# DIV-FF: Dynamic Image-Video Feature Fields For Environment Understanding in Egocentric Videos

Lorenzo Mur-Labadia Josechu Guerrero Ruben Martinez-Cantin
I3A - Universidad de Zaragoza
{lmur, jguerrer, rmcantin}@unizar.es

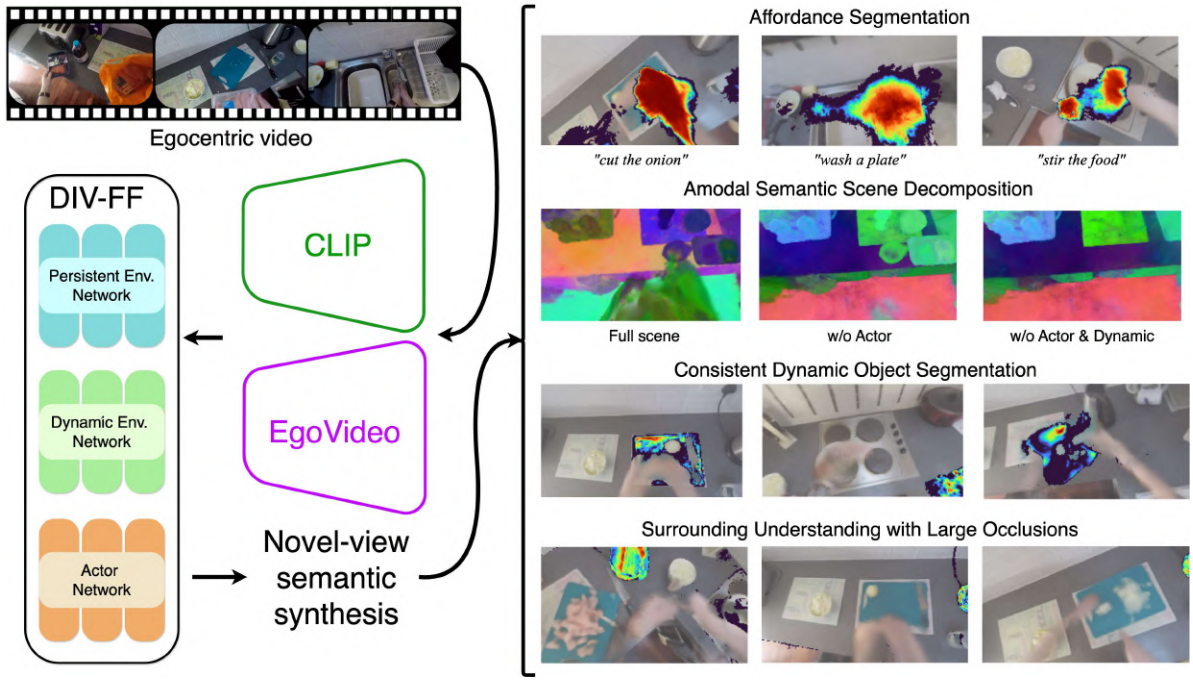


Figure 1. DIV-FF distills image and video language features in a triple stream feature field tailored to egocentric videos with numerous interactions and camera wearer movements. Our approach achieves a deep understanding of the environment, supporting precise affordance segmentation, semantic scene decomposition and consistent segmentation of dynamic objects. With its implicit 3D representation, DIV-FF comprehends not just novel views but also surrounding areas.

## Abstract

*Environment understanding in egocentric videos is an important step for applications like robotics, augmented reality and assistive technologies. These videos are characterized by dynamic interactions and a strong dependence on the wearer's engagement with the environment. Traditional approaches often focus on isolated clips or fail to integrate rich semantic and geometric information, limiting scene comprehension. We introduce Dynamic Image-Video Feature Fields (DIV-FF), a framework that decomposes the egocentric scene into persistent, dynamic, and actor-based components while integrating both image and video-language features. Our model enables detailed segmentation, captures affordances, understands the surroundings and maintains consistent understanding over time. DIV-FF outperforms state-of-the-art methods, particularly in dynamically evolving scenarios, demonstrating its potential to advance long-term, spatio-temporal scene understanding.*

## 1. Introduction

Egocentric videos offer a unique way to understand human activities from a first-person perspective, benefiting applications like mobile robotics, augmented reality and assistive devices. In these videos, an actor continuously moves to interact sporadically with multiple dynamic objects in a static scene, breaking the usual rigid scene assumption. This tight

integration between objects, actions and the dynamic scene introduces both opportunities and challenges for environmental understanding from egocentric videos.

Most existing methods in egocentric environment understanding either consider a short video clip isolated from the physical space [9, 11, 26, 31, 49] or they provide a strong spatial representation but with low semantic understanding [24, 35, 45, 46]. However, when humans interact repeatedly in a fixed environment, we develop a physical and semantic model that integrates the spatial distributions of the elements around us, both *persistent* and *dynamic*. The semantics capture detailed information about *objects* and their *attributes* through natural language descriptions. Additionally, we encode the available action (i.e.: *affordance*) locations in the environment, linking physical zones of interaction to the likely activities they support. Besides, we dynamically update this persistent semantic model as we interact, recording the location and state of dynamic objects at every moment. In that sense, some approaches propose intermediate representations between a pure semantic understanding of the video without explicit representation and a pure geometrical representation. They adopt semantic topological maps [33], local environment state representations [34] or explicit representations [30, 38] for improving the environment understanding of the scene. In this work, we build an implicit (neural network) model that is able to jointly capture the geometry, appearance and semantic understanding encoded in the video, and enable predictions in novel-view points using Neural Radiance Fields (NeRFs).

