Are Visual-Language Models Effective Action Recognition? A Comparative Study

Mahmoud Ali*, Di Yang*, and François Brémond

Inria Center at Université Côte d'Azur, France
{ali.mahmoud, di.yang, francois.bremond}@inria.fr

Abstract. Current vision-language foundation models, such as CLIP, have recently shown significant improvement in performance across various downstream tasks. However, whether such foundation models significantly improve more complex fine-grained action recognition tasks is still an open question. To answer this question and better find out the future research direction on human behavior analysis in-the-wild, this paper provides a large-scale study and insight on current state-of-the-art vision foundation models by comparing their transfer ability onto zero-shot and frame-wise action recognition tasks. Extensive experiments are conducted on recent fine-grained, human-centric action recognition datasets (*e.g.*, Toyota Smarthome, Penn Action, UAV-Human, TSU, Charades) including action classification and segmentation.

Keywords: Video understanding · video foundation model · action recognition · multi-modal learning

1 Introduction

Recent vision-language foundation models [2, 11, 15, 18, 20, 25–27, 30, 34, 35], with a large-scale pre-training, have achieved promising results on many downstream vision tasks due to their impressive generalizability. Among these, the models with visual-language pre-training like CLIP [20] and its successors for video tasks [17,21,31] have revolutionized a myriad of downstream tasks, demonstrating unprecedented versatility and performance.

Despite these successes, the evaluations are primarily focused on general video understanding tasks such as video captioning, video-text retrieval, etc. The capability of these visual-language models to handle more complex and fine-grained action understanding tasks remains under-explored, such as zero-shot action classification and multi-label action segmentation. As these tasks are critical for many applications (*e.g.*, healthcare monitoring and robotic learning), it is essential to understand the current challenges of visual-language models targeting fine-grained human action recognition. Hence, in this paper, we evaluate and compare current state-of-the-art (SoTA) visual-language models, with a particular focus on their performance in zero-shot classification and action segmentation tasks.

---

* Equal contribution.

To further understand how to take good advantage of video-language model for zero-shot action recognition, we firstly compare different kinds of action descriptions obtained from raw action labels and LLMs (*e.g.*, ChatGPT) for action classification to find out which kind of prompt is more appropriate for visual-language models. Secondly, for zero-shot action segmentation in untrimmed video, we apply current video question answering (VQA) models [13, 23] with post-processing for generating frame-level action predictions. We conduct a comparative study on such methods and classical action segmentation methods on more challenging multi-label dataset [8] to fully understand the advantages and limitations of each current approach.

In summary, the contributions of this paper are the following. (i) We perform a large-scale study on evaluating current vision-language foundation models focusing on transfer-learning onto in-the-wild action recognition tasks. (ii) We further provide insight and comparisons on different action description generation strategies for zero-shot action classification, and on different frame-wise action prediction strategies using video question answering (VQA) models for zero-shot action segmentation. (iii) Extensive experiments are conducted using a good number of in-the-wild benchmarks.

2   SoTA Multi-modal Video Foundation Models

Recently, many methods have used language features [20] for video understanding [15,18,25,26,30,34], video captioning [35] and visual question answering [2,27]. However, these methods are designed to handle short temporal videos, and the challenge of handling actions over a long range of time for solving the task of action detection still persists. These models, especially InternVideo [32], aim to understand and generate descriptions of video content, facilitating a multi-modal understanding of visual data.

In this study, we select the most recent and widely used models [17,20,21,23,31] (see Tab. 1) for comparisons and discussions.

CLIP [20] is the first well known vision foundation model via visual-language pre-training. The key idea is to pre-train a transferable vision encoder using natural language supervision. The vision encoder is trained on a large number of image-text pairs by contrastive learning. CLIP successfully demonstrates that semantic information can significantly improve the representation ability of visual encoder for many downstream vision tasks, e.g., image classification, object detection. However, as the visual encoder of CLIP is trained based on images but not videos, CLIP is still limited on performance when transferred onto video-based tasks highly relying on temporal reasoning. In this study, we evaluate CLIP for video-based action recognition tasks as baseline model.

