Ma et al., 2025a), whose synthetic views carry geometric labels from pose fitting and re-rendering. We anneal the minimum crop overlap $o_{\min}$ from 0.7 to 0.1, exposing the model first to near-duplicate views and then to progressively larger viewpoint changes with less shared content.

3.3 Pretrained Multi-View Backbone

As shown in Figure 3, we build on a pretrained VGGT-Omega (Wang et al., 2025a) multi-view transformer. We keep its early blocks frozen and adapt late blocks with LoRA (Hu et al., 2022), preserving the original depth and camera heads at inference. The segmentation heads consume a shared representation $\mathbf{F}_n$ obtained by fusing cached multi-layer tokens. Adapters are initialized to preserve the pretrained backbone at the start of training, enabling two-pass inference with adapters off for reconstruction and on for segmentation. We further evaluate the same adaptation recipe on LingBot-Map (Chen et al., 2026), a streaming reconstruction backbone with a different multi-view attention design, providing evidence that the approach is not specific to VGGT-Omega. Layer choice, fusion, and initialization are detailed in Appendix B.1.

3.4 Promptable Multi-View Segmentation

Architecture. As shown in Figure 3 (*Mask Decoder*), given the sub-images $\{x_n\}$ and a set of prompts (masks, boxes, or labelled points), the model predicts masks for M instances across all N views in a single forward pass. A SAM-style prompt encoder (Kirillov et al., 2023b) encodes each prompt into a dense coverage map $\mathbf{C}_{n,m}$ and a sparse prompt token; during training prompts may be supplied in one or multiple views. The coverage map injects the prompt into the shared representation $\mathbf{F}_n$ from the backbone through a feature-wise linear projection module,

$$\mathbf{F}'_{n,m} = \gamma(\mathbf{C}_{n,m}) \odot \mathbf{F}_n + \beta(\mathbf{C}_{n,m}), \tag{3}$$

where γ and β are learnable projections producing a per-channel scale and shift, yielding an instance-conditioned feature $\mathbf{F}'_{n,m}$ (unprompted view–instance pairs pass through unchanged). The instance-conditioned tokens then pass through cross-view self-attention—which allows prompt-

conditioned information from one view to propagate to the others—and SAM-style two-way attention between image and object tokens, denoted $\mathbf{H}_{n,m} = \mathrm{Attn}(\mathbf{F}'_{n,m})$ with its image and object outputs $\mathbf{H}^{\mathrm{img}}_{n,m}, \mathbf{H}^{\mathrm{obj}}_{n,m}$. A weighted pixel decoder and a per-instance kernel head then produce the mask logits,

$$\mathbf{D}_n = \mathrm{PixelDecoder}(\mathbf{H}^{\mathrm{img}}_{n,m}), \quad \mathbf{k}_{n,m} = \mathrm{KernelHead}(\mathbf{H}^{\mathrm{obj}}_{n,m}), \quad \mathbf{Y}_{n,m} = \langle \mathbf{D}_n, \mathbf{k}_{n,m} \rangle. \tag{4}$$

Auxiliary heads predict mask IoU. Appendix B.2 details the token layout and attention blocks.

Training objective. Every prompt corresponds one-to-one to a ground-truth mask, and targets are warped into every crop—including crops where the prompt was not given—so the objective is a plain weighted sum,

$$\mathcal{L}_{\mathrm{seg}} = \lambda_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}} + \lambda_{\mathrm{iou}}\mathcal{L}_{\mathrm{iou}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}} + \lambda_{\mathrm{pt}}\mathcal{L}_{\mathrm{pt}}, \tag{5}$$

combining: $\mathcal{L}_{\mathrm{bce}}$ and $\mathcal{L}_{\mathrm{dice}}$ for pixel-wise mask supervision (binary cross-entropy and Dice), $\mathcal{L}_{\mathrm{iou}}$ for IoU regression on a per-view confidence head, $\mathcal{L}_{\mathrm{cons}}$ as an auxiliary cross-view token regularizer that pulls object tokens of the same instance toward their cross-view centroid, and $\mathcal{L}_{\mathrm{pt}}$ for point-prompt supervision. Per-term details are in Appendix B.2.

