Quo Vadis, Video Understanding with Vision-Language Foundation Models?

Mahmoud ALI[1] Di Yang[1] Arkaprava Sinha[2] Dominick Reilly[2] Srijan Das[2]

Gianpiero Francesca[3] Francois Bremond[1]

[1]Inria Center at Université Côte d'Azur [2]UNC Charlotte [3]Toyota Motor Europe

{mahmoud.ali, di.yang, francois.bremond}@inria.fr

{asinha13,dreilly1,sdas24}@charlotte.edu

gianpiero.francesca@toyota-europe.com

Abstract

Vision-Language foundation models, including vision-language models (VLMs) and vision-large language models (VLLMs), have been evolving rapidly and have shown good performance on different downstream video understanding tasks, especially on web datasets. However, it is still an open question how much these VLMs and VLLMs perform in more challenging scenarios like Activities of Daily Living (ADL). To answer this, we provide a comprehensive study of VLMs and VLLMs by comparing their zero-shot transfer ability to five downstream tasks including action classification, video retrieval, video description, action forecasting, and frame-wise action segmentation. Extensive experiments are conducted on eleven real-world, human-centric video understanding datasets (*e.g.*, Toyota Smarthome, Penn Action, UAV-Human, EgoExo4D, TSU, Charades) to study these tasks with our insights into the strengths and limitations of these models in zero-shot settings. Moreover, we provide in-deep analysis to find the best setting to improve the model performance in zero-shot action classification tasks. Based on our experiments, we find that these models are still far away from satisfactory performance in all evaluated tasks, particularly in densely labeled and long video datasets.

1 Introduction

Recent vision-language foundation models [22, 18, 36, 26, 27, 38, 1, 21, 28, 32, 8] have witnessed significant attention due to their strong generalization abilities. These models are able to transfer to various downstream tasks without the need for additional fine-tuning thanks to their pre-training on large-scale, multi-modal datasets. By aligning visual and textual features through joint training of a visual encoder and a textual encoder, they have revolutionized numerous tasks in both the vision and language domains. Vision-language foundation models can be broadly categorized into two families. The first category encompasses vision-language models (VLMs), which are pre-trained to align visual and textual features and have shown strong performance in zero-shot transfer tasks such as action classification and video retrieval. The second category consists of vision-large language models (VLLMs). Unlike VLMs, VLLMs are pre-trained to handle more complex language-based tasks by leveraging large-scale language models (LLMs). These models enable zero-shot transfer to video description, action forecasting tasks without the need for task-specific retraining. Due to their strong generalization capabilities, both VLMs and VLLMs are critical for various applications, including healthcare monitoring and human-machine interaction.

While these models have been successful across a wide range of tasks in the image domain, their adaptation to the video domain has primarily been evaluated on general video understanding tasks,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