NeRFs provide a compact implicit representation of the geometry and visual appearance of a scene [28]. The implicit representation of NeRFs can also be used for semantic encoding, supporting multiple applications like robot manipulation [19, 25], navigation [54], or scene editing [21, 44]. For example, Neural Feature Fusion Fields (N3F) [51] extends the NeRFs predictive capabilities in a teacher-student fashion, where a teacher model that predicts semantic features in image space is used to train a NeRF-like student to predict semantic features in 3D space. These semantic capabilities are further extended in Language Embedded Radiance Fields (LERF) [18], enabling natural language query in 3D locations by volume rendering CLIP embeddings. However, LERF assumes a rigid scene which limits its applicability to egocentric videos where the actor is interacting with the environment. Furthermore, semantic distillation is based on single-image semantic features (e.g., CLIP features) which do not capture the dynamic nature of actions or changing elements.

In this work, we propose DIV-FF (Dynamic Image-Video Feature Fields ), the first language embedded feature field capable of decomposing both the geometry and the semantics of the scene for the actor, and also for the persistent and dynamic elements via three different streams. While previous works focus on image-language embeddings, we also introduce video-language embeddings (based on EgoVideo [37]) to understand fine-grained action descriptions. This encodes the environment affordances, possible actions available in the environment for the actor, linking specific activities to physical zones where interactions are likely to occur. A parallel feature field, based on image-language features from CLIP, captures detailed information about objects and their attributes, categorizing them through natural language descriptions rather than fixed semantic tags, even from novel viewpoints. Its implicit representation, similar to NeRFs, ensures that even areas not visible from the egocentric camera remain strongly connected in the environment model. Although this environment model provides a persistent long-term representation, it is dynamically updated as the user interacts, enabling a precise record of the location and state of dynamic objects at every moment. Our main **contributions** are as follows:

- We distill video-language embeddings (from EgoVideo) to understand temporally dependent semantics, such as affordances (available actions), which single-image models like CLIP cannot capture.
- We propose an approach to adapt language embedded feature fields to dynamic egocentric videos by dividing the radiance and feature fields depending on whether they are from the actor, dynamic, or persistent elements.
- We present a robust image-language feature field enhanced by leveraging SAM masks, which also includes the temporal dependency and achieves a consistent segmentation of the dynamic objects over time.
- Our results demonstrate significant improvements in dynamic object (+40.5%) and affordance segmentation (+69.7 %) by using text query relevancy maps. Furthermore, our model effectively connects the egocentric view with the semantics of the surroundings and decomposes the scene into different levels.

## 2. Related works

**Egocentric environment understanding using geometric representations.** Some works that consider the physical layout build semantic explicit representations from videos of indoor scenes using visual SLAM systems. Rhinehart et al. [45] learn 2D maps with the functionality of different actions. Semantic MapNet [16] propose a birds-eye-view spatial memory for mapping, which is updated with recurrent neural networks to remember places visited in the past. Cartillier et al. [4] encode the egocentric frame, project its features, and then decode the semantic labels in a 2D map. Liu et al. [24] recognize and localize activities in an existing 3D voxel map from an egocentric video. The limitations in extracting the camera pose from egocentric video [36] due to the quick camera movements and motion blur have hampered the unification of 3D geometry and video under-

standing. Recently, the arrival of egocentric 3D datasets with camera poses [14, 30, 52, 56] and the improvement of 3D sensors like project ARIA [7] has unlocked the arrival of novel works. Plizzari et al. [38] track active objects through their appearance and spatial consistency in the 3D scene, even when they are out of view. Mur-Labadia et al. [30] extract 2D affordance segmentation maps to build later a point cloud of the environment encoding those labels. Tschernezki et al. [2] proposed a 3D-aware instance object tracking by keeping a long-term consistency. EgoLoc [27] extend episodic memory to 3D by estimating the relative 3D object pose to the user.

**Egocentric environment understanding without geometric representations.** Most egocentric video understanding works just consider a short time window of the video. Although these works obtain a remarkable semantic understanding in multiple tasks like action recognition [26, 49], object segmentation [47, 52], action forecasting [9, 31] or capturing activity threads [40], these approaches ignore the underlying physical space of the scene. Some approaches [33, 34, 43] extract environment-aware features via alternative representations that avoid the geometric reconstruction problems from SLAM in egocentric videos [36]. Ego-Topo [33] builds a topological map, where the nodes represent environment zones with a coherent set of interactions linked by their spatial proximity. EgoEnv [34] encodes the relative directions of the objects to the camera wearer in a local state vector, learning an environment-aware video representation. Ramakrishnan et al. [43] capture the inherent statistics of indoor environments to learn an environment predictive coding, which applies later for navigation.

**Dynamic Radiance Fields.** Neural Radiance Fields (NeRFs) [44] allow capturing and rendering complex 3D scenes from a set of multi-view posed images. Using an implicit function and via differentiable volume rendering, NeRFs map spatial coordinates and viewing directions to colors and densities. Early methods for rendering dynamic scenes [10, 22] use pre-trained motion segmentation methods to mask moving objects, guiding separate NeRFs networks to disentangle motion-based components. Liang et. al [23] leverage DINO features to identify salient foreground regions along spacetime, while Wu et al. [55] decouple moving objects from the static background in a self-supervised manner with two neural radiance fields. NeuralDiff [51] separates the static background, dynamic objects and the actor's body via inductive biases, obtaining a different implicit representation for each part of the scene. Recently, Zhang et al. [58] optimize 3D Gaussians to reconstruct the scene and track the 3D object motions from an egocentric video, but requiring pre-extracted hand-object interaction masks.

