THE FINAL LAYER IS RARELY BEST: LAYER-WISE PROBING OF VISUAL FOUNDATION MODELS FOR 3D RECONSTRUCTION AND SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual foundation models (VFMs) are routinely reused as frozen encoders, yet features are often taken from the final layer without testing whether it is appropriate for a dense downstream task. We study layer selection across 14 VFMs and five probes spanning static geometry, texture, joint Gaussian reconstruction, dynamic reconstruction, and semantic segmentation. Our evaluation uses a common Gaussian representation: static novel-view synthesis follows Feat2GS, while Feat4DGS adds a deformation readout for dynamic scenes and dedicated linear probes for semantics. Across the resulting 70 model–task comparisons, a non-final layer strictly outperforms the final layer in 64 cases. The preferred depth is nevertheless task- and model-dependent: the earliest sampled layer is best for texture in 10 of 14 models, whereas semantic and dynamic optima are distributed across the network. No pretraining objective dominates every probe; the best geometry, texture, joint reconstruction, dynamic reconstruction, and semantic scores come from five different model families. Separate deformation readouts also improve mean dynamic-scene PSNR by 4.26 dB over a shared MLP. These results show that layer choice is a first-order experimental variable and provide task-specific guidance for using frozen VFMs in dense 3D and semantic applications.

1 INTRODUCTION

Visual foundation models (VFMs) (Bommasani et al., 2021) pretrained on large-scale 2D corpora have become standard frozen encoders for dense prediction and 3D reconstruction (Ranftl et al., 2020; Ke et al., 2024; Bae & Davison, 2024; Chen et al., 2024b; Wang et al., 2024; Charatan et al., 2024; Hong et al., 2024; Ye et al., 2024a). Their empirical success suggests that large-scale image pretraining yields representations useful for both reconstruction and recognition. Existing probes study these capabilities through separate tasks and protocols, including depth estimation (El Banani et al., 2024; Ray et al.), correspondence matching (El Banani et al., 2024; Aydemir et al., 2024), and novel view synthesis (NVS) (Chen et al., 2025c). This fragmentation makes it difficult to compare how pretraining objective and representation depth affect reconstruction quality and semantic abstraction under a common evaluation setting.

What is needed is a unified probing protocol that supports side-by-side evaluation of these properties. Feat2GS (Chen et al., 2025c) showed that, for static scenes, regressing 3D Gaussian Splatting parameters from frozen VFM features provides a scalable, label-free probe of reconstruction quality through multi-view photometric consistency. We build directly on that protocol rather than replacing it. Static-scene experiments use Feat2GS unchanged. Feat4DGS extends the probing protocol only where necessary for dynamic scenes, and adds a dedicated linear classifier for semantic segmentation. This unified setup enables dense evaluation of static and dynamic NVS together with semantic prediction under matched multi-view conditions.

This unified probe makes it possible to study representation depth directly. In vision-language and multimodal learning, intermediate features are often preferred over final-layer outputs because deeper layers may overspecialize to pretraining objectives (Zhang et al., 2016; Zheng et al., 2016; Chen et al., 2020; El-Nouby et al., 2024; Yu et al., 2025; Radford et al., 2021; Sun et al., 2025; Bolya et al., 2025). However, this question has not been systematically studied for dense reconstruction and semantic

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.

prediction under a shared protocol. We therefore use Feat4DGS to conduct, to our knowledge, the first systematic empirical layer-wise study across diverse VFMs and pretraining paradigms. Our analysis centers three questions: how pretraining objectives differ, how representation depth affects downstream performance, and what unified probing reveals about the balance between reconstruction quality and semantic abstraction.

Across 70 model–task comparisons, a non-final layer strictly outperforms Final in 64 cases and ties it in two more. This pattern does not reduce to a universal early-to-late spatial–semantic tradeoff: P_{10} is best for texture in 10 of 14 models, while geometry, dynamics, and semantics peak at model-specific depths. Layer selection can even change which encoder appears strongest on a task. These observations make representation depth a first-order experimental variable rather than a secondary implementation choice.

Our contributions are threefold:

1) Unified probing framework. We present Feat4DGS as a unified probing framework for comparing frozen VFM representations across reconstruction and semantic tasks. Static-scene evaluation directly uses Feat2GS, while our extensions are limited to dynamic-scene probing and semantic readout.

2) Large empirical layer-wise study. We evaluate 14 VFMs at five depths on five dense tasks, yielding 70 matched model–task comparisons across diverse architectures and pretraining objectives.

3) Empirically supported guidance for feature selection. We show that non-final features strictly outperform Final in 64 of 70 comparisons, while the preferred depth remains task- and model-dependent. This provides concrete guidance for layer selection and fair encoder comparison in dense 3D and semantic applications.

Overall, the main contribution of this work is the unified empirical study of representation depth across visual foundation models, with Feat4DGS serving as the probing framework that makes this comparison possible.

Sections 2 to 6 present related work, methodology, experiments, results, and limitations, respectively.

2 Related Work

2.1 Measuring 3D Awareness of VFMs

Visual foundation models (VFMs) (Bommasani et al., 2021) have supported a wide range of 3D vision tasks, including geometric cue estimation (Ranftl et al., 2020; Ke et al., 2024; Hu et al., 2024; Bae & Davison, 2024; Fu et al., 2024a; Ye et al., 2024a; Khirodkar et al., 2024), pose estimation (Örnek et al., 2024), tracking (Tumanyan et al., 2024), and spatial reasoning (Chen et al., 2024b; Feng et al., 2024). Prior work has probed this capability through coarse-grained tasks such as multi-view consistency, spatial reasoning, and robot learning (Bonnen et al., 2024; Zuo et al., 2024; Fu et al., 2024b; Linsley et al., 2024; Majumdar et al., 2023; Shang et al., 2024; Zhu et al., 2024), or through fine-grained proxies such as depth, normals, view tokens, matching, and tracking (El Banani et al., 2024; Ray et al.; Aydemir et al., 2024). These probes are informative, but they remain task-specific and heterogeneous in both supervision and representation. Feat2GS (Chen et al., 2025c) provides a dense, label-free static-scene probe by regressing 3D Gaussian Splats from frozen VFM features under multi-view photometric consistency. Our work builds on that protocol rather than replacing it: static-scene experiments use Feat2GS unchanged, while Feat4DGS extends the probing setup only where needed for dynamic scenes and semantic readout. This makes it possible to compare reconstruction and semantic behavior under a common representation and to study how these properties vary across depth.

2.2 Visual Layer Selection

Recent work in vision-language and multimodal systems often uses frozen vision encoders with lightweight adapters (Li et al., 2023; Liu et al., 2023b;a; 2024; Alayrac et al., 2022), which makes visual layer choice an important design decision. Although the final layer is commonly used by default, prior studies across colorization, generative pretraining, masked reconstruction, and contrastive vision-

language learning show that intermediate or penultimate layers often transfer better than the last layer (Zhang et al., 2016; Zheng et al., 2016; Chen et al., 2020; El-Nouby et al., 2024; Yu et al., 2025; Radford et al., 2021; Sun et al., 2025). A recent CLIP study further identifies shallow, middle, and deep representational blocks and shows that representative intermediate layers can outperform final-layer features (Chen et al., 2025a); related analyses reach similar conclusions for PE-style models (Bolya et al., 2025). These findings motivate our layer-wise study, but prior work does not examine dense reconstruction and semantic prediction together under a shared probing protocol.

2.3 NVS from Casual Images with Priors.

Novel view synthesis (NVS) has progressed rapidly with radiance-field and Gaussian-based methods (Mildenhall et al., 2020; Müller et al., 2022; Kerbl et al., 2023). For sparse or casual captures, prior work adds geometric regularization, depth supervision, pretrained visual features, diffusion priors, or feed-forward modeling (Niemeyer et al., 2022; Yang et al., 2023; Deng et al., 2022; Zhu et al., 2023; Jain et al., 2021; Xu et al., 2022; Warburg* et al., 2023; Wu et al., 2024; Yu et al., 2021; Liu et al., 2023c; Sargent et al., 2024; Tung et al., 2024; Charatan et al., 2024; Chen et al., 2024c; Hong et al., 2024; Chen et al., 2024a; Ziwen et al., 2024; Jin et al., 2024), but many approaches still assume known poses (Schönberger & Frahm, 2016) or rely on fragile joint optimization (Wang et al., 2021; Lin et al., 2021; Chen et al., 2023; Jeong et al., 2021; Meng et al., 2021; Truong et al., 2023; Bian et al., 2023; Fu et al., 2024c; Meuleman et al., 2023). Regression-based methods such as DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) instead predict point maps directly, enabling unconstrained reconstruction and initialization of 3DGS (Kerbl et al., 2023); subsequent work extends this paradigm to videos and larger-scale settings (Wang & Agapito, 2024; Zhang et al., 2024; Chen et al., 2025b; Wang et al., 2025; Lin et al., 2025; Team et al., 2025). Feat2GS (Chen et al., 2025c) uses this setting as a static-scene probing protocol by reading out 3DGS parameters from frozen VFM features with a shallow MLP. Our work keeps that static protocol unchanged and extends the probing setup only where necessary for dynamics and semantic readout, rather than proposing a competing NVS algorithm.