X-CLIP [17] presents a novel multi-grained contrastive model for video-text retrieval. To effectively aggregate fine-grained and cross-grained similarity matrices to instance-level similarity, X-CLIP proposes the Attention Over Similarity Matrix (AOSM) module to make the model focus on the contrast between es-

Methods	Training Data	Backbone	Training Strategy
CLIP [20]	CLIP-400-M/LAION-2B	ResNet/ ViT	image-text CL
X-CLIP [17]	CLIP-400M/Kinetics-400	Vit-B	video-text CL
ViCLIP [31]	InternVid-10M-FLT	Vit-L	video-text CL
ViFi-CLIP [21]	CLIP-400M/Kinetics-400	ViT-B	video-text CL
LanguageBind [38]	VIDAL-10M	Vit-L	image/audio/video-text CL

Table 1: A survey of SoTA architectures, CL: Contrastive Learning.

sential frames and words, thus lowering the impact of unnecessary frames and words on retrieval results. With multi-grained contrast and the proposed AOSM module, X-CLIP achieves outstanding performance on video-text retrieval tasks. In this work, we evaluate and compare X-CLIP focusing on fine-grained action recognition tasks with other SoTA CLIP-based approaches.

ViCLIP [31] is a general video foundation model. It applies the Vision Transformer (ViT) [10] with spatio-temporal attention as the video encoder and uses a Transformer-based text encoder following [20]. It develops its capabilities through a mix of self-supervised methods, including masked modeling [29] and cross-modal contrastive learning [19] for in-depth feature representation, allowing for efficient learning of transferable video-language representation. As the video and text encoders are well pre-trained on a web-scale video-language dataset [31] including 7 million videos, corresponding to 234 million clips each with the generated captions, ViCLIP can be used for video and text feature extractions.

ViFi-CLIP [21] explores the capability of a simple baseline called ViFi-CLIP (Video Fine-tuned CLIP) for adapting image pretrained CLIP to the video domain. ViFi-CLIP tackles the challenge of missing temporal relationships of image-based CLIP model, which can effectively improve the video-based downstream tasks. In this paper, we further evaluate this method for more fine-grained tasks.

LanguageBind [38] is a multimodal model that primarily uses language to connect different data types, like videos, infrared images, depth maps, and audio using contrastive learning. It is trained on a large-scale dataset (VIDAL-10M) which contains 10 million samples for all these data types and their corresponding text descriptions. To enhance the model's understanding and the language semantic information, the text descriptions are improved by incorporating metadata, spatial, and temporal information. Additionally, ChatGPT is used to refine the language and create a better semantic representation for each data type.

TimeChat [23] is a time-sensitive multi-modal large language model specifically designed for long video understanding. Timechat is trained on an instruction-tuning dataset, encompassing 6 tasks and a total of 125K instances. This model shows promising zero-shot results on video understanding tasks including dense captioning, temporal grounding, and highlight detection. As this model can directly provide action segmentation prediction by asking related questions without additional training on downstream datasets, we compare this model

to very challenging action segmentation tasks to understand its generalization ability.

UniVTG [13] proposes to Unify the diverse Video Temporal Grounding (VTG) labels and tasks. Thanks to the unified framework, the temporal grounding pre-training is available from large-scale diverse labels and develops stronger grounding abilities *e.g.*, zero-shot grounding. Similar as TimeChat, the zero-shot grounding can provide event boundaries related to action, hence, UniVTG can be used for zero-shot action detection segmentation tasks. In this study, we are the first to provide experimental results using UniVTG for more complex multi-label and frame-wise action segmentation tasks.

All mentioned approaches achieve SoTA performance on many tasks including video-text retrieval, temporal grounding, video captioning, etc. Most tasks are based on web videos and highly relys on video-text alignment quality, while are not focused on daily living action recognition scenarios. It is critical to understand the performance and current challenges of SoTA foundation models for action recognition tasks, so we provide an analysis on this topic to find out more future directions based on the analysis.