3.5 Open-Vocabulary 3D Segmentation

RGB-guided dense feature head. The Feature Decoder in Figure 3 takes the shared representation $\mathbf{F}_n$ from the backbone, mixes the N views with cross-view attention, and upsamples with RGB guidance at each stage to restore image-aligned boundaries, yielding a dense feature map $\widehat{\mathbf{F}}_n$ aligned with the frozen DINOv3 (Siméoni et al., 2025)/Talk2DINO (Barsellotti et al., 2025) space.

Stage 1: single-image pre-training. We first train the head on SA-1B puzzle crops with

$$\mathcal{L}_{\mathrm{stage1}} = \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{warp}}\mathcal{L}_{\mathrm{warp}} + \lambda_{\mathrm{gram}}\mathcal{L}_{\mathrm{gram}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{feat}}$ is the mean pixel-wise cosine distance between the student $\widehat{\mathbf{F}}_n$ and the pulled-back DINOv3 teacher of that view, and the other two terms exploit the exact crop geometry of the shared-parent affine views. The cross-view warping term maps a student feature from view i onto view j over their overlap and pulls it toward a stop-gradient target:

$$\mathcal{L}_{\mathrm{warp}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{1}{|\Omega_{ij}|} \sum_{q\in\Omega_{ij}} \Big(1 - \cos\big(\widehat{\mathbf{F}}_i(\pi_i^{-1} \circ \pi_j(q)),\ \mathrm{sg}[\widehat{\mathbf{F}}_j(q)]\big)\Big), \tag{7}$$

where $\mathcal{P}$ is the set of overlapping view pairs, Ω_{ij} their overlap, q a pixel in view j, and $\pi_i^{-1} \circ \pi_j$ maps q from view j into view i so both cosine arguments sit at the same parent location; $\mathrm{sg}[\cdot]$ blocks the gradient through the target. The Gram anchoring term matches pairwise similarities. We sample K pixels $\{p_k\}$ from a view and stack the L2-normalized student features $\widehat{\mathbf{F}}_n(p_k)$ and the pulled-back teacher features $\mathbf{F}_t(p_k)$ at the same pixels as the rows of $\widehat{\mathbf{S}} = [\widehat{\mathbf{F}}_n(p_1); \ldots; \widehat{\mathbf{F}}_n(p_K)]$ and $\mathbf{S}_T = [\mathbf{F}_t(p_1); \ldots; \mathbf{F}_t(p_K)]$, then match the $K \times K$ Gram matrices

$$\mathcal{L}_{\mathrm{gram}} = \left\| \widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top - \mathbf{S}_T\mathbf{S}_T^\top \right\|_F^2, \tag{8}$$

which is designed to preserve cross-boundary feature dissimilarity and tie different scales to a common similarity geometry. Weights and sampling details are in Appendix B.3.

Stage 2: in-domain text-aligned fine-tuning. Stage 1 aligns $\widehat{\mathbf{F}}_n$ with DINOv3 but not with language. We therefore add per-pixel text-aligned cross-entropy using labels warped from the parent image. For pixel p and class k, $z_{p,k} = \cos(\widehat{\mathbf{f}}_p, \mathbf{e}_k)/\tau$ with $\tau = 0.07$, where $\mathbf{e}_k$ is obtained from a frozen CLIP text tower and mapped into DINOv3 space by frozen Talk2DINO. Stage-1 distillation is retained to preserve DINOv3 alignment, while the new loss introduces explicit supervision for the in-domain semantic classes.

3D lifting and classification. At inference, each RGB frame produces a dense feature map. Depth and camera pose are used to unproject feature pixels into 3D, where features are averaged within a 0.02 m voxel grid to obtain one feature $\mathbf{f}_{\mathrm{voxel}}$ per voxel. Each query point takes its nearest voxel feature and is classified by $\hat{y} = \arg\max_k \cos(\mathbf{f}_{\mathrm{voxel}}, \mathbf{e}_k)$, where $\mathbf{e}_k$ is obtained from the frozen CLIP/Talk2DINO text mapping. The entire process is feed-forward and requires no scene-specific training or test-time optimization.

