# Appendix

## A: Overview

The Supplementary material is organized as follows:

- Section B: Evaluation Datasets
- Section C: Current Challenges on Zero-shot Video Understanding
- Section D: Action Segmentation using VLMs
- Section E: Few-shot Learning for Action Segmentation

## B: Evaluation Datasets

Tab. 1 summarizes the current challenging datasets targeting human behavior analysis. In this paper, we focus on two current challenging tasks, zero-shot classification and frame-wise segmentation tasks. Specifically, we perform the study on real-world scenarios [9, 17, 32, 7, 27] and laboratory scenarios [26, 20] for action understanding including both zero-shot classification and frame-wise segmentation tasks.

**Toyota Smarthome** (Smarthome) [9] is a real-world human-centric daily living action classification dataset. The dataset is challenging as the inter-class variance is small and the activities are fine-grained. It contains 16,115 videos across 31 action classes, offering RGB and skeleton data. We utilize only RGB data, following cross-subject (CS) and cross-view2 (CV2) protocols and we report Top-1 accuracy in this work.

**UAV-Human** [17] features 22,476 UAV-captured human-centric videos, we use the RGB data and follow Cross-subject evaluations (CS1).

**Penn Action** [32] comprises 2,326 sequences of 15 simple sport actions, we use this dataset for action classification using standard train-test splits.

**NTU-RGB+D 60** [26] includes 60 indoor daily living activities and consists of 56,880 RGB-D video sequences with 3D skeletons, captured by the Microsoft Kinect v2 sensor. We only use RGB videos in this work and we follow the cross-subject (CS) evaluation protocol.

**EgoExo4d** [10] is large-scale multimodal multiview video dataset containers totaling 1,286 hours of videos with range between 1 to 42 minutes. It provides ego-centric videos paired with multiple time-synchronized exo-centric video streams, capturing a wide range of skilled human activities. It enriched with extensive annotations including language descriptions, 3D body and hand poses, key steps, procedural dependencies, and proficiency ratings. These densely annotation support various benchmark tasks in video understanding ego-exo relation modeling, action recognition, proficiency estimation, and 3D pose recovery. We only use RGB videos modality of key-step and their correspondence label to evaluate zero-shot action classification.

**LEMMA** [14] consists of a large collection of videos designed to capture multi-agent and multi-task activities from multiple viewpoints. It contains over 324 long video clips that cover diverse activities involving 641 actions and 11,781 action segment . Each video is annotated with detailed information, such as activity labels, agent roles, object interactions, and temporal segmentation. These videos are recorded from various camera angles to provide a comprehensive multi-view perspective, enabling the study of tasks like action recognition and action segmentation.

**Toyota Smarthome Untrimmed** (TSU) [7] extends the action classes and video counts of Smarthome, focusing on frame-wise segmentation tasks. The dataset is very challenging, as each action can be performed multiple times in a video and multiple actions can be performed at the same time. We use TSU for evaluating the generalizability of SoTA models and we report per-frame mAP following Cross-Subject (CS) and Cross-View (CV) evaluation protocols.

**Charades** [27] focuses on fine-grained activities segmentation. It contains many object-oriented activities and variant light conditions. The current methods are still limited to dealing with this dataset, hence, we use this dataset for our study and we report per-frame mAP.

| Dataset | Real-world | 2D | 3D | #Videos | #Actions | Fine-grained | Type | Task |
|---|---|---|---|---|---|---|---|---|
| NTU-RGB+D 60 [26] | × | ✓ | ✓ | 56,880 | 60 | No | Daily living | AC |
| NTU-RGB+D 120 [20] | × | ✓ | ✓ | 114,480 | 120 | No | Daily living | AC |
| Penn Action [32] | ✓ | ✓ | × | 2,326 | 15 | No | Sport | AC |
| UAV-Human [17] | ✓ | ✓ | × | 21,224 | 155 | No | UAV | AC |
| Toyota Smarthome [9] | ✓ | ✓ | ✓ | 16,115 | 31 | Yes | Daily living | AC |
| EgoExo4D [10] | ✓ | ✓ | ✓ | 5035 | 664 | Yes | General video | AC |
| Kinetics [1] | ✓ | × | × | 400,000 | 400 | No | General video | AC |
| LEMMA [] | ✓ | ✓ | ✓ | 324 | 641 | Yes | Daily living | AF |
| PKU-MMD [4] | × | ✓ | ✓ | 1,076 | 51 | No | Daily living | AS |
| Charades [27] | ✓ | × | × | 2,300 | 151 | Yes | Daily living | VD-AS-AF |
| TSU [7] | ✓ | ✓ | ✓ | 536 | 51 | Yes | Daily living | VD-AS-AF |
| Activity-Net [11] | ✓ | - | - | 20k | - | Yes | General video | VR |
| DiDeMo [12] | ✓ | - | - | 10.5K | - | Yes | General video | VR |
| MSR-VTT [30] | ✓ | - | - | 7.2K | - | Yes | General video | VR |