2.4 Foundational Vision Encoders

We evaluate VFMs spanning the main pretraining paradigms used in practice: self-supervised learning (DINOv2, DINOv3, MAE), contrastive vision-language learning (CLIP, PE), hybrid contrastive and masked modeling (TIPS, TIPSv2), geometry-oriented supervision (DUSt3R, MASt3R, MiDaS, VGGT), segmentation-focused training (SAM), generative modeling (DiT), and agglomerative distillation (C-RADIOv3) (Oquab et al., 2023; Siméoni et al., 2025; He et al., 2022; Radford et al., 2021; Bolya et al., 2025; Maninis et al., 2025; Wang et al., 2024; Leroy et al., 2024; Ranftl et al., 2020; Wang et al., 2025; Kirillov et al., 2023; Peebles & Xie, 2023; Ranzinger et al., 2024; Heinrich et al., 2024). Although these encoders are trained for different objectives, they are routinely reused as frozen backbones for dense reconstruction and pixel-level prediction. This diversity makes them a useful testbed for our empirical question: how pretraining objective and representation depth shape reconstruction quality and semantic behavior under a unified probing protocol.

3 Method

Our probing protocol builds directly on Feat2GS (Chen et al., 2025c). Static scenes use its geometry (**G**), texture (**T**), and joint (**A**) modes unchanged. We add only a deformation (**D**) probe for dynamic scenes (Section 3.1) and dedicated linear probes over per-Gaussian features for semantic segmentation (Section 3.2). The resulting framework supports controlled layer-wise comparisons across reconstruction and semantics; it is not a replacement reconstruction algorithm.

3.1 Extending Feat2GS for Dynamic NVS

We show the unified probing setup in Figure 1. For static scenes, we use Feat2GS (Chen et al., 2025c) unchanged. Given frozen VFM features, Feat2GS applies PCA (Halko et al., 2011), bilinearly upsamples the compact features to image resolution, and uses the same shallow readout architecture across models. The readout g_Θ maps each pixel feature $\mathbf{f}_i$, for $i \in \{1, 2, \ldots, n\}$, to a Gaussian

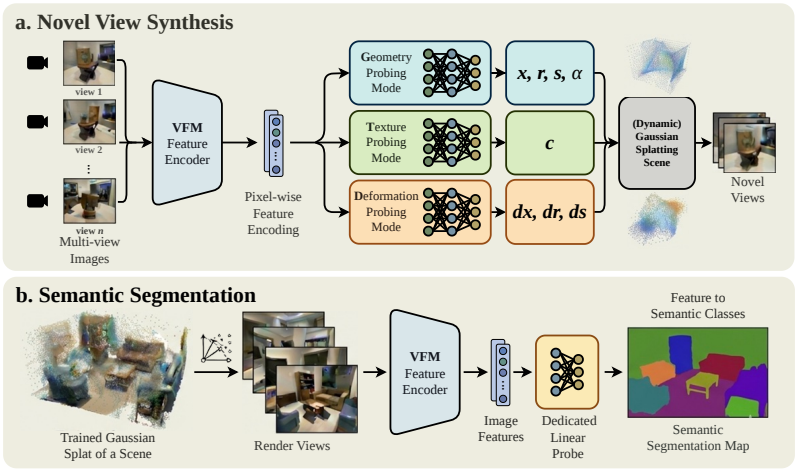


Figure 1: **The Feat4DGS probing framework. (a)** NVS probing, which extends Feat2GS to dynamic scenes through the **D**eformation mode while retaining the static probing modes. **(b)** Semantic probing, where VFM features are projected onto the Gaussian scene representation and a lightweight classifier maps the resulting features to semantic classes. Together, the two probes enable evaluation of spatial reconstruction and semantic representation quality across VFM layers.

primitive:

$$\boldsymbol{G}_i = g_\Theta(\mathbf{f}_i), \tag{1}$$

where each Gaussian $\boldsymbol{G}_i$ is parameterized by position $\boldsymbol{x}_i \in \mathbb{R}^3$, opacity $\alpha_i \in \mathbb{R}$, covariance Σ_i parameterized by scale $\boldsymbol{s}_i$ and rotation $\boldsymbol{r}_i$, and third-order spherical-harmonic coefficients $\boldsymbol{c}_i \in \mathbb{R}^{48}$.

As in Feat2GS, g_Θ is a two-layer MLP with 256 units per layer and a ReLU activation, and the predicted Gaussians are rendered by differentiable rasterization. For casually captured dynamic scenes, Feat4DGS extends this protocol only where static-scene assumptions no longer hold: we initialize camera poses $\boldsymbol{T}$ with Easi3R (Chen et al., 2025b) and jointly update them with the readout parameters using the photometric objective

$$\min_{\Theta, \boldsymbol{T}} \left\| \mathcal{R}(g_\Theta(\mathbf{f}), \boldsymbol{T}) - \mathcal{I} \right\|. \tag{2}$$

Accordingly, Feat4DGS extends the Feat2GS probing protocol to dynamic scenes rather than introducing a separate reconstruction method.

To probe motion representations, we follow the deformation-field formulation of Deformable-GS (Yang et al., 2024). The **D**eformation mode uses separate shallow readouts for position, rotation, and scale offsets conditioned on $t \in [0, 1]$, while directly optimizing the canonical Gaussian parameters $\{\boldsymbol{x}_i, \alpha_i, \boldsymbol{r}_i, \boldsymbol{s}_i, \boldsymbol{c}_i\}$:

$$\delta_{\boldsymbol{x}_i} = g^{(D)}_{\Theta_{\delta_{\boldsymbol{x}}}}(\mathbf{f}_i, t) \tag{3}$$

$$\delta_{\boldsymbol{r}_i} = g^{(D)}_{\Theta_{\delta_{\boldsymbol{r}}}}(\mathbf{f}_i, t) \tag{4}$$

$$\delta_{\boldsymbol{s}_i} = g^{(D)}_{\Theta_{\delta_{\boldsymbol{s}}}}(\mathbf{f}_i, t), \tag{5}$$

where $\Theta = \{\Theta_{\delta_{\boldsymbol{x}}}, \Theta_{\delta_{\boldsymbol{r}}}, \Theta_{\delta_{\boldsymbol{s}}}\}$ are the separate feature readouts.

We use separate readouts because they improve all three reconstruction metrics on every evaluated dynamic scene; the architecture ablation is reported in Figure 3. Training settings are summarized in Section 4.1.5 and detailed in Section A.

3.2 Semantic Gaussians with Arbitrary Visual Encoders

For semantic segmentation, we build on Semantic Gaussians (SeGA) (Guo et al., 2024): per-pixel VFM features are projected onto a pretrained 3D Gaussian representation and subsequently mapped

Table 1: **VFMs for Evaluation.** We evaluate diverse VFMs spanning 2D and 3D pretraining objectives. To reduce confounding from model capacity, we use approximately ViT-L-scale checkpoints where available (∼300M parameters), standardize feature dimensionality to 256 channels via PCA, and use the same probing architecture and evaluation protocol across models.

VFM	Arch.	Channel	Supervision	Dataset
DUSt3R (Wang et al., 2024)	ViT-L/16	1024	Point Regression	3D DUSt3R-Mix
MASt3R (Leroy et al., 2024)	ViT-L/16	1024	Point Regression	3D MASt3R-Mix
MiDaS (Ranftl et al., 2020)	ViT-L/16	1024	Depth Regression	3D MiDaS-Mix
VGGT (Wang et al., 2025)	ViT-L/14	1024	Point and Camera Prediction	3D VGGT-Mix
TIPS (Maninis et al., 2025)	ViT-L/14	1024	Contrastive VLM and MIM	2D WebLI
TIPSv2 (Cao et al., 2026)	ViT-L/14	1024	VLM + MIM + Distillation	2D WebLI-116M
CLIP (Radford et al., 2021)	ViT-L/14	768	Contrastive VLM	2D WIT-400M
PE (Bolya et al., 2025)	ViT-L/14	1024	Contrastive VLM	2D PVD and PE-Mix
DINOv2 (Oquab et al., 2023)	ViT-L/14	1024	Self Distillation	2D LVD-142M
DINOv3 (Siméoni et al., 2025)	ViT-L/16	1024	Self Distillation	2D LVD-1689M
SAM (Kirillov et al., 2023)	ViT-L/16	1024	Segmentation	2D SA-1B
C-RADIOv3 (Ranzinger et al., 2024; Heinrich et al., 2024)	ViT-L/16	1024	Multi-teacher Distillation	2D NV-CC-Img-Text-Dataset
MAE (He et al., 2022)	ViT-L/16	1024	Image Reconstruction	2D ImageNet-1k
DiT (Peebles & Xie, 2023)	DiT-XL/2	1152	Diffusion Generation	2D ImageNet

to semantic classes. To support both text-aligned and non-text-aligned VFMs, we attach a lightweight linear classifier to the projected features.

Dedicated semantic probes. To isolate the semantic information accessible in each representation, we train a linear classifier for every *scene–model–layer* combination. As in NVS, PCA reduces pixel-wise training features to a common dimensionality. For model m, layer ℓ, and scene s, let $\mathbf{f}_i^{(s,m,\ell)} \in \mathbb{R}^d$ denote the PCA-reduced feature at pixel i. We train $h_{s,m,\ell}$ to predict its semantic class:

$$\hat{\mathbf{y}}_i = h_{s,m,\ell}\left(\mathbf{f}_i^{(s,m,\ell)}\right). \tag{6}$$

Each classifier is trained only on its corresponding representation and evaluated on held-out views of the same scene. We hold its architecture, sampling, and optimization fixed across models and layers, so differences measure linear accessibility rather than classifier transfer between incompatible feature spaces.

4 Experiments

Our evaluation asks three questions: **Q1:** Which pretraining objectives transfer best to each probe? **Q2:** How often is the final layer optimal? **Q3:** How stable is layer selection across tasks, scenes, and metrics?

4.1 Experimental Setup

4.1.1 Features

We evaluate 14 VFMs spanning different data types and pretraining objectives, grouped by the supervision categories in Table 1.

