AViON4D: Audio-Visual Open-Vocabulary 4D Egocentric Scene Understanding

Anonymous OpenSUN3D submission

Paper ID 1

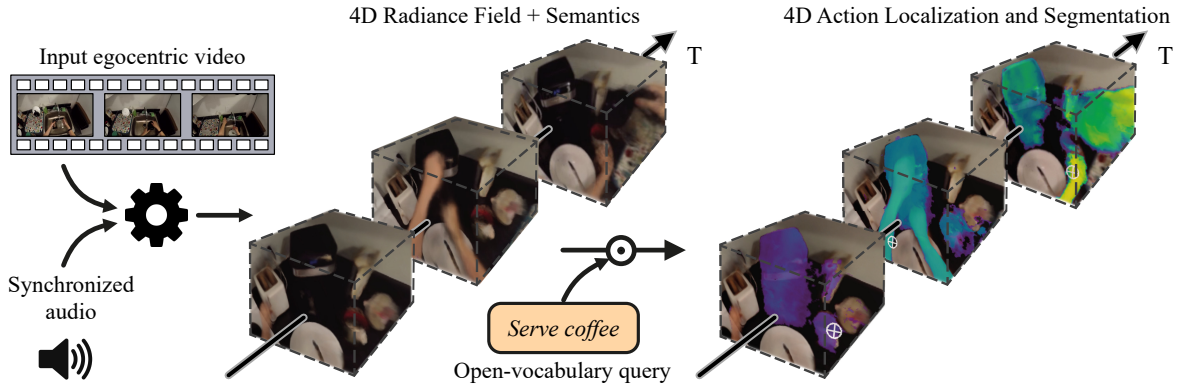

Figure 1. AViON4D lifts video-language features into a dynamic 4D radiance field and integrates them with audio-language features as a complementary temporal signal. From egocentric video with synchronized audio, our method enables open-vocabulary querying of temporal concepts like actions in 3D space and time. For the query "*serve coffee*", we render the 4D semantic radiance field and localize the spatiotemporal (3D + time) region with maximum relevancy for this action.

Abstract

Egocentric videos offer rich, first-person perspectives on human interactions with the world, enabling applications such as activity recognition and assistive technologies. While recent advances in neural radiance fields (NeRFs) and semantic distillation from vision-language models have enabled open-vocabulary 3D scene understanding, these methods are limited to static scenes and cannot encode the temporal semantics and dynamic human-world interactions essential for egocentric spatiotemporal understanding. We introduce AViON4D, the first multimodal NeRF-based method that extends semantic NeRFs from 3D to 4D for egocentric scenes. AViON4D lifts video features into 4D representations, replacing static image embeddings with temporally-aware features that encode action dynamics. To handle the extreme motion characteristic of egocentric video, we design an object-centric feature extraction strategy that maintains spatial and temporal coherence by tracking objects across frames. Furthermore, AViON4D incorporates audio-language features as a complementary temporal signal via late fusion, helping to disambiguate visually similar actions and refine temporal boundaries through distinctive acoustic cues. Extensive experiments on two complementary 4D open-vocabulary tasks demonstrate that AViON4D outperforms existing single-modal and static approaches by up to +6.0% for action localization and +14.90% for action segmentation. Moreover, for the first time, we show that audio cues from egocentric recordings not only enhance performance but also narrow the gap between single-frame and video-based models, offering a more robust and generalizable solution for 4D scene understanding in egocentric settings. Code will be made available upon acceptance.

1. Introduction

The increasing availability of wearable cameras has opened unprecedented opportunities to understand human-world interactions from a first-person perspective. Understanding the temporal dynamics in these egocentric videos is essential for applications like assistive technologies and embodied AI. However, egocentric videos are particularly challenging, characterized by extreme camera motion, frequent occlusions, and dynamic object interactions. Representing videos in 4D (3D space + time) captures this variability and consolidates cross-view information into consistent spatiotemporal representations, enabling novel view synthesis and global

scene reasoning.

Powerful 3D reconstruction methods like Neural Radiance Fields (NeRFs) [28] provide a foundation for building such representations. Recent work [6, 22, 27] has enriched NeRFs with semantic understanding by distilling vision-language models such as CLIP [34] into 3D representations, enabling open-vocabulary object localization and segmentation. However, these semantic NeRFs are constrained to static scenes and cannot capture temporal action semantics essential for egocentric understanding. While recent efforts [47, 52] attempt to tackle egocentric action understanding, they remain constrained to the 2D image plane. More recent works [7, 29] distill video features into 3D representations, but rely on features computed on fixed spatial grids, which are poorly suited to egocentric videos with extreme camera motion. While [7] only focuses on exocentric videos, [29], as a result, requires auxiliary modules like hand-crafted segmenters to compensate.

To overcome these limitations, we propose *AViON4D* (Audio Visual Open-vocabulary Nerf 4D), the first multimodal NeRF-based method for open-vocabulary 4D scene understanding from egocentric videos. Without requiring predefined action categories or dataset-specific training, AViON4D takes a natural language description of an action and localizes it in both 3D space and time. Successfully solving this task requires reasoning simultaneously about spatial location, temporal boundaries, and semantic evolution (e.g., "holding" vs. "pouring"). Our contributions address these challenges and advance 4D egocentric understanding:

- **First multimodal NeRF-based method for 4D egocentric understanding:** We extend semantic NeRFs from static 3D to dynamic 4D scenes, integrating video-language and audio-language features to enable multimodal open-vocabulary egocentric 4D reasoning.
- **Object-centric video-language representation:** Video feature extraction for 4D lifting requires spatially and temporally coherent masks. We track object masks across frames to form tube-based embeddings, producing temporally coherent video features that remain robust under rapid egocentric camera motion.
- **Audio-language integration:** We leverage acoustic cues as a complementary temporal signal, disambiguating visually similar motions and refining temporal boundaries to improve action localization and segmentation.
- **Comprehensive evaluation:** We evaluate five state-of-the-art vision-language encoders on radiance fields and 2D representations, both with and without audio across two tasks, resulting in 40 per-scene models, including ablations that validate our design choices.

Experiments on EPIC-KITCHENS 100 [3, 4] and Ego-Exo4D [14] show that our model significantly outperforms single-modality or single-frame methods. We also demonstrate that incorporating audio allows models trained with only static image features to approach the performance of fully video-based methods, highlighting the power of multimodal temporal cues. Overall, our work demonstrates that the fusion of audio and visual modalities within a temporally aware NeRF-based method enables robust and transferable open-vocabulary action spatiotemporal understanding that generalizes across diverse egocentric scenarios.

2. Related Work

Neural Radiance Fields. NeRFs [28] emerged as a powerful framework for synthesizing photo-realistic novel views of 3D scenes from a sparse set of posed multi-view images by learning an implicit volumetric representation. Leveraging differentiable volume rendering, NeRFs map continuous 3D locations and camera viewing directions to color and density, enabling high-fidelity reconstruction of static scenes. Extending NeRFs to dynamic environments introduces challenges due to temporal variations. Early methods isolated motion from the static background, *e.g.*, Li et al. [25] used optical flow to predict object movement, while Gao et al. [9], Wu et al. [51] trained separate NeRFs for static and dynamic components. Similarly, NeuralDiff [43] employs inductive biases to partition the scene into multiple semantic regions, namely, a static background, dynamic entities, and articulated bodies, each modeled with a distinct implicit representation, leading to improved disentanglement and interpretability. Beyond foreground-background separation, semantic guidance has been explored: Liang et al. [26] leverage DINO features to identify temporally consistent salient regions, enriching NeRFs with semantic awareness. In egocentric perception, Zhang et al. [56] use 3D Gaussian splatting [21] to simultaneously reconstruct scenes and track object motion from first-person videos. Their approach demonstrates the capability to capture fine-grained object trajectories, but requires hand-object interaction masks that are precomputed through external processing steps.

Our approach leverages dynamic NeRFs as backbone but, unlike prior geometry reconstruction-focused methods [26, 43, 56], goes beyond reconstruction by integrating video- and audio-language features for multimodal 4D semantic understanding. Our core contribution is developing versatile multimodal 4D semantic understanding capabilities that are agnostic to the underlying 3D representation.

Feature Distillation in 3D Representations. Recent advances in NeRFs integrate 2D semantic information into 3D scene representations to bridge image-level perception and volumetric understanding. Several works [8, 39, 45, 58] incorporated per-pixel semantic labels into the volumetric optimization process, enabling 3D-aware segmentation across views. A more general alternative distills high-dimensional

2D features into 3D space, rather than discrete labels, typically via a teacher-student paradigm [12, 24, 43], where features from pretrained 2D models (e.g., DINO [2]) supervise a 3D network to embed scene semantics into the radiance field without manual annotations. Building on this, several methods use language-guided supervision by distilling image-text features such as CLIP [35], enabling open-vocabulary querying in 3D. For example, LERF [22] aggregates CLIP features across spatial scales to identify long-tail objects, while N2F2 [1] captures multi-scale semantics by embedding local and global concepts into a shared 3D feature space. LangSplat [33] combines 3D Gaussian splatting [21] with SAM [23] masks, yielding high-quality open-vocabulary instance segmentation. EgoLifter [16] adapts these ideas to egocentric videos but focuses solely on static scene reconstruction, filtering out the camera wearer and dynamic elements to maintain volumetric stability. In contrast, our approach explicitly models dynamic objects and actions, capturing spatiotemporal semantics in 4D.