**Feature Distillation in NeRFs.** Several works extend radiance fields to integrate 2D semantic labels into the 3D space during the optimization [8, 48, 53, 59]. In contrast, the objective of 3D feature distillation methods [12, 21, 51, 57] is to transfer 2D image features from a teacher model (i.e, a self-supervised model like DINO [3]) into a 3D student neural renderer. Expanding on this, 3D language feature fields distill image-text CLIP features [42], enabling querying the 3D student with open-vocabulary text descriptions to obtain relevancy maps. LERF [18] fuses multi-scale patch-level CLIP features conditioned on the scale. N2F2 [1] addresses the need for evaluating the rendering at the different scales by learning a unified feature field, where the different semantic granularities are encoded in a high-dimensional feature space. LangSplat [41] adopts 3D Gaussians [17] and combines CLIP features with multi-scale SAM masks, improving the segmentation quality. EgoLifter [15] augments 3D Gaussian Splatting with instance features from egocentric videos, but it only reconstructs the static part of the scene by filtering out the actor and the dynamic objects.

## 3. Dynamic Image and Video Feature Fields (DIV-FF) from Egocentric Videos

Our approach is to build a language embedding feature field that decomposes the 3D representation in three components (persistent environment, dynamic environment and actor) for accounting the inherent dynamics present in egocentric videos. In addition, we incorporate a second modality stream of embeddings based on video-language models which can capture the action semantics only present in the video modality. Besides, we introduce a time-dependent module on the dynamic and actor stream, capturing the temporal evolution of the feature fields.

### 3.1. Dynamic Neural Radiance Fields

The geometry model [50] captures the dynamic scene by integrating three different radiance fields, illustrated in Figure 2. The *persistent environment network* predicts the color $c_k^p$ and density $\sigma_k^p$ at each point along a ray $r_k$, given a viewing position $g_t$ and unit-norm viewing direction $d_t$. Formally, it is defined as $(c_k^p, \sigma_k^p) = \text{MLP}^p(g_t r_k, d_t)$. To model the dynamic objects in the scene, a second *dynamic environment network* $(c_k^d, \sigma_k^d, \beta_k^d) = \text{MLP}^d(g_t r_k, z_t^d)$ estimates the density $\sigma_k^d$ and the color as a Gaussian distribution $\mathcal{N}(c_k^d, \beta_k^d)$, where $\beta_k^d$ represents the heteroscedastic aleatoric uncertainty associated with the color. It also includes as input a frame-specific code $z_t^d$ that accounts for temporal variations of the dynamic objects, which exhibit sporadic motion relative to the global reference frame. The *actor network* is similar to the dynamic environment network, but since the actor moves continuously linked to the camera, it removes the projection of the ray to the world coordinate system $(c_k^a, \sigma_k^a, \beta_k^a) = \text{MLP}^a(r_k, z_t^a)$. Here, $z_t^a$ is a frame-specific parameter designed to capture the continuous motion of the actor. The predicted material uncertainty

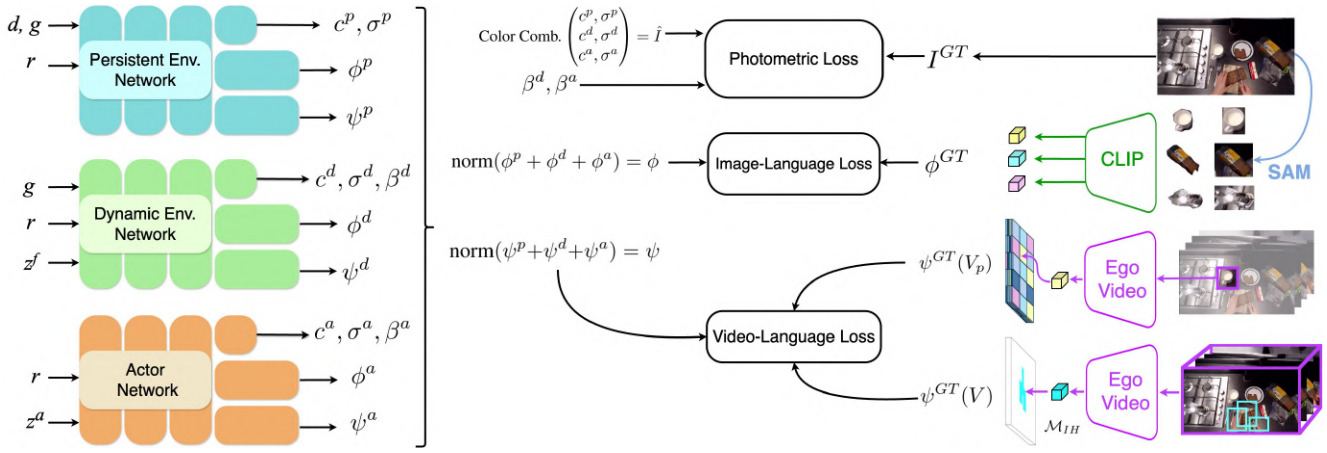


Figure 2. **Overview of DIV-FF.** Our three-stream architecture field predicts the color $c$, the density $\sigma$, the material aleatoric uncertainty $\beta$, the image-language features $\phi$ and the video-language features $\psi$ along a ray $r$ with direction $d$ given the camera viewpoint $g$ and a frame specific code $z$. We first extract SAM masks and bounding boxes from the image, that we leverage to obtain a unique CLIP descriptor $\phi_{GT}$ in all the pixels within the respective mask. We supervise the video-language feature field with local patch features $\psi^{GT}(V_p)$ and a global video embedding $\psi^{GT}(V)$ assigned only to pixels in the interaction hotspot $\mathcal{M}_{IH}$, computed with a pre-trained hand-object detector.

terms $\beta_k^d, \beta_k^a$ indicate the confidence levels associated with each ray $r_k$ for representing the dynamic objects and the actor, respectively. By employing improved color mixing techniques and inductive biases during training, the model accurately reconstructs scene dynamic geometry as a composite of the three radiance fields. For more details on the geometric model, please refer to [50].