4.1.2 Datasets

For NVS, we evaluate six static and six dynamic scenes. The static set comprises Bicycle, Garden, Room, Stump from MipNeRF360 (Barron et al., 2022) and Train, Truck from Tanks and Temples (Knapitsch et al., 2017); the dynamic set comprises Playground from Dynamic Scene (Yoon et al., 2020), Train, Blackswan, Dogs Jump, Drift Straight from DAVIS (Perazzi et al., 2016), and Kid Running from NSFF (Li et al., 2021). These scenes cover indoor and outdoor settings, different scene scales, large camera movement, and rigid, articulated, and non-rigid motion. For semantic segmentation, we evaluate the 12 ScanNet scenes used by Panoptic Lifting (Dai et al., 2017; Siddiqui et al., 2023), which provide calibrated RGB-D trajectories, point clouds, and semantic label maps.

Table 2: **Representation quality across feature depth.** **(a)** Dynamic NVS using the **D**eformation probe, reported as mean PSNR across dynamic scenes. **(b)** Semantic segmentation using independently trained linear probes, reported as mean IoU across 12 ScanNet scenes. Cell colors encode change relative to Final within each model: orange indicates a loss, white parity, and blue a gain. Bold values mark the best evaluated layer.

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	**17.20**	17.09	17.10	17.02	17.13
MASt3R L	17.23	**17.40**	17.01	16.99	16.75
MiDaS L/16	17.04	**17.07**	16.88	16.87	16.95
VGGT 1B	16.90	16.69	17.00	**17.29**	17.26
TIPS L/14	17.02	**17.07**	16.94	16.72	17.07
TIPSv2 L/14	**17.22**	17.02	16.85	16.89	17.01
CLIP L/14	17.02	17.09	**17.28**	17.04	17.05
PE L/14 (336)	16.99	16.94	**17.37**	17.29	16.94
DINOv2 L/14	16.48	16.21	**17.13**	16.80	16.93
DINOv3 L/16	16.49	**17.34**	16.95	16.30	16.08
SAM Large	16.92	**17.51**	17.37	17.31	16.96
C-RADIOv3 L	16.80	**17.11**	16.57	16.51	16.82
MAE L/16	16.63	17.25	17.06	**17.29**	17.03
DiT XL/2	**16.96**	16.87	16.77	15.86	16.16

(a) Dynamic NVS (**D**), PSNR ↑

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	0.538	0.549	0.544	0.550	**0.552**
MASt3R L	**0.541**	0.534	0.526	0.528	0.528
MiDaS L/16	**0.558**	0.522	0.535	0.547	0.548
VGGT 1B	0.530	0.549	0.541	**0.558**	0.553
TIPS L/14	0.532	0.528	**0.534**	0.534	0.530
TIPSv2 L/14	0.520	0.519	0.521	0.524	**0.525**
CLIP L/14	0.533	0.582	0.569	0.555	**0.589**
PE L/14 (336)	0.548	**0.589**	0.583	0.579	0.576
DINOv2 L/14	0.554	0.575	**0.598**	0.595	0.588
DINOv3 L/16	0.541	**0.566**	0.554	0.543	0.537
SAM Large	0.562	**0.576**	0.572	0.575	0.572
C-RADIOv3 L	0.556	**0.585**	0.577	0.552	0.545
MAE L/16	0.536	0.545	**0.546**	0.543	0.543
DiT XL/2	0.539	0.558	0.554	0.564	**0.567**

(b) Semantic Segmentation, mIoU ↑

4.1.3 Metrics

For NVS, we report PSNR as the primary reconstruction metric following Feat2GS (Chen et al., 2025c), together with SSIM and LPIPS. For static scenes, we follow Feat2GS and Splatt3R (Smart et al., 2024) by applying the same validity masks to rendered and test images. For semantic segmentation, we report mean intersection over union (mIoU) over the 12 ScanNet scenes.

4.1.4 Layers

Following Chen et al. (2025a), we sample representative early-, middle-, and late-layer blocks using the 10th (P_{10}), 50th (P_{50}), and 80th (P_{80}) percentiles, together with the penultimate and final layers. This common sampling scheme makes representation depth a primary axis of comparison across model families.

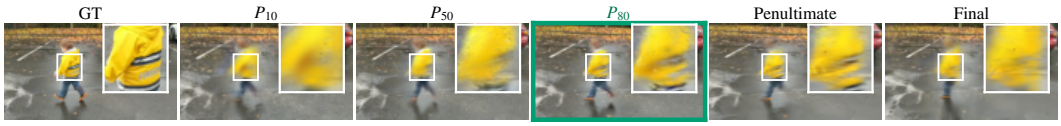
4.1.5 Implementation Details

All probes use a common feature dimensionality and readout budget. Exact preprocessing, optimization schedules, and semantic sampling are provided in Section A, with hardware and runtime information in Section A.4.

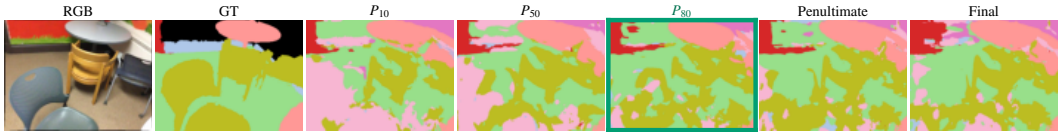
5 Results

5.1 Pretraining Objective Determines the Task Leader

Table 2 shows that no model family dominates all probes. At its best sampled layer, C-RADIOv3 obtains the highest geometry PSNR (19.65 dB), PE the highest texture PSNR (19.26 dB), VGGT the highest joint-reconstruction PSNR (20.47 dB), SAM the highest dynamic PSNR (17.51 dB), and DINOv2 the highest semantic mIoU (0.598); complete static results are in Section B.1. This task-dependent ranking cautions against treating the pretraining objective as a proxy for general-purpose dense representation quality. Geometry-oriented pretraining is competitive, but the dynamic and semantic leaders are segmentation and self-distillation models, respectively.



(a) Dynamic NVS across representation depth using the **D**eformation mode.



(b) Semantic segmentation across representation depth using dedicated linear probes.

Figure 2: **Qualitative comparison across representation depth.** All panels use DINOv3 L/16, fixed before visual inspection. Green outlines mark the best sampled layer for each example. **(a)** On `kid-running`, P_{80} better preserves the jacket boundary and dark stripe than Final. **(b)** On `scene0616_00`, P_{80} yields more coherent regions and cleaner object boundaries. The examples are selected by the representative protocol in Section A.3; they illustrate, but do not determine, the quantitative layer rankings. Additional scenes and views appear in Section C.

5.2 The Final Layer Is Rarely Best

Table 3 compares the best sampled depth with Final for every model and probe. A non-final layer is strictly better in 64 of 70 model–task comparisons and ties Final in two more. The pattern is strongest for static reconstruction: Final is strictly optimal only once across the 42 geometry, texture, and joint-reconstruction comparisons. Texture favors shallow appearance cues—P_{10} is best for 10 of 14 models—whereas the other probes distribute their optima across early, middle, and late encoder blocks. Consequently, neither Final nor a single intermediate percentile is a reliable universal default.

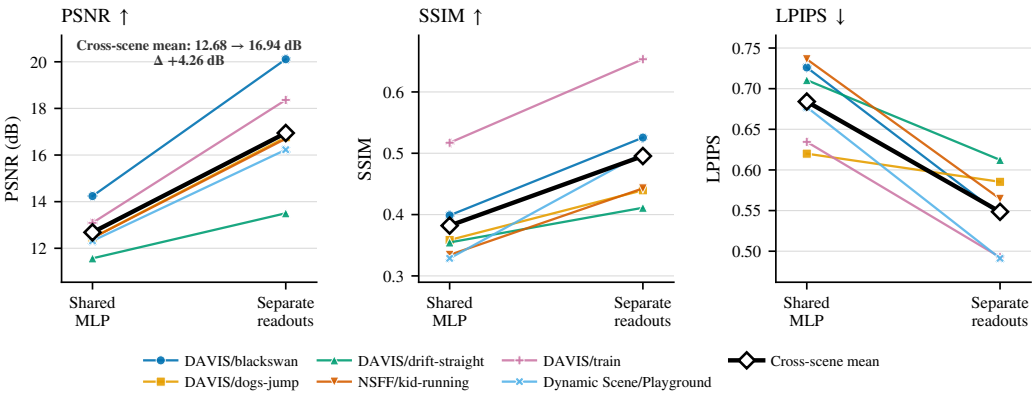
Dynamic reconstruction shows the same dependence on depth (Table 2a). Final is never strictly best: the largest gains are 1.26 dB for DINOv3, 0.80 dB for DiT, and 0.65 dB for MASt3R. The best layer also changes across scenes, with gains over Final reaching 3.29 dB on individual sequences (Section B.2). Moreover, the layer preferred by PSNR can differ from that preferred by SSIM or LPIPS (Section B.3), so selection should follow the target metric rather than a fixed convention. Section B collects the complete NVS analyses.

Table 2b shows smaller but systematic semantic gains: 10 of 14 models improve strictly over Final, and another ties it. C-RADIOv3 shows the largest increase, from 0.545 to 0.585 mIoU at P_{50}, while DINOv2 reaches the highest absolute mIoU, 0.598, at P_{80}. These results do not support a universal monotonic progression from spatial to semantic features; instead, semantic accessibility can peak at different depths for different objectives.

5.3 Qualitative and Architectural Evidence

The representative examples in Figure 2 agree with the quantitative trends. In dynamic NVS, Figure 2a shows that P_{80} preserves the jacket boundary and dark stripe more faithfully than Final. In semantic segmentation, Figure 2b shows that P_{80} produces more coherent regions and cleaner foreground boundaries than the other sampled depths. These examples are illustrative rather than the basis of the ranking; Section A.3 details the selection protocol, and Section C provides additional views for every evaluated scene.