Spatiotemporal Action Understanding. Spatiotemporal action localization involves detecting actions while spatially localizing actors across frames, whereas segmentation further provides dense pixel-wise masks to delineate action boundaries. Early methods relied on frame-level detectors linked over time [11, 37, 48, 49], later evolving to tubelet-based approaches with spatio-temporal classifiers [17, 20, 41]. To enhance precision and consistency, subsequent works integrated contextual reasoning [54], motion cues [38], and tracking techniques [40, 48]. More recently, transformer-based architectures reframe action localization as a direct prediction of spatio-temporal action tubes [15, 53], enabling end-to-end learning and improved performance. However, these methods cannot perceive depth or 3D structure, limiting them to flat image coordinates. By representing videos in 4D (3D space + time), we can handle the variability caused by camera motion and merge information from different viewpoints into consistent spatiotemporal representations. In addition, 4D representations can generate novel views which enable scene navigation from any viewpoint. Achieving 4D action understanding requires reconstructing semantically-aware 4D representations.

Recent advances in neural scene representations, such as NeRFs [28] and 3D Gaussian Splatting [21], enable high-fidelity 3D reconstructions from 2D views. Subsequent efforts extend these models to incorporate semantic understanding by lifting 2D features from vision-language models into 3D space [6, 22, 33]. These semantic 3D representations enable applications such as language-driven object and 3D action localization in reconstructed scenes. For example, Fiebelman et al. [7] lift spatiotemporal features into a Gaussian Splatting framework, demonstrating semantic localization of actions in 3D exocentric video datasets featuring both humans and animals. Our work goes beyond these approaches by targeting egocentric videos and incorporating multi-modal inputs, visual, audio, and language, into a NeRF-based 4D scene representation that allows for open-vocabulary 4D queries. Unlike prior methods relying on 2D features or static 3D reconstructions [6, 22, 33], we integrate video-language features derived from temporally consistent object-centric tubes and audio-language features from audio streams. This multimodal fusion, anchored by stable object-centric tubes, enables robust open-vocabulary spatiotemporal grounding of actions under rapid camera motion and dynamic object interactions. Our results show that incorporating multi-modal supervision significantly improves both semantic understanding and spatial accuracy beyond what uni-modal or single-frame models can achieve.

3. Methods

We address 4D open-vocabulary querying in egocentric scenes by integrating representations in 4D radiance fields with temporal audio cues. Given synchronized video and audio, we encode video features from spatiotemporal crops using a Video-Language Model (VideoLM) and audio features from aligned segments using an Audio-Language Model. We design our architecture to spatially embed visual features in a dynamic neural radiance field while audio remains a separate temporal signal—visual features localize to 3D positions, whereas audio provides temporal context that is not spatially grounded. Figure 2 illustrates our approach.

3.1. Multimodal Feature Extraction

Object-centric Video Crops. Video features require spatiotemporally coherent object masks. We generate these via a two-stage strategy: (1) from the initial frame, we extract object masks using SAM [23] with dense point prompts filtered by confidence and redundancy; (2) we propagate these masks through the video with SAM2 [36] with consistent IDs. Quality checks are performed every K frames, with mask extraction being reapplied when the uncovered area exceeds a threshold (details in Suppl. C). This produces temporally consistent video crops rather than isolated segments (*e.g.*, from SAM [23] alone), enabling robust video feature extraction. Compared to fixed spatial regions where objects are split arbitrarily across spatial boundaries, object-centric crops are crucial for egocentric videos, where rapid camera motion causes objects to appear unstable across frames.

Video Features. Each crop V is passed through the video encoder $\phi_v(\cdot)$ of a VideoLM to obtain high-dimensional semantic embeddings $e_v = \phi_v(V)$. We then assign each embedding to all the pixels inside its corresponding masklet to construct a semantic embedding map for each frame. When multiple masks overlap, we assign the embedding with shorter temporal span to prioritize dynamic components.

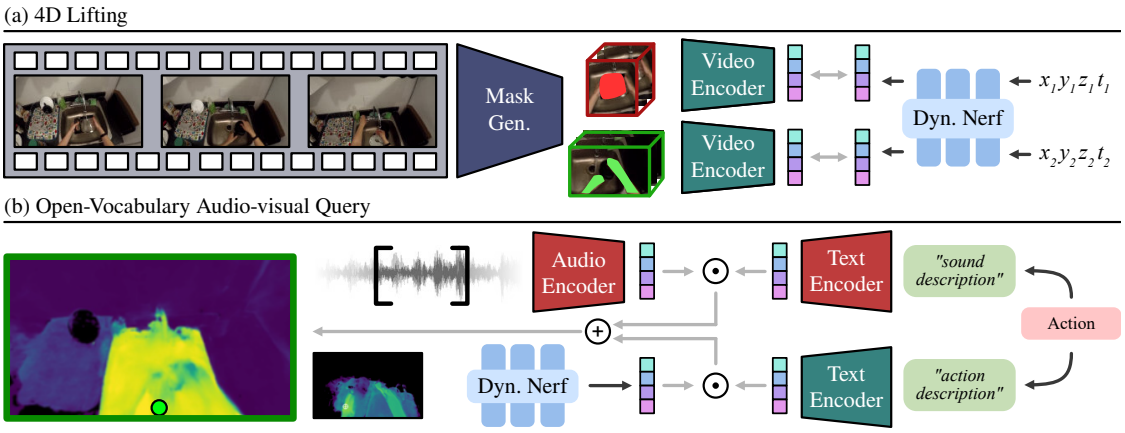

Figure 2. **AViON4D Overview.** We construct spatio-temporal masks tracking objects in egocentric videos, which are further processed with a video encoder to generate video features aligned with text. These features are used for distillation in a radiance field model for egocentric videos, **4D Lifting**. Further, AViON4D processes audio segments with an audio encoder to generate audio features also aligned with text. Once trained, we can perform **Open-Vocabulary Audio-Visual Queries** by computing video and audio similarities to prompts describing the queried action. Our multi-modal approach provides strong spatiotemporal understanding absent in uni-modal approaches.

Audio Features. Complementing visual features, we extract audio features as a separate temporal signal. For each frame at time t, we extract a temporally aligned 2.5-second audio segment A_t centered around the frame timestamp and encode it using ImageBind's audio encoder φ_a to yield embeddings $e_a(t) = \varphi_a(A_t)$. These embeddings capture distinctive acoustic signatures of actions (*e.g.*, chopping, pouring, stirring) that complement visual features and help disambiguate visually similar actions with different acoustic profiles. We maintain this separation rather than distilling audio into the radiance field because visual features are computed for spatially localized object masks that correspond to specific 4D locations in the scene, whereas audio yields the same feature per frame and lacks spatial localization.

3.2. 4D Lifting

We lift spatiotemporal semantics into a 4D radiance field representation. Building on NeuralDiff [42] for dynamic egocentric modeling, we extend it to incorporate video embeddings. We select NeuralDiff for its effectiveness in egocentric scenes, but AViON4D is backbone-agnostic and compatible with other dynamic 3D representations (*e.g.* 4D Gaussian Splatting [50], D-NeRF [32]).

4D Visual-Language Field Representation. Given a set of video frames $V = \{I_t\}_{t=1}^{T}$ with corresponding camera poses $\{P_t\}_{t=1}^{T}$, we construct a 4D radiance field that maps space-time coordinates (x, y, z, t) and viewing direction (θ, ϕ) to color c, density σ, and video embeddings e_v:

$$F : (x, y, z, t, \theta, \phi) \mapsto (c, \sigma, e_v) \tag{1}$$

Dynamic egocentric scenes are modeled through a three-stream neural renderer for background, foreground, and egocentric actors, each processing the same 4D coordinates with separate MLPs and frame codes that specialize into static and dynamic components via inductive bias during training [42].

Visual Feature Distillation. We render color and features using standard volume integration [28]. For each ray r, we sample N points and query the radiance field to obtain color c_i, density σ_i, and video embeddings $e_{v,i}$ at each sample point i. The rendered outputs are:

$$\begin{aligned} \hat{C}(r) &= \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) c_i, \\ \hat{E}_v(r) &= \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) e_{v,i}. \end{aligned} \tag{2}$$

where $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$ is the accumulated transmittance and δ_i is the distance between samples. To supervise the video embeddings, we introduce a feature distillation loss:

$$\mathcal{L}_{\text{v}} = \text{MSE}\left(\frac{\hat{E}_v}{||\hat{E}_v|| + \epsilon}, \frac{e_v}{||e_v|| + \epsilon}\right) \tag{3}$$

where e_v and $\hat{E}_v$ are the ground truth and predicted video embeddings, and ϵ is a small constant for stability. The final training objective supervises all three streams, combining RGB reconstruction, feature distillation, and sparsity regularization:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{v}} + \lambda_\sigma \mathcal{L}_\sigma \tag{4}$$

where $\mathcal{L}_{\text{rgb}}$ is the RGB reconstruction loss, $\mathcal{L}_\sigma$ is density sparsity regularization, and λ_σ is a weighting factor.

