Next Best View Selections for Semantic and Dynamic 3D Gaussian Splatting

Yiqian Li Wen Jiang Kostas Daniilidis

University of Pennsylvania

Abstract

Understanding semantics and dynamics has been crucial for embodied agents in various tasks. Both tasks have much more data redundancy than the static scene understanding task. We formulate the view selection problem as an active learning problem, where the goal is to prioritize frames that provide the greatest information gain for model training. To this end, we propose an active learning algorithm with Fisher Information that quantifies the informativeness of candidate views with respect to both semantic Gaussian parameters and deformation networks. This formulation allows our method to jointly handle semantic reasoning and dynamic scene modeling, providing a principled alternative to heuristic or random strategies. We evaluate our method on large-scale static images and dynamic video datasets by selecting informative frames from multi-camera setups. Experimental results demonstrate that our approach consistently improves rendering quality and semantic segmentation performance, outperforming baseline methods based on random selection and uncertainty-based heuristics.

1. Introduction

Recent advancements in 3D Gaussian Splatting (3DGS) have enabled real-time rendering and semantic scene understanding for both static and dynamic environments. Extensions of 3DGS now incorporate features from large-scale 2D vision-language models for semantic reasoning. To enable dynamic scene representations, recent methods [16, 20, 21, 23, 30, 42] augment Gaussians with temporal modeling to capture changes in geometry and appearance over time. These developments make 3DGS as a promising backbone for applications in robotics, AR/VR, and digital content creation, where efficiency and semantic accuracy are critical. However, the potential of 3DGS in such real-world settings is limited by the substantial resources required for training, especially when scaling to large-scale environments, densely instrumented multi-camera systems, or long-duration dynamic sequences. The key challenge lies in the vast number of candidate viewpoints: although many are redundant or uninformative, choosing which views are most informative is non-trivial. Unlike static reconstruction, dynamic environments introduce additional uncertainty due to evolving geometry, appearance changes, and semantic inconsistencies across time.

A natural solution is **next-best-view** (NBV) selection, which seeks to prioritize views that most improve reconstruction and understanding. Effective NBV selection can significantly improve both geometric reconstruction and semantic consistency, especially in dynamic environments where scene content changes over time. In practice, however, existing strategies often rely on random sampling, uncertainty heuristics [14, 40], or black-box selection models [29, 41], which do not explicitly maximize information gain and therefore can lead to suboptimal reconstruction and semantic coverage. Moreover, prior work predominantly focuses on static scenes, leaving dynamic NBV selection largely unexplored. Under dynamic scenes, these strategies struggle to generalize across scenes with complex motion and appearance variation, often lead to wasted computation, incomplete reconstructions, and semantic drift, leaving us without a principled way to allocate training resources effectively.

In this work, we present a unified framework that combines dynamic and semantic 3DGS and formulates NBV selection as an active learning problem. Building on this formulation, we develop the first NBV algorithm tailored to dynamic semantic 3DGS, jointly capturing geometric, semantic, and temporal deformation information. Unlike FisherRF [11], which considers only geometric information for static scenes, our method systematically leverages Fisher Information to identify views that maximally improve both reconstruction quality and dynamic scene modeling. To handle the computational challenges of large-scale dynamic semantic 3DGS, we propose an efficient diagonal approximation of the Fisher Information matrix and introduce a novel estimator for the Fisher Information of deformation networks using the trace of the gradient outer product. Beyond proposing a new algorithm, our framework integrates NBV selection directly into the 3DGS backbone, greatly extending the scenarios and applications of NBV to semantic and dynamic scene understanding.

We evaluate our framework on large-scale static datasets (Replica [36]) and dynamic multi-camera sequences (Neu3D [20]), showing consistent improvements over baseline methods. Our contributions include:

- The first Fisher Information-driven NBV selection framework for *dynamic semantic 3DGS*, capturing geometric, photometric, semantic, and deformation informativeness.
- An efficient Fisher Information formulation for large-scale semantic and dynamic 3D Gaussian Splatting, which extends the diagonal approximation of FisherRF [11] to semantic Gaussian parameters and deformation networks, enabling tractable NBV selection beyond static radiance fields.
- A novel method to estimate Fisher Information for deformation MLPs, leveraging the trace of the gradient outer product as a proxy.

2. Related works

2.1. Semantic 3D Gaussians Splatting

Recent advancements have significantly improved semantic understanding in 3D scene representations. Early works build upon implicit radiance field representations such as NeRF. Methods like Semantic NeRF [49], Panoptic Lifting [35], and Contrastive Lift [1] pioneered the integration of semantic labels into NeRF-based frameworks, achieving sharp and accurate 3D semantic segmentation. Subsequent approaches—such as Distilled FFD [15], F3F [38], panoptic NeRF [6], NeRF-SOS [4], Interactive Segmentation of Radiance Fields [7], Weakly Supervised 3D Open-Vocabulary Segmentation [22], FFNF [24], VL-Fields [37], Feature-nerf [45], Distilled Feature Fields [34], and LERF [13] leverage semantic features from pretrained vision-language models like CLIP [31], LSeg[17], or DINO[25] to enable open-vocabulary, pixel-level semantic understanding.