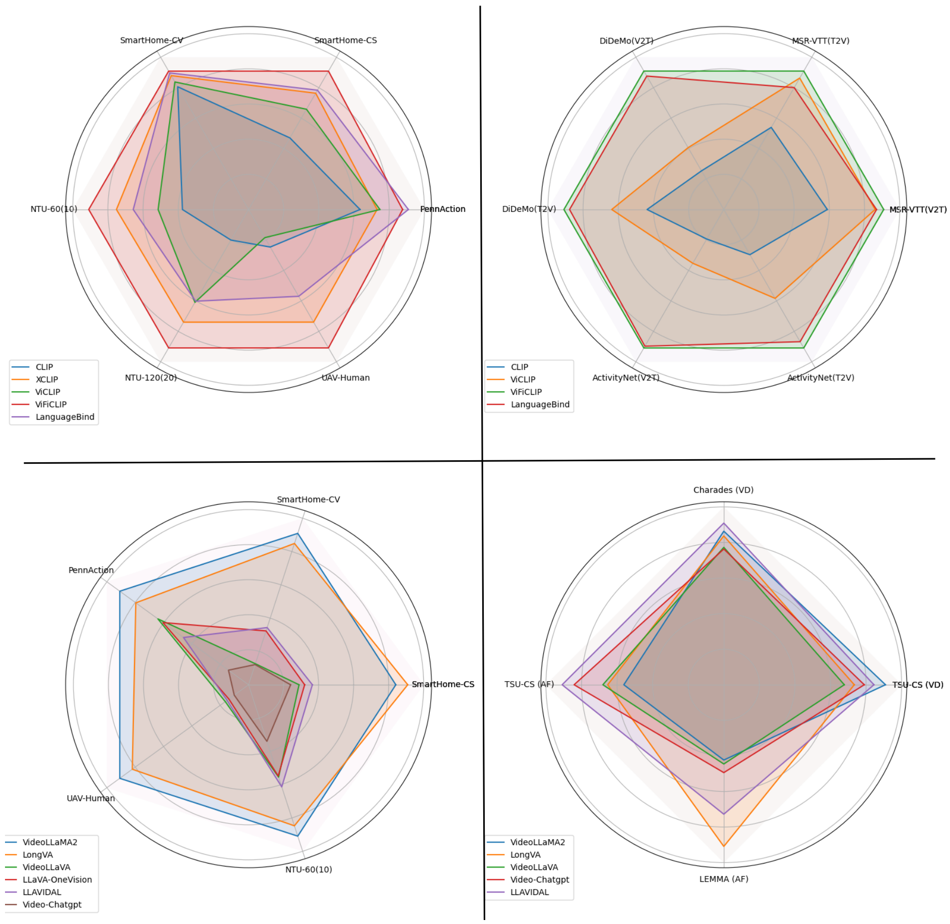

Figure 1: Comparisons and statistics of the SoTA models on different datasets for zero-shot video understanding tasks, *i.e.*, action classification with VLMs (top left), video retrieval with VLMs (top right), action classification with VLLMs (bottom left), video description and action forecasting with VLLMs (bottom right).

such as captioning and classification. The zero-shot transfer capability of these VLMs to handle more complex and fine-grained action understanding tasks remains under-explored. In real-world applications, compositional activities can be performed at the same time and viewpoints, subject appearances, objects may change largely with time, hence, it is crucial to understand the generalization ability to current challenges of vision-language foundation models targeting video understanding in more complex scenarios [6, 29, 14, 7] such as action segmentation in untrimmed videos with multi-label annotations.

To address this gap, this paper provides a comprehensive comparison and ablation study of state-of-the-art (SoTA) vision-language foundation models focusing on their performance in fine-grained zero-shot tasks. Specifically, we evaluate vision-language models [4, 13, 15, 3, 41, 20] on action classification and video retrieval tasks using the action labels. To further understand the performance of vision-large language models, we evaluate [4, 13, 15, 3, 41, 20]on fine-grained LLM-based tasks including video description, action forecasting, and frame-wise action segmentation. The experimental comparisons on their performance for different tasks are summarized in Fig. 1.

In summary, the key contributions of this paper are as follows: (i) We conduct a comprehensive large-scale evaluation of the generalization capabilities of VLMs and VLLMs, with a focus on their transfer ability to challenging downstream video understanding tasks. (ii) Through detailed experiments, we identify optimal video sampling strategies for these models and analyze how different forms of action label descriptions influence zero-shot action classification. (iii) We investigate the impacts of fine-grained labels and diverse viewpoints on vision-language alignment. We provide insights and comparisons of various frame-wise action prediction techniques, using video question answering (VQA) models to investigate zero-shot action segmentation. (iv) Extensive experiments are conducted across eleven benchmark datasets covering five core zero-shot tasks *i.e.*, action classification, action segmentation, action forecasting, video retrieval, video description. Our experimental analysis reveals current model limitations and suggests future research directions.