Table 1: A survey of recent datasets for in-the-wild human action classification (top), action segmentation (bottom).

| Methods | Training Data | Type | Task |
|---|---|---|---|
| CLIP [23] | CLIP-400-M/LAION-2B | ILM | AC-VR |
| X-CLIP [21] | CLIP-400M/Kinetics-400 | VLML | AC-VR |
| ViCLIP [28] | InternVid-10M-FLT | VLM | AC-VR |
| ViFi-CLIP [24] | CLIP-400M/Kinetics-400 | VLM | AC-VR |
| LanguageBind [33] | VIDAL-10M | ILM/VLM | AC-VR |
| Video-LLaMA2 [3] | Webvid-2M /LLaVA-CC3M | VLLM | AC-VD-AF |
| LongVA [31] | V-NIAH | VLLM | AC-VD-AF |
| Video-LLaVA [18] | LAION-CC-SBU/Valley/LLaVA-mixed/Video-ChatGPT | VLLM | AC-VD-AF |
| LLaVA-OneVision [16] | LLaVA-Hound-255K | VLLM | AC-VD-AF |
| LAVIDAL [2] | ADL-X | VLLM | AC-VD-AF |
| Video-Chatgpt [22] | VideoChatGPT | VLLM | AC-VD-AF |
| UniVTG [19] | Ego4D/VideoCC/CLIP teacher | VLLM | AS |
| TimeChat [25] | TimeIT | VLLM | AS |
| VTimeLLM [13] | LCS-558K/InternVid-10M-FLT/VideoInstruct100K | VLLM | AS |

Table 2: A survey of SoTA architectures, AC:Action Classification, VR: Video Retrieval, VD: Video Description, AF: Action Forecasting, AS: Action Segmentation

The mentioned datasets are different from the datasets of web videos used for training video foundation models. Our selected evaluated datasets can further reflect the generalization ability of video foundation models on daily living scenarios.

# C: Current Challenges on Zero-shot Video Understanding

In this work, we provide an analysis of the performance of current vision-language foundation models with five challenging video-based tasks to study to study the transfer ability performance of video representation and their alignments with language. The five tasks are: zero-shot action classification, video-text retrieval, video description, action forecasting, and frame-wise temporal action segmentation. The evaluation and comparisons are performed on real-world datasets.

**Action Classification** Zero-shot action classification is to pre-train an action classification model and then transfer this model onto an unseen dataset. Unlike traditional methods that rely on extensive action labels, zero-shot approaches aim to generalize knowledge from known actions to unknown ones. Specifically, the semantic information, such as textual descriptions of the action labels, and the videos in the dataset are embedded using CLIP-based methods [29, 21, 28, 24]. Subsequently, given a video embedding, we search for its closest semantic information as the action prediction. We select such tasks as it highly relays on video-text alignment but has not been fully evaluated by current research.

In real-world video understanding applications, the ability to recognize actions without the need for specific training data is invaluable. However, visual features are often low-level, such as shapes, colors, and motions, while action descriptions are more abstract, this makes the model difficult to accurately match the two types of features. Additionally, current zero-shot learning models are still limited to dealing with variations in camera angles, lighting conditions, etc. Hence, this study aims to evaluate and compare the CLIP-based vision language foundation models including VLMs and VLLMs on such tasks focusing on real-world scenarios.