More recently, 3D Gaussian Splatting (3DGS) [12] has emerged as a compelling explicit alternative to NeRF due to its faster training, real-time rendering capabilities, and high-quality scene reconstruction. Semantic extensions of 3DGS have shown great promise. Semantic Gaussians [8] and ConceptFusion [9] project features from 2D pretrained encoders onto 3D Gaussians, enabling open-vocabulary understanding. OpenGaussian [43] and OpenScene [28] achieve point-level semantic segmentation and produce sharp and accurate semantic segmentation results. Feature 3DGS [50], CLIP-FO3D [47], and PLA [2] further distill semantic features from 2D foundation models into the 3D domain using differentiable rasterization. Thanks to the real-time rendering and optimization capabilities of 3DGS, notable examples like SGS-SLAM [19], SemGauss-SLAM [51], and GS³LAM [18] also integrate semantic features into 3DGS-based SLAM pipelines for robust and scalable scene reconstruction.

Motivated by the strong generalization and ease of training of Feature 3DGS [50], we adopt its framework as the foundation for our semantic Gaussian Splatting module.

2.2. Dynamic Gaussians Splatting

Several recent works have extended 3D scene representation to dynamic scenes by introducing temporal modeling, enabling novel view synthesis in dynamic environments. Based on NeRF, early approaches such as D-NeRF [30] introduced time as an explicit input to model non-rigid deformations, while Nerfies [26] leveraged deformation fields with elastic regularization to animate casually captured scenes. HyperNeRF [27] advanced this by modeling topological changes via a hyperspace embedding, handling complex non-rigid transformations. STaR [46] decomposes scenes into static and dynamic components using a self-supervised rigid-body model.

As for 3DGS-based models, 4D Gaussian Splatting (4D-GS) [42] introduces a unified representation that combines 3D Gaussians with 4D neural voxels with a lightweight MLP to efficiently predict Gaussian deformations over time. Gaussian-Flow [21] alternatively models temporal attribute deformations using time and frequency domain parameterizations. Dynamic 3D Gaussians [23] uses dense 6-DOF tracking and dynamic reconstruction through enforcing local rigidity constraints to enable Gaussians to maintain persistent attributes when they move and rotate over time. Explicit sampling strategies like Fully Explicit Dynamic Gaussian Splatting [16] have also been proposed to separate static and dynamic components during training to represent continuous motion and reduce computational cost.

In our work, we adopt the 4D-GS [42] framework as the foundation for our dynamic Gaussians module, leveraging its efficient modeling of dynamic scenes and real-time rendering capabilities.

2.3. Uncertainty and view selection

Driven by the need to reduce training costs and enable active training, a paradigm where models selectively acquire data based on informativeness, many recent works have focused on extracting useful cues from the trained model, existing data, and candidate viewpoints. A central theme across these methods is the estimation and exploitation of uncertainty. Magic Moments [3] introduces higher moments of rendering equations to quantify the uncertainty of radiance fields and show their relationship with rendered error and their usage as powerful NBV selection criteria. CSS [40] utilizes the expectation and variance of categorical distributions to probabilistically update semantic maps within the 3D Gaussian Splatting (3DGS) framework. Similarly, [14] proposes a colorized surface voxel to interpret color uncertainty and surface information. While these methods primarily adopt Bayesian formulations to explicitly quantify uncertainty, oth-

ers infer uncertainty by learning data distributions via alternative models. DNRSelect [41] employs a reinforcement learning-based view selector trained on rasterized images to find optimal views for deferred neural rendering. ActiveInit-Splat [29] defines a black-box objective function based on density and occupancy metrics and employs a Gaussian process as the surrogate model to predict informative views.

Beyond uncertainty estimation, several methods directly target maximizing information gain. GauSS-MI [44] utilizes Shannon Mutual Information to select informative viewpoints. FisherRF [11] introduces a Fisher Information-based metric to guide view selection in radiance fields without reliance on ground-truth data. AG-SLAM [10] further extends this idea by integrating it into an active SLAM framework, enabling online scene reconstruction and autonomous exploration via information-driven trajectory planning.

In this work, we adopt Fisher Information as the unified selection criterion for both semantic Gaussian Splatting and deformation-based dynamic modeling, facilitating view selection that balances scene semantics and motion-aware updates.

3. Method

In this section, we describe our active learning algorithm with Fisher Information for NBV selection in dynamic semantic 3DGS. We begin by reviewing the necessary preliminaries on semantic 3DGS and dynamic 3DGS, which form the foundation of our unified backbone. We then introduce our formulation for computing Fisher Information, first for Gaussian parameters, which capture both geometry and semantics features in Section 3.2 and subsequently for the deformation network, which captures temporal dynamics in Section 3.3. This decomposition allows us to quantify the informativeness of candidate views across all key aspects of the scene representation, enabling principled and efficient NBV selection.

3.1. Preliminaries

In this Section, we first introduce the two core components of our representation framework. Section 3.1.1 details the formulation of Semantic 3D Gaussian Splatting, which distills semantic features from 2D vision-language models into a 3D explicit representation. Section 3.1.2 extends this framework to 4D Gaussian Splatting, incorporating a deformation network to capture dynamic scene changes over time.

3.1.1 Semantic 3D Gaussians Splatting

Following the work of feature 3DGS [50], the semantic 3D Gaussian Splatting (3DGS) represents a scene using a collection of Gaussians $\{\mathcal{G}_i\}_{i=1}^N$, where each Gaussian $\mathcal{G}_i$ is defined by its location $x_i \in \mathbb{R}^3$, rotation $q_i \in \mathbb{R}^4$, scale $s_i \in \mathbb{R}^3$, opacity $\alpha_i \in \mathbb{R}$, color $c_i \in \mathbb{R}^3$, and a semantic feature vector $f_i \in \mathbb{R}^d$. The full parameter set is:

$$\Theta_i = \{x_i, q_i, s_i, \alpha_i, c_i, f_i\}.$$

To project the 3D Gaussians into image space, each covariance matrix Σ_i is transformed to camera space:

$$\Sigma_i' = JW\Sigma_i W^\top J^\top, \tag{1}$$

where W is the world-to-camera transformation and J is the Jacobian of the projection function. The original 3D covariance is decomposed as:

$$\Sigma_i = R_i S_i^2 R_i^\top, \tag{2}$$

where R_i is the rotation matrix, which can be converted from quaternions q_i) and S_i is a diagonal matrix of scales s_i. In this way, the Σ_i can be guaranteed to be positive semi-definite during optimization.