3   Current Challenges on Action Recognition

In this work, we provide an analysis of the performance of current vision foundation models with two challenging video-based tasks: zero-shot action classification and frame-wise temporal action segmentation. The evaluation and comparisons are performed on real-world datasets.

3.1   Zero-shot 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 [17, 21, 31, 33]. Subsequently, given a video embedding, we search for its closest semantic information as the action prediction. We select such tasks as it highly relys 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 foundation models on such tasks focusing on real-world scenarios.

Dataset	Real-world	2D	3D	#Videos	#Actions	Fine-grained	Type
NTU-RGB+D 60 [24]	×	✓	✓	56,880	60	No	Daily living
NTU-RGB+D 120 [14]	×	✓	✓	114,480	120	No	Daily living
Penn Action [37]	✓	✓	×	2,326	15	No	Sport
UAV-Human [12]	✓	✓	×	21,224	155	No	UAV
Toyota Smarthome [9]	✓	✓	✓	16,115	31	Yes	Daily living
Kinetics [4]	✓	×	×	400,000	400	No	General video
PKU-MMD [5]	×	✓	✓	1,076	51	No	Daily living
Charades [28]	✓	×	×	2,300	151	Yes	Daily living
TSU [8]	✓	✓	✓	536	51	Yes	Daily living

Table 2: A survey of recent datasets for in-the-wild human action classification (top), action segmentation (bottom).

3.2   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, 7] 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 [17, 20, 21, 31] by evaluating their features on temporal action segmentation tasks.

3.3   Evaluation Datasets

Tab. 2 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 [8, 9, 12, 28, 37] and laboratory scenarios [14, 24] 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 [12] features 22,476 UAV-captured human-centric videos, we use the RGB data and follow Cross-subject evaluations (CS1).

Penn Action [37] 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 [24] 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.

NTU-RGB+D 120 [14] extends the number of action classes and videos of NTU-RGB+D 60 to 120 classes 114,480 videos. We follow the cross-subject (CS) evaluation protocols.

Toyota Smarthome Untrimmed (TSU) [8] 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 [28] 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.

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.

4   Experimental Analysis and Discussion

We conduct extensive experiments to evaluate the mentioned foundation models [17, 20, 21, 31] on both action classification and segmentation tasks. We study their generalization ability by quantifying the performance improvement obtained by zero-shot learning on real-world action classification (see Sec. 4.1) and action segmentation (see Sec. 4.2) datasets after visual-text pre-training. Subsequently, we evaluate the generalization ability of [17, 20, 21, 31] by providing more analysis.

4.1   Comparisons on Zero-shot Action Classification

One of the advantages of visual-language model [20] compared to classical model [1, 4] is the application for zero-shot classification on unseen datasets. As zero-shot classification can effectively evaluate the alignment of visual and textual features, in this section, we compare the feature quality of various SoTA models [17, 20, 21, 31, 38] which are trained via visual-language alignment for real-world zero-shot action classification tasks. Specifically, given a video embedding, we search its closest textual embedding (extracted using raw action labels) as the action prediction in a close-world setting.

The results reported in Tab. 3 suggest that the original image-based CLIP [20] model struggles with video-based tasks due to a lack of temporal consistency in the features. X-CLIP [17] and ViCLIP [31] extend the CLIP model by incorporating video encoder and trained on video tasks, such as video-text retrieval [17] and video classification [31] with very general video data, bring improvements but

Methods	Smarthome		Penn	UAV	NTU-10	NTU-20
	CS (%)	CV2 (%)	Top-1 (%)	CS (%)	CS (%)	CS (%)
CLIP [20]	10.1	13.6	63.1	1.6	13.8	5.1
X-CLIP [17]	16.5	14.8	72.7	4.8	27.6	18.8
ViCLIP [31]	14.1	14.2	74.3	1.2	18.9	15.5
ViFi-CLIP [21]	**19.6**	**15.3**	87.1	**5.9**	**33.4**	**23.1**
LanguageBind [38]	16.9	15.1	**90.4**	3.7	24.1	15.3

