GEOMETRY MEETS VISION:
REVISITING PRETRAINED SEMANTICS IN DISTILLED FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic distillation in radiance fields has spurred significant advances in open-vocabulary robot policies, e.g., in manipulation and navigation, founded on pretrained semantics from large vision models. While prior work has demonstrated the effectiveness of visual-only semantic features (e.g., DINO and CLIP) in Gaussian Splatting and neural radiance fields, the potential benefit of geometry-grounding in distilled fields remains an open question. In principle, visual-geometry features seem very promising for spatial tasks such as pose estimation, prompting the question: *Do geometry-grounded semantic features offer an edge in distilled fields?* Specifically, we ask three critical questions: *First, does spatial-grounding produce higher-fidelity geometry-aware semantic features?* We find that image features from geometry-grounded backbones contain finer structural details compared to their counterparts. Secondly, *does geometry-grounding improve semantic object localization?* We observe no significant difference in this task. Thirdly, *does geometry-grounding enable higher-accuracy radiance field inversion?* Given the limitations of prior work and their lack of semantics integration, we propose a novel framework **SPINE** for inverting radiance fields without an initial guess, consisting of two core components: (i) coarse inversion using distilled semantics, and (ii) fine inversion using photometric-based optimization. Surprisingly, we find that the pose estimation accuracy *decreases* with geometry-grounded features. Our results suggest that visual-only features offer greater versatility for a broader range of downstream tasks, although geometry-grounded features contain more geometric detail. Notably, our findings underscore the necessity of future research on effective strategies for geometry-grounding that augment the versatility and performance of pretrained semantic features.

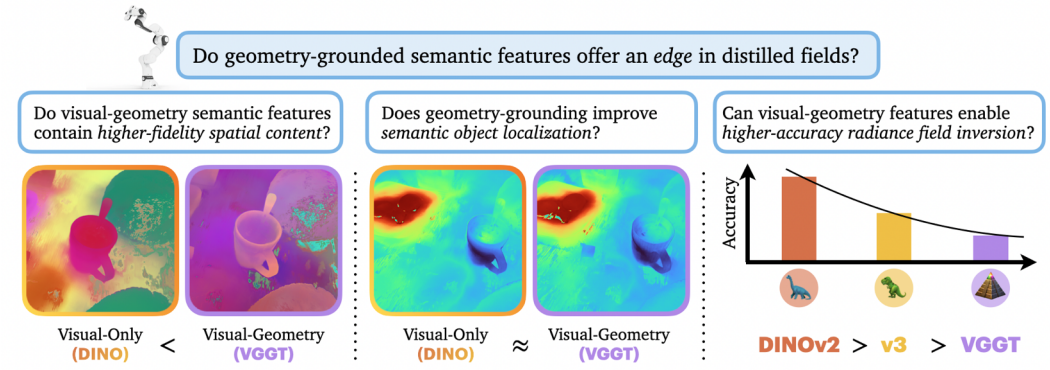

Figure 1: We revisit pretrained semantics in distilled radiance fields, asking three critical questions to compare *visual-geometry* semantic features against *visual-only* features. We find that while visual-geometry features retain richer spatial fidelity, they do not improve performance in downstream tasks such as semantic localization or radiance field inversion, suggesting the greater versatility of visual-only semantic features.

1 INTRODUCTION

Large foundation models have driven rapid advances in open-vocabulary robot policies, enabling robots to perform complex, multi-stage tasks, entirely from natural-language instructions; see (Firoozi et al., 2025) for a detailed review. Through semantic distillation, prior work blends photorealistic novel-view synthesis from radiance fields, such as Gaussian Splatting (Kerbl et al., 2023) and neural radiance fields (NeRFs) (Mildenhall et al., 2021), with generalizable pretrained semantics from foundation models to endow robots with semantic task understanding capabilities grounded in the real-world. This synergy underlies the success of many language-conditioned robot policies in manipulation (Ze et al., 2023; Shen et al., 2023) navigation (Chen et al., 2025).

In general, large vision backbones (CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023), DINOv3 (Siméoni et al., 2025)) have been limited to visual feature learning without spatial grounding, potentially impeding the emergence of 3D-aware features. However, recent work VGGT (Wang et al., 2025) demonstrates that large vision backbones can be trained to produce visual-geometry features by grounding these models with a 3D reconstruction task objective, achieving superior performance in many downstream tasks, such as keypoint tracking. Although visual-geometry features seem promising for spatial tasks such as pose estimation, the actual performance of *visual-geometry* semantic features relative to *visual-only* semantic features remains unknown, particularly in distilled radiance fields. To address this gap, we revisit pretrained semantics in distilled fields, asking: *Do visual-geometry semantic features offer an edge in distilled fields?* To rigorously explore the relative performance between visual-only features (DINOv2 and DINOv3) and visual-geometry features (VGGT), we ask three critical questions, each focused on important downstream applications of distilled radiance fields in robotics, illustrated in Figure 1.

First, *do visual-geometry semantic features contain higher-fidelity spatial content?* We examine the information content of visual-only and visual-geometry features, mapping these features from their respective semantic spaces to a lower-dimensional visual space. We find that *visual-geometry semantics provide finer spatial content, with more prominent structural detail, e.g., sharper edges, more accurate subpart decomposition, etc*. We note that high-fidelity geometric detail could be essential in some robotics applications, e.g., fine and dexterous manipulation.