2 SoTA Vision-Language Foundation Models for Video Understanding

Recently, many methods have used language features [22] for video understanding [18, 36, 26, 27, 21, 32], video captioning [38] and visual question answering [1, 28]. These models, like InternVideo [35], aim to understand and generate descriptions of video content, facilitating a multi-modal understanding of visual data. In this section we study the related work on SoTA image-language models, video-language models and Video-Large Language Models.

Image-Language Models (ILMs) Image-language models like CLIP [22], SigLIP [40], and EVA-CLIP [31] are multi-modal models that align visual and textual data to create common space representations, enabling tasks such as image classification, captioning, and retrieval. CLIP, trained using large scale image-text pairs using a dual-encoder architecture and contrastive learning to align images with text. SigLIP [40] enhances CLIP by replace the loss function with a simple pairwise sigmoid loss which is more memory efficient and enabling training using large batch sizes without requiring additional resources, while EVA-CLIP [31] taken the benefit of flash attention to reduce the training cost as well incorporates novel training strategies, such as enhanced data augmentation and more efficient optimization techniques, to improve learning efficiency and representation quality. Despite their strengths, these models still limited to understand the temporal consistency, struggles with video understanding tasks.

Video-Language Models (VLMs) Video-language models such as XCLIP [19], ViCLIP [34], ViFiCLIP [23], and LanguageBind [43] are designed to align and understand video content with textual information, leveraging both visual and language encoders to capture the complex interplay between temporal sequences and language semantics. These models extend image-language models by incorporating temporal dynamics to understand video content, aligning sequences of frames with language to capture spatial and temporal information. XCLIP [19] enhances CLIP by proposes the Attention Over Similarity Matrix (AOSM) module to make the model focus on the contrast between essential frames. ViCLIP [34] extend CLIP image encoder to video encoder by adding spatiotemporal attention modules and masking video during the the pretraining. ViFiCLIP [23] fine-tune image and text encoder of CLIP model for video domain by simple frame-level late feature aggregation via temporal pooling. LanguageBind [43] extend video-language models by improved the text descriptions using incorporating metadata, spatial, and temporal information then alignment language with diffrent modalities like videos, infrared images, depth maps, and audio using contrastive learning. These models still struggle to recognize a fine-grained action recognition task especially in long videos as well their limitation to generalize on different scenarios like ADL.

Video-Large Language Models (VLLMs) Video-large language models (VLLMs) like VideoLLaMA [4], VideoLLaVA [15], VideoChatGPT [20], LongVA [41], and LLaVIDAL [3] integrate large language models (LLMs) with video understanding and fine-tuned on instructional language-vision data to process and generate text based on video content, enhancing tasks like video question answering, summarizing, and interactive dialogue. VideoLLaMA [4] uses attention mechanisms for temporal video comprehension, while VideoLLaVA [15] improves video-language alignment by unify visual representation of image and videos into the language feature space. VideoChatGPT [20] combines video encoders with ChatGPT for real-time video dialogue, LongVA [41] handles long-form video content with hierarchical modeling, and LLaVIDAL [3] uses additional cues like 3D pose and objects jointly with visual-text embeddings and all these modalities project to LLM to enhance understanding ability of human actions. The transfer-ability limitations of these models appears in densely labeled dataset as well the fine-grained temporal discrimination tasks like temporal localization. Recently models like UniVTG [16], TimeChat [25], VtimeLLM [12] handled the tasks that demand precise timing and action recognition, such as frame-wise action segmentation, video question answering, action forecasting by integrating visual and textual data align with time information within a large language model framework. UniVTG [16] proposes to Unify the diverse