Table 1: **Promptable multi-view segmentation on NVOS and SPIn-NeRF.** Per-scene methods fit each test scene; generalization methods use one forward pass. Ours uses only SA-1B-style masks for task-specific segmentation training.

Category	Method	Train	NVOS		SPIn	
			mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)
Per-scene opt.	SPIn-NeRF (Mirzaei et al., 2023)	NeRF	–	–	90.7	98.8
	SA3D (Cen et al., 2023)	NeRF+SAM	91.1	98.4	92.4	98.8
	SAGA (Cen et al., 2025a)	3DGS+SAM	92.6	98.6	93.7	99.2
	SA3D-GS (Cen et al., 2025b)	3DGS+SAM	92.7	98.5	93.4	99.1
	OmniSeg3D (Ying et al., 2024)	3DGS+SAM	92.8	98.6	94.5	99.3
General.	SAM2 (Ravi et al., 2025)	SA-V (video)	88.7	94.6	86.6	93.6
	MV-SAM (Jeong et al., 2026)	SA-1B (single image)	92.1	97.5	**92.9**	97.1
	Ours	SA-1B (single image)	**93.2**	**98.3**	88.5	**97.7**

4 Experiments

Feature Puzzles extends feed-forward reconstruction with scene understanding. We evaluate promptable segmentation on NVOS and SPIn-NeRF (Section 4.1), instance tracking on ScanNet++ (Section 4.2), OOD tracking on nuScenes-mini (Section 4.2), and open-vocabulary 3D segmentation following EmbodiedSplat (Section 4.3).

4.1 Promptable Multi-View Segmentation

We evaluate promptable multi-view segmentation on NVOS (Ren et al., 2022) and SPIn-NeRF (Mirzaei et al., 2023) following the SAGA protocol, using a single reference-view mask prompt. Task-specific training uses only single-image SA-1B-style masks, with no video or real multi-view supervision. As shown in Table 1, our feed-forward model achieves the best mIoU on NVOS (93.2%) and outperforms SAM2-Video on SPIn-NeRF (88.5% vs. 86.6%), while remaining below MV-SAM and the strongest per-scene methods there. Across both datasets, inference requires no scene-specific optimization and generalizes directly to unseen scenes.

4.2 Instance Spatial Tracking

ScanNet++ evaluation. We follow the IGGT (Li et al., 2026) protocol on ScanNet++ (Yeshwanth et al., 2023), propagating a reference mask to all frames in one forward pass over 19 instances from 7 scenes. IGGT is trained on ScanNet++ RGB-D and is therefore in-distribution, whereas our task-specific segmentation training uses only SA-1B and is OOD to ScanNet++. As shown in Table 2, our method achieves 63.60% mIoU and 87.50% T-SR, approaching IGGT (73.02% / 98.90%) while outperforming SAM2* and SpaTracker+SAM.

OOD evaluation on nuScenes-mini. We adapt the same protocol to nuScenes-mini (Caesar et al., 2020) (9 scenes, 6 surround cameras), using projected 3D boxes as pseudo-masks. Here, all methods are OOD: IGGT is trained on ScanNet++ RGB-D, while our task-specific segmentation training still uses only SA-1B. Despite wide baselines and minimal frustum overlaps across the six cameras, our method in one forward pass achieves 4.78% T-mIoU$_{\text{all}}$, 20.97% T-mIoU$_{\text{vis}}$, and the highest T-SR$_{\text{vis}}$ of 88.89%. IGGT obtains higher T-SR$_{\text{all}}$ through HDBSCAN matching but lower mIoU. See Appendix E for detailed visual comparisons on ScanNet++ and nuScenes-mini.