### 3.2. Image-Language Feature Field

We extend the three-stream geometry model to distill image-language semantic features from CLIP [42]. Since the CLIP image encoder is a global image descriptor, it lacks pixel-aligned embeddings. To address this, LERF [18] extracts multi-scale patch-level features, which often fail to encompass the target object or add excessive contextual information. It results in blurred object boundaries and noise, requiring DINO [3] for regularization.

As shown in Figure 2, our model incorporates pixel-aligned CLIP features by leveraging accurate object masks generated by Segment Anything Model (SAM) [20] inspired by recent works [1, 41]. Specifically, we extract CLIP features per each segmented mask region $\phi_{\mathcal{M}}^{GT}$ and its respective bounding box $\phi_{\mathcal{B}}^{GT}$. We assign the same weighted descriptor $\phi^{GT} = 0.75 \cdot \phi_{\mathcal{M}}^{GT} + 0.25 \cdot \phi_{\mathcal{B}}^{GT}$ to all the pixels within the mask. This balanced approach achieves pixel-level alignment while preserving semantic context. Furthermore, the use of precise semantic masks with sharp object boundaries eliminates the need for DINO regularization used in previous works [18].

### 3.3. Video-Language Feature Field

While the CLIP image-language features contain fine-grained and accurate details of the objects, they ignore interaction semantics present in egocentric videos as they require temporal information. Therefore, we incorporate in parallel a video-language feature field to capture dynamic semantics, such as *affordances* and potential interactions. We leverage Video-Language Pre-trained (VLP) models [37, 39], which offer richer and action-oriented context by pairing narrative descriptions with video using contrastive learning. We select Ego-Video [37], the state-of-the-art in multiple Ego4D [13] challenges, for this task. Similar to CLIP, the video encoder of Ego-Video outputs a single descriptor from a video patch, not pixel-aligned features. In this case we cannot use object masks as in Section 3.2, because our goal is to identify *interaction hotspot* regions, including both the hands and the object parts (e.g. "knife edge", "spatula handle"), not just the entire object. While SAM's small masks could localize these areas, their limited size loses essential action context.

Therefore, we distill the video-language feature field with patch and global-level embeddings. We first precompute video descriptors $\psi^{GT}(V_p)$ from medium-sized video patches $V_p$, balancing fine-grained details with action context. Second, we derive a global descriptor $\psi^{GT}(V)$ for the entire video, assigned solely to the pixels within the interaction hotspot area $\mathcal{M}_{IH}$ [32]:

$$\mathcal{L}_V = \left\| \hat{\psi} - \psi^{GT}(V_p) \right\|^2 + \mathcal{M}_{IH} \left\| \hat{\psi} - \psi^{GT}(V) \right\|^2 \quad (1)$$

This improves the feature field's capability to capture relevant interaction regions. We obtain the interaction hotspot

| Method | S01 | S02 | S03 | S04 | S05 | S06 | S07 | S08 | S09 | S10 | Average mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LERF | 22.1 | 10.1 | 11.7 | 9.7 | 13.2 | 18.6 | 12.5 | 6.2 | 19.0 | 5.5 | 12.8 |
| NeuralDiff + OWL-ViT | 9.4 | 10.2 | 13.2 | 15.4 | 9.4 | 13.3 | 14.5 | 23.2 | 23.7 | 28.9 | 16.1 |
| NeuralDiff + OWL-ViT + SAM | 8.7 | 12.6 | 23.2 | 23.9 | 13.8 | 15.9 | 17.8 | **28.0** | **32.9** | **41.1** | 21.7 |
| DIV-FF (CLIP in patches) | 26.9 | 21.7 | 18.3 | 16.8 | 18.1 | 24.9 | 17.3 | 12.3 | 17.9 | 23.6 | 19.8 |
| DIV-FF (CLIP in SAM masks) | 30.7 | 19.3 | 29.6 | 24.9 | **31.3** | 26.1 | 28.8 | 14.8 | 23.8 | 35.1 | 26.2 |
| DIV-FF (full model, video infer.) | 16.1 | 15.4 | 9.3 | 9.5 | 21.8 | 20.7 | 10.7 | 18.5 | 17.9 | 20.6 | 16.6 |
| DIV-FF (full model, image infer.) | **40.3** | **30.4** | **37.4** | **29.8** | 29.5 | **32.6** | **30.6** | 15.1 | 25.1 | 33.6 | **30.5** (+40.5%) |

Table 1. **Dynamic Object Segmentation by CLIP image-language feature field.** Compared with LERF, DIV-FF considers a dynamic scene in the geometric reconstruction. Our full model assigns the same descriptor to all the pixels within a SAM mask. This descriptor is a weighted average between the CLIP of the mask and the bounding box. We compute relative improvement against the best baseline model.

mask $\mathcal{M}_{IH}$ as the union of the hands and active objects bounding box, pre-extracted with an existing hands-object detector [47]. Additionally, the training of this feature field is regularized with pixel-aligned DINO [3] features thanks to its object decomposition properties [18].