| Methods | TSU | | Charades |
|---|---|---|---|
| | CS(%) | CV(%) | mAP(%) |
| PDAN [6] w/ CLIP [23] | 16.3 | 10.0 | 15.9 |
| PDAN [6] w/ ViCLIP [28] | 21.5 | 13.4 | 16.2 |
| PDAN [6] w/ ViFi-CLIP [24] | **28.6** | **15.9** | **16.4** |
| MS-TCT [5] w/ CLIP [23] | 5.3 | 5.7 | 12.7 |
| MS-TCT [5] w/ ViCLIP [28] | 15.8 | 8.2 | 16.3 |
| MS-TCT [5] w/ ViFi-CLIP [24] | **21.3** | **17.3** | **16.9** |
| MS-TCT [5] w/ I3D [1] (SoTA) | **33.7** | - | **25.4** |

Table 3: Frame-level mAP on TSU and Charades for comparison of SoTA vision foundation models with SoTA temporal modeling methods for action segmentation.

| Methods | Label | TSU | | Charades |
|---|---|---|---|---|
| | | CS(%) | CV(%) | mAP(%) |
| PDAN [6] w/CLIP [23] | 5% | **6.2** | 4.3 | 8.7 |
| PDAN [6] w/ViCLIP [28] | 5% | 3.5 | 3.3 | 10.1 |
| PDAN [6] w/ViFi-CLIP [24] | 5% | 5.6 | **5.7** | **11.1** |
| PDAN [6] w/CLIP [23] | 10% | 4.4 | 4.7 | 11.1 |
| PDAN [6] w/ [28] | 10% | 4.0 | 3.5 | **11.6** |
| PDAN [6] w/ViFi-CLIP [24] | 10% | **6.1** | **5.8** | 11.3 |

Table 4: Frame-level mAP on TSU and Charades with randomly selected **5% (top)** and **10% (bottom)** for action segmentation.

**Video-Text Retrieval** Video-text retrieval is considered as another type of zero-shot task on a different dataset format where each video in this dataset has a unique description. Its goal is to search and retrieve relevant video content based on a given text query and vice versa.These tasks are commonly used to evaluate how well vision-language models can generalize their learned representations to connect video content with descriptive text.

**Video Description** Following [22], we conduct a comprehensive evaluation of Video-Large Language Models (VLLMs) based on their text generation capabilities, specifically focusing on their ability to produce dense, informative descriptions for input videos. The generated descriptions are assessed in comparison to ground truth annotations using five key metrics: Correctness of Information, Detail Orientation, Contextual Understanding, Temporal Understanding, and Consistency. This evaluation is crucial for benchmarking the model's ability to comprehend visual content and generate meaningful, contextually appropriate text, a key requirement for tasks like automated video captioning, summarization, and human-computer interaction. Following [2], TSU videos are trimmed into 1-minute clips and are input to the VLLMs. Thereafter, the clip-level descriptions are concatenated and summarized into a single video-level description using GPT-3.5 turbo. For Charades, descriptions are obtained directly from each video.

**Action Forecasting** Action forecasting evaluates an agent's ability to predict an action before it happens. Given a human action video and the corresponding actions that occur in the video, the agent's goal is to choose the action that immediately follows the observed sequence of actions. This task was popularized by challenges such as EPIC-KITCHENS [8] and Breakfast [15] to measure the action concept reasoning abilities of vision models. In this work, we follow the protocol proposed in [2], in which action forecasting is evaluated in a MCQ manner on the Toyota Smarthome Untrimmed [7] and LEMMA [14] datasets.

**Frame-wise Action Segmentation in Untrimmed Videos** Temporal Action Segmentation focuses on per-frame activity classification in untrimmed videos. The main challenge is how to model long-term relationships among various activities at different time steps. Specifically, action segmentation entails the automatic partitioning of untrimmed video sequences into distinct segments, each corresponding to a coherent action. Current methods [6, 5] have two steps, they firstly extract visual features on top of the temporal segments of a long-term video using a strong video encoder. Secondly, they design temporal modeling to process the features. Hence, the performance of the temporal modeling highly relies on the video encoder from current video foundation models. In this study, we compare SoTA vision foundation models [23, 21, 28, 24] by evaluating their features on temporal action segmentation tasks.