4.3 Open-Vocabulary 3D Segmentation

We follow the EmbodiedSplat protocol (Lee et al., 2026) on ScanNet, ScanNet200, and ScanNet++, using 11 ScanNet and 4 ScanNet++ scenes with stride 10. Features are voxel-averaged at 0.02 m, lifted with ground-truth geometry, and classified using Talk2DINO text embeddings. Thus, the benchmark measures semantic feature quality under fixed geometry. We compare with per-scene 3DGS methods (Qin et al., 2024; Shi et al., 2024; Wu et al., 2024; Cheng et al., 2025; Kim et al., 2025; Li et al., 2025b; Katragadda et al., 2025) and feed-forward EmbodiedSplat. See Appendix E for coherent semantic reconstruction across diverse large-scale environments.

Table 2: **Instance spatial tracking on ScanNet++ and nuScenes-mini.** IGGT is in-distribution on ScanNet++ but OOD on nuScenes. Ours is OOD with respect to task-specific training on both benchmarks and uses only SA-1B for task-specific segmentation training. T-SR denotes temporal success rate; nuScenes reports all-frame (all) and visible-frame (vis) metrics.

Method	Training Data	ScanNet++		nuScenes-mini			
		mIoU (%)	T-SR (%)	T-mIoU$_{\text{all}}$	T-mIoU$_{\text{vis}}$	T-SR$_{\text{all}}$	T-SR$_{\text{vis}}$
SpaTracker+SAM (Li et al., 2026)	Kubric synth. (video)	16.15	23.68	–	–	–	–
CoTracker3+SAM (Karaev et al., 2025)	Kubric synth. (video)	–	–	4.99	18.15	0.00	17.41
SAM2* (Ravi et al., 2025)	SA-V (video)	44.16	57.89	4.48	20.71	0.00	28.70
IGGT (Li et al., 2026)	ScanNet++ RGB-D	**73.02**	**98.90**	2.34	9.46	59.26	78.89
Ours	SA-1B (single image)	63.60	87.50	**4.78**	**20.97**	**47.13**	**88.89**

Table 3: **3D semantic segmentation on the EmbodiedSplat benchmark.** † zero-shot evaluation with SA-1B task-specific training only; ‡ closed-set evaluation after in-domain fine-tuning on the evaluated semantic classes. "cls" denotes class. Bold = best, underline = second best.

Method	ScanNet						ScanNet200		ScanNet++		Recon. time	Per-Scene/ Gen.	on/ off
	10 cls		15 cls		19 cls		70 cls		20 cls				
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc			
In-domain / benchmark-specific methods													
LangSplat (Qin et al., 2024)	6.52	20.11	3.25	13.16	1.34	7.44	0.72	4.39	2.21	10.18	∼6 hr	Per-Scene	Offline
LEGaussians (Shi et al., 2024)	6.79	18.13	4.13	15.49	2.53	5.67	1.39	5.45	2.93	9.34	∼6 hr		
Online-LangSplat (Katragadda et al., 2025)	7.13	21.56	3.89	14.52	3.45	8.97	2.45	4.12	4.51	11.34	5.4 min		Online
OpenGaussian (Wu et al., 2024)	29.50	44.61	23.74	39.14	22.52	35.02	15.15	25.66	25.65	37.03	∼2.5 hr	Per-Scene	Offline
Occam's LGS (Cheng et al., 2025)	42.14	70.28	35.04	63.71	30.49	57.91	20.32	40.49	34.08	61.19	∼2 hr		
Dr. Splat (Kim et al., 2025)	39.21	66.66	31.84	60.58	28.38	55.85	19.29	33.84	39.85	58.34	∼2 hr		
InstanceGaussian (Li et al., 2025b)	29.77	52.32	28.79	50.07	26.57	48.63	23.20	38.32	29.98	47.47	∼3 hr		
EmbodiedSplat (Lee et al., 2026)	57.41	82.45	55.18	**80.27**	52.12	75.66	34.75	52.36	44.03	66.27	8 min	Generalizable	Online
EmbodiedSplat-fast (Lee et al., 2026)	51.05	80.15	46.92	77.15	43.89	72.73	32.43	58.14	51.09	**78.68**	1 min 10 s		
Ours (SN200, SN++)‡	**68.03**	**83.32**	**64.38**	80.11	**60.95**	**78.90**	**47.79**	**67.74**	**54.68**	66.97	∼25.9 s	Generalizable	Online
Zero-shot / Open-vocabulary													
Ours (SA-1B)†	41.09	58.55	35.87	50.73	30.50	45.93	13.79	23.79	30.12	49.23	∼25.9 s	Generalizable	Online