Table 3: Zero-shot transfer results (Top-1 accuracy) without re-training on 2D action classification benchmarks of Smarthome, Penn Action, UAV-Human, NTU-RGB+D 60 (only 10 classes in the test set are used) and NTU-RGB+D 120 (only 20 classes in the test set are used).

remain limited to handle fine-grained tasks (*e.g.*, on Smarthome and UAV-Human). In contrast, ViFi-CLIP benefits from specific fine-tuning on Kinetics [3], which includes many actions found in the evaluation datasets, and can enhance the performance in fine-grained action classification. However, the performance is still far away from satisfactory. The variability in viewpoints, subjects, and environmental conditions can affect the quality of visual features.

We also observe that performances are better with laboratory datasets, and are even better with the Penn-action dataset, as this is a small dataset with very few action labels. So, these results suggest that Vision Language Foundation models are good with basic actions (similar to the web action classes), but struggle with fine-grained actions, as it is difficult to distinguish between two similar actions, just based on their labels. It will be also interesting to perform experiments in open-world settings to verify whether performances are still good on the Penn-action dataset.

To go deeply analyze the models, in Tab. 6, we list the Smarthome classes that benefit the most and the least from the evaluated models. We find that for the actions that have very similar motions (*e.g.*, Uselaptop vs. Readbook, Walk vs. Enter), compositional motions (*e.g.*, Cook.Stir), and large viewpoints variations (*e.g.*, for cross-view evaluation), the SoTA models are still limited. We can deduce from the results that more modalities (*e.g.*, skeleton data that represents human motion) and more pre-training data are needed to further improve action recognition performance.

4.2   Comparisons on Action Segmentation

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 [20, 21, 31] using their visual features with two recent temporal models [6, 7] respectively. The results in Tab. 4 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
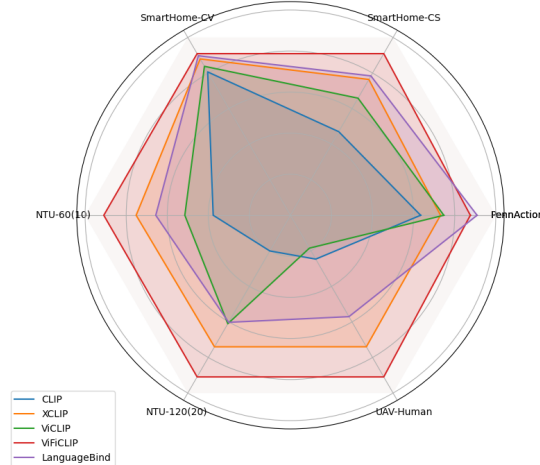
Fig. 1: Statistics of the results on different datasets.

Methods	TSU		Charades
	CS(%)	CV(%)	mAP(%)
PDAN [7] w/ CLIP [20]	16.3	10.0	15.9
PDAN [7] w/ ViCLIP [31]	21.5	13.4	16.2
PDAN [7] w/ ViFi-CLIP [21]	**28.6**	**15.9**	**16.4**
MS-TCT [6] w/ CLIP [20]	5.3	5.7	12.7
MS-TCT [6] w/ ViCLIP [31]	15.8	8.2	16.3
MS-TCT [6] w/ ViFi-CLIP [21]	**21.3**	**17.3**	**16.4**
MS-TCT [6] w/ I3D [4] (SoTA)	**33.7**	-	**25.4**

Table 4: Frame-level mAP on TSU and Charades for comparison of SoTA vision foundation models with SoTA temporal modeling methods for action segmentation.

not at the level of State-of-the-art action detection methods [6]. 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.