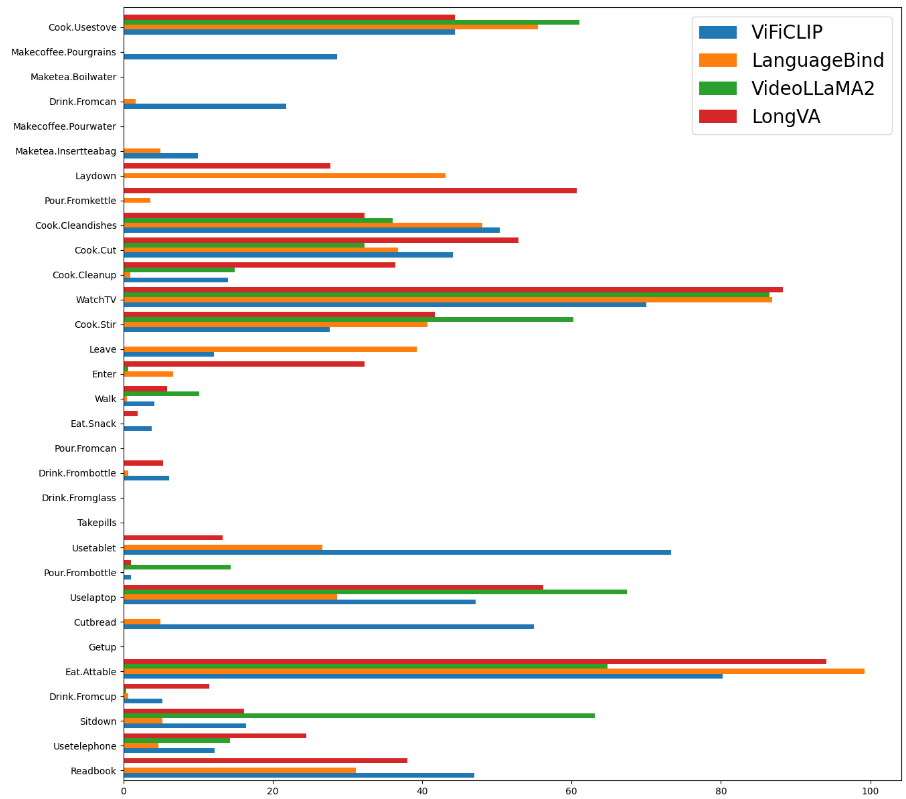


Figure 1: Per-action zero-shot classification analysis on Smarthome-CS for VLMs and VLLMs

# D: Action Segmentation using VLMs

In this section, we compare the performance of the visual-language models in action segmentation tasks. As current methods for action segmentation tasks adopt a temporal model to process the continuous pre-extracted visual features on top of the untrimmed video, this experiment is to compare the representation ability of a single visual encoder of SoTA models [23, 28, 24] using their visual features with two recent temporal models [6, 5] respectively. The results in Tab. 3 show that similar to zero-shot action classification, the visual representation of ViFi-CLIP is more effective than other models for segmentation tasks. We also observe that the performances of Vision Language Foundation models are still not at the level of State-of-the-art action detection methods [5]. This can be explained by the fact that these Foundation models have been trained on web videos, which are quite different from Activity of Daily Living (ADL) Videos, such as TSU or Charades.

# E: Few-shot Learning for Action Segmentation

learning is commendable and enables obtaining good accuracy with limited labeled data. This highlights the model practicality in real-world applications where data scarcity is prevalent. The few-shot transfer ability of our evaluated CLIP-based models on top of temporal modeling [5] is shown in Tab. 4. The results are consistent with previous evaluation, ViFi-CLIP [24] has mostly the best visual representation ability.