Second, *does geometry-grounding improve semantic object localization?* Many state-of-the-art robot policies rely on semantic object localization in distilled radiance fields for successful task execution, underscoring its immense relevance. We find that *visual-geometry semantic features perform similarly to visual-only features*, which suggests that spatial grounding does not boost object-class semantic feature content. However, effective co-supervision of semantic features with geometry and vision can lead to improved performance, presenting an important area for future research.

Third, *can visual-geometry features enable higher-accuracy radiance field inversion?* Unlike rendering/rasterization, radiance field inversion remains a challenging problem, with existing methods limited by need for good initial estimates. To address this problem and further facilitate semantics evaluation, we introduce a novel framework *SPINE* for inverting radiance fields without any camera pose. SPINE directly leverages embedded pretrained semantics to compute: (i) coarse pose estimates using a co-trained semantic field which maps semantic features to a distribution over camera poses and (ii) fine pose estimates by refining the coarse solution through novel-view synthesis in radiance fields and robust perspective-n-point optimization. Surprisingly, we observe that visual-geometry semantic features underperform visual-only features, despite their greater spatial content.

Our findings highlight that *visual-only semantic features are more versatile than visual-geometry features*, even in spatial tasks like pose estimation. Importantly, our results underscore the need for additional research on effective strategies for geometry-grounding to augment the versatility and performance of pretrained semantic features.

2 RELATED WORK

Radiance Fields marked a notable breakthrough in 3D scene reconstruction, achieving photorealistic image rendering and novel-view synthesis entirely from RGB images. NeRFs learn separate density and color fields parameterized by multi-layer perceptrons, mapping a 3D point at a specified camera direction to its volume density (associated with its opacity) and radiance (color). Given its generality

to different sensing modalities, NeRFs have been widely applied in robotics, e.g., robot planning (Adamkiewicz et al., 2022; Chen et al., 2024), localization (Yen-Chen et al., 2021; Maggio et al., 2022), and manipulation (Kerr et al., 2022; Weng et al., 2022). In contrast to implicit representation of NeRFs, Gaussian Splatting (GS) (Kerbl et al., 2023) utilizes explicit ellipsoidal primitives, each parameterized by a mean, covariance, opacity, and color, to represent non-empty space. This design choice enables the use of fast tile-based rasterization for faster training and real-time rendering speeds. Like NeRFs, GS has found widespread applications in robotics, e.g., robot planning and localization (Chen et al., 2025; Michaux et al., 2025) and manipulation (Lu et al., 2024).

Distilled Semantics in Radiance Fields. Foundation models, e.g., CLIP and DINO learn robust visual features from images/language that encapsulate open-world semantics through pretraining on internet-scale data. Increasingly, robot policies have embedded semantics from CLIP and DINO into radiance fields to enable language-conditioned robot manipulation (Rashid et al., 2023; Shen et al., 2023; Shorinwa et al., 2024b), mapping (Shorinwa et al., 2025), and object localization (Yin et al., 2025). In general, these methods combine visual-only image features, e.g., DINO, with vision-language semantics from CLIP to support language conditioning, particularly for semantic localization, which plays an essential role in downstream robotics tasks. Recent work trains foundation models for 3D reconstruction to learn visual-geometry features with potentially useful applications to spatial tasks. However, no existing work has examined the integration of these features in distilled radiance fields. In this work, we explore visual-geometry semantic features in distilled radiance fields, quantifying the performance of these features relative to visual-only features in important robotics applications.

3 Preliminaries

We review important technical concepts, necessary for understanding the discussion in this paper. We introduce NeRFs, GS, and discuss semantic distillation in radiance fields in Appendix A.1.

NeRFs. NeRFs learn implicit volumetric color and density fields, encoding the occupancy and radiance of each point in the scene, given a set of images $\mathcal{I}$ and corresponding camera poses, which is typically computed via structure-from-motion (Schonberger & Frahm, 2016). Specifically, the color field $\mathbf{c} : \mathbb{R}^3 \times \mathbb{S}^2 \mapsto \mathbb{R}^3$ maps a 3D point $\mathbf{x} \in \mathbb{R}^3$ and a camera viewing direction $d \in \mathbb{S}^2$ to an RGB color $\mathbf{c}(\mathbf{x}, d)$. Likewise, the density field $\rho : \mathbb{R}^3 \mapsto \mathbb{R}_+$ maps $\mathbf{x}$ to a non-negative volume density $\rho(\mathbf{x})$, representing the differential probability of a light ray terminating at a particle located at $\mathbf{x}$. Using ray-marching to render images from $\mathbf{c}$ and ρ, NeRFs utilize stochastic gradient descent to optimize the parameters of $\mathbf{c}$ and ρ (represented as MLPs), minimizing the photometric error between the rendered and ground-truth images.