Finally, Figure 3 shows that the dynamic result depends on a probe with enough structure to expose motion information. Quantitatively, Figure 3a shows that replacing one shared deformation MLP with separate position, rotation, and scale readouts improves every evaluated scene on PSNR, SSIM, and LPIPS, raising mean PSNR from 12.68 to 16.94 dB (+4.26 dB). Figure 3b shows the corresponding reduction in foreground distortion. We therefore use separate readouts throughout.



(a) Quantitative comparison of deformation readout architectures.



(b) Qualitative comparison of deformation readout architectures.

Figure 3: **Ablation of the deformation probe architecture. (a)** Separate position, rotation, and scale readouts improve every evaluated scene on PSNR, SSIM, and LPIPS relative to a shared deformation MLP; mean PSNR rises from 12.68 to 16.94 dB. **(b)** The separate readouts reduce severe foreground distortion in the Playground scene.

Table 3: **Task-specific layer selection.** Each cell reports the best evaluated layer and its improvement over Final (PSNR for G/T/A/D; mIoU for Semantic). Darker blue badges indicate larger within-task gains.

Model	G	T	A	D	Semantic
DUSt3R L	P_{50} (+0.58)	P_{10} (+0.30)	P_{80} (+0.47)	P_{10} (+0.07)	Final (+0.000)
MASt3R L	Penult. (+0.13)	P_{10} (+0.21)	Final (+0.00)	P_{50} (+0.65)	P_{10} (+0.014)
MiDaS L/16	P_{80} (+0.14)	P_{10} (+0.17)	P_{50} (+0.23)	P_{50} (+0.12)	P_{10} (+0.010)
VGGT 1B	P_{80} (+0.86)	P_{50} (+0.06)	P_{80} (+0.24)	Penult. (+0.03)	Penult. (+0.005)
TIPS L/14	P_{80} (+0.41)	P_{10} (+0.06)	P_{50} (+1.01)	P_{50}/ Final (+0.00)	P_{80} (+0.004)
TIPSv2 L/14	P_{50} (+0.35)	P_{10} (+0.12)	P_{50} (+0.17)	P_{10} (+0.21)	Penult. / Final (+0.000)
CLIP L/14	P_{80} (+0.09)	P_{80} (+0.04)	P_{50} (+0.29)	P_{80} (+0.23)	Final (+0.000)
PE L/14 (336)	Penult. (+0.05)	Penult. (+0.51)	P_{80} (+0.31)	P_{80} (+0.43)	P_{50} (+0.013)
DINOv2 L/14	P_{10} (+1.08)	P_{10} (+0.16)	P_{50} (+0.48)	P_{80} (+0.20)	P_{80} (+0.009)
DINOv3 L/16	P_{10} (+0.38)	P_{10} (+0.30)	P_{50} (+1.53)	P_{50} (+1.26)	P_{50} (+0.029)
SAM Large	P_{50} (+0.39)	P_{10} (+0.09)	P_{50} (+0.15)	P_{50} (+0.55)	P_{50} (+0.004)
C-RADIOv3 L	P_{50} (+0.16)	P_{80} (+0.20)	P_{50} (+0.21)	P_{50} (+0.29)	P_{50} (+0.040)
MAE L/16	P_{80} (+0.43)	P_{10} (+0.35)	Penult. (+0.51)	Penult. (+0.27)	P_{80} (+0.003)
DiT XL/2	P_{10} (+0.54)	P_{10} (+0.17)	P_{10} (+1.96)	P_{10} (+0.80)	Final (+0.000)

5.4 Task-Specific Layer-Selection Takeaways

Layer selection changes not only absolute scores but also which encoder appears strongest. Using only Final, VGGT would lead dynamic NVS at 17.26 dB; after selecting each model's best sampled

depth, SAM instead leads at 17.51 dB, while VGGT reaches 17.29 dB. Similarly, Final favors CLIP over DINOv2 for semantics (0.589 versus 0.588 mIoU), whereas DINOv2 becomes the overall leader at 0.598 mIoU with P_{80}. Because the gains from layer selection vary substantially by model, fixed-depth evaluation can confound encoder quality with the compatibility of one output layer and the downstream probe. Fair comparisons should therefore tune depth per model on validation data or report layer-wise results under the same selection budget.

6 Limitations and Future Work

Our empirical conclusions are bounded by the evaluated models, tasks, scenes, and representation depths. NVS covers a small set of static and dynamic scenes, and the sampled layers do not constitute a dense sweep or evaluate multi-layer fusion. The observed trends therefore provide empirical guidance within this benchmark rather than universal or causal conclusions.

The probing setup also inherits assumptions from its underlying protocols. Static NVS uses Feat2GS Chen et al. (2025c) unchanged and thus relies on camera poses and point clouds initialized by unconstrained stereo reconstruction (Leroy et al., 2024; Wang & Agapito, 2024; Zhang et al., 2024; Chen et al., 2025b); the dynamic extension retains analogous initialization and assumes limited temporal and illumination variation. Future studies could examine VFM-based pose and geometry initialization (Örnek et al., 2024; Brachmann et al., 2024) and unconstrained capture settings (Snavely et al., 2006; Tung et al., 2024; Martin-Brualla et al., 2021; Chen et al., 2022; Xu et al., 2024; Zhang et al., 2025; Kulhanek et al., 2024). Finally, the semantic classifier is trained per scene–model–layer to isolate representation quality, so cross-scene generalization remains outside the present evaluation.

7 Conclusion

We presented a layer-wise study of 14 frozen VFMs across static geometry, texture, joint reconstruction, dynamic reconstruction, and semantic segmentation. A common Gaussian representation enables controlled comparison: static NVS uses Feat2GS unchanged, while Feat4DGS adds separate deformation readouts for dynamic scenes and dedicated linear probes for semantics.

The principal result is simple: Final is strictly suboptimal in 64 of 70 model–task comparisons. Yet there is no universal replacement. P_{10} dominates texture readout, while geometry, joint reconstruction, dynamics, and semantics favor task- and model-specific depths. The absolute leader also changes for every probe, and dynamic-scene and metric-level analyses reveal additional variation hidden by aggregate means.

These findings make layer selection a first-order design and reporting choice for frozen VFMs. Practitioners should tune depth on the target dense task rather than default to the final representation, and comparisons between encoders should control layer choice explicitly. The conclusions remain specific to the evaluated models and scenes, but the protocol provides a reproducible basis for extending the study to broader encoders, tasks, and capture conditions.

AI use statement

In this work, we used generative AI tools for implementing methods, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, and cleaning and reformatting datasets. Additionally, we used generative AI tools for managing experiments, creating or modifying scientific figures or images, creating or editing software code, creation of artifacts, drafting parts of a research paper, editing a research paper to improve readability, and proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. We thoroughly critiqued and iterated on figures, writing, and code to better understand and validate everything. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf. 2

Görkay Aydemir, Weidi Xie, and Fatma Güney. Can visual foundation models achieve long-term point tracking? *arXiv preprint arXiv:2408.13575*, 2024. 1, 2

Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022. 5

Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4160–4169, 2023. 3

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=INqBOmwIpG. 1, 3, 5

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 2

Tyler Bonnen, Stephanie Fu, Yutong Bai, Thomas O'Connell, Yoni Friedman, Nancy Kanwisher, Joshua B. Tenenbaum, and Alexei A. Efros. Evaluating multiview object consistency in humans and image models, 2024. URL https://arxiv.org/abs/2409.05862. 2

Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 9

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29325–29335, June 2026. 5

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 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*, pp. 19457–19467, 2024. 1, 3

Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision (ECCV)*, 2024a. 3

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024b. 1, 2

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6677–6695, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.339. URL https://aclanthology.org/2025.emnlp-main.339/. 3, 6

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20s.html. 1, 3

Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12952, 2022. 9

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9158–9168, October 2025b. 3, 4, 9

Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8264–8273, 2023. 3

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6348–6361, June 2025c. 1, 2, 3, 6, 9, 18

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *ECCV*, 2024c. 3

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12882–12891, 2022. 3

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024. 1, 2

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Ángel Bautista, Vaishaal Shankar, Alexander T Toshev, Joshua M. Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12371–12384, 2024. 1, 3

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2093–2103, 2024. 2

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024a. 2

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. 2024b. 2

Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20796–20805, June 2024c. 3

Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting, 2024. URL https://arxiv.org/abs/2403.15624. 4

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288, 2011. 3

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022. 3, 5

Greg Heinrich, Mike Ranzinger, Hongxu, Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2.5: Improved baselines for agglomerative vision foundation models, 2024. URL https://arxiv.org/abs/2412.07679. 3, 5

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *ICLR*, 2024. 1, 3

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 2

Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5885–5894, October 2021. 3

Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5846–5854, 2021. 3

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias, 2024. URL https://arxiv.org/abs/2410.17242. 3

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024. 1, 2

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/. 3

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024. 2

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980. 18

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. 3, 5

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 5

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024. 9

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 3, 5, 9

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html. 2

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5741–5751, 2021. 3

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025. URL https://arxiv.org/abs/2511.10647. 3

Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. *arXiv preprint arXiv:2406.04138*, 2024. 2

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a. 2

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf. 2

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/. 2

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023c. 3

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36: 655–677, 2023. 2

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. Tips: Text-image pretraining with spatial awareness, 2025. URL https://arxiv.org/abs/2410.16512. 3, 5

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7210–7219, 2021. 9

Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. GNeRF: GAN-based Neural Radiance Field without Posed Camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16539–16548, 2023. 3

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127. 3

Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5480–5490, 2022. 3

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 5

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. Foundpose: Unseen object pose estimation with foundation features. *ECCV*, 2024. 2, 9