Methods	Frames	Smarthome		Penn	UAV	NTU-10
		CS (%)	CV2 (%)	Top-1 (%)	CS (%)	CS (%)
CLIP [22]	16	10.1	13.6	63.1	1.6	13.8
X-CLIP [19]	32	16.5	14.8	72.7	4.8	27.6
ViCLIP [34]	8	14.1	14.2	74.3	1.2	18.9
ViFi-CLIP [23]	32	19.6	15.3	87.1	5.9	**33.4**
LanguageBind [43]	8	16.9	15.1	**90.4**	3.7	24.1
Video-LLaMA2 [4]	16	21.0	**18.0**	85.9	**7.2**	24.9
LongVA [41]	256	**22.7**	16.8	75.2	6.5	23.2
Video-LLaVA [15]	8	7.2	2.5	60.5	1.3	15.2
LLaVA-OneVision [13]	8	8.0	6.4	57.1	1.1	15.0
LAVIDAL [3]	100	9.1	6.8	43.4	1.2	16.8
Video-Chatgpt [20]	16	6.0	2.4	13.4	0.8	9.3

Table 1: **Zero-shot** transfer results and comparisons without re-training on action classification benchmarks of Smarthome (Top-1 accuracy) and Penn Action.

Actions	Smarthome											
	Video-LLaMA2		*LongVA*		*Video-LLaVA*		*LLaVA-OneVision*		*LAVIDAL*		*Video-Chatgpt*	
	CS (%)	CV2 (%)	CS (%)	CV2 (%)	CS (%)	CV2 (%)	CS (%)	CV2 (%)	CS (%)	CV2 (%)	CS (%)	CV2 (%)
Eat.Attable	64.8	43.2	94.1	96.9	3.2	0	0	0	30.1	14.0	19.1	5.5
WatchTV	86.5	-	88.3	-	13.1	-	0	-	25.6	-	13.5	-
Cleandishes	36.1	-	32.3	-	1.5	-	0 7	-	0	-	0	-
Uselaptop	67.4	94.2	56.2	57.7	5.1	3.8	0	0	0	0	0	0
Readbook	0	0	38.1	14.4	9.9	0	12.1	0	0	0	10.8	0
Cook.Stir	60.3	-	41.7	-	20.6	-	0	-	4.5	-	0	-
Sitdown	63.1	80.3	16.2	0	21.2	18.7	86.6	62.2	16.2	5.7	11.2	1.6
Drink.Fromcup	0.4	0	11.5	1.6	4.6	2.2	0		2.1	2.8	0.3	0.6
Walk	10.2	3.4	6.1	1.8	9.2	0.2	0.3	0	2.3	0.5	0	0
Enter	0.8	0	32.3	0	0	0	0	0	0	0	0	0

Table 2: Analysis on different actions of Smarthome using VLLMs.

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. TimeChat [25] is a time-sensitive multi-modal large language model specifically designed for long video understanding. It utilizes a sliding video Q-Former, which dynamically adjusts to different video token lengths, optimizing the extraction and compression of video features for improved long video processing. VTimeLLM [12] is Video Large Language Model with boundary-awareness, specifically designed to improve temporal reasoning and video comprehension. Its three-stage training approach starts with aligning features using large-scale image-text data, then incorporates multi-event video-text data paired with temporal question answering to develop time boundary awareness, and finally uses instruction tuning on high-quality dialogue datasets to enhance its temporal reasoning abilities. Despite their promising results on certain datasets, these models still far from satisfactory performance and are limited to short videos and simpler datasets. Such datasets often lack fine-grained actions and are not densely labeled, which restricts the model's ability to handle more complex and detailed video content.

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 relies 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 video understanding tasks, so we provide an analysis on this topic to find out more future directions based on the analysis. In this study, we select the most recent and representative Vision-Language foundation models [22, 19, 34, 23, 43, 4, 41, 3, 13, 15**?** , 16, 12] (see Tab.2 in the supplementary).

3 Experimental Analysis and Discussion

We conduct extensive experiments to evaluate the performance of various vision-language foundation models, including both VLMs [22, 19, 34, 23, 43] and VLLMs [4, 3, 15, 20, 13, 41], across different tasks. Specifically, we examine their generalization abilities by measuring the improvements in zero-shot learning within real-world scenarios. We evaluate VLMs on tasks such as action classification (see Sec. 3.1.1) and video retrieval (see Sec. 3.2), while VLLMs are assessed on more tasks including action classification (see Sec. 3.1.2), video description (see Sec. 3.3), action forecasting (see Sec. 3.4), and action segmentation (see Sec. 3.5). Furthermore, we provide additional analysis to assess the

Actions	Smarthome									
	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 3: Analysis on different actions of Smarthome using VLMs.

generalization capabilities of both VLMs and VLLMs in these tasks. See the supplementary for more experiments, analysis, and datasets comparisons.