Gaussian Splatting. Gaussian Splatting (GS) utilizes 2D (Huang et al., 2024) or 3D (Kerbl et al., 2023) Gaussian primitives to represent non-empty space, each defined by a mean $\mu \in \mathbb{R}^3$, covariance $\Sigma \in \mathbb{S}^3_{++}$, opacity $\alpha \in \mathbb{R}_+$, and spherical harmonics (for view-dependent color) parameters. Like NeRFs, GS optimizes these parameters by minimizing the photometric error between rendered and ground-truth images, initialized from a sparse point cloud generally computed using structure-from-motion. Notably, GS employs a tile-based rasterization procedure to efficiently project the Gaussian primitives to the image plane, circumventing the expensive volumetric ray-marching procedure used by NeRFs, for faster training and rendering. Moreover, the explicit scene representation enables relatively easier theoretical analysis compared to NeRFs, in addition to providing more accurate depth estimation and mesh extraction.

4 Visual-Geometry Semantics in Distilled Radiance Fields

We present our approach to distilling spatially-grounded semantics from vision foundation models into radiance fields. We utilize the state-of-the-art Visual Geometric Grounded Transformer (VGGT) as the vision backbone, given its effectiveness across a broad range of geometric scene-understanding tasks, e.g., camera intrinsics/extrinsics estimation, multi-view depth estimation, dense point cloud reconstruction, and multi-frame point tracking.

Extracting Pretrained Visual-Geometry Semantic Features. We extract ground-truth pretrained semantic embeddings for each image from the depth and point heads of VGGT and its intermediate

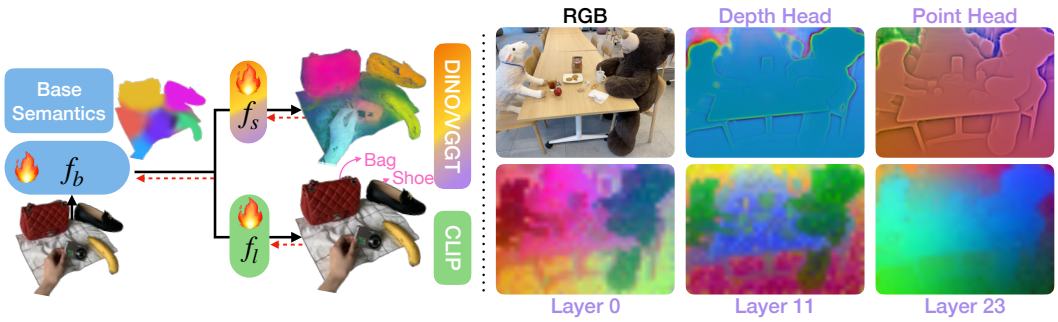

Figure 2: (left) **Semantics distillation architecture**, showing co-supervision of CLIP with DINO/VGGT via base semantics module. (right) **VGGT's semantic embeddings from different heads**, showing high-fidelity geometric content of the point head.

layers, which were trained for depth estimation and dense point cloud reconstruction, respectively. We visualize these semantic embeddings on the right side of Figure 2 using the first three principal components. We find that the point head produces features with the highest-fidelity spatial detail, compared to the depth head and other intermediate layers. Given a query image $\mathcal{I} \in \mathbb{R}^{H \times W \times C}$, the VGGT encoder outputs a semantic embedding $f \in \mathbb{R}^{H \times W \times d_s}$, where $d_s = 128$. For computational efficiency, we preprocess the entire dataset prior to training.

Distilling Semantics into Radiance Fields. We learn a semantic field $f_s : \mathbb{R}^3 \mapsto \mathbb{R}^{d_s}$, which maps a 3D point $\mathbf{x}$ to visual-geometry features $f_s(\mathbf{x})$ alongside a semantic field $f_l : \mathbb{R}^3 \mapsto \mathbb{R}^{d_l}$ that maps 3D points to the shared image-language embedding space of CLIP. Note that d_s and d_l denote the dimension of the embedding space of VGGT and CLIP, respectively, which varies depending on the specific VGGT head and CLIP model, e.g., 128 for VGGT's depth/point head and 768 and 1024 for CLIP's ViT and ResNet models. We train a semantic field for CLIP features to enable downstream open-vocabulary tasks. For effective co-supervision of both semantic fields, the VGGT and CLIP semantic fields share the same hashgrid encodings (i.e., base semantics), associating their semantic embeddings with the same visual and geometric features, illustrated in the left side of Figure 2.

Given a dataset $\mathcal{D}$ of images and associated camera poses, we render images in the semantic space by back-projecting RGB images using the estimated depth from the radiance field to reconstruct a point cloud in the local camera frame. We subsequently compute the VGGT and CLIP semantic features for the constituent points with f_s and f_l, respectively.

During training, we optimize the parameters of f_s and f_l simultaneously with the visual attributes of the radiance field using the loss function: $\mathcal{L} = \mathcal{L}_r + \sum_{\mathcal{I} \in \mathcal{D}, c \in \{s,l\}} \|I_{f,c} - \hat{I}_{f,c}\|_F^2 - \sum_{\mathcal{I} \in \mathcal{D}, c \in \{s,l\}} \text{csim}(I_{f,c} - \hat{I}_{f,c})$ where $\mathcal{L}_r$ denotes the RGB loss components of the base radiance field, $I_{f,c}$ and $\hat{I}_{f,c}$ denote the ground-truth and rendered semantic features, respectively, with s and l denoting the spatial and language components, and csim represents the cosine-similarity function. Although the Frobenius-norm term is not strictly required in the loss function, we retain it to improve the numerical stability of the cosine-similarity term, which is undefined for vectors of zero norm. In our experiments, we apply the same distillation procedure to train DINO-embedded distilled radiance fields.