Rendering is performed using volumetric α-compositing in front-to-back order. The color C and semantic feature F_s at a pixel are computed as:

$$C = \sum_{i\in\mathcal{N}} c_i \alpha_i T_i, \quad F_s = \sum_{i\in\mathcal{N}} f_i \alpha_i T_i, \tag{3}$$

where $\mathcal{N}$ is the set of Gaussians overlapping the pixel (sorted by depth), and the transmittance T_i is:

$$T_i = \prod_{j=1}^{i-1}(1-\alpha_j). \tag{4}$$

In order to supervise the semantic features, we distill from the pretrained 2D foundation model Lseg [17] with a feature dimension of 512. To reduce training and rendering cost, we keep low-dimensional features $f_i \in \mathbb{R}^{128}$ at each 3D Gaussian and employ a lightweight convolutional decoder to upsample the rendered feature map $F_s \in \mathbb{R}^{H\times W\times 128}$ to $F_s' \in \mathbb{R}^{H\times W\times 512}$. The overall loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \gamma \mathcal{L}_{\text{feat}}, \tag{5}$$

with:

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\|I-\hat{I}\|_1 + \lambda\mathcal{L}_{\text{D-SSIM}}, \quad \mathcal{L}_{\text{feat}} = \|F_t - F_s'\|_1,$$

where $\hat{I}$ is the rendered image and I is the ground truth.

3.1.2 Dynamic 3DGS

4D-GS [42] extends the static 3D Gaussian Splatting framework to dynamic scenes by introducing a deformation network that models the motion and shape changes of 3D Gaussians over time. At any timestamp t, the Gaussians are transformed via a deformation network:

$$G_t' = G + \Delta G_t = \{X_i + \Delta X_i(t), s_i + \Delta s_i(t), r_i + \Delta r_i(t)\}, \tag{6}$$

where $\Delta X_i(t)$, $\Delta s_i(t)$, and $\Delta r_i(t)$ are the time-dependent predicted deformations.

The deformation network F_ϕ consists of two components: a spatial-temporal structure encoder H and a multi-head deformation decoder D. The encoder H uses a multi-resolution voxel grid decomposition, inspired by K-Planes [5], over six 2D plane projections $(x, y), (x, z), (y, z), (x, t), (y, t), (z, t)$. Each Gaussian's position and time (X_i, t) are queried from these planes via bilinear interpolation to produce a feature embedding:

$$F_i = \phi_d \left(\bigcup_{(i,j)} \text{interp}(R^{(l)}_{(i,j)}) \right), \tag{7}$$

where $R^{(l)}_{(i,j)}$ denotes the voxel planes at resolution level l, and ϕ_d is a small MLP used to fuse the plane features.

The decoder $D = \{\phi_x, \phi_r, \phi_s\}$ consists of three heads that output the deformation for position, rotation, and scale:

$$\Delta X_i(t) = \phi_x(F_i), \quad \Delta r_i(t) = \phi_r(F_i), \quad \Delta s_i(t) = \phi_s(F_i). \tag{8}$$

During rendering, the pretrained 3DGS are first deformed through $\{G'_i\}_{i=1}^N = F\left(\{G_i\}_{i=1}^N\right)$. Then, the deformed Gaussians G'_t are rendered via differentiable splatting to produce the final image:

$$\hat{y}_t = S(M_t, G'_t), \tag{9}$$

where M_t is the camera pose at time t and S denotes the splatting renderer.

3.2. Fisher information for semantic 3DGS

In this section, we will describe how to qualify Fisher information for Gaussian parameters for NBV selection.
Following the instruction of FisherRF [11], let the radiance field be parameterized by G, representing the set of attributes for all 3D Gaussians in the scene, including location x, scale s, opacities α, colors c, and semantic feature f. At a given camera pose $M \in \text{SE}(3)$, the rendering function $S(M, G)$ produces a synthetic image. We consider the negative log-likelihood of the observed image y under the rendered prediction:

$$\mathcal{L}(y; G) = -\log p(y|M, G) = \|y - S(M, G)\|_2^2. \tag{10}$$

Under standard regularity conditions, the *observed* Fisher Information with respect to the parameters G at pose M is given by the Hessian of this loss:

$$\mathcal{I}(G; M) = \nabla_G S(M, G)^\top \nabla_G S(M, G) + \lambda I, \tag{11}$$

where λI is a log-prior regularization term.

However, due to the large number of parameters, computing the full Fisher Information matrix is intractable in practice. Following the approximation proposed in FisherRF [11], we apply a Laplace approximation by retaining only the diagonal elements of the Hessian and adding a prior regularization term:

$$\mathcal{I}(G; M) \approx \text{diag}\left(\nabla_G S(M, G)^\top \nabla_G S(M, G)\right) + \lambda I, \tag{12}$$

This diagonal approximation significantly reduces computational cost and memory overhead, while preserving the essential structure of the Fisher Information needed for evaluating local sensitivity.

To select the most informative viewpoint from a candidate pool $\{M_i\}_{i=1}^N$, we pick the view that maximize the *Expected Information Gain* (EIG), defined as:

$$\text{EIG}(M_i) = \text{tr}\left(\mathcal{I}(G; M_i) \cdot \mathcal{I}_{\text{train}}^{-1}\right), \tag{13}$$

where $\mathcal{I}_{\text{train}}$ is the accumulated Fisher Information of current training views. The trace measures how much the candidate view would sharpen the current uncertainty landscape over parameters.