Zero-shot. As shown in Table 3, our Stage-1 model uses only SA-1B task-specific training with no in-domain labels, outperforming the listed per-scene 3DGS methods on ScanNet-15/19 while remaining below EmbodiedSplat trained on in-domain RGB-D scans.

In-domain fine-tuning. Table 3 also reports results after fine-tuning the same representation on ScanNet200/ScanNet++ vocabularies (‡). Because these classes are explicitly supervised during fine-tuning, this setting is closed-set rather than zero-shot open-vocabulary. In this setting, it exceeds EmbodiedSplat on mIoU on all five subsets (e.g., SN200 47.79 *vs.* 34.75, SN++ 54.68 *vs.* 44.03), showing that the SA-1B-pretrained features provide an effective initialization for in-domain training.

Efficiency. For the speed benchmark, we measure the model's full feed-forward pipeline, including its own reconstruction and semantic prediction, without using ground-truth depth. On a 363-image scene, inference takes $\sim$ 25.9 s. This is substantially faster than the reported scene-processing times of per-scene 3DGS methods ($\sim$ 2–6 hr) and about 18$\times$ faster than EmbodiedSplat (8 min).

5 Ablation Study

Speed and memory. We study how a single forward pass of the promptable model scales with the number of input views (Figure 4a–b). We report a 6-view operating point from the view-count sweep and a long-view stress test on SceneNN-MV scene 005 (40–248 frames), whose longest contiguous sequence contains 248 frames. At 6 views the model achieves 97.8% precision and 84.2% recall. On the long sequence, precision falls steadily as absent-view false positives accumulate, from 78.3% at 40 views to 22.8% at 248 views, while recall stays comparatively stable (69.5–83.1%). GPU memory and runtime both grow approximately linearly with the view count, from 5.40 GB / 1.68 s at 6 views to 32.11 GB / 16.68 s at 248 views. Although all 248 frames fit in a single forward pass, the substantial precision degradation at large view counts indicates that perceptual scalability, rather than memory alone, remains a limitation.

Component ablations. We evaluate the contribution of individual components under a short 2,500-step budget (Figure 4c–d; Appendix C). For open-vocabulary segmentation, using only the last

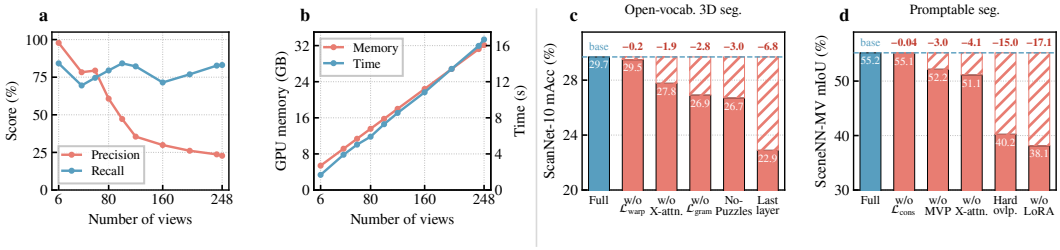

Figure 4: **Ablation study. (a, b)** Precision/recall, memory, and runtime versus view count. The long-view stress test shows near-linear memory/runtime scaling, while precision degrades at large view counts. **(c, d)** Component ablations for open-vocabulary (ScanNet-10 mAcc) and promptable segmentation (SceneNN-MV mIoU); numbers indicate the drop from the full model.

Table 4: **Multi-backbone generalization.** The same head and task-specific SA-1B training are applied to an offline and a streaming reconstruction backbone, without in-domain fine-tuning.