4.3   More Study

In this section, we provide further analysis based on the main results.

Can Augmenting Action Labels Improve Zero-shot Results? As raw action labels are too simple to fully express video content and have insufficient semantic information, we manually enrich the expression of the action labels in two levels, the augmented label and action description. Then we re-evaluate the zero-shot action classification on Smarthome and a subset of NTU-RGB+D,
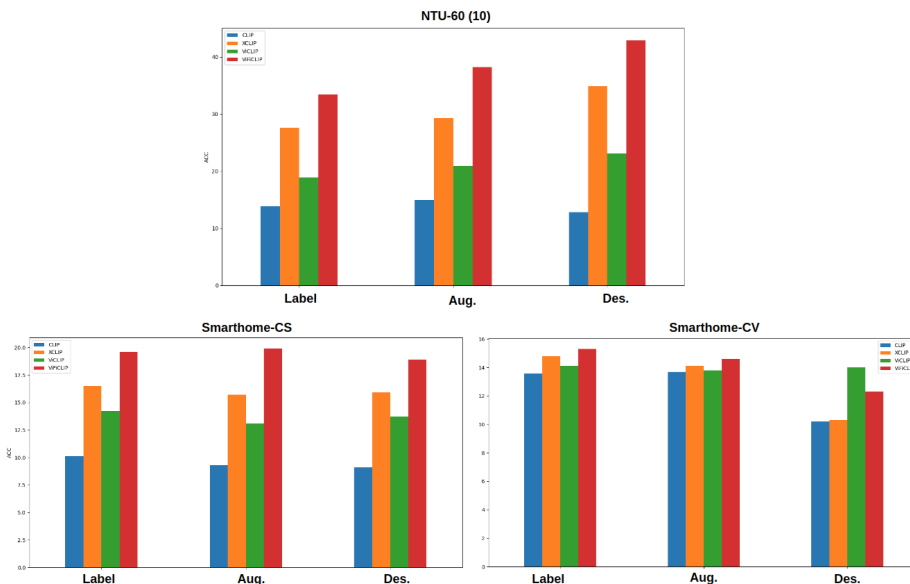
Fig. 2: Comparisons of text information using raw action labels, augmented action labels (Aug.) and full action description (Des.) on NTU and Smarthome.

| Methods | Smarthome | | | | | | NTU-10 | | |
| | CS(%) | | | CV(%) | | | Top-1(%) | | |
	Label	Aug.	Des.	Label	Aug.	Des.	Label	Aug.	Des.
CLIP [20]	10.1	9.3	9.1	13.6	13.7	10.2	13.8	15.0	12.8
X-CLIP [17]	16.5	15.7	15.9	14.8	14.1	10.3	27.6	29.3	34.9
ViCLIP [31]	14.2	13.1	13.7	14.1	13.8	14.0	18.9	20.9	23.1
ViFi-CLIP [21]	19.6	**19.9**	18.9	**15.3**	14.6	12.3	33.4	38.2	**42.9**

Table 5: Ablation study on zero-shot action classification benchmarks of Smarthome and NTU-10 with different text embeddings: original label (Label), augmented label (Aug.), and action description (Des.).

named NTU-10 [14] with 10 selected actions. The results in Tab. 5 and Fig. 2 suggest that the CLIP-based models are sensitive to the text embedding on NTU-10 and action description can improve the text features for zero-shot classification. However, a dataset like Smarthome, where the original labels include most information (*e.g.*, people make coffee on the table), does not benefit from the augmentation of action labels.