3.3. Fisher Information for Deformation Network

In this section, we introduce a novel estimator for the Fisher Information of multilayer perceptrons (MLPs), which we employ to quantify the informativeness of our deformation module.

For the deformation network F_ϕ, the informativeness of a candidate view $(\mathbf{x}^{acq}, \mathbf{y}^{acq})$ is quantified by the Fisher Information with respect to deformation network parameters ϕ and the current training data D^{train}:

$$\begin{aligned}\mathcal{I}[\phi; \mathbf{y}^{acq} \mid \mathbf{x}^{acq}, D^{train}] &= H[\phi \mid D^{train}] \\ &\quad - H[\phi \mid \mathbf{y}^{acq}, \mathbf{x}^{acq}, D^{train}].\end{aligned} \tag{14}$$

Following FisherRF [11], this gain can be upper bounded by the curvature of the log-likelihood:

$$\begin{aligned}\mathcal{I}[\phi] \;\le\; \tfrac{1}{2}\,\text{tr}\Big(&\mathbf{H}''_\phi\big[\log p(\mathbf{y}^{acq} \mid \mathbf{x}^{acq}, \phi)\big] \\ &\cdot\; \mathbf{H}''_\phi\big[\log p(\phi \mid D^{train})\big]^{-1}\Big).\end{aligned} \tag{15}$$

where $\mathbf{H}''_\phi$ denotes the Hessian with respect to ϕ.

For view selection, the prior Hessian term $\mathbf{H}''_\phi[\log p(\phi \mid D^{train})]^{-1}$ is constant across candidate views and can be dropped. Thus our acquisition criterion simplifies to maximizing

$$\mathcal{S}(\mathbf{x}^{acq}) = \text{tr}\big(\mathbf{H}''_\phi\big[\log p(\mathbf{y}^{acq} \mid \mathbf{x}^{acq}, \phi)\big]\big).$$

Finally, we estimate the trace efficiently with Hutchinson's estimator:

$$\mathcal{S}(\mathbf{x}^{acq}) \;\approx\; \mathbb{E}_{\mathbf{v}\sim\mathcal{N}(0,I)}\big[\|\mathbf{H}''_\phi\,\mathbf{v}\|^2\big],$$

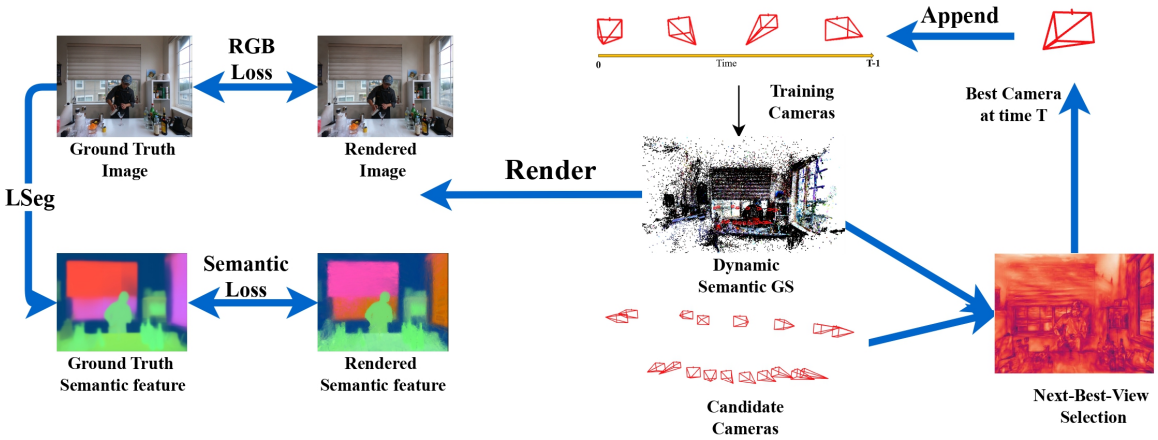

Figure 1. **Overview of our work.** At each timestep, we select the NBV from a set of candidate camera views using a Fisher Information-based criterion. The selection process evaluates each candidate view based on its expected contribution to both semantic feature learning and dynamic deformation modeling. Once selected, the chosen view is incorporated into the training process of our Dynamic and Semantic 3DGS. For rendering and supervision, the Gaussians are first deformed by an MLP deformation network based on the current timestep, then projected into RGB images and semantic feature maps. These outputs are compared against ground-truth images and semantic labels to jointly optimize both the dynamic deformation and semantic representation.

where $\mathbf{v}$ is a random probe vector and Hessian-vector products are computed via automatic differentiation.

This formulation explicitly ties Fisher Information to the deformation network parameters ϕ, ensuring that the selected views maximize temporal consistency and improve the modeling of dynamic scene evolution.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed method. We first describe the experimental setup, including the datasets, baseline methods, and evaluation metrics. We then report results on two key scenarios: selecting a limited number of candidate camera poses from large-scale semantic datasets for static semantic 3DGS, and selecting informative frames from multi-camera sequences for dynamic and semantic 3DGS. Finally, we conduct ablation studies to analyze the impact of the log-prior regularization parameter and different feature choices for NBV selection, providing deeper insights into the design and effectiveness of our approach.