3.1 Zero-shot Action Classification

Video-language models (VLMs) [22, 19, 34, 23, 43] and video-large language models (VLLMs) [4, 3, 15, 20, 13, 41] adopt different approaches to zero-shot action classification evaluation. In this section, we assess the performance of each category of models across five public datasets under various scenarios to understand their strengths and limitations in handling diverse and complex action recognition tasks.

3.1.1 Vision-Language Models (VLMs)

For Vision-Language Models (VLMs) [22, 19, 34, 23, 43], we begin by extracting text embedding for all action labels. For each query video, we sample select frames and extract their corresponding visual features using a pre-trained vision encoder. Finally, we compute cosine similarity between the video's visual features and the text embedding to retrieve the action label with the highest similarity score. In the first row of Tab.1, we present the performance results of these models. The original image-based CLIP [22] model struggles with video tasks due to its lack of temporal consistency. In contrast, models like X-CLIP [19],ViCLIP [34], and languagebind [43] demonstrate improvements by extending CLIP with a video encoder, specifically trained on tasks such as video-text retrieval and video classification. However, despite these enhancements, they remain limited in handling fine-grained tasks (e.g., Smarthome and UAV-Human datasets). Fine-tuning ViFiCLIP on the Kinetics dataset [2] improves performance in fine-grained action classification, as the dataset contains actions similar to those in the evaluation sets, aiding generalization. Still, the performance is not fully satisfactory, as these models are typically trained and fine-tuned on web-scale data, which differs significantly from real-world scenarios and activities of daily living.

3.1.2 Vision-Large Language Models (VLLMs)

Zero-shot action classification in Video-Large Language Models (VLLMs),typically uses a video question answering (VQA) approach. A natural language query is constructed asking the model to identify the action depicted in the video. For instance, the query could be, "Which action from this list [actions label] matches the video content?". By mapping both the video features and the query into a shared semantic space, the model can predict the most relevant action. However, some models may require post-processing to refine the predictions because the initial response might not provide a precise action label. This post-processing helps to enhance the alignment between actions and video content and improve the accuracy of zero-shot classification.

We observe from our experiments that the performances of VLMs and VLLMs are better with laboratory datasets, and are even better with the Penn-action dataset [42] comparing to challenging dataset like smarthome dataset [7] which contains fine-grained actions. As shown in Fig. 2 the action features in the Smarthome dataset exhibit significant overlap, making it difficult to distinguish between them. In contrast, the action features in the PennAction dataset are more clearly separated, indicating better-defined and distinct feature representations for each action. This demonstrates that Smarthome's fine-grained actions pose more challenges for the model in terms of feature

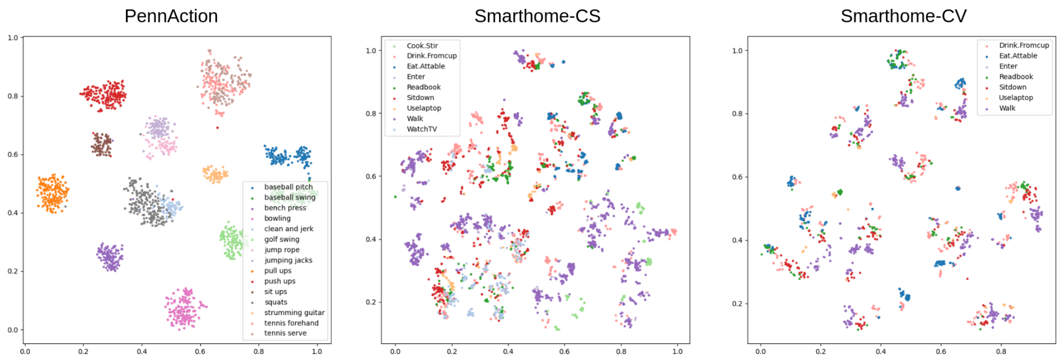

Figure 2: TSNE visualization of ViFiCLIP on PennAction and 10 classes in Smarthome