## 4. Experimental Settings

**Implementation details.** We extend the three stream architecture of NeuralDiff [50] by incorporating 4-layer, 256-width MLPs for the image $\phi$ and video language $\psi$ feature fields, respectively. Both the coarse and fine models use 64 samples, while we select the best 32 samples for the feature distillation. The representations are summed and normalized post-rendering. We use an Adam optimizer with a learning rate of 5 $\times 10^{-4}$ and a cosine annealing scheduler. We train the geometry for 10 epochs with a batch size of 1024, then distill semantic features in two phases: training only the semantic heads for 5,000 iterations, followed by the full model for 3 epochs on an NVIDIA 4090.

**Feature extractors.** To extract the CLIP image embeddings, we utilize the OpenCLIP ViT-B/16 model [5] trained on the LAION-2D dataset following [18] for fair comparison. We prompt SAM [20] with a 32 $\times$ 32 grid, filtering redundant masks by 0.7 IoU, 0.85 stability score, and 0.7 overlap rate. We reduce the dimensionality of CLIP descriptors by training a scene-wise language auto-encoder [41] to reduce the memory cost. We sample 4 video frames at 60 fps with a temporal stride of 15. Local-patch video features are extracted using a patch size of 33% the image size and a stride factor of 0.5. For masking the global video descriptor, we employ a hand-object detector [47], specifically finetunned for egocentric sequences.

**Dataset.** We conduct our experiments on the EPIC-Diff subset [50] of the EPIC-Kitchens [6] dataset. On average, each sequence comprises of 900 calibrated frames, spanning 14 minutes of egocentric video, featuring multiple viewpoints and a large number of manipulated objects. Our evaluation encompasses both our method and the baselines on the test set, which includes frames not utilized during model training. This set facilitates assessments of new-view synthesis and segmentation capabilities.

**Baselines.** We compare against the following baselines:

- **LERF** [18] assumes a static scene, using a single stream for geometry and semantics. The image-language field is distilled from multi-scale patch-level CLIP features.
- **OWL** [29]. We apply this open-vocabulary object detector on the novel-view rendered images produced by Neural-Diff [50].
- **OWL+SAM** [20, 29] obtains the object's masks from the bounding box coordinates provided by the OWL baseline.

**Ablations.** We compare different versions of DIV-FF.

- **DIV-FF (CLIP in patches)** keeps the CLIP patch features from LERF $\phi^{GT} = \phi^{GT}_{\mathcal{P}}$, but it introduces the dynamic geometry model from [50].
- **DIV-FF (CLIP in SAM masks)** substitutes CLIP patch features by embeddings from SAM masks $\phi^{GT} = \phi^{GT}_{\mathcal{M}}$.
- **DIV-FF (full model, image inference)** incorporates the bounding box to obtain the CLIP descriptor $\phi^{GT} = 0.75 \cdot \phi^{GT}_{\mathcal{M}} + 0.25 \cdot \phi^{GT}_{\mathcal{B}}$, where $\phi^{GT}_{\mathcal{B}}$.
- **DIV-FF (full model, video inference).** In the full model, we render from the parallel video-language feature field.

## 5. Results

Once trained, our DIV-FF model predicts the color $c$, density $\sigma$, CLIP $\phi$ and EgoVideo $\psi$ semantic features of a novel-view in an specific time-step and separates the actor and the dynamic elements from the persistent environment. We evaluate this comprehensive spatio-temporal semantic understanding in different downstream tasks.

### 5.1. Dynamic Object Segmentation

In each scene, we identify a subset of objects that move throughout the video and evaluate in the novel-views the relevancy maps originated by the text queries in the $\phi_{img}$ image-language feature field. Following the method proposed by LERF [18], we compute the relevancy score as: $\min_i \frac{\exp(\phi_{img} \cdot \phi_{quer})}{\exp(\phi_{img} \cdot \phi_{quer}) + \exp(\phi_{img} \cdot \phi^i_{can})}$. This formula evaluates how closely the rendered embedding $\phi_{img}$ matches the query embedding $\phi_{quer}$ compared to a set of predefined

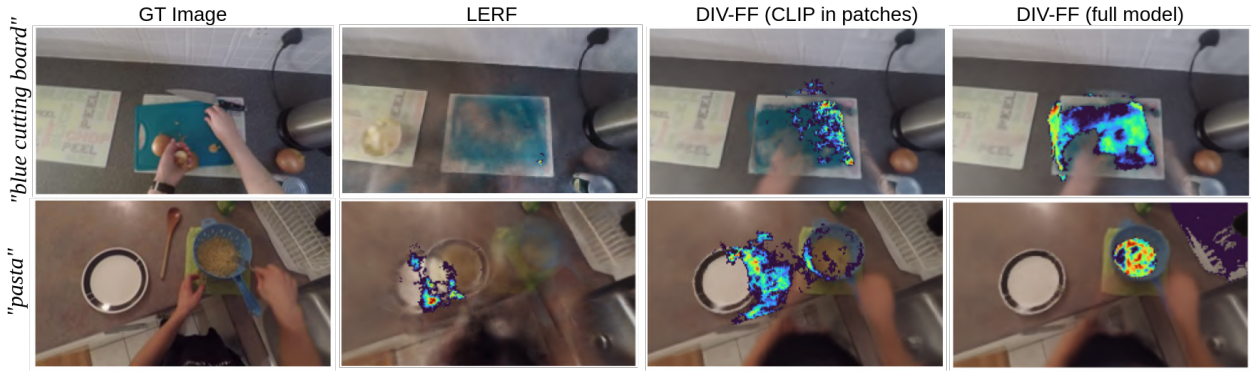


Figure 3. **Ablations on the image-language feature field.** Treating the egocentric video as a dynamic scene enhances geometric reconstruction, while utilizing SAM masks further improves object segmentation accuracy.