4.1. Experimental Set-up

Dataset For semantic 3DGS active training, we choose Replica [36], which composes various indoor environments with semantic labels. We follow the dataset setting of Feature 3DGS [50] to pick 75-80 images for each scene and leave 15 of them for testing.
For dynamic and semantic 3DGS active training, we choose Neu3D [20], which includes 15-21 synchronous camera videos. From each view, we sample 40 temporally synchronized frames. The central view, `cam00`, is reserved as the test view, and the remaining views are used for training. To gain the ground truth semantic features, we apply LSeg [17] to semantic segmentation and attain pixel-level features for each image. We utilize Colmap [33] [32] to get sparse point clouds for training initialization.

Baselines We compare our method against random view selection and semantic uncertainty-based selection. Following the approach of Magic Moments [3], we use feature covariance as a measure of uncertainty and implement the computation efficiently in CUDA. To estimate the second moment of semantic features, we first square each feature vector associated with a Gaussian and accumulate the results to compute the per-pixel second moment. The covariance is then obtained by subtracting the squared mean (i.e., the square of the rendered semantic output) from the second moment.

Metrics We evaluate our rendering quality with Structural Similarity Index Measure (SSIM) [39], Peak-signal-to-noise ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [48], and semantic segmentation quality with mean intersection-over-union (mIoU) and mean accuracy (mAcc). To evaluate semantic segmentation, we first project Gaussian semantic features into rendered images to get per-pixel features. Then, we project feature maps into class probabilities by computing dot products between features

Method	SSIM (↑)	PSNR (↑)	LPIPS (↓)	mAcc (↑)	mIoU (↑)
Random	0.9077	29.6600	0.1477	0.9189	0.7120
Magic Moments [3]	0.9615	33.5200	0.0591	0.9256	0.7593
FisherRF [11]	0.9593	32.1700	0.0614	0.9230	0.7441
Ours	0.9626	33.6000	0.0572	0.9289	0.7648

Table 1. Average performance on Replica dataset under different view selection strategies. The **best**, **second best**, and **third best** results are highlighted with light red, orange, and yellow backgrounds, respectively.

and text features from predefined labels and assigning every pixel to the class with the highest logit per pixel.

4.2. Static semantic 3DGS active training

For active training in static semantic Gaussian Splatting, we first start our training with two images. Then after 200 iterations, we compute Fisher Information for all the remained camera poses and add the one that yields the highest expected information gain to the training set, until adding 10 images every 200 iterations. The model is then trained for a total of 7,000 iterations. All experiments are conducted on an NVIDIA L40 GPU, requiring approximately 4GB of memory and 45 minutes of training per scene. For each NBV selection, it takes approximately 7 seconds.

Table 1 summarizes the average performance of different view selection strategies on the Replica dataset. Our method, which incorporates Fisher Information computed with semantic information, achieves the best results across all metrics, indicating improved visual quality and semantic accuracy.

Removing the semantic information from the Fisher Information calculation leads to a consistent decline in performance, highlighting the importance of incorporating semantic sensitivity when estimating view informativeness. The semantic covariance baseline Magic Moments [3], which relies on second-order statistics of categorical features, underperforms compared to our Fisher Information-based method with semantic information. The random view selection baseline results in the weakest performance across all metrics, further emphasizing the advantage of principled, information-driven view selection.

4.3. Dynamic and semantic 3DGS active training

For active training in dynamic and semantic 3DGS, we begin by training a static semantic model using the first frame from all cameras for 3000 iterations. Subsequently, to train the deformation network for dynamic, for each future timestep, we compute the Fisher Information across all candidate camera views and iteratively select the view with the highest information gain every 200 iterations. This process continues until all 39 remaining timesteps are each assigned a single informative camera frame. The dynamic training phase runs

Method	SSIM (↑)	PSNR (↑)	LPIPS (↓)	mAcc (↑)	mIoU (↑)
Random	0.8831	26.7361	0.1933	0.8848	0.7284
Magic Moments [3]	0.9184	28.3795	0.1581	0.8890	0.7350
FisherRF [11]	0.9186	27.6743	0.1605	0.8863	0.7547
Ours	0.9239	28.6191	0.1553	0.8963	0.7604

Table 2. Average performance across Neu3D under different view selection strategies.

Method	SSIM (↑)	PSNR (↑)	LPIPS (↓)	mAcc (↑)	mIoU (↑)
Ours (w/o reg.)	0.9173	28.5853	0.1589	0.8929	0.7413
Ours ($\lambda = 10^{-7}$)	0.9184	28.0638	0.1594	0.8881	0.7387
Ours ($\lambda = 10^{-6}$)	0.9239	28.6191	0.1553	0.8963	0.7604
Ours ($\lambda = 10^{-5}$)	0.9186	28.5295	0.1584	0.8898	0.7374

Table 3. Average performance on Neu3D dataset under our view selection strategy with and without regularization.

Method	SSIM (↑)	PSNR (↑)	LPIPS (↓)	mAcc (↑)	mIoU (↑)
Geom [11]	0.9186	27.6743	0.1605	0.8863	0.7547
Geom+Sem	0.9200	28.2788	0.1596	0.8885	0.7571
Geo+Def	0.9198	28.1501	0.1578	0.8905	0.7367
Ours	0.9239	28.6191	0.1553	0.8963	0.7604

Table 4. Average performance across Neu3D under different view selection strategies.

14000 iterations. For regularization, here we set $\lambda = 10^{-6}$. Training is conducted on an NVIDIA L40 GPU, requiring approximately 8GB of memory and 2 hours per scene. For each NBV selection, it takes approximately 20 seconds.