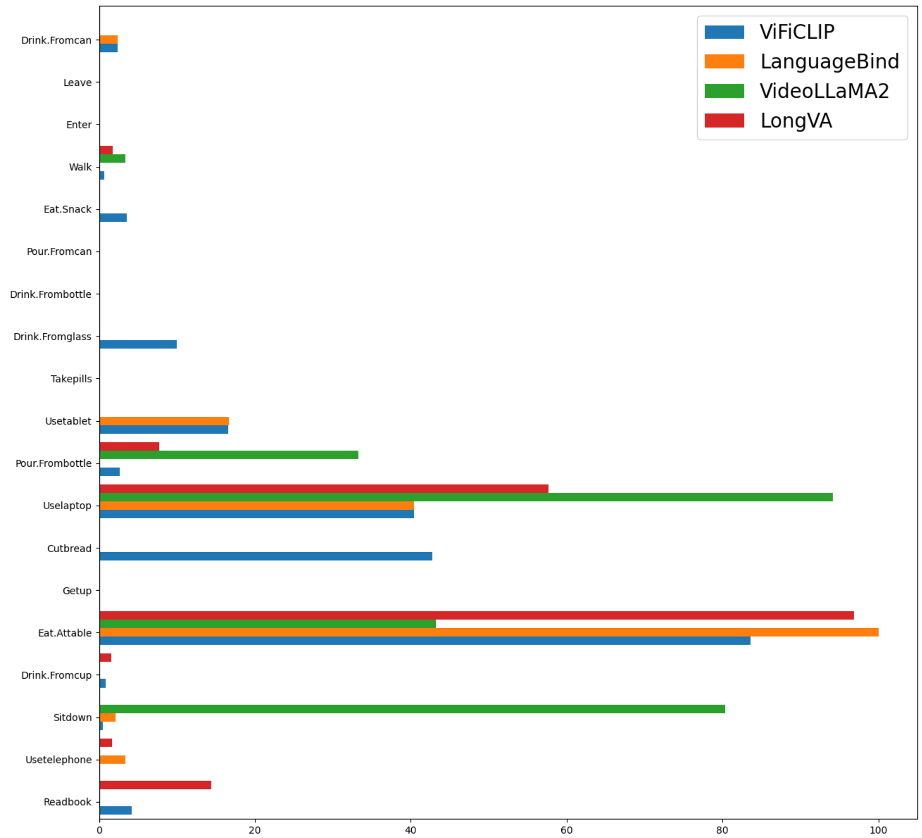


Figure 2: Per-action zero-shot classification analysis on Smarthome-CV for VLMs and VLLMs

ChatGPT

**NTU-10:**

*Label*: Falling

*Augmented Label:* A video of person {falling}

*Description:* A person falling involves the individual losing their balance or stability, leading to a sudden descent to the ground or a lower surface due to various factors such as slipping, tripping, or experiencing a loss of equilibrium.

**Smarthome:**

*Label:* Pour from bottle

*Augmented Label:* A video of person {pour from bottle}

*Description:* Pour from bottle:A person is holding a bottle and slowly tipping it over to let the liquid inside flow out in a controlled manner.

**PennAction:**

*Label:* squats

*Augmented Label:* A video of person {squatting}

*Description:* Pour from bottle:A person performing squats by bending their knees and hips to lower their body and then returning to a standing position.

Figure 3: Different label formats for PennAction, NTU60 and Smarthome generated by GPT-3.5 .

```
"Cooking and Food Preparation":
  {
   "Cutting and Chopping": ["Cut green chilies", "Cut cucumber", "Cut cabbage", "Cut carrots", "Cut beef", "Cut bell peppers", ...],
   "Adding Ingredients": ["Add cheese to the dish", "Add lemon to the recipe", "Add radish to the salad", "Add celeries to the mixture", ...],
   "Getting Ingredients": ["Get oil from the pantry", "Get cinnamon stick from the spice rack", "Get cilantro from the fridge", ...],
   "Washing": ["Wash knife in the sink", "Wash cherry tomatoes under running water", "Wash lettuce in the colander", "Wash pot or saucepan in the sink",..],
   "Pouring and Mixing": ["Pour the coffee into a cup or mug", "Pour hot water into the cup", "Pour milk into the glass", "Pour cold milk into the bowl"],
   "Peeling": ["Peel coriander leaves from the stem", "Peel garlic cloves from the bulb", "Peel cucumber with a peeler", "Peel onions with a knife", ...],
   "Heating and Cooking": ["Heat the saucepan on the stove", "Turn on the stove to preheat", "Simmer the tea on low heat", "Turn off the stove when done"]
   "Putting Away Items": ["Put away lemon juice in the fridge", "Put away pot holder in the drawer", "Put away napkin in the laundry", ...]
  },
"Bicycle Repair and Maintenance":
  {
   "Tire and Brake Maintenance": ["Remove the wheel carefully", "Pull the tire lever around the rim", "Release the brakes", ...],
   "Chain and Gear Maintenance": ["Oil the chain thoroughly", "Loosen the bolt of the chain tensioner", "Lubricate the bike chain", ...],
   "Cleaning and General Maintenance": ["Clean the bike frame","Clean the wheel hubs","Wipe off the excess oil", "Tighten the loose bolts on the bike",.]
  },
"COVID-19 Testing":
  {
   "Sample Collection and Preparation": ["Collect saliva sample", "Rotate and swirl swab", "Slowly extract swab", "Dip swab in testing tube",...],
   "Testing Process": ["Place the test strip into the testing tube", "Cover the test tube", "Check the test kit", "Fold the instructions", ...],
   "Patient Interaction": ["Confirm patient consciousness", "Tap patient to confirm consciousness", "Do artificial respiration",...]
  }
  .
  .
  .
```