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 18

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387. 3, 5

F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 5

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 1, 3, 5

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1, 2, 3, 5

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12490–12500, June 2024. 3, 5

Arijit Ray, Dina Bashkirova, Reuben Tan, Kuo-Hao Zeng, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. R2d3: Imparting spatial reasoning by reconstructing 3d scenes from 2d images. 1, 2

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9420–9429, 2024. 3

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ylZHvlwUcI. 2

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9043–9052, 2023. 5, 18

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, 2025. URL https://arxiv.org/abs/2508.10104. 3, 5

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 6

Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pp. 835–846. 2006. 9

Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23199–23209, October 2025. 1, 3

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. 2025. URL https://arxiv.org/abs/2511.16624. 3

Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4190–4200, 2023. 3

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision (ECCV)*, 2024. 2

Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 3, 9

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3, 9

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 (CVPR)*, pp. 5294–5306, June 2025. 3, 5

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, June 2024. 1, 3, 5

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3

Frederik Warburg*, Ethan Weber*, Matthew Tancik, Aleksander Hołyński, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *International Conference on Computer Vision (ICCV)*, 2023. 3

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21551–21561, 2024. 3

Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections. *arXiv preprint arXiv:2407.12306*, 2024. 9

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. 2022. 3

Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8254–8263, 2023. 3

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20331–20341, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.01922. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.01922. 4, 18

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024a. 1, 2

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024b. URL https://arxiv.org/abs/2409.06765. 18

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 3

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. 1, 3

Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision*, pp. 341–359. Springer, 2025. 9

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 3, 9

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1, 3

Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good practice in cnn feature transfer, 2016. URL https://arxiv.org/abs/1604.00133. 1, 3

Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. Spa: 3d spatial-awareness enables effective embodied representation. *arXiv preprint arXiv:2410.08208*, 2024. 2

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting, 2023. 3

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint 2410.12781*, 2024. 3

Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? *arXiv preprint arXiv:2410.10799*, 2024. 2

A Implementation and Evaluation Details

A.1 Probe Configuration

We implement all probes in PyTorch (Paszke et al., 2019) and gsplat (Ye et al., 2024b). Images are resized to 512 pixels on the longer side. VFM features are reduced to 256 channels with PCA and bilinearly upsampled to image resolution. The static readout is a two-layer ReLU MLP with 256-dimensional hidden units, optimized with Adam (Kingma & Ba, 2015) together with the Gaussians and cameras. Following Feat2GS (Chen et al., 2025c), a 1K-iteration warm start decays the readout learning rate from 10^{-2} to 10^{-4}, followed by 7K optimization iterations with the same decay ratio.

The **D**eformation mode follows the Deformable-GS (Yang et al., 2024) schedule: a 3K-iteration deformation warmup precedes 21K total iterations, and the deformation-readout learning rate decays from 8×10^{-4} to 1.6×10^{-6}. Adam uses $\beta = (0.9, 0.999)$. Separate readouts predict position, rotation, and scale offsets; Figure 3 validates this choice.

A.2 Semantic Probe Training

For each scene–model–layer probe, training views are ranked by their summed inverse class frequency. We collect at most 5K pixels per class per view and 100K pixels per class overall. A linear classifier is trained for 50 epochs with cross-entropy, Adam with learning rate 10^{-3} and weight decay 10^{-4}, and a batch size of 4096. The same architecture, sampling limits, and optimization budget are used for every representation.

Table 1 lists the evaluated VFMs and pretraining objectives.

A.3 Qualitative Example Selection

We fixed DINOv3 L/16 as the shared qualitative model before inspecting rendered or predicted images. For dynamic NVS, we recomputed per-view foreground PSNR directly from the raw render and ground-truth image, evaluating only pixels with `mask == 255`. For each task, we measured each view's improvement from Final to its best sampled layer. We summarized each scene by the median improvement over its views, ranked scenes by absolute distance to the median across scene medians, and ranked views by absolute distance to their own scene median.

The main dynamic NVS panel uses the closest-ranked scene and view. Two semantic scenes tied exactly in distance to the cross-scene median. We first fixed each tied scene's closest-to-median view, then selected the main panel as the one with the larger visible region in which the best intermediate layer was correct and Final was wrong. The appendix shows the three closest-to-median views for every NVS scene and the two closest-to-median views for every semantic scene. These are representative examples, not maximum-gain cases.

A.4 Compute Resources

All experiments use one NVIDIA A100/H100 80GB GPU and 117 GB of RAM.

For NVS, one experiment is a model, layer, probe, and scene combination. A static-scene run takes approximately 10–15 minutes; a dynamic **D**eformation mode run takes approximately 45–70 minutes. For semantics, one experiment evaluates one model and layer over all 12 ScanNet scenes specified by Siddiqui et al. (2023). Training the first classifier for each scene and running inference takes approximately 4–5 hours in total; after scene assets are cached, subsequent model–layer evaluations take approximately 50 minutes.

B Additional NVS Results

B.1 Probe Modes

Table 4, comprising Tables 4a to 4c, reports complete layer-wise static NVS results for the geometry (**G**), texture (**T**), and joint (**A**) Feat2GS probes. The task leaders differ: C-RADIOv3 leads G, PE

Table 4: **Representation quality across feature depth for static NVS.** Mean PSNR across the six canonical static scenes for the G, T, and A probes. Cell colors encode change relative to the Final layer within each model: orange indicates a loss, white indicates parity, and blue indicates a gain. Bold values indicate the best evaluated layer.

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	18.98	**19.41**	19.18	19.03	18.83
MASt3R L	19.30	18.72	19.18	**19.43**	19.30
MiDaS L/16	19.28	19.33	**19.49**	19.31	19.36
VGGT 1B	19.46	19.31	**19.49**	19.39	18.63
TIPS L/14	18.56	18.86	**18.99**	18.43	18.57
TIPSv2 L/14	19.23	**19.35**	19.13	19.13	19.00
CLIP L/14	19.19	19.32	**19.50**	19.27	19.41
PE L/14 (336)	18.10	18.40	19.20	**19.22**	19.16
DINOv2 L/14	**19.61**	19.59	19.45	18.44	18.53
DINOv3 L/16	**18.97**	18.93	18.72	17.78	18.59
SAM Large	19.22	**19.35**	19.30	19.19	18.95
C-RADIOv3 L	19.45	**19.65**	19.53	19.44	19.49
MAE L/16	18.49	18.17	**19.42**	18.52	18.99
DiT XL/2	**19.18**	19.17	18.81	18.77	18.64

(a) Geometry (G), PSNR ↑

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	**18.92**	18.87	18.76	18.73	18.63
MASt3R L	**18.94**	18.88	18.84	18.57	18.73
MiDaS L/16	**18.92**	18.86	18.87	18.82	18.75
VGGT 1B	18.89	**18.94**	18.90	18.84	18.88
TIPS L/14	**18.93**	18.81	18.88	18.77	18.87
TIPSv2 L/14	**18.97**	18.92	18.95	18.87	18.85
CLIP L/14	18.90	18.99	**19.01**	18.98	18.98
PE L/14 (336)	18.72	18.90	18.84	**19.26**	18.75
DINOv2 L/14	**19.02**	18.96	18.97	18.99	18.86
DINOv3 L/16	**18.93**	18.79	18.57	18.68	18.63
SAM Large	**18.95**	18.88	18.89	18.94	18.86
C-RADIOv3 L	18.95	18.92	**19.14**	18.70	18.94
MAE L/16	**18.86**	18.62	18.73	18.65	18.51
DiT XL/2	**19.00**	18.93	18.83	18.75	18.84

(b) Texture (T), PSNR ↑

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	19.00	18.59	**19.74**	19.13	19.27
MASt3R L	19.49	19.34	19.43	19.50	**19.87**
MiDaS L/16	19.71	**20.19**	19.99	19.89	19.97
VGGT 1B	20.24	20.45	**20.47**	20.28	20.23
TIPS L/14	18.63	**19.72**	19.27	18.72	18.71
TIPSv2 L/14	19.66	**20.01**	19.86	18.38	19.83
CLIP L/14	18.62	**19.92**	19.70	19.71	19.63
PE L/14 (336)	19.39	19.75	**19.93**	19.69	19.62
DINOv2 L/14	19.85	**20.25**	17.30	19.96	19.77
DINOv3 L/16	19.37	**19.90**	18.34	17.92	18.37
SAM Large	19.69	**19.99**	19.82	19.63	19.84
C-RADIOv3 L	19.91	**20.30**	20.01	19.95	20.09
MAE L/16	19.26	18.83	19.44	**19.64**	19.13
DiT XL/2	**19.97**	19.58	19.11	17.65	18.01

(c) All (A), PSNR ↑

Table 5: **Per-scene dynamic NVS layer selection.** For each model and dynamic scene, we report the best evaluated layer and its PSNR improvement over Final in parentheses. Darker blue badges indicate larger gains within each scene. Scene-level results expose the variation underlying the aggregate dynamic-NVS means.