Visualizing the Distilled Semantic Features. We examine the content of the distilled semantic features extracted by the DINO and VGGT backbones via principal component analysis (PCA),resulting in three-dimensional features which we visualize as images. Concretely, given a matrix $\hat{I}_f \in \mathbb{R}^{m \times d}$ of d-dimensional semantic features, we compute the singular value decomposition (SVD) of A to obtain the tuple (U, S, V), such that: $\hat{I}_f = U \Sigma V^T$, where $U \in \mathbb{R}^{m \times p}$, $V \in \mathbb{R}^{n \times p}$, and $\Sigma \in \mathbb{R}^{p \times p}$ denotes the left , right , and singular values, respectively. In practice, we compute the low-rank SVD for computational efficiency. Subsequently, we map the semantic features in I_v to the RGB image space using the first-three principal components via the transformation: $I_v = \hat{I}_f V_{[:,:3]}$, with $I_v \in \mathbb{R}^{m \times 3}$, which is reshaped into a 3-channel 2D image for visualization (i.e., $I_v \in \mathbb{R}^{W \times H \times 3}$).

To quantify the geometric content of the distilled semantic features, we introduce the *geometric fidelity factor* (GFF), which captures the edge information present in the semantic features relative to the physical scene, as determined by the RGB image. To do so, we apply the Sobel–Feldman operator (Duda & Hart, 1973) to the RGB image, which approximates the gradient of the image intensity through simple convolutions with the image using 3x3 kernels. The Sobel–Feldman operator suffices for this use-case, although more sophisticated gradient estimators could be applied in more complex problems. We apply hard-thresholding to the norm of the gradients to produce edges at varying resolutions. We post-process the semantic image I_v and RGB image to obtain the binary edge masks $I_{e,\text{sem}} \in \mathbb{R}^{W\times H\times 3}$ and $I_{e,\text{rgb}} \in \mathbb{R}^{W\times H\times 3}$, respectively. After extracting edges from the semantic and RGB images, we compute the GFF using:

$$\text{GFF} := \sum_{(i,j,k)} I_{e,\text{sem}}[i,j,k] / \sum_{(i,j,k)} I_{e,\text{rgb}}[i,j,k], \tag{1}$$

representing the fraction of edges retained by the distilled features. We examine the relative geometric content of the semantic features at different gradient thresholds in Section 7.2.

5 Semantic Localization

The distilled semantic features enable open-vocabulary object localization within the radiance field given a query: e.g., "find me a mug." For semantic localization, we compute the semantic embeddings of the language query ϕ_{query} using CLIP and subsequently compute the cosine similarity between the query and all points in the radiance field to identify candidate matches. For increased robustness, we utilize the semantic relevancy score (Kerr et al., 2023) given by: $\nu(\phi_{\text{query}}) = \min_i \frac{\exp(\phi_{\text{query}}\cdot\phi_f)}{\exp(\phi_f\cdot\phi^i_{\text{canon}})+\exp(\phi_{\text{query}}\cdot\phi_f)}$, where ϕ_f represents the rendered semantic embeddings from the radiance field and ϕ_{canon} represents the semantic embedding of *canonical* prompts, i.e., generic or negative prompts to better distinguish between confident localization matches from non-confident ones. We use canonical prompts such as: "object," "stuff," and "things." The semantic relevancy score can be viewed as the pairwise softmax over a set of positive and negative queries with respect to the rendered semantic embeddings. For conservative results, we take the minimum over all pairwise softmax distributions to define the semantic relevancy score.

To evaluate semantic localization accuracy, we map the ground-truth segmentation mask to the Euclidean space, assigning the value zero to pixels outside the map and one otherwise. Likewise, we normalize the semantic relevancy score generated by the radiance field at each view to values in $[0, 1]$ to obtain a relevancy mask. Afterwards, we map the ground-truth mask and the relevancy mask to the RGB space using a colormap. Thereafter, we compute the objective localization accuracy using widely-used perceptual metrics, such as the structural similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS), and peak-signal-to-noise ratio (PSNR). We do not use metrics that require binary masks, e.g., mean intersection-over-union (mIoU), as these metrics would require the selection of a similarity threshold for each rendered semantic relevancy image. Identifying an optimal similarity threshold for DINO and VGGT features individually presents computational challenges, especially since the optimal values are generally view-dependent. Moreover, approximating the optimal thresholds could lead to confounding results.

6 Inverting Radiance Fields

We explore the application of visual-geometry semantics to the inversion of radiance fields. The inverse problem is particularly challenging compared to the well-posed forward problem of image rendering in radiance fields, especially without any simplifying assumptions. In fact, existing methods (Yen-Chen et al., 2021; Chen et al., 2025) struggle with camera pose estimation in radiance fields without a good initial guess. As a result, we introduce SPINE, a novel algorithm for inverting radiance fields using distilled semantics for camera pose recovery without an initial guess. Moreover, SPINE enables us to comprehensively evaluate the relative performance of spatially-grounded features in radiance field inversion problems. Using a semantics-conditioned inverse model, SPINE directly maps semantic embeddings to a distribution over camera poses, which is subsequently refined to compute high-accuracy pose estimates using robust optimization, described in the following discussion.