3.3. Open-Vocabulary Audio-Visual Query

AViON4D enables audio-visual querying for 4D scene understanding by integrating spatiotemporal visual features from a dynamic NeRF with temporally aligned audio features. We evaluate AViON4D on two complementary tasks: *4D action localization*, which identifies the precise spatiotemporal coordinates (x, y, z, t) where an action occurs, and *4D action segmentation*, which segments all spatiotemporal regions involved in the action (*e.g.*, hands and objects being manipulated). Given a text query, we encode it into the corresponding embedding spaces, compute visual and audio relevance scores, and combine these to localize and segment the most relevant spatiotemporal regions.

As shown in Figure 2(b), given an action name as our text query q, we first derive the action (video) description q_v and the sound (audio) description q_a (of the action). We encode q_v using VideoLM's text encoder $\phi_l(\cdot)$, to obtain $e_{v,l} = \phi_l(q_v)$, and encode q_a using audio text encoder $\varphi_l(\cdot)$ to obtain $e_{a,l} = \varphi_l(q_a)$. We then compute visual similarity at each point as the normalized dot product, $s_{vl}(x, y, z, t) = \hat{E}_v(x, y, z, t) \cdot e_{v,l}$, and audio similarity at each timestep as $s_{al}(t) = e_a(t) \cdot e_{a,l}$. Finally, we combine these scores as $s(x, y, z, t) = \alpha s_{vl}(x, y, z, t) + (1 - \alpha)s_{al}(t)$.

For 4D action localization, we identify the action at the point (x, y, z, t) with maximum combined score, following [22]. For 4D action segmentation, we segment all regions where the score exceeds a threshold $s(x, y, z, t) > \tau$.

4. Experiments

We present a comprehensive evaluation across models, tasks, and egocentric scenes. Our primary benchmark is EPIC-KITCHENS 100 (EK-100) [3, 4], using seven sequences (approx. 32,000 frames). We evaluate five visual encoders in eight baselines (three 2D and five semantic field representations), training a separate NeRF for each scene-model combination, resulting in 35 networks. This follows reference works in the field on semantic NeRFs [6, 22], which used five and eight scenes, respectively, for quantitative evaluation with a single visual encoder, and balances comprehensive assessment with computational feasibility. Additionally, and to further assess robustness and generalization, we extend the evaluation to two Ego-Exo4D scenes [14] with two visual encoders. Models are evaluated on two tasks—4D action localization and segmentation—both with and without audio. We also perform four ablations to study spatiotemporal crops, audio description types, audio extraction window size, and robustness to background noise, bringing the total to 40 per-scene NeRF reconstructions.

4.1. Experimental Setup

Datasets. We evaluate our approach on EK-100 [3, 4], a widely-used benchmark in egocentric scene understanding that contains video recordings of humans performing various kitchen tasks in real-world noisy conditions. Unlike other datasets like Ego4D [13], Ego-Exo4D [14] or Assembly101 [30], EK-100 uniquely provides all essential components needed for our evaluation: (1) camera poses estimated by EPIC Fields [44] for NeRF reconstruction, (2) dense object and hand segmentation masks from VISOR [5] for spatial evaluation, (3) natural language narrations aligned with video for open-vocabulary querying, and (4) synchronized audio with sound descriptions from EPIC Sounds [18].

Following standard practice in semantic NeRFs [6, 19, 22], we evaluate on 7 scenes from EK-100, that provide coverage of diverse kitchen activities, camera motions, and acoustic conditions. Our selection criteria ensure diverse action types, varying scene complexity, and complete multimodal annotations across VISOR masks, EPIC Fields poses, and EPIC-Sounds audio. Inherent to NeRF, each combination of visual encoder and scene requires a separate NeRF reconstruction and optimization, making our evaluation of 5 encoders across 7 scenes computationally intensive, in line with prior semantic NeRF benchmarks [6, 19, 22]. For each sequence, frames are divided into training and testing sets. The training set is used for reconstruction and feature distillation, while the testing set validates predictions in unseen views, i.e., GT action labels are only used for validation, not for optimizing the representations. Additional details on preprocessing and dataset statistics are available in Suppl. B.1.

To demonstrate cross-domain generalization beyond our primary EK-100 evaluation, we also validate on basketball and bike repair scenes from [14], manually annotated for this work. We select these domains as they differ significantly in motion patterns, object types, and acoustic signatures, providing a rigorous test of our method's transferability without domain-specific training. Details on the labeling process and dataset statistics are in Suppl. B.2.

Training Since our framework requires no large-scale training as it leverages pre-trained encoders, the only training component is NeRF scene reconstruction, which involves per-scene optimization. Crucially, we use the pre-trained vision-language models (CLIP, LaViLa, EgoVideo, ImageBind) in their original form without any fine-tuning on EK-100 or Ego-Exo4D action vocabularies or any task-specific training. This ensures true open-vocabulary capability, as our system was never exposed to the specific action classes it must localize during evaluation.

Evaluation Protocol and Metrics. For 4D action localization, we follow LERF [22] and define success when the top-ranked pixel lies within the target region, extended to 4D using spatiotemporal mask tubelets. Evaluation is performed on 2D rendered frames at novel viewpoints: we select the highest-scoring pixel as the predicted location

following [22], with predictions within a 10-pixel radius considered correct to account for boundary imprecision in the GT masks. We report Top-1 and Top-5 temporal and spatiotemporal accuracies, where temporal requires predictions to be within the correct frame range, whilst spatiotemporal additionally requires the predicted 2D pixel location within the correct hand/object masks. For 4D action segmentation, we compute mean Intersection over Union (mIoU) per frame on the rendered 2D masks, following [46]. To ensure fair comparison across encoder score distributions, we report results at each configuration's optimal threshold in [0.1-0.9].

Baselines. We evaluate two groups of baselines: 4D NeRF-based (RF) and 2D representations, with five language-vision models (three image-based and two video-based). As image-based, we use CLIP [34], SigLIP [55], which has shown improved performance over CLIP, and ImageBind [10], which provides features aligned with multiple modalities. For video-based, we evaluate models trained on egocentric video data: LaViLa [57] and EgoVideo [31], which incorporate temporal understanding not present in image-based models.

For 4D NeRF-based baselines, we distill the extracted features into the RF representation, enabling 4D consolidation and novel view synthesis. For 2D baselines, we apply the same encoders directly to video frames without NeRF consolidation, computing prompt similarity directly with video frames. Since our evaluation uses novel views rendered from the 4D representation (following standard 3D semantic practice [6, 22]), and 2D methods cannot generate such views, we provide these 2D baselines with perfect original frames, which favors them in this validation setup. Despite this advantage, this comparison aims to demonstrate the value of NeRF feature consolidation for dynamic egocentric scenes.

Finally, to validate the effect of the proposed multi-modal approach, we evaluate all baselines with and without audio features processed by ImageBind's audio encoder [10].

Comparison to SOTA. Since there is no model available for the task of open-vocabulary 4D action localization and segmentation on radiance fields from egocentric videos, we compare, in our experiments, to a set of variants of our framework. However, our vanilla CLIP RF baseline can be viewed as an adaptation of widely cited 3D semantic NeRF approaches, such as LERF [22] and OpenNeRF [6], for spatiotemporal understanding in egocentric settings.

4.2. 4D Action Localization on EK-100

Table 1 compares these approaches, revealing insights on: (1) the impact of temporal modeling in dynamic vs. static models, (2) the benefits of 4D consolidation in NeRF *vs*. 2D, and (3) multimodal audio-visual fusion gains.

Static vs. dynamic models Overall, models pre-trained on video data consistently outperform static image-based models. This is expected, as video models are inherently capable of capturing temporal dependencies, which are absent in static image-based models. Among the image-based models, ImageBind achieves the highest performance with a Top-1 accuracy of 15.1% and Top-5 accuracy of 22.4%. In contrast, the best-performing video model, LaViLa, attains 18.0% Top-1 and 32.7% Top-5 accuracies, demonstrating the advantage of temporal modeling.

RF vs. 2D representations Neural representations that consolidate frame-level features into continuous 4D representations yield significant performance improvements. This approach mitigates noise through multi-view aggregation, reducing individual frame errors even in dynamic egocentric scenes with estimated camera poses. CLIP substantially benefits from this consolidation, improving by $+6.1\%$ Top-1 and $+5.7\%$ Top-5 accuracy points. ImageBind shows a $+1.7\%$ improvement in Top-1, though the Top-5 gain is less pronounced. LaViLa demonstrates consistent improvement with $+3.0\%$ in Top-1 and $+7.4\%$ in Top-5 accuracy, further validating the 4D feature consolidation across encoders.

Visual-only vs. audiovisual inputs The incorporation of audio leads to consistent improvements across both image and video models. This is especially notable for image-based models, where the addition of audio provides a form of temporal grounding at a relatively low computational cost. For instance, CLIP improves from 13.5% to 19.5% in Top-1 accuracy when audio is included, outperforming the best single-modality video model (LaViLa, 18.0% Top-1). Video-based models also benefit from multi-modal fusion. LaViLa improves its Top-1 accuracy by 1.5%, while EgoVideo gains 3.1% in Top-1. Interestingly, multi-modal fusion leads to divergent effects on Top-5 accuracy: while EgoVideo improves to a Top-5 accuracy of 38.7%, the best among all evaluated models, LaViLa experiences a slight drop. These findings highlight the potential of audio-visual integration in real-world conditions, such as the noisy environments of EK-100, while also suggesting that its impact may vary depending on model architecture and training paradigms.

