Audio-Visual Open-Vocabulary Egocentric Spatio-Temporal Action Localization with NeRFs

Egocentric video recordings from wearable devices capture detailed human-world interactions, creating valuable information for activity recognition and assistive technology development. The rapidly changing viewpoints, frequent occlusions, and continuous motion inherent to egocentric recordings create substantial obstacles for 3D scene understanding. Although neural radiance fields (NeRFs) [1] combined with semantic feature distillation from vision-language models [2] have achieved impressive view-consistent 3D semantic reconstructions, current approaches process individual frames in isolation and fail to leverage the temporal continuity present in video sequences.

In this work, we propose the first multimodal NeRF-based framework for 4D egocentric scene understanding, extending semantic understanding from static 3D to dynamic 4D scenes. This enables open-vocabulary spatiotemporal action localization by integrating audio-visual cues within a unified 4D representation—capabilities beyond existing 2D or static 3D methods. Our framework is component-agnostic and modular, demonstrated across 5 visual encoders (CLIP, SigLIP, ImageBind, LaViLa, EgoVideo) with systematic ablations validating each design choice.

We introduce object-centric tube-based video features and leverage audio-language models to capture temporal dynamics, allowing the system to ground object and action representations across space and time. By extending dynamic neural radiance field architectures [3] to jointly model visual-language features in 4D space while processing audio features as separate temporal signals, we create a queryable multimodal 4D representation.

Our framework processes egocentric videos through a two-stage pipeline: generating temporally consistent object-centric crops using SAM 2 [4] for spatio-temporal segmentation, then encoding these crops with video-language models to obtain semantic embeddings. Simultaneously, we extract audio features from temporally aligned audio segments using audio-language models such as ImageBind.

Extensive experiments on the EPIC-KITCHENS [5] dataset show that our multimodal framework provides absolute gains of up to 6.0% (44.4% relative improvement) over single-modality baselines. For the first time, we demonstrate that leveraging audio cues from egocentric videos not only enhances performance but also narrows the gap between single-frame and video-based models. Our 4D representation addresses instances where objects become temporarily occluded or actions remain partially observable due to rapid viewpoint changes by consolidating multi-modal signals across time and space, maintaining coherent object/action representations that 2D frame-based models cannot achieve. Ablation studies reveal that object-centric spatio-temporal crops significantly outperform grid-based extraction for egocentric videos due to continuous object motion, and optimal audio window lengths provide temporal context without noise.

This work demonstrates that multimodal fusion within a temporally aware NeRF framework enables robust open-vocabulary action localization in challenging egocentric environments, providing new possibilities for human activity analysis in 4D.

References

[1] B. Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." ECCV, 2020.
[2] J. Kerr et al. "LERF: Language Embedded Radiance Fields." ICCV, 2023.
[3] V. Tschernezki et al. "NeuralDiff: Segmenting 3D objects that move in egocentric videos." 3DV, 2021.
[4] N. Ravi et al. "SAM 2: Segment Anything in Images and Videos." ICLR, 2025.
[5] D. Damen et al. "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset." ECCV, 2018.