Model	**Playground**	**Train**	**Kid Running**	**Dogs Jump**	**Drift Straight**	**Blackswan**
DUSt3R L	P_{50} (+0.35)	P_{50} (+0.15)	P_{80} (+0.14)	P_{10} (+0.78)	P_{50} (+0.21)	P_{80} (+0.78)
MASt3R L	P_{10} (+0.55)	P_{10}/ Penult. (+0.11)	P_{50} (+0.27)	P_{50} (+2.70)	Penult. (+0.50)	Penult. (+0.65)
MiDaS L/16	P_{50} (+1.09)	P_{80} (+0.54)	P_{10} (+0.17)	Final (+0.00)	P_{10} (+0.43)	P_{50} (+0.27)
VGGT 1B	Penult. (+0.42)	Penult. (+0.15)	Penult. (+0.43)	Final (+0.00)	P_{80} (+0.04)	P_{10} (+1.05)
TIPS L/14	P_{10} (+0.74)	Final (+0.00)	P_{10} (+0.56)	Final (+0.00)	P_{10} (+0.27)	P_{50} (+0.93)
TIPSv2 L/14	P_{50} (+0.76)	P_{10} (+0.11)	P_{10} (+0.23)	P_{10} (+0.71)	P_{50} (+0.07)	P_{50} (+0.49)
CLIP L/14	Final (+0.00)	P_{50} (+0.20)	P_{80} (+0.35)	Penult. (+1.77)	Penult. (+0.30)	P_{50} (+1.38)
PE L/14 (336)	Penult. (+0.52)	P_{80} (+1.07)	Final (+0.00)	P_{10} (+0.62)	P_{80} (+0.59)	Penult. (+0.67)
DINOv2 L/14	Penult. (+0.05)	P_{80} (+0.04)	P_{80} (+0.29)	P_{80} (+0.52)	P_{80} (+0.08)	Penult. (+0.66)
DINOv3 L/16	P_{80} (+1.27)	P_{10} (+0.72)	P_{50} (+1.58)	P_{50} (+0.93)	P_{50} (+0.70)	P_{50} (+2.73)
SAM Large	P_{50} (+0.42)	Final (+0.00)	P_{50} (+0.13)	P_{80} (+2.60)	P_{50} (+0.33)	Penult. (+0.41)
C-RADIOv3 L	P_{80} (+0.39)	Final (+0.00)	P_{50} (+0.08)	P_{50} (+3.29)	Final (+0.00)	Final (+0.00)
MAE L/16	P_{50} (+0.56)	Penult. (+0.15)	Penult. (+0.44)	Penult. (+0.58)	P_{80} (+0.28)	P_{50} (+0.68)
DiT XL/2	P_{10} (+0.88)	P_{50} (+2.01)	P_{50} (+0.79)	P_{80} (+0.92)	P_{10} (+0.20)	P_{10} (+1.87)

leads T, and VGGT leads A. Layer preference also changes by probe. Most notably, P_{10} is best for T in 10 of 14 models, while G and A more often favor middle layers.

B.2 Scene-Level Analysis

Table 5 decomposes dynamic PSNR by scene. Aggregate layer choices mask substantial variation: for example, the best non-final layer improves Dogs Jump by 3.29 dB for C-RADIOv3, 2.70 dB for MASt3R, and 2.60 dB for SAM, while the same models prefer different depths on other sequences.

B.3 Additional Reconstruction Metrics

Table 6, with SSIM and LPIPS detailed in Tables 6a and 6b, confirms that the main PSNR conclusion is not metric-specific: non-final layers frequently improve both metrics. However, the optimum can change with the metric. For DUSt3R, for example, P_{10} maximizes PSNR, P_{50} maximizes SSIM, and P_{80} minimizes LPIPS. Layer selection should therefore be aligned with the intended quality criterion.

Table 6: **Dynamic NVS quality across feature depth.** Mean SSIM and LPIPS across the six canonical dynamic scenes for the Deformation probe. Cell colors encode performance change relative to the Final layer within each model: orange indicates worse performance, white indicates parity, and blue indicates better performance. Bold values indicate the best evaluated layer; higher SSIM and lower LPIPS are better.

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	0.501	**0.501**	0.501	0.492	0.499
MASt3R L	0.500	**0.505**	0.490	0.498	0.494
MiDaS L/16	**0.498**	0.495	0.494	0.486	0.495
VGGT 1B	0.498	0.490	0.496	**0.503**	0.499
TIPS L/14	0.495	**0.499**	0.495	0.494	0.499
TIPSv2 L/14	0.500	0.499	0.490	0.499	**0.501**
CLIP L/14	0.497	0.490	0.497	0.496	**0.499**
PE L/14 (336)	0.499	0.494	**0.509**	0.501	0.498
DINOv2 L/14	0.483	0.477	**0.496**	0.493	0.494
DINOv3 L/16	0.486	**0.503**	0.502	0.490	0.486
SAM Large	0.492	0.506	0.504	**0.508**	0.495
C-RADIOv3 L	0.486	**0.494**	0.486	0.487	0.492
MAE L/16	0.488	**0.506**	0.501	0.500	0.494
DiT XL/2	**0.502**	0.491	0.493	0.480	0.483

(a) SSIM ↑

Model	P_{10}	P_{50}	P_{80}	Penultimate	Final
DUSt3R L	0.542	0.542	**0.536**	0.545	0.544
MASt3R L	0.542	**0.541**	0.550	0.545	0.541
MiDaS L/16	**0.539**	0.544	0.542	0.547	0.543
VGGT 1B	0.547	0.549	0.544	**0.543**	0.548
TIPS L/14	**0.542**	0.545	0.553	0.561	0.562
TIPSv2 L/14	0.546	**0.541**	0.554	0.556	0.558
CLIP L/14	0.551	**0.543**	0.546	0.544	0.544
PE L/14 (336)	0.546	0.544	0.539	**0.538**	0.547
DINOv2 L/14	0.560	0.572	0.544	**0.544**	0.546
DINOv3 L/16	0.548	**0.540**	0.550	0.577	0.578
SAM Large	0.546	**0.538**	0.544	0.542	0.541
C-RADIOv3 L	0.562	0.548	0.546	0.544	**0.540**
MAE L/16	0.550	**0.540**	0.544	0.541	0.545
DiT XL/2	0.548	**0.547**	0.574	0.593	0.569

(b) LPIPS ↓

C Additional Qualitative Results

Section A.3 describes how these examples were selected. All panels use the same DINOv3 L/16 features as Figure 2. Rather than showing maximum-gain cases, they show the views nearest each scene's median best-vs-Final improvement.

C.1 Dynamic Novel-View Synthesis

The following panels show the three closest-to-median views for each dynamic NVS scene: Figures 4 to 9. Foreground PSNR is evaluated on pixels with `mask == 255`.

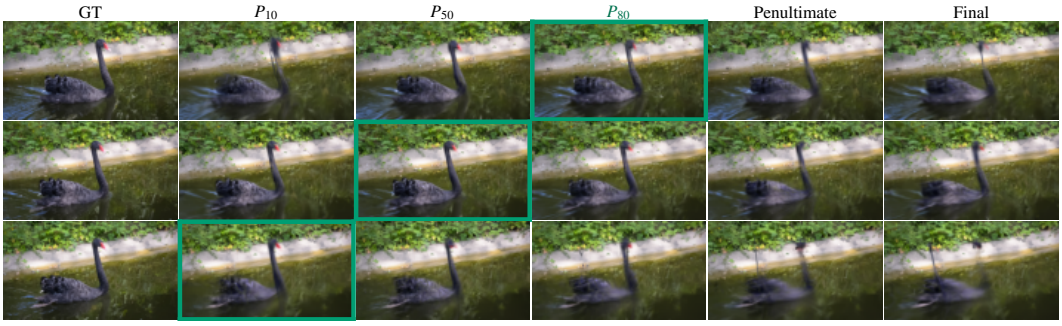


Figure 4: Representation-depth comparison for `DAVIS/blackswan`. An articulated swan moves against reflective water and foliage. Thin neck boundaries, changing reflections, and non-rigid motion stress both geometry and appearance reconstruction. The three rows are the views closest to the scene-median best-vs-Final foreground PSNR improvement. The best observed layers are P_{80}, P_{50}, and P_{10}; green borders mark the best layer in each row.

C.2 Semantic Segmentation

The following panels show the two closest-to-median views for each ScanNet scene: Figures 10 to 21.

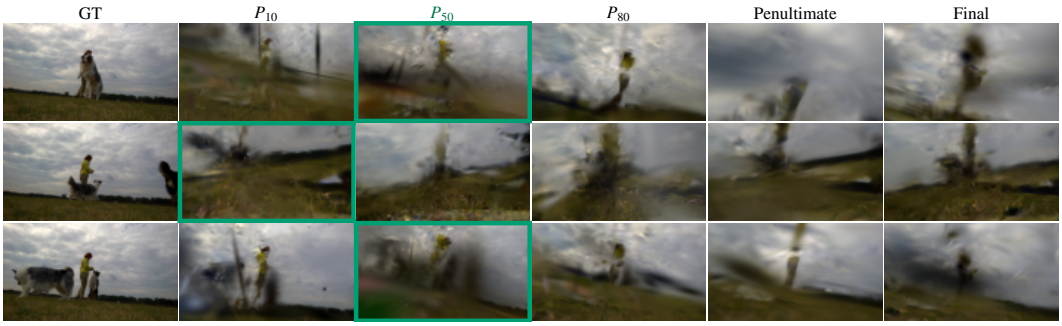


Figure 5: Representation-depth comparison for `DAVIS/dogs-jump`. Multiple non-rigid subjects move independently and occlude one another, making deformation and appearance reconstruction difficult. The full frame is retained without an inset to preserve the global multi-object motion. The three rows are the views closest to the scene-median best-vs-Final foreground PSNR improvement. The best observed layers are P_{50} and P_{10}; green borders mark the best layer in each row.