Table 2 presents the average performance across five dynamic scenes from the Neu3D dataset using different view selection strategies. The proposed method based on Fisher Information achieves the best performance across all evaluation metrics, presenting superior visual quality and semantic consistency. In contrast, the semantic covariance baseline consistently underperforms the Fisher-based method. The random selection baseline yields the weakest performance, especially in perceptual quality and segmentation accuracy, further validating the effectiveness of Fisher Information-guided view selection for dynamic and semantic 3DGS.

We also analyze the attention patterns of the three selection methods. As shown in Fig. 4, which visualizes the interest heatmaps, the subject is in the process of closing the kitchen tongs. Magic Moments [3] distributes attention evenly across the entire kitchen tongs. FisherRF [11] emphasizes the connection between the handles and the gripping ends. In contrast, our method focuses more on the handles, likely due to their dynamic motion during the closing action.

4.4. Ablation study

Ablation on choice of log-prior regularization parameter. We conduct an ablation study to evaluate the effect of the log-prior regularization parameter λ in the Fisher In-

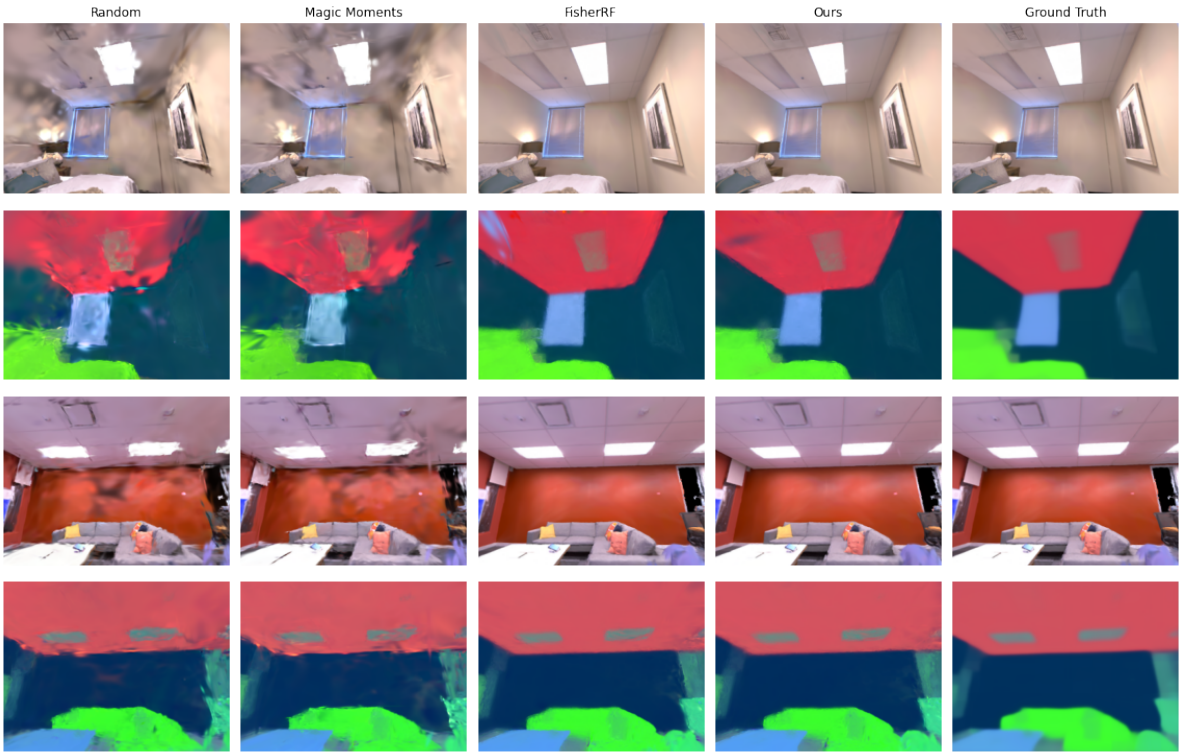

Figure 2. **Semantic segmentation results on Replica dataset.** The first and third rows of images are the rendered results. The second and forth rows of images are the results of visualizing the feature maps of test set after being trained with 12 training views. All the methods are based on Feature 3DGS [50] except for different view selection methods to augment the training data.

formation computation (Eq. 11). This term is critical for stabilizing the diagonal approximation of the Hessian, especially in dynamic and semantic scenes. Table 3 compares four variants of our method on the Neu3D dataset: no regularization ($\lambda = 0$), weak regularization ($\lambda = 10^{-7}$), and our default setting ($\lambda = 10^{-6}$), and strong regularization ($\lambda = 10^{-5}$). The results show that the full model with $\lambda = 10^{-6}$ consistently outperforms the other three across all metrics. Without regularization, the Fisher scores become noisier and less reliable, leading to degraded view selection and semantic accuracy.

Ablation on Features Used for NBV Selection. We conduct an ablation study to evaluate the contribution of different features in the Fisher Information computation for NBV selection in 4DGS. Specifically, we compare four variants on the Neu3D dataset: **Geom**, which uses Fisher information only on geometric parameters following *FisherRF* [11]; **Geom+Sem**, which additionally incorporates semantic features; **Geom+Def**, which combines geometric and deformation parameters; and **Ours**, which jointly considers geometric, semantic, and deformation parameters. As shown in Table 4, adding semantic information improves fine-detail reconstruction and high-frequency texture fidelity, while adding deformation information enhances temporal consistency in dynamic regions. Our full model, which integrates all three components, achieves the best performance across all metrics, confirming that semantic and deformation cues provide complementary benefits to purely geometric NBV selection.