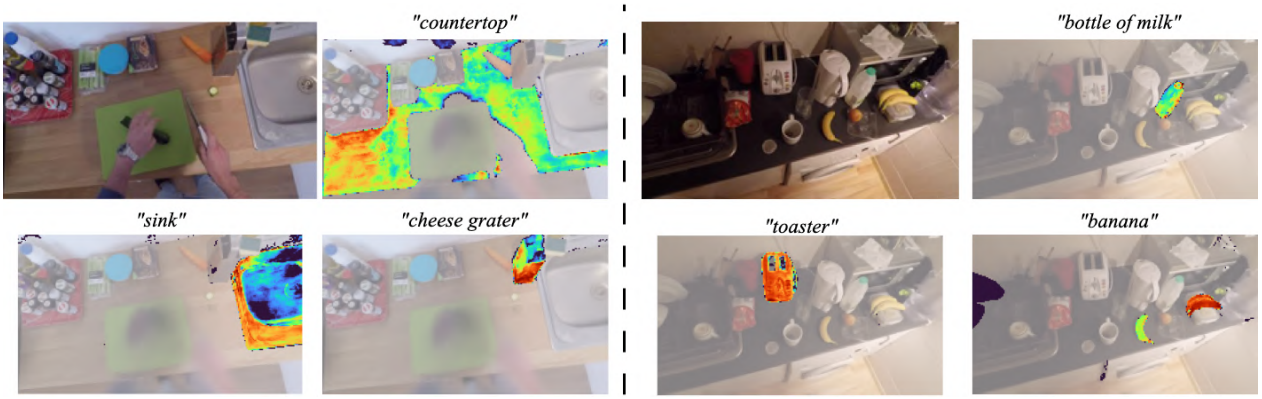


Figure 4. **DIV-FF Image-Language relevancy maps in novel-views.** We can see the performance of various text queries for dynamic object segmentation. We can see how the object contours are well defined as we used masks during training.

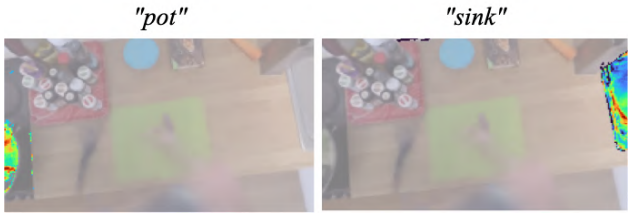


Figure 5. **Surrounding Understanding.** DIV-FF understands the novel view and the surrounding environment, enabling segmentation of objects at the image's edges with limited observability.

canonical phrases $\phi^i_{can}$ (*"object", "thing", "stuff", "texture", "hands"*). Segmentation masks are generated for relevance scores above a specified threshold. For the evaluation, we leverage existing annotations from [51] and report the mean intersection over union (mIoU). We visualize the text query relevancy maps by normalizing from 50 % to the maximun relevancy.

Table 1 presents quantitative results on EPIC-Diff scenes. The full version of DIV-FF achieves the best performance (30.5 mIoU), surpassing the OWL+SAM detector (21.7 mIoU) by +40.5%, illustrating that distilling semantic features outperforms traditional open vocabulary object detection from novel views, since the OWL model fails due to artifacts and the blurry hand effects in the novel view rendered. The CLIP patch-level version of DIV-FF (19.8 mIoU) significantly improves upon LERF by explicitly considering the dynamics parts in the semantic and geometric fields with the triple stream architecture of DIV-FF. This leads to sharper reconstructions, particularly for moving objects as shown in Figure 3. Subsequently, leveraging SAM to extract object-level CLIP features further improves performance (26.2 mIoU), and generates more accurate and consistent semantic renderings compared to CLIP patch-level embeddings. Finally, the introduction of contextual information from the object bounding boxes ultimately yields to the best performance (30.5 mIoU).

Figure 4 showcases novel-view renderings for various text queries in two scenes, effectively capturing fine-grained details like the *"countertop"* borders. The uniform assignation of the same CLIP descriptor across all object pix-


Figure 6. **Consistent Dynamic Object Segmentation along different time-steps in novel views:** The dynamic and actor streams contain respective frame-specific codes $z_t^f$ and$z_t^a$. This time encoding is also propagated to the semantic feature field, obtaining consistent segmentations despite the continuous movement of the *"spatula"* and *"blue cutting board"*.

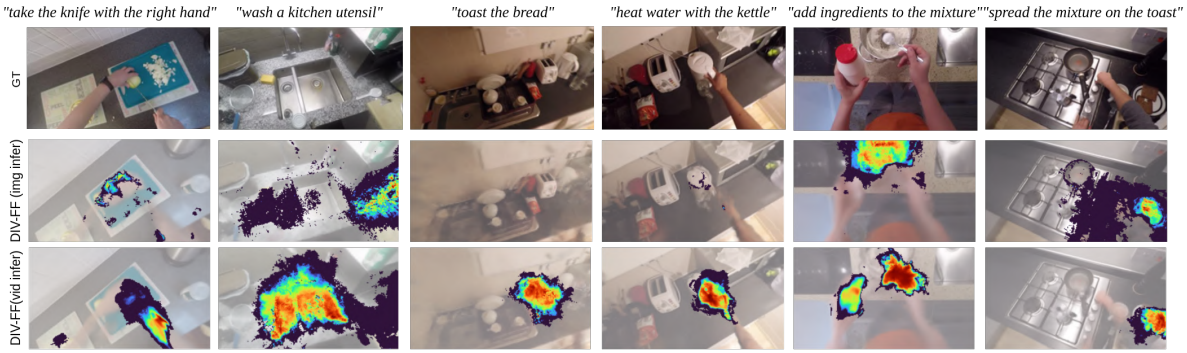


Figure 7. **Affordance Segmentation qualitative examples.** We compare the relevancy maps produced by the image-language field against those from the video-language field of DIV-FF, based on a detailed action description text query.