Few-shot Learning for Action Segmentation. Few-shot transfer-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 [6] is shown in Tab. 7. The results are

			Smarthome							
Actions			CS(%)					CV(%)		
	CLIP	XCLIP	ViCLIP	ViFiCLIP	L-Bind	CLIP	XCLIP	ViCLIP	ViFiCLIP	L-Bind
Eat.Attable	96.4	91.3	100.0	80.2	99.2	100.0	97.3	100.0	83.7	100
WatchTV	100.0	55.7	98.7	70.0	86.9	-	-	-	-	-
Cleandishes	6.8	68.4	51.2	50.4	48.1	-	-	-	-	-
Uselaptop	0.0	44.4	53.9	47.2	28.7	2.0	50.0	30.8	40.4	40.4
Readbook	0.0	54.2	0.0	47.0	31.1	0.0	0.0	0.0	4.2	0.0
Cook.Stir	0.0	0.0	19.1	27.6	40.7	-	-	-	-	-
Sitdown	0.0	5.9	0.0	16.4	5.3	0.0	0.0	0.0	0.5	2.1
Drink.Fromcup	0.3	0.1	0.7	5.2	0.7	0.0	0.0	0.0	0.9	0.0
Walk	0.08	1.2	58	4.7	0.5	0.0	0.0	0.0	0.7	0.0
Enter	0.0	0.0	0.0	0.0	6.8	0.0	0.0	0.0	0.0	0.0

Table 6: Analysis on different actions of Smarthome using SoTA foundation models.

Methods	**Label**	**TSU**		**Charades**
		CS(%)	CV(%)	mAP(%)
CLIP [20]	5%	**6.2**	4.3	8.7
ViCLIP [31]	5%	3.5	3.3	10.1
ViFi-CLIP [21]	5%	5.6	**5.7**	**11.1**
CLIP [20]	10%	4.4	4.7	11.1
ViCLIP [31]	10%	4.0	3.5	**11.6**
ViFi-CLIP [21]	10%	**6.1**	**5.8**	11.3

Table 7: Transfer learning results by on Toyota Smarthome Untrimmed (TSU) and Charades with randomly selected **5% (top)** and **10% (bottom)** of labeled training data after pre-training on different foundation models.

consistent with previous evaluation, ViFi-CLIP [21] has mostly the best visual representation ability.

Can Visual-language Model Used for Zero-shot Action Segmentation? For zero-shot frame-wise action segmentation, one of the solutions is to apply the zero-shot action classification on each frame, which is much more complex. In this section, we propose to leverage current VQA methods [13, 23] to directly generate predictions of the action boundary for a given video by asking the specific questions about the actions. We compare TimeChat [23] and UniVTG [13] on Charades with event-level IoU accuracy (see Tab. 9) and we find that UniVTG [13] is more effective. To further compare the performance of VQA model to the temporal modeling methods using CLIP-based features like PDAN [7] w/ ViFi-CLIP [21], we convert the action boundary to frame-level prediction and use mAP for fair comparison. The results in Tab. 8 demonstrate the UniVTG, even without the need for re-training, can achieve better accuracy than ViFi-CLIP for Charades. However, for more complex scenarios like TSU, where multiple actions can be performed in the same video and they can be overlapped, the UniVTG model is still challenging to deal with, the two stage approaches using ViFi-CLIP features.

Methods	TSU		Charades
	CS(%)	CV(%)	mAP(%)
TimeChat [23]	2.5	3.4	14.7
UniVTG [13]	2.4	3.2	**17.7**
PDAN w/ ViFi-CLIP [21]	**28.6**	**15.9**	16.4

Table 8: Frame-level mAP on TSU and Charades for comparison of VQA methods with zero-shot action segmentation.

Methods	Charades			
	R@0.3	R@0.5	R@0.7	mIoU
TimeChat [23]	42.4	23.1	9.4	25.0
UniVTG [13]	**55.8**	**29.2**	**10.6**	**31.8**

Table 9: R1@IOU on Charades for comparison of VQA methods with zero-shot action segmentation.

4.4   Discussions and Novel Direction

From our study, we find that current SoTA visual-language foundation model still has challenge for action recognition, the semantic gap between visual features and action descriptions makes it difficult to capture fine-grained details. To solve the problem, we suggest to use more modalities (*e.g.*, audio [22] and geometry [36]) to complement visual information, and to design more effective temporal modeling to capture long-term temporal reasoning to improve action segmentation. Additionally, we can also take the advantage of Large Language Models to enhance the understanding of action descriptions and to improve zero-shot classification. Finally, we believe that more comprehensive datasets from real-world and video generative model [16] can cover a broader range of actions.