Learning an Inverse Model. SPINE learns a neural field $p_\psi : \mathbb{R}^d \mapsto \mathcal{P}$ which maps semantic (image) embeddings $f(\mathcal{I}) \in \mathbb{R}^d$ to a distribution over candidate poses, where $\mathcal{P}$ denotes the space of valid distributions. For VGGT, we use the camera embeddings as input to p_ψ; whereas for DINO, we use the class token as input. We decompose the camera pose $P \in \mathrm{SE}(3)$ into its translation $\mathbf{t} \in \mathbb{R}^3$ and orientation $\mathbf{R} \in \mathrm{SO}(3)$. Note that optimizing over the space of orientations is non-trivial, given that the orthogonality constraint in $\mathrm{SO}(3)$. To circumvent this challenge, SPINE parameterizes the camera orientation using the corresponding Lie algebra $\mathfrak{so}(3)$, the vector space of three-dimensional skew-symmetric matrices. Leveraging the isomorphism between $\mathfrak{so}(3)$ and $\mathbb{R}^3$, we represent the camera rotation by $\mathbf{r} \in \mathbb{R}^3$. Note that we can construct a skew-symmetric matrix from $\mathbf{r}$ and subsequently map elements of $\mathfrak{so}(3)$ to $\mathrm{SO}(3)$ using the exponential map, i.e., $\exp : \mathfrak{so}(3) \mapsto \mathrm{SO}(3)$.

We jointly optimize the parameters of the GMM and the visual and semantic attributes of the radiance field, detaching the gradients between both fields to simultaneously learn the forward and inverse maps of the radiance field without compromising visual fidelity. Moreover, we make no additional assumptions beyond those made by the underlying radiance field and train SPINE entirely on the same inputs as the radiance field using the mean-squared-error (MSE).

Camera pose estimation. Given a query image, we compute the semantic embedding of the image, which is mapped to $\mathcal{P}$ using p_ψ. In general, the estimated camera pose is not always of sufficiently high accuracy. Consequently, we render images from the radiance field at the estimated coarse camera pose, using novel-view synthesis to generate an RGB-D image. Subsequently, we match image features from the query image to the rendered image, associated to a point-cloud in the radiance field through the rendered depth. Given the set of corresponding matches $\mathcal{C}$, we solve the perspective-n-point (PnP) problem to refine the coarse pose estimates: $(\hat{\mathbf{t}}, \hat{\mathbf{R}}) = \arg\min_{\mathbf{t}\in\mathbb{R},\mathbf{R}\in\mathrm{SO}(3)} \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{C}} \|\mathbf{p} - \mathbf{K}(\mathbf{R}\mathbf{q} + \mathbf{t})\|_2^2$, where $(\mathbf{p}, \mathbf{q})$ denote a corresponding pair with homogeneous image coordinates p (in the query image) and 3D point q (in the radiance field), $\mathbf{K}$ denotes the camera intrinsic matrix, and $\hat{\mathbf{t}}$ and $\hat{\mathbf{R}}$ represent the estimated camera translation and rotation, respectively. For robustness to spurious correspondences, we solve the PnP problem using RANSAC, mitigating the effects of outliers on the estimated pose.

We evaluate the accuracy of the pose estimates using the $\mathrm{SE}(3)$ error, composed of the translation and rotation error between the ground-truth and estimated camera poses. We compute the translation error directly as the ℓ_2 norm of the difference between the ground-truth and estimated translation and leverage the trace property of rotation matrices: $\mathrm{trace}\,(\mathbf{R}) = 1 + 2\cos(\theta)$ to compute the smallest rotation angle required to align the ground-truth and estimated camera orientations, where θ denotes the angle associated with rotation matrix $\mathbf{R}_\delta$. Specifically, the absolute rotation error θ is given by: $\theta = \arccos\left(\frac{\mathrm{trace}(\mathbf{R}_\delta)-1}{2}\right)$, where $\mathbf{R}_\delta = \hat{\mathbf{R}}^\mathsf{T}\mathbf{R}_{\mathrm{gt}}$ denotes the relative rotation matrix between the the ground-truth and estimated rotation matrices, $\hat{\mathbf{R}}$ and $\mathbf{R}_{\mathrm{gt}}$, respectively.

7 Experiments

We examine the performance of visual-geometry semantic features compared to visual-only features in distilled radiance fields. Via extensive experiments, we explore the following questions, spanning the core applications of distilled radiance fields in robotics: (i) *Do visual-geometry semantic features contain higher-fidelity spatial content?* (ii) *Does geometry-grounding improve semantic object localization?* (iii) *Can visual-geometry features enable higher-accuracy radiance field inversion?* The full results for all datasets are provided in the Appendix.