Methods	MSR-VTT		DiDeMo		ActivityNet	
	T2V	V2T	T2V	V2T	T2V	V2T
CLIP [22]	29.0	25.8	11.5	19.1	8.3	12.2
ViCLIP [34]	42.4	41.3	18.4	27.9	15.1	24.0
ViFi-CLIP [23]	**44.8**	**43.5**	**41.2**	**39.8**	**38.9**	**37.4**
LanguageBind [43]	42.8	38.3	39.7	38.4	38.4	35.7

Table 4: Zero-shot Video retrieval using VLMs.

Method	TSU CS(%)	LEMMA
Video-LLaMA2 [4]	16.7	30.6
LongVA [41]	19.4	**65.7**
Video-LLaVA [15]	20.2	32.2
VideoChatGPT [20]	25.0	35.7
LLAVIDAL [3]	**27.0**	52.6

Table 5: Action Forecasting performance using VLLMs.

discrimination, while PennAction allows for more effective differentiation between action categories, as this is a small dataset with very few action labels. See Fig.6,7and 8 in the supplementary for more VLMs analysis.These results indicate that Vision-Language Foundation Models perform well on basic actions, particularly those similar to common web action classes. However, they struggle when it comes to fine-grained actions, where distinguishing between similar actions based solely on their labels is challenging. This suggests that while these models are adept at recognizing broad or generic actions, they have limitations in more nuanced tasks. Further experimentation in open-world settings, particularly with datasets like PennAction, would be valuable to assess whether these models maintain their performance in less controlled environments.

We deeply analyze the results of VLMs in Tab. 3 and VLLMs in Tab. 2 we list the Smarthome classes that benefit the most and the least from the evaluated models. See the full analysis for all actions in Fig.1 and Fig.2 in the supplementary for Cross-subject and Cross-view evaluation. 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.

3.2 Zero-shot Video-Text Retrieval

The primary goal of video retrieval is to identify and rank videos from a large dataset that best match an input query, such as a text description, effectively bridging the semantic gap between video content and language. In our experiments Tab. 4, we evaluate Vision-Language Models (VLMs) on both text-to-video (T2V) and video-to-text (V2T) retrieval tasks, reporting (R@1) on three public datasets: MSR-VTT [37], DiDeMo [11], and ActivityNet [10]. First, we extract visual and textual features using pre-trained encoders of the VLMs, which have been aligned within a shared embedding space through contrastive learning during training. Next, we compute a similarity score, typically using cosine similarity, between the query and candidate videos for T2V or V2T tasks. Finally, based on these similarity scores, the videos and texts are ranked, and the most relevant ones are retrieved, with R@1 indicating the performance of the models in retrieving the top relevant result. The results show that the features extracted by ViFi-CLIP [23] with Kinetics fine-tuning, have the best generalization ability on such task for all the datasets.

Method	TSU CS(%)						Charades					
	CI	*DO*	*CU*	*TU*	*Con*	**Average**	*CI*	*DO*	*CU*	*TU*	*Con*	**Average**
Video-LLaMA2 [4]	45.8	**52.0**	**59.6**	42.8	**58.8**	**51.8**	44.4	50.8	**50.4**	40.6	41.6	45.6
LongVA [41]	32.2	41.2	49.6	33.6	52.2	41.8	37.6	47.0	49.2	35.4	**51.8**	44.2
Video-LLaVA [15]	37.8	33.8	40.2	40.4	39.6	38.6	38.2	44.4	44.0	37.4	40.2	40.8
Video-Chatgpt [20]	43.0	45.8	41.4	43.0	50.0	45.0	35.8	44.2	41.6	42.2	37.8	40.3
LLAVIDAL [3]	**46.0**	48.6	42.2	**45.8**	58.0	48.1	**51.8**	**54.2**	44.0	**49.2**	41.8	**48.0**