Backbone	Multi-view architecture	Open-vocabulary 3D segmentation								Instance spatial tracking			
		ScanNet						ScanNet200		nuScenes-mini			
		10 cls		15 cls		19 cls		70 cls					
		mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	T-mIoU$_{\text{all}}$	T-mIoU$_{\text{vis}}$	T-SR$_{\text{all}}$	T-SR$_{\text{vis}}$
LingBot-Map	Online, streaming	27.38	40.26	19.45	29.88	15.69	27.21	5.67	11.38	3.68	13.21	41.98	78.90
VGGT-Omega (default)	Offline, all-view attention	**41.09**	**58.55**	**35.87**	**50.73**	**30.50**	**45.93**	**13.79**	**23.79**	**4.78**	**20.97**	**47.13**	**88.89**

backbone layer (*Last layer*) costs 6.8 mAcc, training without puzzles (*No-Puzzles*) 3.0, removing Gram anchoring 2.8, and removing cross-view attention (*X-attn.*) 1.9. The No-Puzzles variant jointly removes the puzzle-based multi-view formulation and its associated cross-view processing, so this comparison measures their combined contribution rather than isolating the shared-parent supervision alone. For promptable segmentation, removing LoRA costs 17.1 mIoU, training only on low-overlap crops (*Hard ovlp.*) 15.0, removing cross-view attention 4.1, and removing multi-view prompting (*MVP*) 3.0. The Hard-ovlp. result shows that beginning directly with difficult low-overlap views is detrimental under this training budget, rather than isolating the effect of the curriculum schedule itself. $\mathcal{L}_{\text{warp}}$ and $\mathcal{L}_{\text{cons}}$ have little effect, indicating that these auxiliary consistency losses are not the primary source of the measured performance gains.

Backbone architecture. We apply the same recipe to two backbones with different multi-view designs (Table 4): VGGT-Omega (Wang et al., 2026a), an offline model where all views attend to each other, and LingBot-Map (Chen et al., 2026), a causal streaming model where each frame attends only to past context. Both use the same head, cached layers, and SA-1B training. The recipe transfers to LingBot-Map without modification, but VGGT-Omega is better on all metrics (e.g., +13.7 mIoU on ScanNet-10). Since LingBot-Map fits the DINOv3 targets better (0.816 *vs.* 0.766 cosine), the gap is unlikely to arise from poorer DINOv3 feature fitting alone and may partly reflect differences in cross-view context, including LingBot-Map's causal attention.

6 Conclusion

We presented Feature Puzzles, a simple recipe that turns a single-image dataset into synthetic multi-view supervision by cropping a parent image into overlapping sub-images and pulling labels back, so supervision targets are exactly consistent across overlapping affine crop views. A pretrained VGGT-Omega backbone with layer-selective LoRA feeds two heads—promptable multi-view masks and open-vocabulary 3D features—whose task-specific adaptation requires no real multi-view or video supervision. Despite this task-specific supervision requirement, a single forward pass achieves the best mIoU on NVOS, transfers to instance spatial tracking on ScanNet++ and to the out-of-distribution nuScenes-mini setting, and enables zero-shot open-vocabulary 3D segmentation from SA-1B supervision. With additional in-domain semantic fine-tuning, the same representation further exceeds EmbodiedSplat in mIoU across all evaluated subsets. These results are consistently achieved through efficient feed-forward inference without scene-specific per-scene optimization.

REFERENCES

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *ICCV*, 2025.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with NeRFs. In *NeurIPS*, 2023.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. In *AAAI*, 2025a.

Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields. *International Journal of Computer Vision (IJCV)*, 2025b.

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

Xuweiyi Chen, Tian Xia, Sihan Xu, Jianing Yang, Joyce Chai, and Zezhou Cheng. SAB3R: Semantic-augmented backbone in 3d reconstruction. *arXiv:2506.02112*, 2025.

Jiahuan Cheng, Jan-Nico Zaech, Luc Van Gool, and Danda Pani Paudel. Occam's lgs: An efficient approach for language gaussian splatting. In *BMVC*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3d. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40212–40229, 2024.