7.1 Evaluation Setup

We discuss the evaluation setup briefly, and provide additional details in Appendix A.2. We evaluate the visual-only semantics from DINOv2 and DINOv3 , and visual-geometry features from VGGT in nine scenes from three benchmark datasets, namely: *Ramen*, *Teatime*, and *Waldo_kitchen* in the LERF dataset (Kerr et al., 2023); *Bed*, *Covered Desk*, and *Table* in the 3D-OVS dataset (Liu et al., 2023); and *Office*, *Kitchen*, and *Drone* in the robotics dataset (Shorinwa et al., 2025). For each scene, we train a semantic GS and NeRF representation and compute the following metrics across 100 camera poses: geometric fidelity factor (GFF) metric for semantic content analysis (Section 4); SSIM, PSNR,

and LPIPS for semantic object localization and radiance field inversion (Section 5); and rotation and translation error in degrees and meters, respectively (Section 6).

7.2 Semantic Content of Distilled Features

As described in Section 4, we project the distilled semantic features into a three-dimensional subspace using the first-three principal components to aid visualization. Figure 3 shows the PCA visualization of the semantic features in the *Teatime* scene, highlighting the object-level composition of the scene. In the top row in Figure 3, we observe that the DINOv2 and DINOv3 features for the bear and the sheep are strongly distinct from the table and chairs, underscoring its focus on object-level decomposition. In contrast, VGGT features emphasize the geometric details of the scene, evidenced by the prominent edges of the bear, sheep, table, and chair, although some object-level features are visible. Likewise, in the bottom row, we see that VGGT highlights the structure (outline) of the mug and plate, unlike DINOv2 and DINOv3, although the DINO-based features provide more detailed object-level information. However, the wood grain on the table surface are more pronounced in the DINO-based features compared to those of VGGT. These findings suggest that visual-geometry semantic features provide detailed geometric information compared to visual-only features; however, visual-only features may provide more consistent object-level semantics (e.g., object class). We observe similar findings in the semantic content of other scenes, discussed in Appendix A.3.

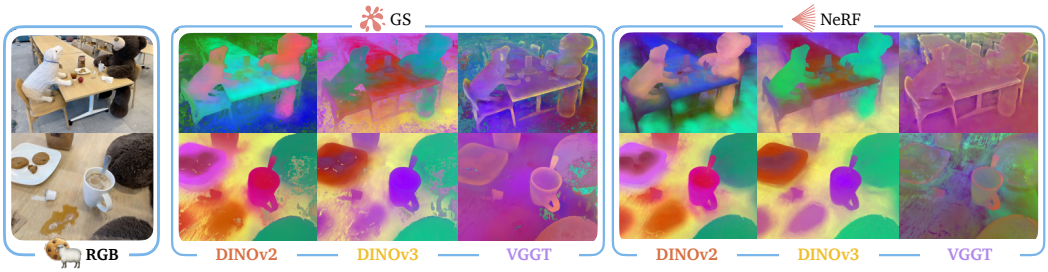

Figure 3: **Semantic content of distilled features.** Whereas visual-only features provide object-level information, visual-geometry features provide more structural details, such as an object's contour.

Next, we use GFF Equation (1) to quantitatively assess the geometric content of distilled features. We apply the Sobel-Feldman filter to the semantic images and extract the edges contained in these images at different resolutions, by varying the threshold of the edge gradient. In Figure 4, we visualize the results for the *Teatime* scene with thresholds of 0.1 and 0.3. Even at the lowest threshold of 0.1, we observe more prominent geometry in the VGGT features Increasing the gradient threshold leads to an overall decrease in the number of edges contained in the spatially-grounded and visual-only features. However, VGGT still provides the most structural content. Surprisingly, we see that DINOv2 provides stronger geometric information compared to DINOv3 in this scene, although not in all scenes.

We aggregate the quantitative results for all scenes and plot the GFF against gradient thresholds in Figure 4. For GS, we see that VGGT's features have the most edges at lower gradient thresholds, with DINOv2's features having the least, consistent with our qualitative observations. Moreover, we observe that the GFF of DINOv2 and DINOv3 remains almost constant across different thresholds, suggesting a lack of diversity in their geometric content, unlike VGGT. For NeRFs, the GFF remains similar across all semantic features, with DINOv3 having fewer edges at higher thresholds.

7.3 Semantic Object Localization

We examine the performance of spatially-grounded vs. visual-only features in semantic object localization using the procedure described in Section 5. In each scene, we use CLIP to encode the natural-language queries and subsequently generate the continuous relevancy mask. We use GroundingDINO (Liu et al., 2024) and SAM-2 (Ravi et al., 2024) to annotate the ground-truth segmentation mask, used in computing the segmentation accuracy metrics: SSIM, PSNR, and LPIPS.