Qualitative Results Figure 3 presents qualitative results for CLIP, ImageBind, SigLIP, and EgoVideo, for different scenes and queries. The examples show that, in many instances, relying only on visual features can lead to incorrect object localization. For the prompt *"wash plate"*, ImageBind selects a frame where the plate is present but is not part of the action. The model's lack of temporal understanding causes this incorrect localization. Incorporating audio helps disambiguate through strong auditory cues: a plate resting produces no noise, while washing produces characteristic

Table 1. **4D Action Localization in EK-100**. Mean top-1 and top-5 accuracy comparing image-based encoders (CLIP, ImageBind, SigLIP) that extend semantic 3D NeRF methods to dynamic scenes, and video-based encoders (LaViLa, EgoVideo) with temporal understanding. RF rows show NeRF-based methods; 2D rows show direct frame processing. Our complete method combines video encoders (LaViLa or EgoVideo) with audio in the RF setting. Temporal accuracy requires prediction within action intervals; spatiotemporal accuracy additionally requires pixel-level localization. Values in parentheses show audiovisual improvements over visual-only (δ= audiovisual - visual). **Key findings:** (1) Audio-visual fusion consistently improves performance across all methods. (2) Video encoders outperform image-based approaches due to temporal modeling. (3) NeRF consolidation improves over 2D through multi-view aggregation and temporal consistency.

	Model	Type	Visual				Audiovisual			
			Spatiotemporal		Temporal		Spatiotemporal		Temporal	
			Acc	Acc5	Acc	Acc5	Acc	Acc5	Acc	Acc5
2D	CLIP [34]	Image	7.4	19.5	13.5	35.7	12.1 (+4.7)	24.6 (+5.1)	19.6 (+6.1)	37.0 (+1.3)
2D	ImageBind [10]	Image	13.4	25.4	19.7	38.6	13.4 (0.0)	23.9 (-1.5)	16.6 (-3.1)	35.7 (-2.9)
2D	LaViLa [57]	Dynamic	15.0	25.4	31.3	59.7	15.0 (0.0)	27.0 (+1.6)	33.0 (+1.7)	62.8 (+3.1)
RF	CLIP [34]	Static	13.5	25.2	15.0	32.8	19.5 (+6.0)	24.0 (-1.2)	22.2 (+7.2)	34.2 (+1.4)
RF	SigLIP [55]	Static	13.4	28.6	19.5	37.4	16.5 (+3.1)	32.8 (+4.2)	19.5 (0.0)	40.4 (+3.0)
RF	ImageBind [10]	Static	15.1	22.4	16.7	31.1	16.5 (+1.4)	24.4 (+2.0)	18.0 (+1.3)	32.7 (+1.6)
RF	LaViLa [57]	Dynamic	18.0	32.7	28.7	50.9	19.5 (+1.5)	29.7 (-3.0)	28.3 (-0.4)	49.5 (-1.4)
RF	EgoVideo [31]	Dynamic	13.5	37.2	28.2	56.5	16.6 (+3.1)	38.7 (+1.5)	32.7 (+4.5)	58.1 (+1.6)

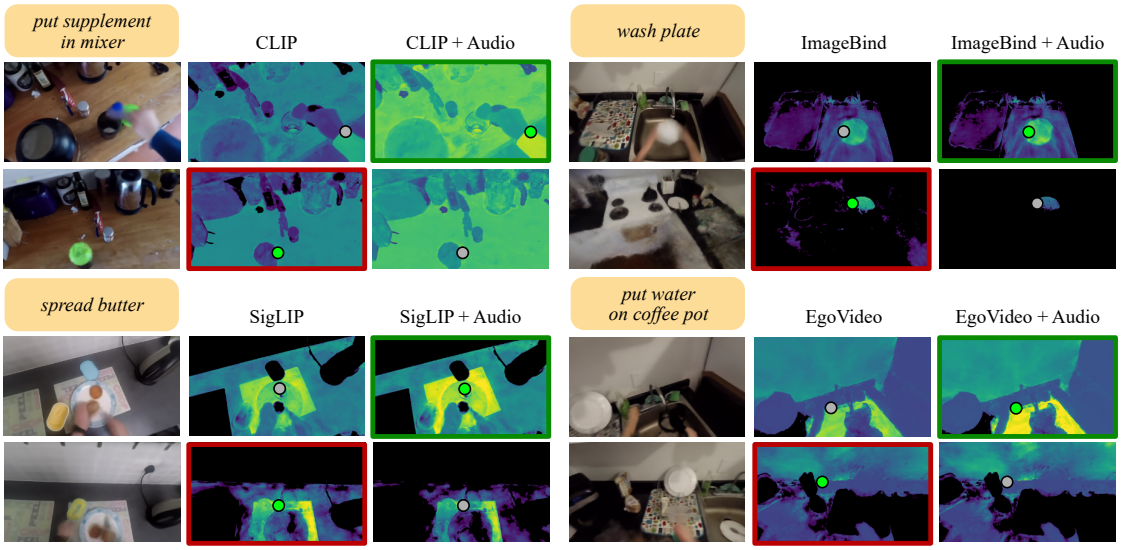

Figure 3. **Qualitative examples for action localization across various action prompts, scenes, and visual encoders.** For each query: Top/bottom rows show correct/incorrect frames; left to right shows rendered novel view, visual-only similarity map, audiovisual similarity map. Results highlight that relying solely on visual features leads to incorrect localization (framed in red). Incorporating audio as an additional modality improves localization (in green) accuracy by leveraging action-specific auditory cues for image and video models.

water sounds. Additionally, AViON4D's multi-modality can also help alleviate localization errors present due to noise in the feature field learned by the radiance field model. The example presented for EgoVideo illustrates a case where the model selects the incorrect frame due to small artifacts in the regions connecting two surfaces. These effects are diminished when aggregated with auditory features, resulting in the correct frame selected afterwards. Additional qualitative results are presented in Suppl. D.

4.3. 4D Action Segmentation on EK-100

Segmentation results presented in Table 2 mirror the localization findings. Dynamic models achieve overall superior performance, with LaViLa reaching 51.01% mIoU. Among static models, SigLIP leads at 49.52%, while ImageBind (35.99%) and CLIP (31.86%) show substantially lower performance. Audio integration provides consistent improvements across all models. Static models show larger gains: CLIP improves by +14.90% and ImageBind by +13.49%. Dynamic models, although with higher visual baselines, also experience increases: EgoVideo (+6.98%) and LaViLa

Table 2. **4D Action Segmentation in EK-100**. mIoU of static and dynamic encoders for 4D action segmentation. Dynamic encoders outperform static approaches, with audio-visual fusion providing substantial gains for weaker visual baselines (values in parentheses show audiovisual improvements *i.e.* δ = audiovisual - visual).

Model	Type	mIoU	
		Visual	Audiovisual
CLIP [34]	Static	31.86	46.76 (+14.90)
SigLIP [55]	Static	49.52	53.27 (+3.75)
ImageBind [10]	Static	35.99	49.48 (+13.49)
LaViLa [57]	Dynamic	51.01	54.00 (+2.99)
EgoVideo [31]	Dynamic	47.02	54.00 (+6.98)

Table 3. **4D Action Localization on Diverse Ego-Exo4D**. Spatiotemporal accuracy (Acc) and top-5 accuracy (Acc5) comparison across basketball (UB_03) and bike repair (FB_06) scenes. LaViLa models demonstrate superior baseline performance, with audio-visual fusion providing consistent improvements (represented by values in parentheses *i.e.*, δ = audiovisual - visual).

Scene	Model	Visual		Audiovisual	
		Acc	Acc5	Acc	Acc5
FB_06	CLIP [34]	0.00	0.14	0.14 (+0.14)	0.14 (0.0)
FB_06	LaViLa [57]	0.14	0.43	0.29 (+0.15)	0.43 (0.0)
UB_03	CLIP [34]	0.00	0.33	0.17 (+0.17)	0.33 (0.0)
UB_03	LaViLa [57]	0.33	0.33	0.33 (0.0)	0.50 (+0.17)

(+2.99%). With audio-visual fusion, both dynamic models reach 54.00% mIoU, representing the highest segmentation performance across all evaluated architectures.

4.4. Generalization to Diverse Scenarios

To evaluate zero-shot capability beyond kitchen domains, we manually annotate two scenes from Ego-Exo4D [14] for evaluation. We select scenes with distinct characteristics: (1) fair_bike_06_10 (FB_06), capturing repair activities in a bike workshop and (2) unc_basketball_03-31-23_01_7 (UB_03), representing dynamic sports activities in a basketball court. These scenes feature significantly different scenarios, motions, actions, objects, and audio compared to EK-100, providing a strong test of cross-domain generalization. As shown in Table 3, results are consistent with EK-100 patterns: LaViLa outperforms CLIP in both scenes, and audio integration provides substantial improvements for both models, demonstrating that our approach generalizes across diverse activity domains in zero-shot settings.

4.5. Ablations

We conduct ablations to validate our design choices. This section reports our main ablation on spatiotemporal crops, while Suppl. E covers audio ablations on description types, window size, and noise robustness.

Table 4. **Ablation on spatiotemporal crops.** Effect of different cropping strategies for video feature extraction. Row with gray background (Mask) indicates our proposed strategy.