els allows DIV-FF to segment objects of any size, such as *"sink"* or *"banana"*. As Figure 6 shows, our approach also segments consistently dynamic objects across multiple novel views in different time-steps of the sequence due to the combined impact of object-level CLIP features and the temporal encoding in the frame-specific codes. Unlike egocentric methods limited to short time windows, our environment understanding extends beyond the current view to the surrounding regions. Figure 5 illustrates this capability, showing how our 3D semantic implicit model segments the *"pot"* and *"sink"*, despite being almost occluded in the edge of the image.

## 5.2. Affordance Segmentation

We identify affordable actions in each scene and generate Ego-Video [37] text queries $\psi_{quer}$ describing the interaction, which are more complex than simple object labels as they capture nuanced action dynamics. We compute the relevancy score from the video-text feature field $\psi$ against a different set of canonical phrases $\psi^i_{can}$ (*"general task", "indistinct movement", "unclear action", "background"*). We manually annotate affordance segmentation masks for five affordable actions per sequence, resulting $\approx 700$ masks. We report mIoU.

Table 2 demonstrates the effectiveness of video-language features in capturing actions. Previous methods that rely on single-image CLIP features miss the dynamic action context in egocentric videos. Consequently, the video-language feature field of DIV-FF excels in the affordance segmentation, achieving 20.7 mIoU (+69.7 %), benefiting from training on video narrations, unlike CLIP models that use static image captions. We visualize these differences in Figure 7, showing the relevancy scores for text queries detailing specific actions. The image-language model performs well when actions are explicitly linked to objects, such as *"cut the onion"* or *"add ingredients to the mixture"*. However, it struggles with verbs or semantic contexts that imply a location, like *"wash a kitchen utensil"*—which sug-

| Method | S01 | S02 | S03 | S04 | S05 | S06 | S07 | S08 | S09 | S10 | Average mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| OWL-ViT | 4.8 | 4.2 | 1.4 | 5.6 | 4.8 | 2.3 | 13.2 | 4.0 | 5.4 | 4.4 | 5.0 |
| OWL-ViT + SAM | 4.6 | 5.4 | 1.9 | 4.6 | 5.8 | 1.1 | 8.6 | 4.5 | 7.7 | 8.3 | 5.3 |
| LERF | 18.2 | 17.4 | 6.8 | 11.5 | 11.9 | 18.4 | 11.7 | 7.5 | 15.2 | 4.2 | 12.2 |
| DIV-FF (CLIP in patches) | 17.1 | 15.6 | 7.1 | 9.4 | 12.9 | 19.7 | 12.4 | 11.3 | 15.3 | 12.6 | 13.3 |
| DIV-FF (full m., image infer.) | 17.3 | 13.7 | 6.2 | 13.7 | 19.1 | 8.1 | 18.5 | 7.1 | 11.1 | 3.6 | 11.8 |
| DIV-FF (full m., video infer.) | **20.6** | **19.9** | **14.4** | **22.4** | **30.1** | **22.3** | **20.1** | **16.8** | **17.1** | **23.1** | **20.7** (+69.7%) |

Table 2. **Affordance Segmentation.** We compare the segmentation masks of a set of affordable actions in the scene. The full version of DIV-FF is composed by two parallel semantic feature fields, image (CLIP + SAM + boxes descriptors) and video (Ego-Video) respectively. We compute relative improvement against the best baseline model.

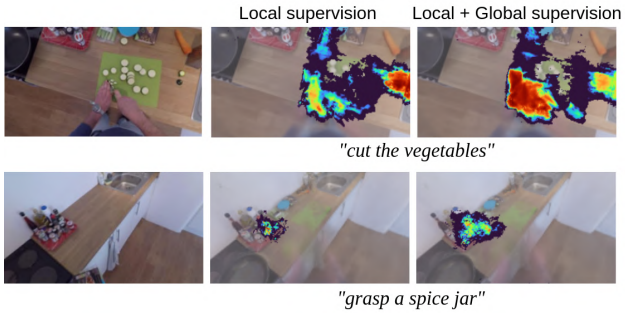


Figure 8. **Video-Language Loss ablation.** Including the global supervision term in the interaction hotspot mask produces sharper relevancy maps compared to just having the patch-level (local) term of the loss.