5. Conclusion

To address the critical challenge of reducing training data requirements while preserving both rendering fidelity and semantic accuracy across static and dynamic scenes, we introduce a unified 3DGS backbone that jointly models semantic and temporal representations, together with an active learning framework based on Fisher Information for Next-Best-View selection. By quantifying the informativeness of candidate viewpoints with respect to geometric and semantic Gaussian parameters and the deformation network, our method identifies the most informative viewpoints to guide the training process. Our experimental results on Replica and Neu3D validate the effectiveness of the approach. These findings highlight that view selection guided by explicit information measures can lead not only to faster convergence but also to more reliable modeling of dynamic and semantic scenes. By reducing the reliance on large amounts of training data, our method enables more efficient and scalable training

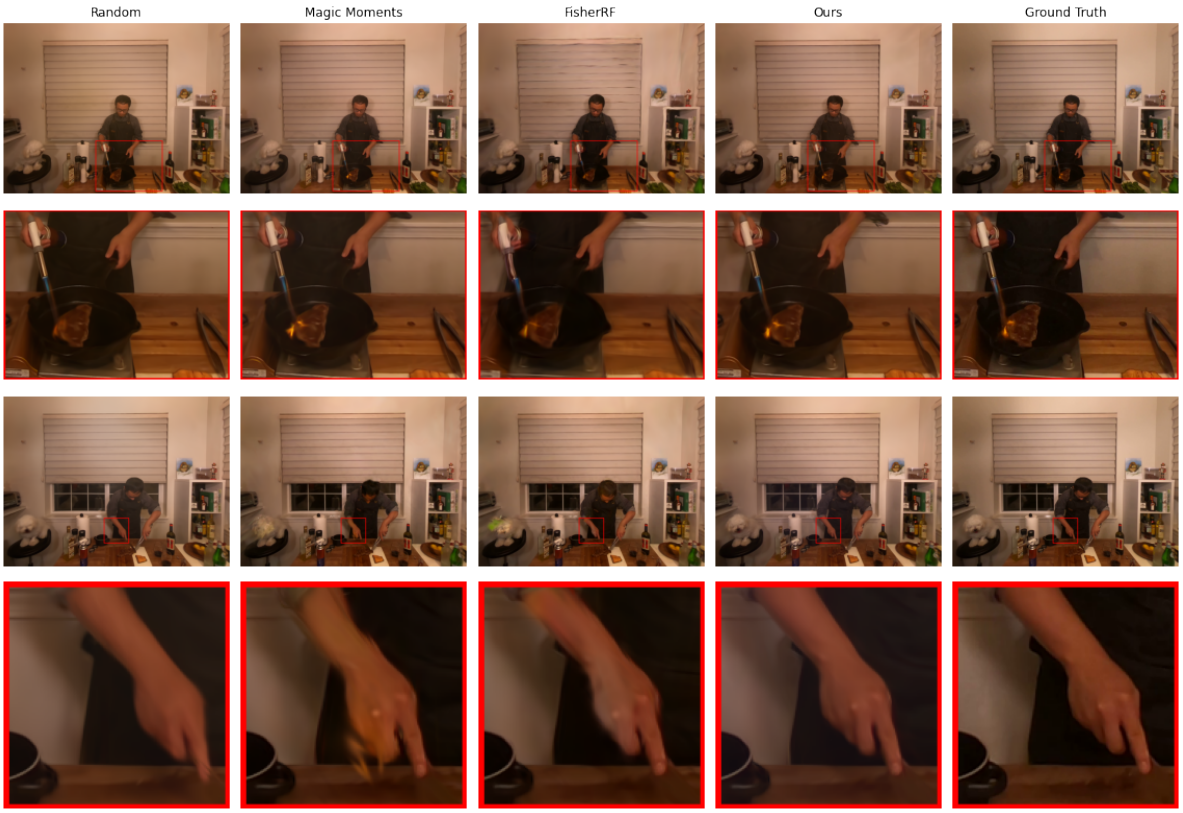

Figure 3. **Zoomed-in qualitative study of our method on Neu3D dataset.** The second and fourth rows are zoom-in figures. Visualizations are the results of the test set after being trained with 39 training views. All the methods are based on the same dynamic and semantic 3DGS, except for different view selection methods to augment the training data.

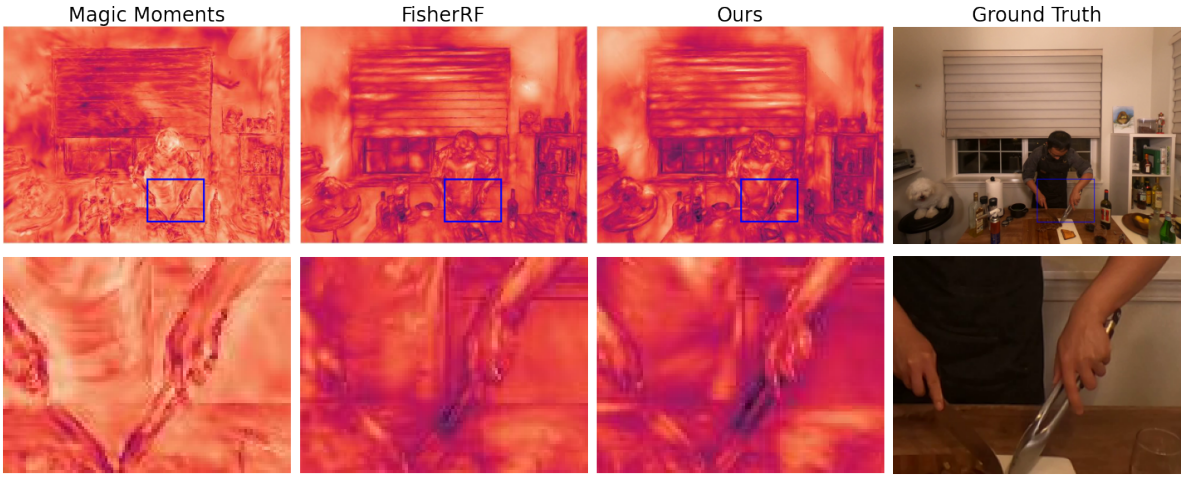

Figure 4. **Heatmap for selection method interest on Neu3D dataset.** The second row is zoom-in figures.

of high-quality 3DGS models.