	Spatiotemporal		Temporal	
	Acc	Acc5	Acc	Acc5
Grid	6.0	9.0	10.6	18.1
Mask	18.0	32.7	28.7	50.9

Spatiotemporal Crops. Even though [7, 29] also use video models to generate features for NeRF distillation, these methods partition frames into regular grids, forming temporal prisms to extract video features and produce dense 2D semantic maps. In contrast, we use spatiotemporal crops guided by temporally coherent segmentation masks, resulting in object-centric feature representations. In our ablation study, we compare the methodology of feature extraction employed by these works with our approach, using the LaViLa video encoder with 16-frame clips for all 7 scenes of EK-100. For grid extraction, we follow [29] and use square patches sized at 33% of the image height.

Table 4 demonstrates that mask-based extraction substantially outperforms grid-based extraction, with this advantage being especially critical for egocentric videos. Rapid camera motion causes objects to constantly move across fixed grid boundaries, creating temporally inconsistent features that degrade performance. In contrast, our segmentation-based approach maintains object-centric features that remain robust despite challenging egocentric dynamics.

5. Conclusions

We presented AViON4D, a novel multimodal method for 4D egocentric scene understanding. Unlike existing methods limited to static 3D scenes, AViON4D captures spatiotemporal action semantics through complementary video-language and audio-language features. Conceived as a component-agnostic method, AViON4D accepts free-form natural language queries without dataset-specific training, enabling zero-shot generalization across diverse domains. As a result, the method exhibits remarkable versatility and generality, as proven by its successful application to egocentric scenarios of widely different nature. Through comprehensive evaluation across five state-of-the-art vision-language encoders, 2D and 4D representations, unimodal and multimodal settings, totaling 40 per-scene NeRF reconstructions, we demonstrate significant and consistent improvements across different scenes, datasets, and tasks.

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

[2] 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

[3] 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. 2, 5

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2, 5

[5] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 5

[6] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. Open-nerf: Open set 3d neural scene segmentation with pixel-wise features and rendered novel views. In *ICLR*, 2024. 2, 3, 5, 6

[7] Gal Fiebelman, Tamir Cohen, Ayellet Morgenstern, Peter Hedman, and Hadar Averbuch-Elor. 4-legs: 4d language embedded gaussian splatting. *Computer Graphics Forum (Proceedings of Eurographics)*, 2025. 2, 3, 8

[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. 2

[9] 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. 2

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 6, 7, 8

[11] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015. 3

[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. 5

[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. 2, 5, 8

[15] Alexey A Gritsenko, Xuehan Xiong, Josip Djolonga, Mostafa Dehghani, Chen Sun, Mario Lucic, Cordelia Schmid, and Anurag Arnab. End-to-end spatio-temporal action localisation with video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18373–18383, 2024. 3

[16] 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

[17] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017. 3

[18] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 5

[19] Sungwon Hwang, Junha Hyung, Daejin Kim, Min-Jung Kim, and Jaegul Choo. Faceclipnerf: Text-driven 3d face manipulation using deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3469–3479, 2023. 5

[20] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 4405–4413, 2017. 3

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. on Graphics (SIGGRAPH)*, 2023. 2, 3

[22] 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, 5, 6

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[24] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Ad-*

vances in Neural Information Processing Systems, 35:23311–23330, 2022. 3

[25] 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. 2

[26] 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. 2

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

[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. In *ECCV*, 2020. 2, 3, 4

[29] Lorenzo Mur-Labadia, Josechu Guerrero, and Ruben Martinez-Cantin. Div-ff: Dynamic image-video feature fields for environment understanding in egocentric videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 8

[30] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12999–13008, 2023. 5

[31] 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. 6, 7, 8

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 4

[33] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, pages 20051–20060, 2024. 3

[34] 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*, 2021. 2, 6, 7, 8

[35] 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

[36] 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. In *ICLR*, 2025. 3

[37] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016. 3

[38] Suman Saha, Gurkirt Singh, and Fabio Cuzzolin. Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4414–4423, 2017. 3

[39] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 2

[40] Gurkirt Singh, Suman Saha, and Fabio Cuzzolin. Tramnet-transition matrix network for efficient action tube proposals. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 420–437. Springer, 2019. 3

[41] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11995, 2019. 3

[42] 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. 4

[43] 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

[44] 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. 5

[45] 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. 2

[46] Zifu Wang, Maxim Berman, Amal Rannen-Triki, Philip Torr, Devis Tuia, Tinne Tuytelaars, Luc V Gool, Jiaqian Yu, and Matthew Blaschko. Revisiting evaluation metrics for semantic segmentation: Optimization and evaluation of fine-grained intersection over union. *Advances in Neural Information Processing Systems*, 36:60144–60225, 2023. 6

[47] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918, 2024. 2

[48] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015. 3

[49] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 284–293, 2019. 3

[50] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 4

[51] 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. 2

[52] Tao Wu, Shuqiu Ge, Jie Qin, Gangshan Wu, and Limin Wang. Open-vocabulary spatio-temporal action detection. *arXiv preprint arXiv:2405.10832*, 2024. 2

[53] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 3

[54] Zhenheng Yang, Jiyang Gao, and Ram Nevatia. Spatio-temporal action detection with cascade proposal and location anticipation. *arXiv preprint arXiv:1708.00042*, 2017. 3

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 6, 7, 8

[56] 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. 2

[57] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 6, 7, 8

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

Supplementary Material: Appendix

A. Overview

This appendix provides additional details and analyses to support the findings presented in the main manuscript. We organize this material as follows: B describes the specification of the subsets and annotations used for evaluating egocentric spatiotemporal action understanding on EPIC-KITCHENS and Ego-Exo4D. Section C presents detailed information about the implementation and experimental configurations. Later, Section D showcases additional qualitative results and visual examples. Section E presents ablation studies on description types for audio features, audio extraction window size and audio robustness to background noise. Sections F and G discuss the limitations of our approach and broader societal implications. Finally, we conclude with Section H, which provides a complete list of all assets used in this work.

B. Data Preprocessing

B.1. EPIC-KITCHENS subset

We evaluate on scenes from the EPIC-KITCHENS (EK-100) dataset [1, 2] that meet the following criteria: (1) availability of segmentation masks from VISOR [3], (2) released estimated camera poses from EPIC Fields [15], and (3) existing corresponding audio annotations from EPIC-Sounds [6]. These criteria ensure comprehensive multimodal ground truth for our analysis. We select 7 representative scenes that provide diverse kitchen environments, action types, camera motions, and acoustic conditions.

B.1.1. Training and Test Splits

Each video is split into training and test sets. The training set is used for reconstruction and feature distillation, while the test set evaluates generalization to unseen views. To construct a reliable test set based on manually verified segmentation masks (VISOR GroundTruth-SparseAnnotations), we apply a two-stage sampling strategy:

1. **Action coverage**: we sample frames with manually annotated masks from different actions, ensuring at least one frame per action (*within action*).
2. **Background coverage**: When possible, we sample twice as many frames from non-action segments of the video (*outside of action*). In action-dense scenes (>80% action coverage), we proportionally reduce background sampling while maintaining temporal diversity where possible, which ensures balanced training data and fair evaluation across all test frames.

The remaining frames are allocated to the training set. Some frames are later excluded from both splits due to missing camera pose annotations in EPIC Fields [15]. Similarly, actions without corresponding manual segmentation masks in VISOR [3] are removed from the test set.

To support reproducibility and future benchmarking, we release our training and test splits alongside the source code. Detailed dataset statistics are provided in Table 1.

B.1.2. Annotation Refinements

To associate actions with sound descriptions, we use annotations from EPIC-Sounds [6]. An action is paired with a sound description if its timestamp aligns. In cases where multiple sound descriptions align with a single action, we compute text feature embeddings for each and use their average representation during retrieval.

To ensure the reliability of the ground truth in our evaluations, we perform minimal refinements to address objective annotation inconsistencies visible in the video frames. The adjustments were made only when necessary to enable accurate interpretation of the depicted actions.

Refined ambiguous narrations. As shown in Figure 1 (left), we refined ambiguous narrations in several instances. For example, narration ID `P18_01_41`, originally described as "put back in toaster", was refined to "put back bread slices in toaster". Similarly, for `P18_01_36`, the narration "place it on the plate" was modified to "place knife on the plate".

Consolidated action segments. We also consolidated redundant action segments. As shown in Figure 1 (center), segments `P18_02_13` ("wash coffee pot") and `P18_02_14` ("still washing coffee pot") were merged into a single, continuous action to better reflect the uninterrupted activity.

Extended temporal range. In a few cases, we also refined the temporal range of the actions. For instance, in segment `P18_02_23`, shown in Figure 1 (right), the action "spread peanut butter onto toast" was initially annotated to end too early. We extended the segment to include all relevant frames, such as the bottom frame, in which the action is still visibly ongoing.

In total, we refined the annotations of 18 out of 112 actions (approximately 16%), keeping the changes minimal and limited to essential corrections. The 112 action segments (2-33 per scene) average 3.32 s (226.35 frames) each, combining 20 verbs and 33 nouns into 94 unique narrations. The action distribution shows most narrations (78) occur once, 15 twice, and "open drawer" 4 times as the most frequent. All updated annotations will be released to ensure fair and consistent comparisons in future work.