Figure 4: EGOEXO4D Coarse-grained labels generated by GPT3.5.

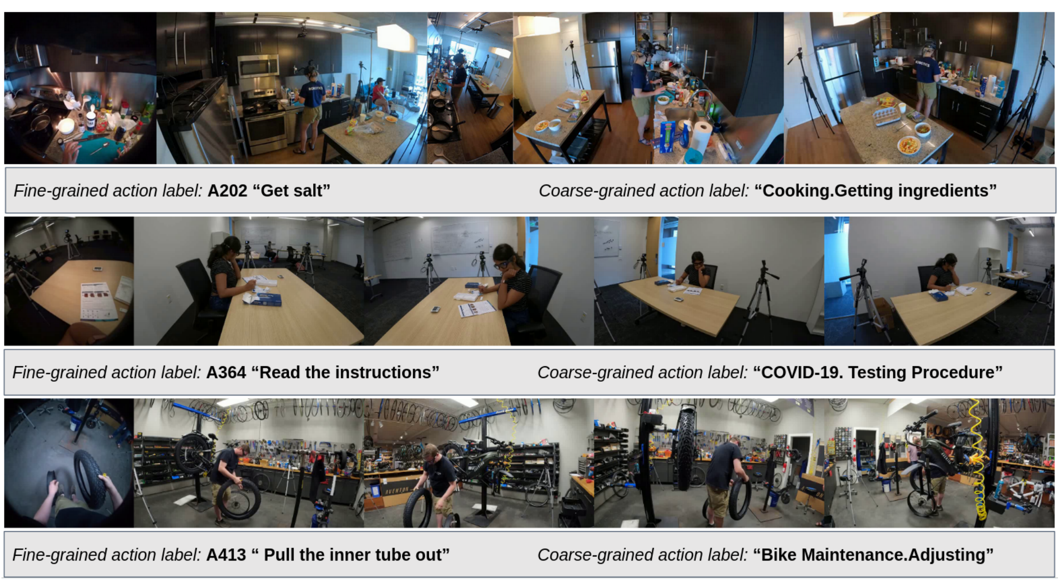


Figure 5: EGOEXO4D samples. For each video there are five view (ego, exo1, exo2, exo3, exo4), Fine-grained action label and Coarse-grained labelss.