5   Conclusion

In this study, we evaluate state-of-the-art visual-language models for fine-grained action recognition, focusing on zero-shot action classification and action segmentation. While models like ViFi-CLIP, fine-tuned on the Kinetics dataset, demonstrated the best performance, and the VQA model UniVTG, show attracted results for zero-shot action segmentation. Our results highlight current challenges in handling complex actions and long-term temporal consistency. The findings suggest that incorporating additional modalities, such as skeleton data, could enhance model accuracy and robustness. Future research should explore integrating multi-modal data and fine-tuning strategies to improve action recognition performance.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. ICCV (2021)
2. Cadene, R., Ben-younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: CVPR (June 2019)
3. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. CoRR (2019)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
5. Chunhui, L., Yueyu, H., Yanghao, L., Sijie, S., Jiaying, L.: Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv:1703.07475 (2017)
6. Dai, R., Das, S., Kahatapitiya, K., Ryoo, M., Bremond, F.: MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In: CVPR (2022)
7. Dai, R., Das, S., Minciullo, L., Garattoni, L., Francesca, G., Bremond, F.: Pdan: Pyramid dilated attention network for action detection. In: WACV (2021)
8. Dai, R., Das, S., Sharma, S., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. IEEE TPAMI (2022)
9. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: ICCV (2019)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)
12. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: CVPR (2021)
13. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: ICCV (2023)
14. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3D human activity understanding. IEEE TPAMI (2020)
15. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
16. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
17. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP:: End-to-end multi-grained contrastive learning for video-text retrieval. In: ACMMM (2022)
18. Mezghani, L., Bojanowski, P., Alahari, K., Sukhbaatar, S.: Think before you act: Unified policy for interleaving language reasoning with actions. arXiv:2304.11063 (2023)
19. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. In: arXiv:1807.03748 (2018)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)

21. Rasheed, H., khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Finetuned clip models are efficient video learners. In: CVPR (2023)
22. Recasens, A., Luc, P., Alayrac, J.B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., Grill, J.B., van den Oord, A., Zisserman, A.: Broaden your views for self-supervised video learning. In: ICCV (2021)
23. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR (2024)
24. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3D human activity analysis. CVPR (2016)
25. Shukor, M., Dancette, C., Cord, M.: ep-alm: Efficient perceptual augmentation of language models. In: ICCV (2023)
26. Shukor, M., Dancette, C., Rame, A., Cord, M.: Unified model for image, video, audio and language tasks. In: ICCVW (2023)
27. Shukor, M., Dancette, C., Rame, A., Cord, M.: UnIVAL: Unified model for image, video, audio and language tasks. Transactions on Machine Learning Research (2023)
28. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
29. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS (2022)
30. Wang, A.J., Ge, Y., Yan, R., Yuying, G., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., Shou, M.Z.: All in one: Exploring unified video-language pre-training. In: CVPR (2023)
31. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., Qiao, Y.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: ICLR (2024)
32. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: Internvideo: General video foundation models via generative and discriminative learning. arXiv:2212.03191 (2022)
33. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR (2019)
34. Xu, H., Ghosh, G., Huang, P.Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., Zettlemoyer, L.: Vlm: Task-agnostic video-language model pre-training for video understanding. arXiv preprint arXiv:2105.09996 (2021)
35. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: CVPR (2023)
36. Yang, D., Wang, Y., Dantcheva, A., Kong, Q., Garattoni, L., Francesca, G., Bremond, F.: Lac - latent action composition for skeleton-based action segmentation. In: ICCV (2023)
37. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: ICCV (2013)
38. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, C.W., Li, Z., Liu, W., Yuan, L.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: ICLR (2024)