B.2. Ego-Exo4D subset

To demonstrate our framework's generalizability across different activity domains, we construct an additional evaluation subset from Ego-Exo4D [5], a large-scale multimodal dataset capturing diverse human activities from both egocentric and exocentric viewpoints. We selected two representative egocentric scenes with complementary characteristics: (1) `unc_basketball_03-31-23_01_7` (48s, 35 origi-

Table 1. **EK-100 Subset Statistics.** Statistics per scene for the EK-100 subset used in our experiments. **KID**: EK-100 video ID. **Length (s)**: full video duration in seconds. **Frames**: total number of original frames. **w/ Poses**: frames with registered camera poses from EPIC Fields. **w/ GT Masks**: frames annotated with GT VISOR segmentation masks. **Train/Test**: number of frames included in our training and test splits. **Actions**: action segments with at least one frame with ground truth VISOR annotations. **A w/ Sound**: number of these actions that also have corresponding audio annotations from EPIC-Sounds [6].

ID	KID	Length (s)	Frames	w/ Poses	w/ GT Masks	Train	Test	Actions	A w/ Sound
01	P09_07	55	3325	3013	32	2931	48	10	7
02	P32_07	29	1745	1728	6	1710	17	3	3
03	P18_01	166	9992	9832	69	9627	162	33	26
04	P18_02	190	11435	11342	44	11215	90	24	13
05	P04_26	37	2219	2084	4	2074	11	2	2
06	P03_11	53	3164	2654	29	2578	52	10	9
07	P06_11	77	4624	4106	36	4008	65	17	7

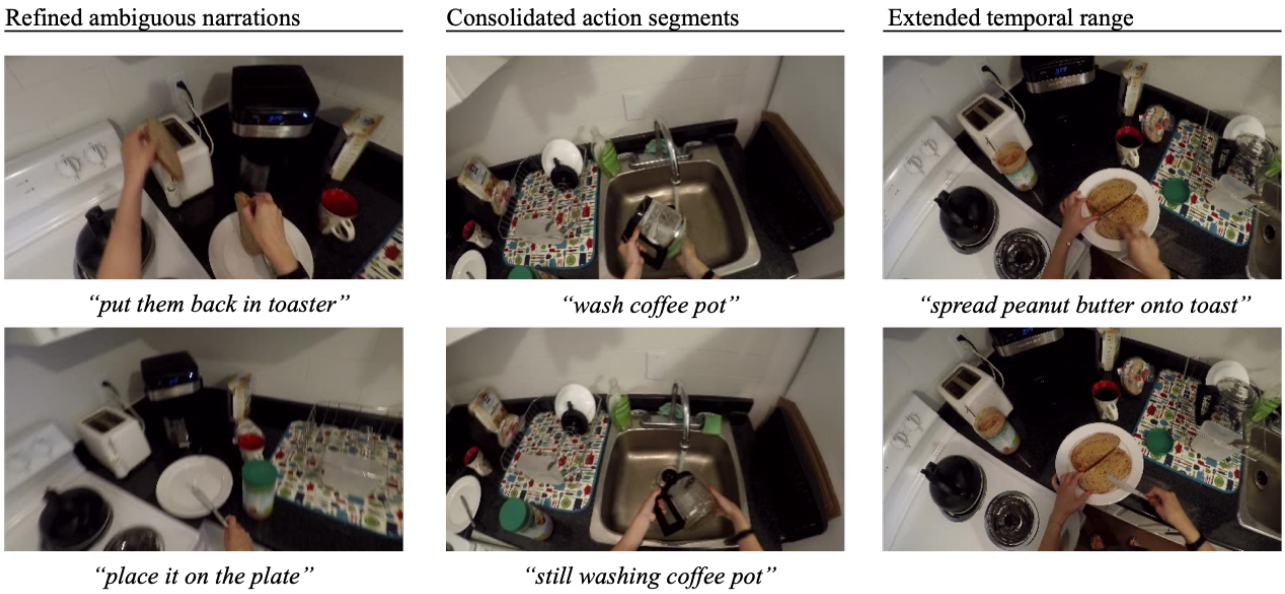

Figure 1. **Annotation Refinements. Left column:** Examples of clarified narrations. Top: `P18_01_41:frame_0000008680` was refined from "put back in toaster" to "put back *bread slices* in toaster". Bottom: `P18_01_36:frame_0000007662` was refined from "place *it* on the plate" to "place *knife* on the plate". **Middle column:** Consolidated action segments. Top frame (`P18_02_13:frame_0000005760`) labeled "wash coffee pot" and bottom frame (`P18_02_14:frame_0000007395`) labeled "still washing coffee pot" were merged into a single continuous action with a single annotation. **Right column:** Extended temporal range. Action segment `P18_02_23` originally included the top frame (`frame_0000010407`) for "spread peanut butter onto toast" but excluded the bottom frame (`frame_0000010740`), where the action was still ongoing. This frame was added to accurately reflect the full action duration.

nal narrations, 6 unique actions) capturing dynamic sports activities in a basketball court, and (2) `fair_bike_06_10` (35s, 9 original narrations, 7 unique actions) documenting technical repair activities in a bike workshop. These scenes complement our EK-100 evaluation by providing different environmental conditions, activity types, and motion dynamics.

B.2.1. Manual Annotation Process

Unlike EK-100, which leveraged existing multi-source annotations, the Ego-Exo4D scenes required comprehensive manual annotation across all modalities to provide suitable ground truth for our evaluation. Our annotation process included:

1. **Action boundaries and narrations**: We refined and consolidated the narrations provided by Ego-Exo4D to match the concise style of EK-100, mainly containing [verb +

object], as the original descriptions were overly detailed for our task requirements. For example, the original Ego-Exo4D provides highly granular descriptions such as "C shoots a jump shot at the hoop with both hands" and "The basketball bounces off the rim," which we consolidated into simpler actions like "shoots basketball at the hoop" and "catches basketball." Additionally, we manually annotated action boundaries as only timestamps were provided in the original annotations.

2. **Segmentation masks**: We combined existing segmentation masks provided by Ego-Exo4D with additional masks generated using SAM2 [13] to ensure comprehensive coverage, as not all frames and relevant objects had annotations in the original dataset. Notably, hand segmentation was not included in the original annotations, requiring us to generate these critical masks for egocentric activity understanding.
3. **Audio annotations**: We manually annotated corresponding sound descriptions to enable multimodal evaluation consistent with our EK-100 analysis.

All annotations were carefully reviewed to maintain consistency with our evaluation framework and ensure high-quality ground truth for reliable assessment. In total, we processed and refined annotations for 44 narrations across both scenes (35 from the basketball scene, 9 from the bike scene), consolidating them into 13 unique actions.

To match the conditions of our EK-100 evaluation, we estimated camera poses for the Ego-Exo4D scenes using the EPIC Fields approach [15], enabling consistent 3D reconstruction and neural rendering across both datasets.

B.2.2. Training and Test Splits

Similar to the EK-100 subset, each video is split into training and test sets following our two-stage sampling strategy. We apply the same approach of ensuring action coverage by sampling frames from different actions, and background coverage by including frames from non-action segments where possible. Detailed dataset statistics are provided in Table 2.

To support reproducibility and future benchmarking, we will release our annotations, training and test splits alongside the source code.

C. Implementation Details

C.1. Mask and Crop Generation.

We compute dense feature maps for every third frame and extract object-centric crops via SAM [9] and SAM2 [13], building on the implementation of Zrporz [18]. SAM is applied using a 32×32 grid of point prompts. The resulting masks are filtered based on the following criteria: Intersection-over-Union (IoU) with previously selected masks less than 0.7, stability score of at least 0.85, and overlap ratio below 0.7. Each retained mask is assigned a unique masklet ID. These masks are then propagated across subsequent frames using SAM2, ensuring the consistency of masklet IDs. Every 20 propagated frames, we compute the uncovered area ratio. If this ratio increases by more than 1% relative to the baseline established at the previous keyframe, we treat this as a significant scene change. At these change points, SAM is reapplied to the current frame to detect new objects and update the set of active masklets accordingly.

To generate the final crops, we compute the largest bounding box that encloses all masks associated with each masklet instance over a fixed-length video clip. For practical reasons, we define the clip duration as the temporal input length of the downstream VideoLM. For each crop, we center this bounding box on the current position of the object and pad it to form a square. This ensures consistent spatial framing and contextual integrity while aligning with the VideoLM input requirements.

C.2. Feature Extractors.

We extract features using the following pretrained models:

- CLIP [7]: Visual features extracted using OpenCLIP ViT-B/32.
- SigLIP [16]: Visual features extracted using SigLIP-B/16 with 256 image patches.
- ImageBind [4]: Visual and audio features extracted using ImageBind-H.
- LaViLa [17]: Zero-shot visual features extracted using LaViLa TSF-B.
- EgoVideo [11]: Video features extracted using the EgoVideo model with 4-frame input segments.

C.3. NeRF Training

We extend the three-stream architecture of NeuralDiff [14] by adding a two-layer MLP to each feature stream. The output dimension of each MLP matches the corresponding feature embedding size: 256 for LaViLa, 512 for CLIP and EgoVideo, 768 for SigLIP, and 1024 for ImageBind. We train separate models for each feature type on a per-scene basis. Each experiment takes approximately 10 hours on a single NVIDIA A6000 GPU.

Training is conducted using hierarchical sampling with separate coarse and fine networks. We use a batch size of 1024 and optimize with Adam for 10 epochs. The initial learning rate is set to 5×10^{-4} and decays following a cosine annealing schedule. All feature embeddings are normalized after the rendering stage.