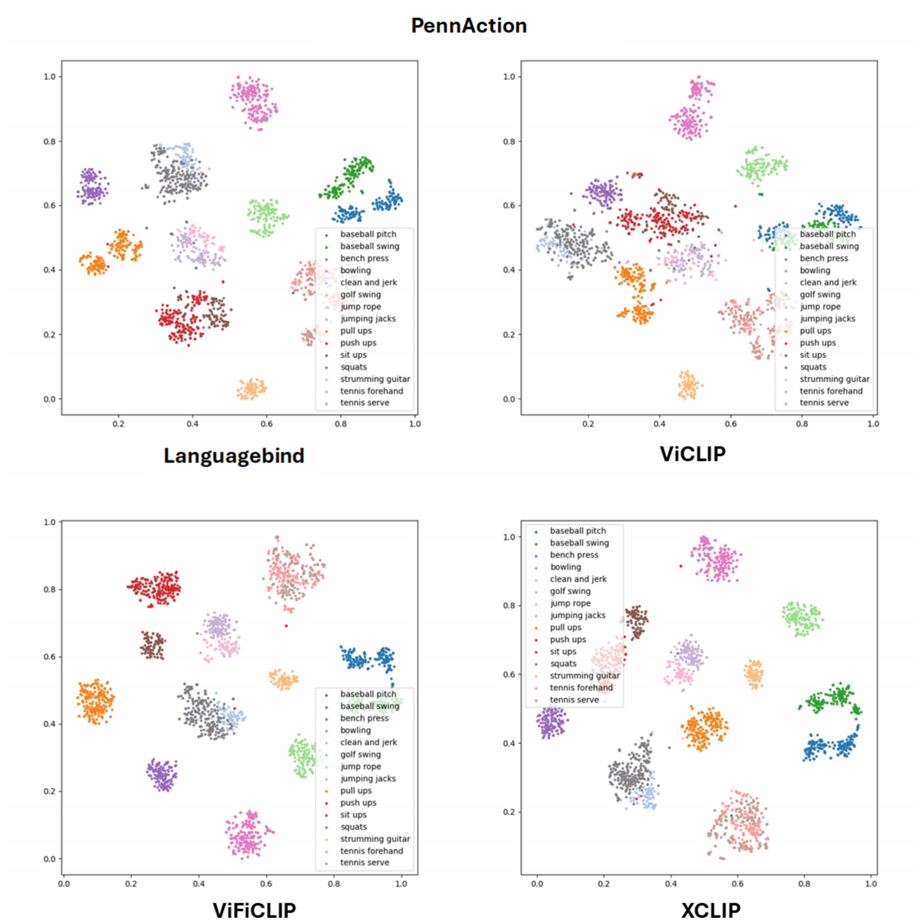


Figure 6: TSNE visualization of VLMs features for PennAction dataset .

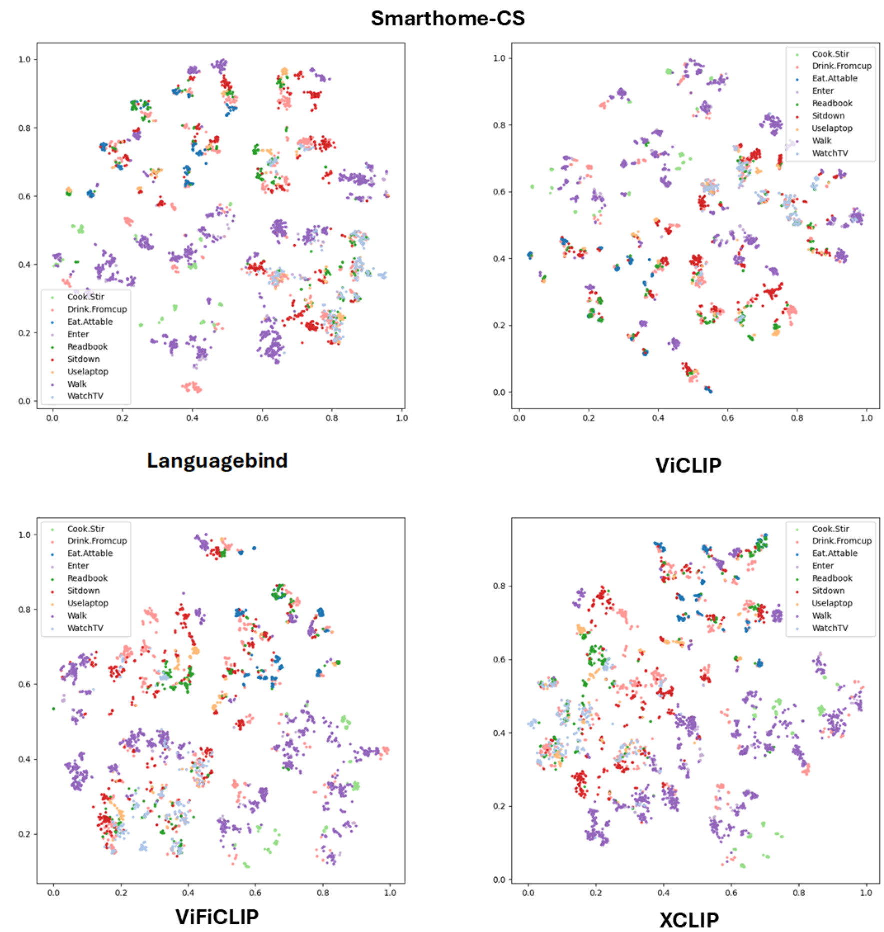


Figure 7: TSNE visualization of VLMs features for Smarthome-CS dataset .