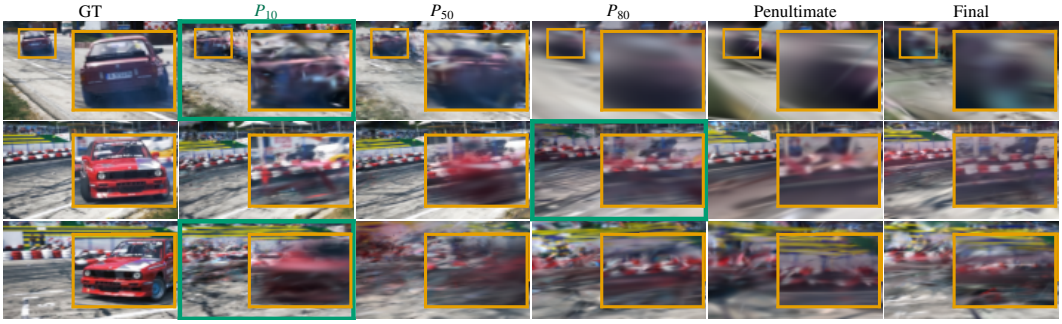


Figure 6: Representation-depth comparison for `DAVIS/drift-straight`. A small, fast car is observed under very large camera motion. A tight orange source box and a substantially enlarged inset expose local car shape and texture differences that are difficult to see globally. The three rows are the views closest to the scene-median best-vs-Final foreground PSNR improvement. The best observed layers are P_{10} and P_{80}; green borders mark the best layer in each row.

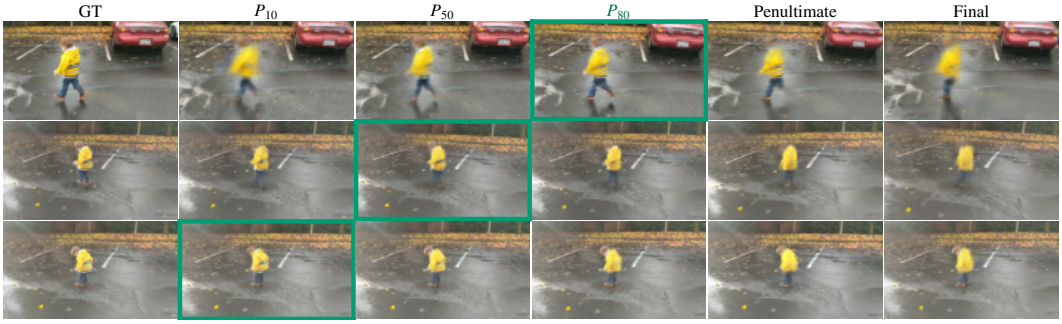


Figure 7: Representation-depth comparison for `DAVIS/kid-running`. A single articulated child crosses wet pavement. The foreground is large enough to inspect directly, so no inset is used; the comparison shows how depth changes jacket, limb, and silhouette fidelity. The three rows are the views closest to the scene-median best-vs-Final foreground PSNR improvement. The best observed layers are P_{80}, P_{50}, and P_{10}; green borders mark the best layer in each row.

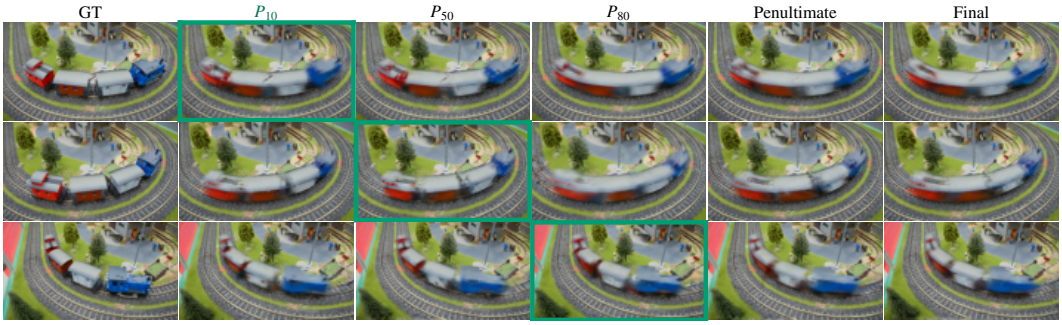


Figure 8: Representation-depth comparison for `DAVIS/train`. A fast rigid train moves through a detailed miniature environment. Motion blur, repeated track geometry, and fine scene structure make appearance reconstruction challenging. The three rows are the views closest to the scene-median best-vs-Final foreground PSNR improvement. The best observed layers are P_{10}, P_{50}, and P_{80}; green borders mark the best layer in each row.

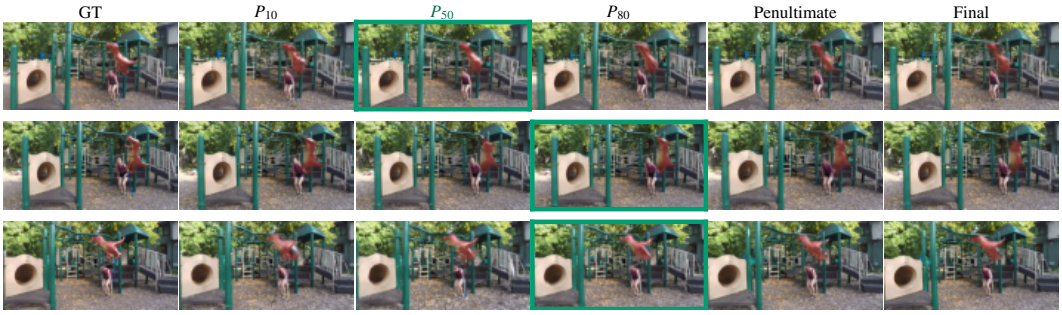


Figure 9: Representation-depth comparison for `Dynamic_Scene/Playground`. A large articulated person moves through cluttered playground structures. Because the moving foreground already occupies a large part of the image, a full-frame comparison is more informative than an inset. The three rows are the views closest to the scene-median best-vs-Final foreground PSNR improvement. The best observed layers are P_{50} and P_{80}; green borders mark the best layer in each row.

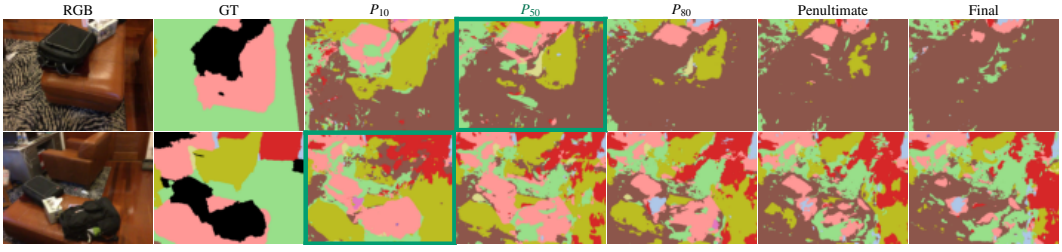


Figure 10: Semantic depth comparison for ScanNet scene `scene0050_02`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by floor, table, chair, and door. The best observed layers are P_{50} and P_{10}. Relative to Final, the winning intermediate layers most often correct floor, table, and door pixels. Green borders mark the best layer in each row.

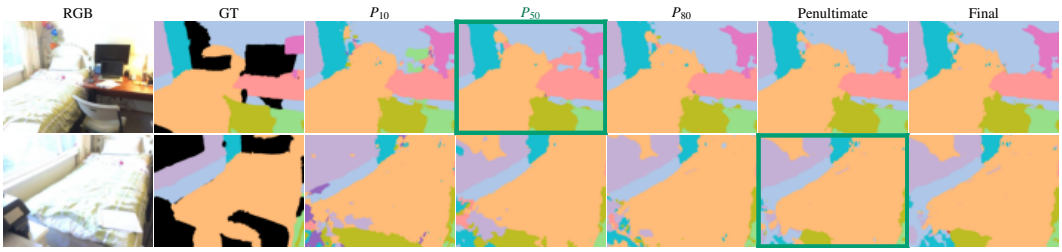


Figure 11: Semantic depth comparison for ScanNet scene `scene0144_01`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by bed, wall, window, and table. The best observed layers are P_{50} and Penultimate. Relative to Final, the winning intermediate layers most often correct bed, wall, and floor pixels. Green borders mark the best layer in each row.

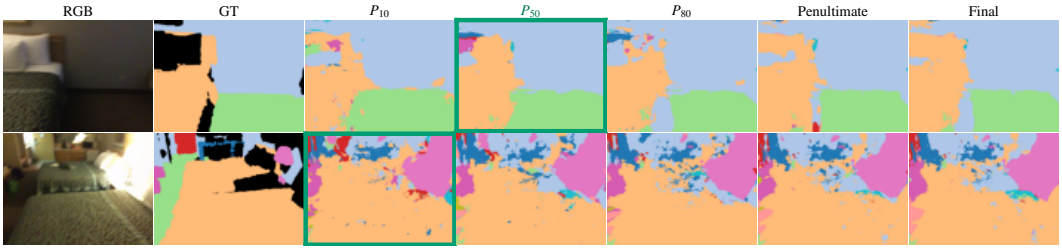


Figure 12: Semantic depth comparison for ScanNet scene `scene0221_01`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by bed, wall, floor, and door. The best observed layers are P_{50} and P_{10}. Relative to Final, the winning intermediate layers most often correct bed, floor, and wall pixels. Green borders mark the best layer in each row.

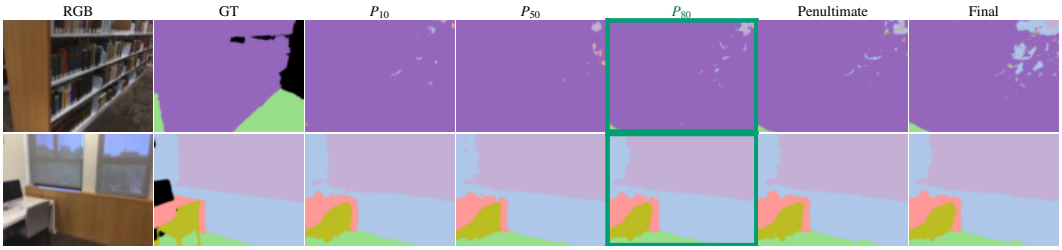


Figure 13: Semantic depth comparison for ScanNet scene `scene0300_01`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by shelves, window, wall, and floor. The best observed layers are P_{80}. Relative to Final, the winning intermediate layers most often correct shelves, wall, and table pixels. Green borders mark the best layer in each row.