Figure 5 summarizes our results. We find no significant difference in the localization accuracy of visual-only vs. visual-geometry features across GS and NeRF, suggesting that both semantic features

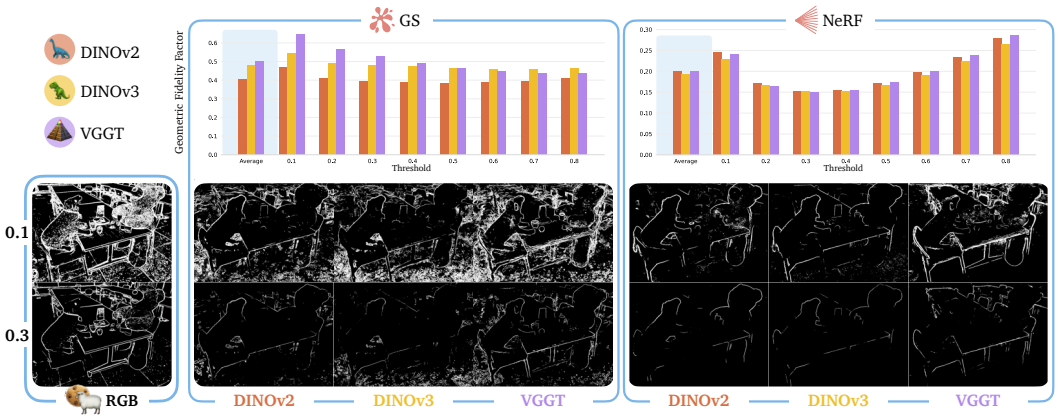

Figure 4: **Geometric fidelity factor (GFF) of visual-geometry and visual-only features.** VGGT's features contain prominent object edges, unlike visual-only semantic features.

are effective in co-supervising CLIP for open-vocabulary localization. However, we observe marginal degradation in performance with geometry-grounded features (VGGT). In addition, we visualize the ground-truth RGB and segmentation mask and the relevancy masks in the *Teatime* scene, highlighting the effectiveness of both kinds of semantic features in localizing the cookies, sheep, and bear. We provide additional results in Appendix A.4.

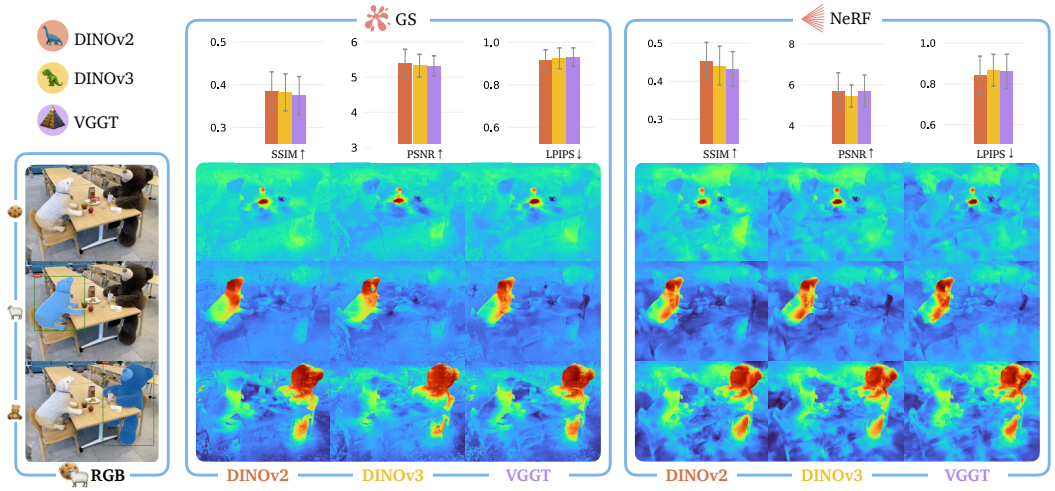

Figure 5: **Semantic object localization.** Both visual-only features (DINOv2/DINOv3) and visual-geometry features (VGGT) achieve similar localization accuracies (*Teatime* scene visuals).

7.4 Inverting Radiance Fields

We evaluate the accuracy of visual-only and spatially-grounded features in radiance field inversion. Surprisingly, we find that visual-geometry features underperform visual-only features, summarized in Figure 6. Specifically, DINOv2 achieves the lowest rotation and translation errors, while VGGT computes the least accurate pose estimates. These results suggest that existing methods for geometry-grounding may degrade the versatility of semantic features as general-purpose image features, constituting an interesting area for future work.

Further, we compare SPINE to existing baseline methods for radiance field inversion. Particularly, we compare DINOv2-based SPINE with (Chen et al., 2025) and (Yen-Chen et al., 2021) for pose estimation in GS and NeRFs, respectively. Since the baselines require an initial guess, we assess the performance of the baselines across two initialization domains, defined by the magnitude of the initial

rotation and translation error, R_{err} and T_{err}, respectively: (ii) low initial error with $R_{\text{err}} = 30\,\text{deg}$, $T_{\text{err}} = 0.5\text{m}$, and (iii) medium initial error with $R_{\text{err}} = 100\,\text{deg}$, $T_{\text{err}} = 1\text{m}$. We reiterate that SPINE does not utilize any initial guess.

Figure 6 summarizes the results. We observe that the baselines struggle without a good initial guess. Unlike these methods, SPINE computes more accurate pose estimates using semantics in the coarse phase, without any initial guess. Moreover, via photometric optimization, SPINE improves the accuracy of the coarse estimates. However, we note that the success of fine inversion depends on the optimizer used.