The average rendering time per frame varies by feature type, ranging from approximately 3 minutes for LaViLa to 10 minutes for ImageBind. Rendering is performed with a chunk size of 1024 to limit memory usage, measured on a single NVIDIA A6000 GPU. Since AViON4D is agnostic to the underlying 3D representation, it can be integrated

Table 2. **Ego-Exo4D Subset Statistics.** Statistics per scene for the Ego-Exo4D subset used in our experiments. **Scene ID**: Ego-Exo4D video identifier. **Length (s)**: video duration in seconds. **Frames**: total number of original frames. **Train**: number of frames included in our training split. **Test**: number of frames included in our test split. **Actions**: unique action types with ground truth annotations.

ID	Scene ID	Length (s)	Frames	Train	Test	Actions
01	unc_basketball_03-31-23_01_7	48	1511	1463	44	6
02	fair_bike_06_10	35	1059	1027	32	7

with more efficient rendering techniques, such as Gaussian Splatting [8], to significantly reduce rendering times.

C.4. Audio-Visual Relevance Score Fusion

To account for differences in score distributions and dynamic ranges across models, we use model-specific α values to fuse visual and audio similarity scores. A uniform α would be suboptimal, as the relative contributions of visual and audio features vary significantly between models. These variations influence both the scale and the effectiveness of cross-modal fusion, directly impacting localization performance. We determine the optimal α values via grid search, resulting in the following settings: LaViLa (0.5), CLIP (0.5), SigLIP (0.9), EgoVideo (0.8), and ImageBind (0.8). This adaptive weighting ensures a balanced contribution from both modalities, leading to consistently improved localization performance across all models.

D. Qualitative Results

The sequence in Figure 2 shows a person washing their hands at a kitchen sink (1), followed by washing a plate (2), and finally washing a coffee pot (3, 4) before placing it to dry (5, 6). The corresponding heatmaps below track temporal actions, with brightness indicating similarity scores. The model successfully identifies and localizes coffee pot washing while distinguishing it from other washing activities. Spatial predictions are accurate, with activations consistently focused on the sink area and intensifying during coffee pot washing.

Figure 3 demonstrates the advantage of video-based models (EgoVideo, LaViLa) over image-based models (CLIP, SigLIP, ImageBind) for temporally opposite actions. Video models correctly distinguish "take bread slices out of toaster" from "put back bread slices in toaster", while image-based models confuse these visually identical but temporally opposite actions, despite some (SigLIP, ImageBind) retrieving semantically related content.

Additional qualitative results (see Figure 4) demonstrate that relying solely on visual features can cause inaccurate spatio-temporal action localization, while incorporating audio enhances accuracy by providing action-specific auditory cues for both image and video models. For example, the query "wash spoon" is initially localized by CLIP to a static scene, but when the sound of running water is included, the action is correctly identified. For LaViLa when dealing with the query "take peas", audio helps suppress noise artifacts, enabling precise spatio-temporal localization.

E. Additional ablations

In this section, we evaluate different components of AViON4D, complementing the main ablation in the manuscript. For these experiments, we use all seven scenes of EK-100 subset, with CLIP, LaViLa, and EgoVideo models.

E.1. Audio Window.

We evaluate audio window lengths of 1, 2.5, and 3.5 seconds for extracting audio segments per frame (Table 3). The 1-second window falls below ImageBind's requirement of approximately 2-second audio clips sampled at 16 kHz (32,000 samples), triggering zero-padding and degrading audio feature quality. On the other hand, the 3.5-second window introduces excessive context, potentially overlapping information between unrelated frames, which also negatively affects performance. The 2.5-second window achieves the best trade-off as it is long enough to avoid padding while remaining temporally focused, leading to the highest accuracy across all three models.

E.2. Text Prompts.

AViON4D computes a relevancy score by comparing audio embeddings with textual descriptions. These queries can be either the same action descriptions used for visual similarity or dedicated descriptions of the associated sounds. In our setup, we use paired sound descriptions from the EPIC-Sounds Dataset. In this ablation, we compare this approach to using sound descriptions automatically generated by a Large Language Model (LLM) such as GPT-4 [10]. Specifically, we prompt GPT-4 with: "Which is the sound produced by *[action]*". Our results show that sound-focused descriptions lead to the greatest performance improvements, and that LLM-generated descriptions closely approximate the effectiveness of human-annotated ones.

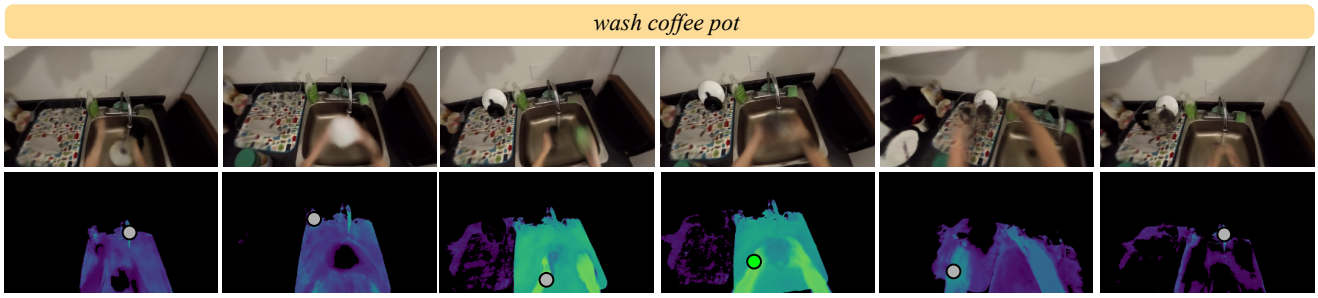

Figure 2. **Spatio-temporal localization for "wash coffee pot".** Top: Rendered video frames showing hand washing (1), plate washing (2), and coffee pot washing (3-4), then placing to dry (5-6). Bottom: Similarity heatmaps show the model correctly distinguishes coffee pot washing from other washing activities, with spatial predictions consistently focused on the sink area.

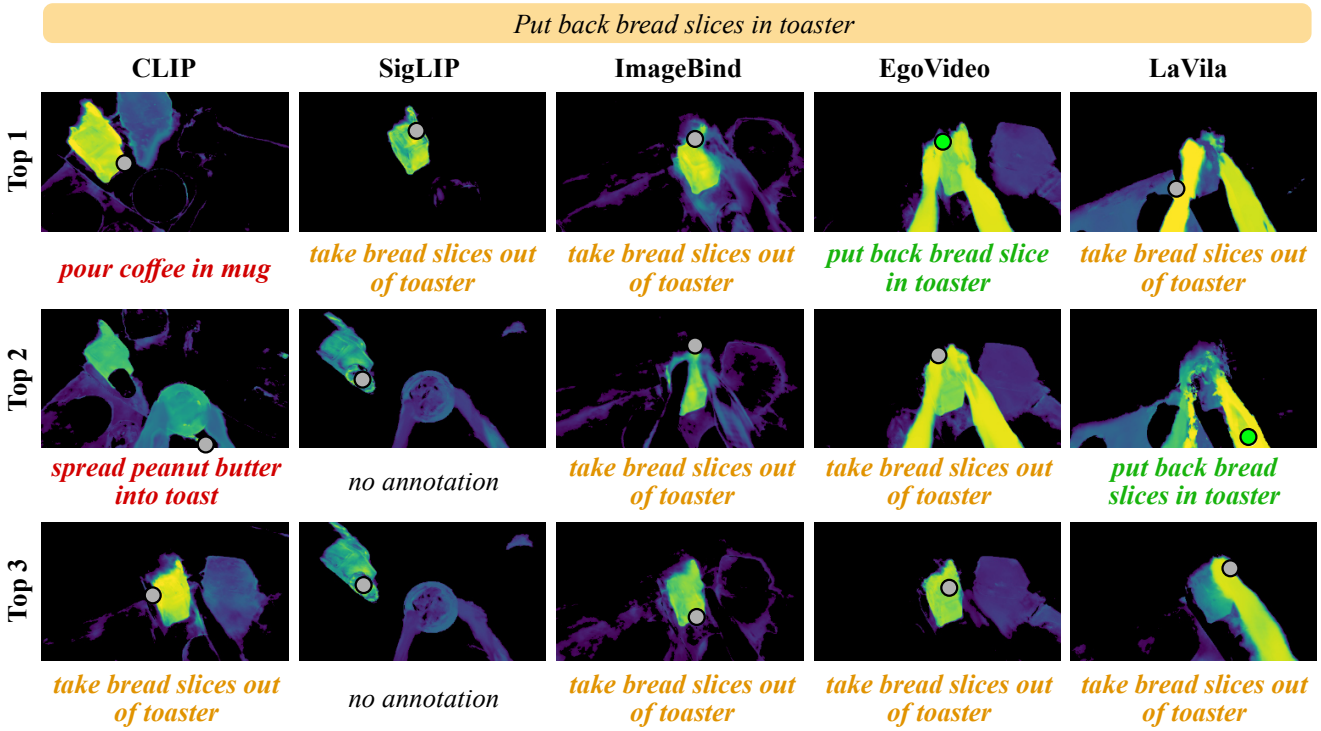

Figure 3. **Comparison of top-3 similarity maps for "put back bread slices in toaster".** Results across CLIP, SigLIP, ImageBind, EgoVideo, and LaViLa visual encoders. Ground truth annotations are shown below each prediction. Video models (EgoVideo, LaViLa) correctly retrieve the action (text in green), whereas image-based models either confuse it with the opposite action ("take bread slices out of toaster" in orange) or mistake it for a different action (text in red).