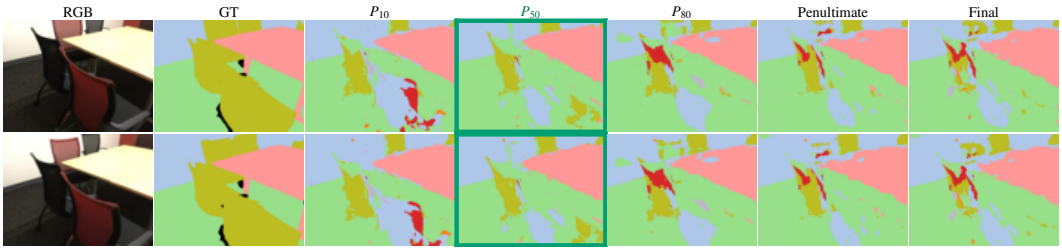


Figure 14: Semantic depth comparison for ScanNet scene `scene0354_00`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by chair, floor, table, and wall. The best observed layers are P_{50}. Relative to Final, the winning intermediate layers most often correct chair, table, and floor pixels. Green borders mark the best layer in each row.

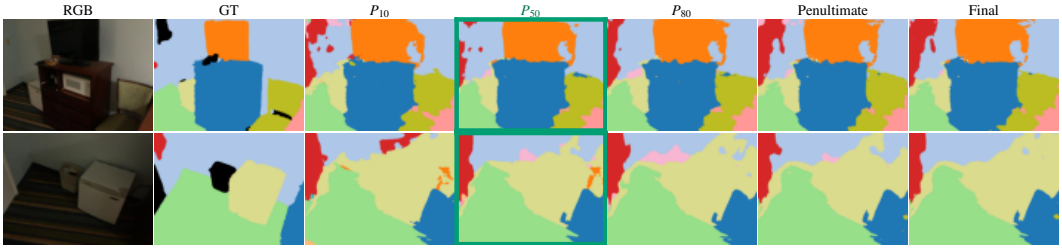


Figure 15: Semantic depth comparison for ScanNet scene `scene0389_00`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by wall, floor, cabinet, and refrigerator. The best observed layers are P_{50}. Relative to Final, the winning intermediate layers most often correct wall, floor, and cabinet pixels. Green borders mark the best layer in each row.

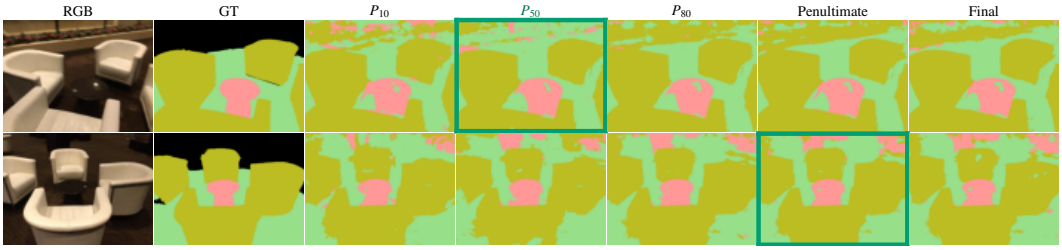


Figure 16: Semantic depth comparison for ScanNet scene `scene0423_02`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by chair, floor, and table. The best observed layers are P_{50} and Penultimate. Relative to Final, the winning intermediate layers most often correct chair, floor, and table pixels. Green borders mark the best layer in each row.

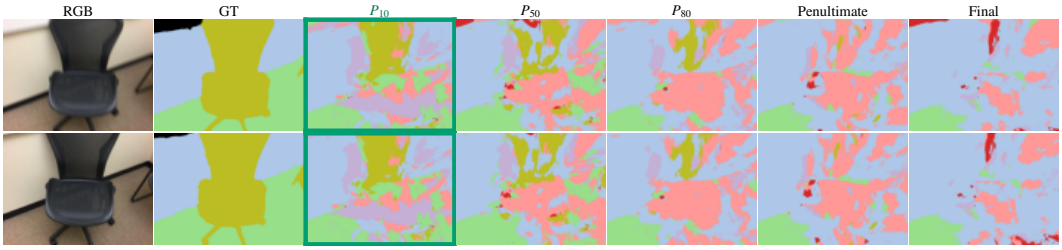


Figure 17: Semantic depth comparison for ScanNet scene `scene0427_00`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by chair, wall, and floor. The best observed layers are P_{10}. Relative to Final, the winning intermediate layers most often correct chair, floor, and wall pixels. Green borders mark the best layer in each row.

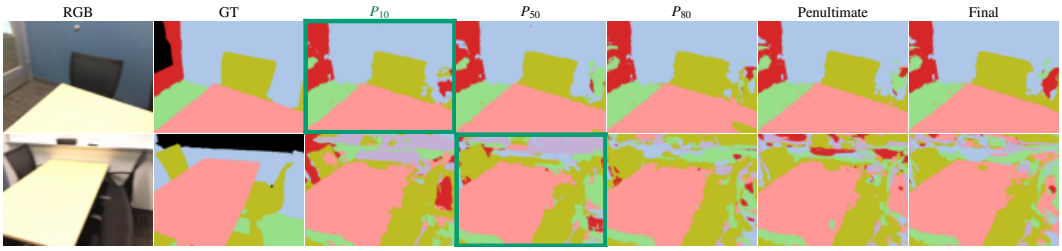


Figure 18: Semantic depth comparison for ScanNet scene `scene0494_00`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by table, wall, chair, and floor. The best observed layers are P_{10} and P_{50}. Relative to Final, the winning intermediate layers most often correct wall, floor, and chair pixels. Green borders mark the best layer in each row.

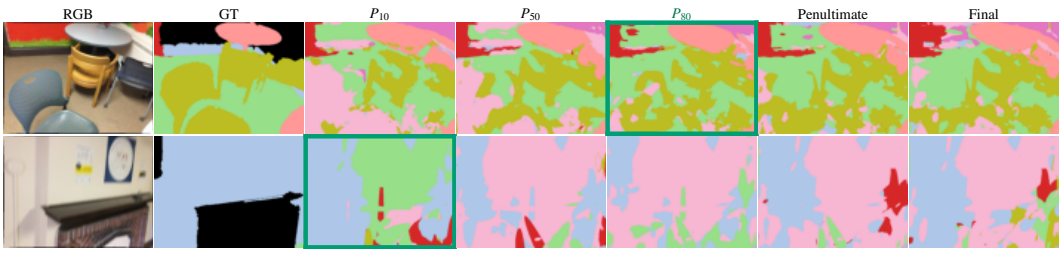


Figure 19: Semantic depth comparison for ScanNet scene `scene0616_00`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by wall, chair, floor, and table. The best observed layers are P_{80} and P_{10}. Relative to Final, the winning intermediate layers most often correct chair, wall, and floor pixels. Green borders mark the best layer in each row.

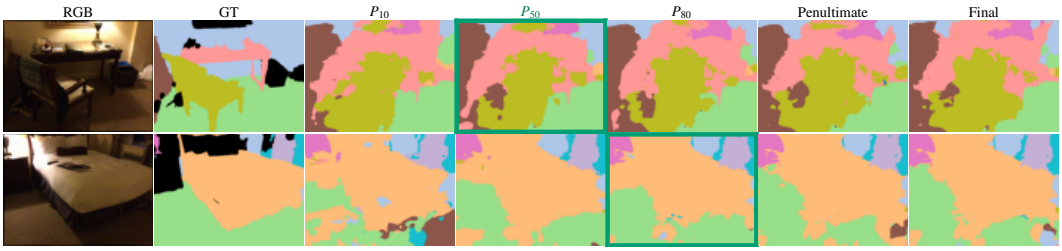


Figure 20: Semantic depth comparison for ScanNet scene `scene0645_02`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by floor, bed, wall, and chair. The best observed layers are P_{50} and P_{80}. Relative to Final, the winning intermediate layers most often correct floor, wall, and table pixels. Green borders mark the best layer in each row.

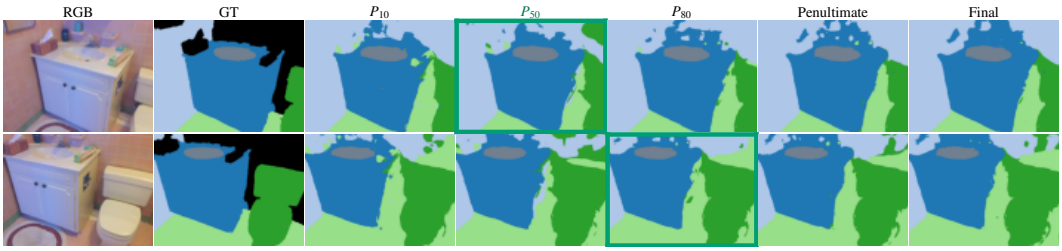


Figure 21: Semantic depth comparison for ScanNet scene `scene0693_00`. The two rows are the views closest to the scene-median best-vs-Final mIoU improvement. The visible labels are dominated by cabinet, wall, toilet, and floor. The best observed layers are P_{50} and P_{80}. Relative to Final, the winning intermediate layers most often correct cabinet, floor, and wall pixels. Green borders mark the best layer in each row.