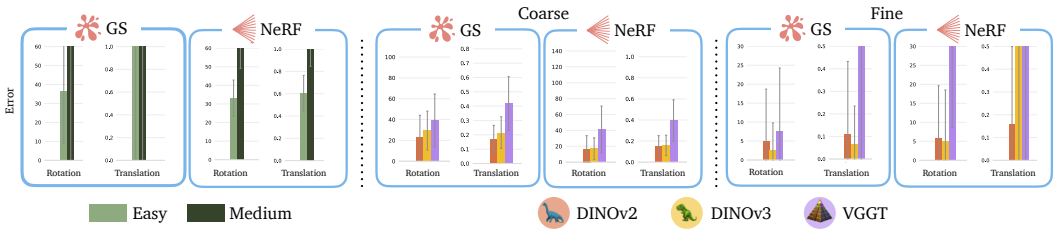

Figure 6: **Inverting radiance field.** (Left) Existing RF inversion methods struggle in the absence of a good initial guess. (Center) In contrast, SPINE computes better coarse pose estimates using the distilled semantics. (Right) Via photometric optimization, SPINE refine the coarse estimates for higher accuracy.

8 Conclusion

We explore the relative performance of visual-geometry semantic features compared to visual-only features in distilled radiance fields, across three critical areas for state-of-the-art robot policies. First, our work examines the semantic content of these image features, finding that spatially-grounded features provide more prominent geometric detail compared to visual-only features, which could be useful in fine-grained robotics task, e.g., dexterous manipulation. Second, we evaluate the accuracy achieved by these features in open-vocabulary semantic localization, and observe no significant difference in their performance. Third, we derive a novel method for inverting radiance fields and compare the performance of visual-only and visual-geometry features in this task. We demonstrate the effectiveness of our method compared to existing baselines, without requiring an initial guess, unlike the baselines. Moreover, we find that visual-only features outperform spatially-grounded features in radiance field inversion.

9 Limitations and Future Work

Self-Supervised Geometry-Grounding. Our findings suggest that existing solutions for geometry-grounding may impair the versatility of pretrained features, e.g., in radiance field inversion. We believe that this limitation may be due to the fully-supervised approach used by prior work. Future work will explore self-supervised approaches for spatial-grounding to eliminate inductive biases, improve adaptability, and enable larger-scale pre-training.

Synergy between Geometry and Vision. In addition, our experiments revealed that visual-geometry semantics did not improve the semantic object localization accuracy, despite its more significant structural content, suggesting the lack of sufficient synergy between the geometric and visual contents. Future work will introduce more effective strategies for establishing synergy between the geometry-oriented and visual-oriented semantic features for more robust scene understanding.

Efficient Inference. Existing geometry-grounded vision backbones require notable compute overhead compared to ungrounded backbones, amplified by the absence of lightweight variants. Future work will examine more efficient architectures for spatially-grounded vision backbones to enable their use in real-time applications, e.g., in robot manipulation.

REFERENCES

Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022.

Timothy Chen, Preston Culbertson, and Mac Schwager. Catnips: Collision avoidance through neural implicit probabilistic scenes. *IEEE Transactions on Robotics*, 40:2712–2728, 2024.

Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. *IEEE Transactions on Robotics*, 2025.

Richard O Duda and Peter E Hart. Pattern classification and scene analysis. *A Wiley-interscience publication*, 1973.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

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.

Justin Kerr, Letian Fu, Huang Huang, Yahav Avigal, Matthew Tancik, Jeffrey Ichnowski, Angjoo Kanazawa, and Ken Goldberg. Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In *6th annual conference on robot learning*, 2022.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19729–19739, 2023.

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.

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.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision*, pp. 349–366. Springer, 2024.

Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. *arXiv preprint arXiv:2209.09050*, 2022.

Jonathan Michaux, Seth Isaacson, Challen Enninful Adu, Adam Li, Rahul Kashyap Swayampakula, Parker Ewen, Sean Rice, Katherine A Skinner, and Ram Vasudevan. Let's make a splan: Risk-aware trajectory optimization in a normalized gaussian splat. *IEEE Transactions on Robotics*, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

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.

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 (ICML)*, pp. 8748–8763. PMLR, 2021.

Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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

William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

Ola Shorinwa, Jiankai Sun, and Mac Schwager. Fast-splat: Fast, ambiguity-free semantics transfer in gaussian splatting. *arXiv preprint arXiv:2411.13753*, 2024a.

Ola Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe Kennedy III, and Mac Schwager. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. *arXiv preprint arXiv:2405.04378*, 2024b.

Ola Shorinwa, Jiankai Sun, Mac Schwager, and Anirudha Majumdar. Siren: Semantic, initialization-free registration of multi-robot gaussian splatting maps. *arXiv preprint arXiv:2502.06519*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–12, 2023.

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

Thomas Weng, David Held, Franziska Meier, and Mustafa Mukadam. Neural grasp distance fields for robot manipulation. *arXiv preprint arXiv:2211.02647*, 2022.

Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1323–1330. IEEE, 2021.

Tenny Yin, Zhiting Mei, Tao Sun, Lihan Zha, Emily Zhou, Jeremy Bao, Miyu Yamane, Ola Sho, and Anirudha Majumdar. Womap: World models for embodied open-vocabulary object localization. In *Proceedings of the Conference on Robot Learning*, 2025.

Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on robot learning*, pp. 284–301. PMLR, 2023.

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*, pp. 15838–15847, 2021.