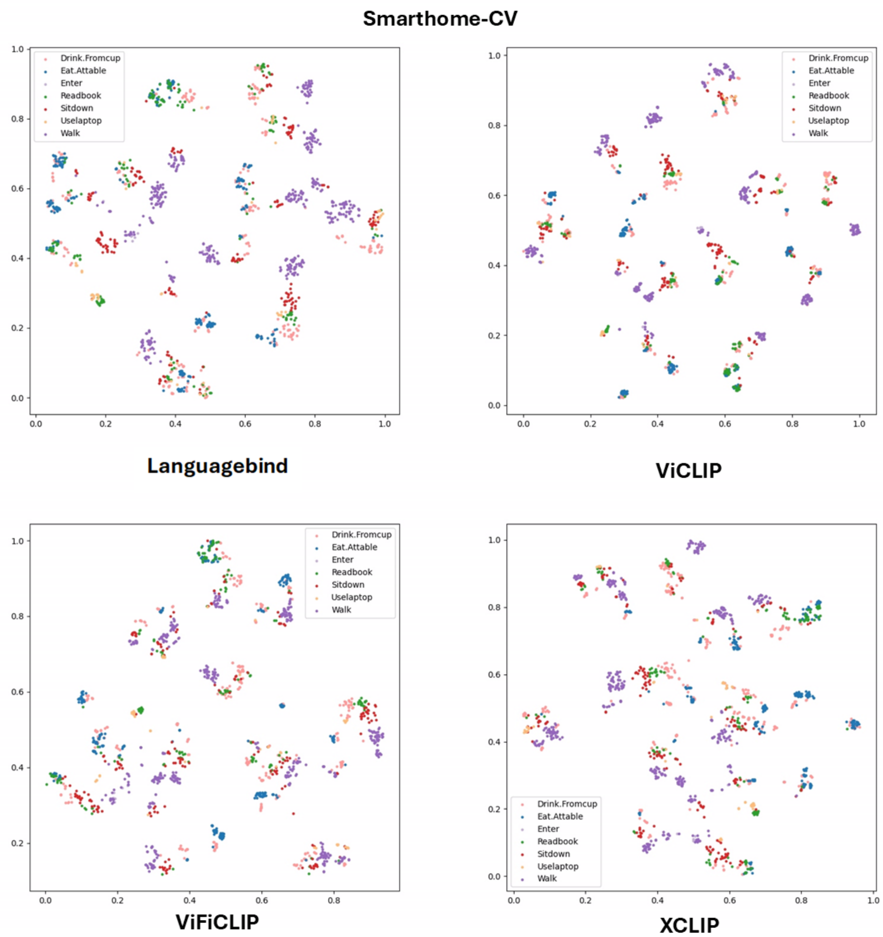


Figure 8: TSNE visualization of VLMs features for Smarthome-CV dataset .

# References


[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[2] Rajatsubhra Chakraborty, Arkaprava Sinha, Dominick Reilly, Manish Kumar Govind, Pu Wang, Francois Bremond, and Srijan Das. Llavidal: Benchmarking large language vision models for daily activities of living. *arXiv preprint arXiv:2406.09390*, 2024.

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[4] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv:1703.07475*, 2017.

[5] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael Ryoo, and Francois Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In *CVPR*, 2022.

[6] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Pdan: Pyramid dilated attention network for action detection. In *WACV*, 2021.

[7] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE TPAMI*, 2022.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[9] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *ICCV*, 2019.

[10] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[12] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

[13] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, 2024.

[14] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *ECCV*, 2020.

[15] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[17] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, 2021.

[18] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[19] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2023.

[20] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3D human activity understanding. *IEEE TPAMI*, 2020.

[21] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP:: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*, 2022.

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[24] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *CVPR*, 2023.

[25] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024.

[26] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3D human activity analysis. *CVPR*, 2016.

[27] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[28] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024.

[29] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019.

[30] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[31] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[32] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013.

[33] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024.