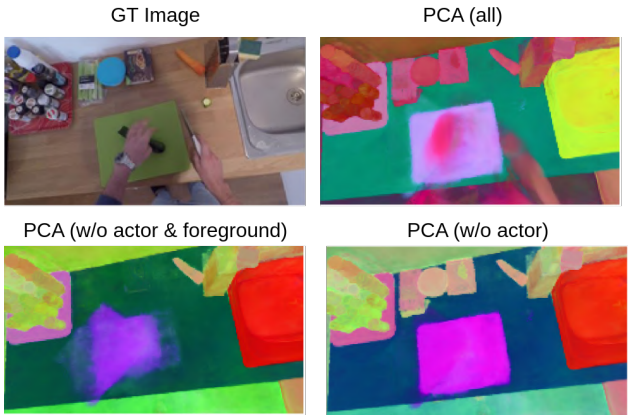


Figure 9. **Amodal Scene Understanding.** We visualize the PCA components obtained from the different composition of the image-text feature fields, showing accurate decomposition of the objects contours due to the SAM masks regularizing effect.

gests the sink—or *"toast the bread"*, associated with the toaster. In these instances, the video-language model distinctly outperforms, accurately identifying the action's interaction hotspot. We also highlight that the localization of these fine-grained areas is due to the additional global supervision to the local medium-size Ego-Video patch features, as Figure 8 shows. The joined effect of the two losses improves the relevancy maps by explicitly guiding the optimization toward the interaction hotspot regions. Table 2 also shows that for single-image models, patch-based methods (LERF, CLIP in patches) outperform the full model using object masks, as we suggested in Section 3.3.

### 5.3. Amodal Scene Understanding.

Our DIV-FF model comprises three distinct levels of geometry and semantics, representing different scene levels as illustrated in Figure 9. This introduces significant versatility in the environment understanding. For example, we can remove the actor's hands to reveal the dynamic objects without occlusions. Additionally, eliminating both the actor and dynamic elements exposes only the persistent parts of the scene. Our intuition is that this static spatial-semantic representation contains strong priors that can be exploited when the user revisits the scene at another time.

## 6. Conclusions

We proposed Dynamic Image-Video Feature Fields (DIV-FF) to address the limitations of existing egocentric video understanding methods. By decoupling the scene into persistent, dynamic, and actor streams and integrating video-based semantics, our approach achieves robust and consistent semantic segmentation over time. The model's ability to perceive and reason about both persistent and evolving scene elements marks a significant improvement in affordance and dynamic object understanding. Experimental results highlight DIV-FF's effectiveness in representing the rich and dynamic nature of egocentric environments, setting a promising direction for future work in spatial-temporal scene modeling and interaction-aware perception.

## Acknowledgments

This work was supported by projects PID2021-125209OB-I00, TED2021-129410B-I00, TED2021-131150B-I00 and Aragon Government DGA T45-23R.

## References


[1] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. *arXiv preprint arXiv:2403.10997*, 2024. 3, 4

[2] Yash Bhalgat, Vadim Tschernezki, Iro Laina, João F Henriques, Andrea Vedaldi, and Andrew Zisserman. 3d-aware instance segmentation and tracking in egocentric videos. *arXiv preprint arXiv:2408.09860*, 2024. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3, 4, 5

[4] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 964–972, 2021. 2

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 5

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 5

[7] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 3

[8] 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. 3

[9] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 6252–6261, 2019. 2, 3

[10] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 3

[11] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 2

[12] 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. 3

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 4

[14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 3

[15] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *European Conference on Computer Vision*, pages 382–400. Springer, 2025. 3

[16] Joao F Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018. 2

[17] 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. 3

[18] 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, 3, 4, 5

[19] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. *arXiv preprint arXiv:2409.18121*, 2024. 2

[20] 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*, pages 4015–4026, 2023. 4, 5

[21] 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, 3

[22] 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*, pages 6498–6508, 2021. 3

[23] Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, and James Tompkin. Semantic attention flow fields for monocular dynamic scene decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21797–21806, 2023. 3

[24] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity

recognition and localization on a 3d map. In *European Conference on Computer Vision*, pages 621–638. Springer, 2022. 2

[25] Haozhe Lou, Yurong Liu, Yike Pan, Yiran Geng, Jianteng Chen, Wenlong Ma, Chenglong Li, Lin Wang, Hengzhen Feng, Lu Shi, et al. Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation. *arXiv preprint arXiv:2408.14873*, 2024. 2

[26] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 2, 3

[27] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 45–57, 2023. 3

[28] 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. 2

[29] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[30] Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5238–5249, 2023. 2, 3

[31] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Josechu Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. *arXiv preprint arXiv:2406.01194*, 2024. 2, 3

[32] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 4

[33] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 2, 3

[34] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egoenv: Human-centric environment representations from egocentric video. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[35] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016. 2

[36] Suvam Patra, Kartikeya Gupta, Faran Ahmad, Chetan Arora, and Subhashis Banerjee. Ego-slam: A robust monocular slam for egocentric videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 31–40. IEEE, 2019. 2, 3

[37] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 2, 4, 7

[38] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. *arXiv preprint arXiv:2404.05072*, 2024. 2, 3

[39] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 4

[40] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13770–13779, 2022. 3

[41] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 3, 4, 5

[42] 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*, pages 8748–8763. PMLR, 2021. 3, 4

[43] Santhosh Kumar Ramakrishnan and Tushar Nagarajan. Environment predictive coding for visual navigation. *ICLR 2022*, 2022. 3

[44] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8931–8940, 2024. 2, 3

[45] Nicholas Rhinehart and Kris M Kitani. Learning action maps of large environments via first-person vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–588, 2016. 2

[46] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 2

[47] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 3, 5

[48] 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. 3

[49] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9954–9963, 2019. 2, 3

[50] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. Neuraldiff: Segmenting 3d objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*, pages 910–919. IEEE, 2021. 3, 4, 5

[51] 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. IEEE, 2022. 2, 3, 6

[52] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[53] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 3

[54] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 2

[55] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in neural information processing systems*, 35:32653–32666, 2022. 3

[56] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018. 3

[57] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *European Conference on Computer Vision*, pages 57–74. Springer, 2025. 3

[58] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. *arXiv preprint arXiv:2406.19811*, 2024. 3

[59] 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. 3