Our framework offers a foundation for adaptive data acquisition. In robotics and SLAM, principled NBV selection can help agents actively choose viewpoints to improve mapping accuracy and semantic awareness under limited sensing budgets. In AR/VR, efficient view selection accelerates scene capture and dynamic updates. These impacts highlight the potential of our approach to enable efficient, scalable, and intelligent scene understanding across diverse domains.

Acknowledgment We greatly acknowledge financial support by the grants NSF FRR 2220868, NSF IIS-RI 2212433, and ONR N00014-22-1-2677.

References

[1] Yash Bhalgat, Iro Laina, João F Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In *NeurIPS*, 2023. 2

[2] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7010–7019, 2023. 2

[3] Parker Ewen, Hao Chen, Seth Isaacson, Joey Wilson, Katherine A Skinner, and Ram Vasudevan. These magic moments: Differentiable uncertainty quantification of radiance field models. *arXiv preprint arXiv:2503.14665*, 2025. 2, 5, 6

[4] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. NeRF-SOS: Any-view self-supervised object segmentation on complex scenes. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[5] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12488, 2023. 4

[6] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 2

[7] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4201–4211, 2023. 2

[8] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *CoRR*, abs/2403.15624, 2024. 2

[9] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Subramanian Iyer, Nikhil Varma Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso M de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. In *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023. 2

[10] Wen Jiang, Boshu Lei, Katrina Ashton, and Kostas Daniilidis. Ag-slam: Active gaussian splatting slam. *arXiv preprint arXiv:2410.17422*, 2024. 3

[11] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 4, 6, 7

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[13] 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*, pages 19729–19739, 2023. 2

[14] Hyunseo Kim, Hyeonseo Yang, Taekyung Kim, YoonSung Kim, Jin-Hwa Kim, and Byoung-Tak Zhang. Active neural 3d reconstruction with colorized surface voxel-based view selection. *arXiv preprint arXiv:2405.02568*, 2024. 1, 2

[15] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022. 2

[16] Junoh Lee, Changyeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic gaussian splatting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[17] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3, 5

[18] Linfei Li, Lin Zhang, Zhong Wang, and Ying Shen. GS^{3}LAM: Gaussian semantic splatting SLAM. In *ACM Multimedia 2024*, 2024. 2

[19] Mingrui Li, Shuhong Liu, Heng Zhou, Guohao Zhu, Na Cheng, Tianchen Deng, and Hongyu Wang. Sgs-slam: Semantic gaussian splatting for neural dense slam. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXI*, page 163–179, Berlin, Heidelberg, 2024. Springer-Verlag. 2

[20] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022. 1, 2, 5

[21] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21136–21145, 2024. 1, 2

[22] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 2

[23] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809, 2024. 1, 2

[24] Kirill Mazur, Edgar Sucar, and Andrew J Davison. Feature-realistic neural fusion for real-time, open set scene understanding. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8201–8207. IEEE, 2023. 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, 2024. 2

[26] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5865–5874, 2021. 2

[27] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 2

[28] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 2

[29] Konstantinos D Polyzos, Athanasios Bacharis, Saketh Madhuvarasu, Nikos Papanikolopoulos, and Tara Javidi. Activeinitsplat: How active image selection helps gaussian splatting. *arXiv preprint arXiv:2503.06859*, 2025. 1, 3

[30] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10313–10322, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 1, 2

[31] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[33] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[34] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *Conference on Robot Learning*, pages 405–424. PMLR, 2023. 2

[35] 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*, pages 9043–9052, 2023. 2

[36] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 5

[37] Nikolaos Tsagkas, Oisin Mac Aodha, and Chris Xiaoxuan Lu. Vl-fields: Towards language-grounded neural implicit spatial representations. *CoRR*, 2023. 2

[38] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453, 2022. 2

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[40] Joey Wilson, Marcelino Almeida, Min Sun, Sachit Mahajan, Maani Ghaffari, Parker Ewen, Omid Ghasemalizadeh, Cheng-Hao Kuo, and Arnie Sen. Modeling uncertainty in 3d gaussian splatting through continuous semantic splatting. *arXiv preprint arXiv:2411.02547*, 2024. 1, 2

[41] Dongli Wu, Haochen Li, and Xiaobao Wei. Dnrselect: Active best view selection for deferred neural rendering. *arXiv preprint arXiv:2501.12150*, 2025. 1, 3

[42] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 1, 2, 3

[43] 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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[44] Yuhan Xie, Yixi Cai, Yinqiang Zhang, Lei Yang, and Jia Pan. Gauss-mi: Gaussian splatting shannon mutual information for active 3d reconstruction. *arXiv preprint arXiv:2504.21067*, 2025. 3

[45] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8962–8973, 2023. 2

[46] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13144–13152, 2021. 2

[47] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2048–2059, 2023. 2

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep

features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[49] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2

[50] 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 (CVPR)*, pages 21676–21685, 2024. 2, 3, 5, 7

[51] Siting Zhu, Renjie Qin, Guangming Wang, Jiuming Liu, and Hesheng Wang. Semgauss-slam: Dense semantic gaussian splatting slam. *CoRR*, abs/2403.07494, 2024. 2