E.3. Audio Robustness to Background Noise.

We evaluate the robustness of our audio-visual fusion approach under noisy conditions by adding synthetic white noise to the audio tracks (Table 5). We added white noise with a Signal-to-Noise Ratio (SNR) of 15-20 dB to evaluate performance degradation under realistic background noise conditions. Prior analysis of the original audio revealed challenging baseline conditions: under-leveled recordings ($-$ 31.9 dB RMS), substantial ambient noise and poor frequency balance.

Despite these already degraded audio conditions, adding synthetic noise caused minimal performance degradation, typically 0% to 1.6% across all encoders. Importantly, audio-visual fusion continues to outperform visual-only baselines even with added noise. For instance, CLIP achieves 19.1% with noisy audio versus 13.5% with vision only, highlighting the robustness of our fusion strategy in noisy real-world conditions.

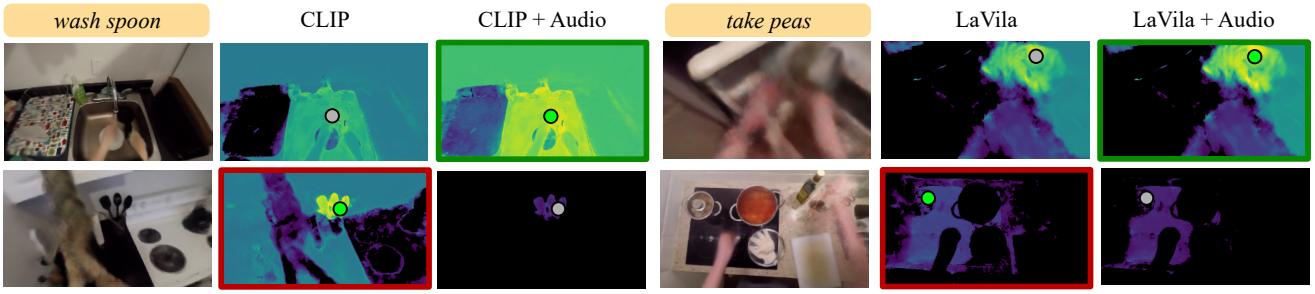

Figure 4. **Audio improves spatio-temporal action localization: additional results.** Left: Query "wash spoon" with CLIP. Without audio (middle), CLIP localizes spoons in a static scene; with audio (right), the action is correctly identified. Right: Query "take peas" with LaViLa. Audio helps suppress noise artifacts (middle vs. right), enabling precise spatio-temporal localization.

Table 3. **Audio Window Ablation.** Impact of window length on per-frame audio encoding. Spatiotemporal mean accuracies (Acc, Acc5) across CLIP, LaViLa, and EgoVideo models with different audio window sizes (1s, 2.5s, 3.5s). Row with gray background indicates the configuration used in our final method.

	CLIP		LaViLa		EgoVideo	
	Acc	Acc5	Acc	Acc5	Acc	Acc5
1	12.0	22.5	19.5	35.7	13.6	35.7
2.5	19.5	24.0	19.5	29.7	16.6	40.5
3.5	13.6	28.5	18.0	35.7	13.6	35.7

Table 4. **Text Prompt Ablation.** Impact of description type for audio feature relevance. Baseline: without audio. Action Desc.: using action descriptions. GPT Desc.: GPT-4 generated descriptions. Sound Desc.: sound-specific descriptions. Row with gray background indicates the configuration used in our final method.

	CLIP		LaViLa		EgoVideo	
	Acc	Acc5	Acc	Acc5	Acc	Acc5
Baseline	13.5	25.2	18.0	32.7	13.5	37.2
Action Desc.	6.1	22.6	19.5	27.1	13.6	35.9
GPT Desc.	12.0	22.4	16.4	26.8	13.5	36.9
Sound Desc.	19.5	24.0	19.5	29.7	16.6	40.5

Table 5. **Audio Robustness to Background Noise.** Impact of synthetic white noise (SNR 15–20 dB) on spatiotemporal mean spatiotemporal accuracies. Visual: vision-only baseline, Audio-visual (orig.): original audio conditions, Audio-visual + Noise: with added synthetic noise.

	CLIP		ImageBind		LaViLa	
Condition	Acc	Acc5	Acc	Acc5	Acc	Acc5
Visual	13.5	25.2	15.1	22.4	18.0	32.7
Audio-visual (orig.)	19.5	24.0	16.5	24.4	19.5	29.7
Audio-visual + Noise	19.1	24.0	15.1	24.1	19.5	29.7

F. Limitations

While AViON4D demonstrates promising results for egocentric open-vocabulary 4D understanding, it still faces limitations.

Object-Centric Masking and Spatial Precision. We extract object-centric video crops using square bounding boxes around segmentation masks, which inevitably include surrounding contextual regions. This is particularly evident for surface-type masks (*e.g.*, countertops, tables), where objects placed on these surfaces are unintentionally included in the crop. While temporal localization remains accurate, spatial precision is reduced in these cases.

Limitations of Foundation and 3D Models. AViON4D inherits the limitations of pretrained foundation models for both visual and audio feature extraction and 3D reconstruction. Semantic understanding quality is bounded by encoder training data and architectural choices. Reconstruction quality depends on the 3D representation backbone. Although NeuralDiff achieves state-of-the-art results in EK-100 egocentric scenes [15], like all NeRF-based methods, it can still produce artifacts in regions with extreme motion. These dependencies are common to all methods leveraging pretrained foundation models and 3D neural representations.

Per-Scene Optimization Overhead. Our current implementation, building on NeRF-based 3D representations, requires training a separate network for each video, taking approximately 10 hours per scene on a single NVIDIA A6000 GPU. This per-scene optimization is inherent to NeRF-based methods and limits real-time deployment.

These limitations stem primarily from the current state of foundation models and 3D representations rather than our core contribution. AViON4D is designed to be modular and representation-agnostic, enabling direct integration of alternative 3D representations (e.g., Gaussian Splatting) and improved encoders as they become available. Notably, these

limitations do not affect the validity of our experimental findings or diminish the demonstrated benefits of multimodal 4D scene understanding. As these technologies mature, we expect the practical impact of these limitations to diminish substantially.

G. Broader and Societal Impacts

Our multi-modal framework enables fine-grained understanding of actions in egocentric videos, offering promising applications in domains such as assistive technologies, human-robot interaction, and activity monitoring. However, these capabilities also raise critical societal considerations.

Privacy Risks. By localizing and interpreting user behavior in 3D space over time, the system may inadvertently expose sensitive personal information, particularly in private or domestic settings. The inclusion of audio further compounds these risks by capturing ambient sounds, which may reveal conversations or background noise not intended for analysis. Such concerns highlight the need for strong safeguards around data collection, storage, and usage.

Bias Propagation. Our method relies on pre-trained vision and language models, which are known to encode societal biases based on the data they were trained on. These biases may manifest in the interpretation or prioritization of actions, potentially leading to unfair or inaccurate outcomes. Without explicit mitigation strategies, such issues could be perpetuated or amplified in downstream applications.

Responsible Deployment. As egocentric devices and multimodal AI systems become more widespread, their deployment must be guided by principles of transparency, accountability, and fairness. This includes ensuring informed consent, enabling opt-out mechanisms, and actively working to reduce bias and protect user privacy.

H. List of Assets

We provide the implementation sources and model weights for all methods used in this work:

1. OpenCLIP [7, 12] (https://github.com/mlfoundations/open_clip). The open-source implementation of OpenAI's CLIP model is primarily licensed under the MIT license.
2. SigLIP [16] (https://huggingface.co/google/siglip-base-patch16-224). The code is released under an Apache-2.0 license.
3. ImageBind [4] (https://github.com/facebookresearch/ImageBind). The code and model weights are released under the CC-BY-NC 4.0 license.
4. LaViLa [17] (https://github.com/facebookresearch/LaViLa). The code and model weights are released under a MIT License.
5. EgoVideo [11] (https://github.com/OpenGVLab/EgoVideo. The repository by OpenGVLab does not explicitly state a license. The accompanying research paper is published under the CC-BY-NC 4.0 license.
6. AutoSeg-SAM2 [18] (https://github.com/zrporz/AutoSeg-SAM2). The code is released under a MIT License.
7. SAM [9] (https://github.com/facebookresearch/segment-anything). The source code and model weights are released under the Apache License 2.0.
8. SAM2 [13] (https://github.com/facebookresearch/sam2). The source code and pre-trained model weights are released under the Apache License 2.0.
9. NeuralDiff [14] (https://github.com/dichotomies/epic-fields-rendering). The official PyTorch implementation is licensed under the MIT License. The checkpoints are provided alongside the source code covered under the same MIT License.

References

[1] 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. 1

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1

[3] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 1

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3, 7

[5] 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. 1

[6] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023. 1, 2

[7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3, 7

[8] 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. 4

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 7

[10] OpenAI. Chatgpt: Language models are few-shot learners. https://chat.openai.com, 2023. Accessed: 2025-05-15. 4

[11] 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. 3, 7

[12] 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*, 2021. 7

[13] 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. In *ICLR*, 2025. 3, 7

[14] 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, 7

[15] 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. 1, 3, 6

[16] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3, 7

[17] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 3, 7

[18] Zrporz. Autoseg-sam2, 2024. Automated image segmentation tool based on Segment Anything Model (SAM). 3, 7