Yulu Gao, Bohao Zhang, Zongheng Tang, Jitong Liao, Wenjun Wu, and Si Liu. VGGT-Segmentor: Geometry-enhanced cross-view segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yoonwoo Jeong, Cheng Sun, Yu-Chiang Frank Wang, Minsu Cho, and Jaesung Choe. MV-SAM: Multi-view promptable segmentation using pointmap guidance. *arXiv:2601.17866*, 2026.

Aleksandar Jevtić, Christoph Reich, Felix Wimbauer, Oliver Hahn, Christian Rupprecht, Stefan Roth, and Daniel Cremers. Feed-forward SceneDINO for unsupervised semantic scene completion. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *ICLR*, 2025.

Saimouli Katragadda, Cho-Ying Wu, Yuliang Guo, Xinyu Huang, Guoquan Huang, and Liu Ren. Online language splatting. In *ICCV*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3d reconstruction. In *International Conference on 3D Vision*, 2026.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

Jun-Seong Kim, GeonU Kim, Yu-Ji Kim, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *CVPR*, 2025.

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

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

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

Seungjun Lee, Zihan Wang, Yunsong Wang, and Gim Hee Lee. EmbodiedSplat: Online feed-forward semantic 3dgs for open-vocabulary 3d scene understanding. In *CVPR*, 2026.

Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. Instancegaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14078–14088. IEEE, 2025a.

Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. InstanceGaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In *CVPR*, 2025b.

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

Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Haotong Lin, Sili Chen, Jun Hao 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.

Jiahao Ma, Miaomiao Liu, David Ahmedt-Aristizaba, and Chuong Nguyen. Hashpoint: Accelerated point searching and sampling for neural rendering. *arXiv preprint arXiv:2404.14044*, 2024.

Jiahao Ma, Lei Wang, Miaomiao Liu, David Ahmedt-Aristizabal, and Chuong Nguyen. Puzzles: Unbounded video-depth augmentation for scalable end-to-end 3d reconstruction. In *Advances in Neural Information Processing Systems*, 2025a.

Jiahao Ma, Tianyu Wang, Miaomiao Liu, David Ahmedt-Aristizabal, and Chuong Nguyen. Dchm: Depth-consistent human modeling for multiview detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7731–7740. IEEE, 2025b.

Jiahao Ma, Qiang Zhang, Peiran Liu, Zeran Su, Pihai Sun, Gang Han, Wen Zhao, Wei Cui, Zhang Zhang, Zhiyuan Xu, et al. Robotpan: A 360° surround-view robotic vision system for embodied perception. *arXiv preprint arXiv:2604.13476*, 2026.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, and Alex Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023.

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 *CVPR*, 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, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G. Schwing, and Oliver Wang. Neural volumetric object selection. In *CVPR*, 2022.

Johannes L. Schönberger 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, 2016.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *CVPR*, 2024.

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

Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, Xinggang Wang, and Eunbyung Park. Uni3R: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33280–33290, 2026.

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

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

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025c. URL https://arxiv.org/abs/2507.02546.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations*, 2026b.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. OpenGaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *NeurIPS*, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

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

Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. OmniSeg3D: Omniversal 3d segmentation via hierarchical contrastive learning. In *CVPR*, 2024.

Qiang Zhang, Jiahao Ma, Peiran Liu, Shuai Shi, Zeran Su, Zifan Wang, Jingkai Sun, Wei Cui, Jialin Yu, Gang Han, et al. Meshmimic: Geometry-aware humanoid motion learning through 3d scene reconstruction. *arXiv preprint arXiv:2602.15733*, 2026.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024.

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

Zhengyu Zou, Hao Li, Kuixuan Jiao, Liu Liu, Tingyang Xiao, Xiaolin Zhou, Fangzhou Hong, Zhizhong Su, Dingwen Zhang, and Ziwei Liu. Iggt4d: Streaming 4d instance-grounded geometry transformer. *arXiv preprint arXiv:2607.19228*, 2026.

Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. PanSt3R: Multi-view consistent panoptic segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2025.