Table 6: Video Description performance using VLLMs. [CI: Correctness, DO: Detail Orientation, CU: Contextual Understanding, TU: Temporal Understanding, Con: Consistency]

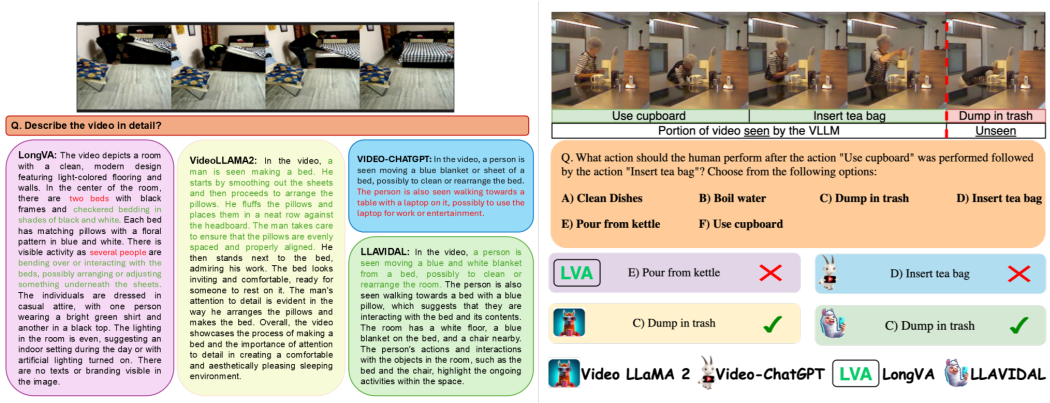

Figure 3: Comparison of VLLMs on Video Description [left] using sample video from Charades and Action Forecasting Task [Right] using sample video from TSU.

3.3 Video Description

Tab. 6 presents a comparison of Video-Large Language Models (VLLMs) in video description generation, evaluated using the five metrics introduced in [20]. On the TSU dataset, VideoLLaMA-2 outperforms its counterparts, showcasing its strength in capturing spatio-temporal features. Conversely, LLAVIDAL excels on the Charades dataset, largely due to its enhanced ability to comprehend human-object interactions, which are central to this dataset. No single VLLM consistently outperforms across all metrics across datasets, underscoring the necessity of selecting models tailored to the specific characteristics of the task and data at hand.

3.4 Action Forecasting

In Tab. 5 we report the results of action forecasting on the TSU and LEMMA datasets. We find that VideoChatGPT [20] and LLAVIDAL [3] excel on TSU due to their respective enhanced temporal instruction tuning and daily activity instruction tuning. In contrast, on the LEMMA dataset LongVA performs best, likely due to its ability to analyze frames at a larger scale using a grid-based approach. This is beneficial for LEMMA where actions consist of object-verb pairs and the size of objects appear small in the videos. These results highlight the necessity to tailor the architecture and instruction tuning data of VLMs to the task at hand.

3.5 Zero-shot Action Segmentation

For zero-shot frame-wise action segmentation, one solution is to apply zero-shot action classification to each individual frame, but this method adds significant complexity to the process and miss the temporal information. Instead, we utilizing recent Video Question Answering (VQA) approaches such as TimeChat [25], UniVTG [16],and VTimeLLM [12], which can directly from a query question predict action boundaries by answering targeted questions about the actions occurring in the video. We conducted a comparison between these models on the Charades dataset, using event-level Intersection over Union (IoU) accuracy (see Tab.8), and found that UniVTG[16] outperforms TimeChat and VTimeLLM in terms of accuracy. To provide a fair comparison with temporal models that use CLIP-based features, such as PDAN [5] with ViFi-CLIP [23], we convert the action boundaries

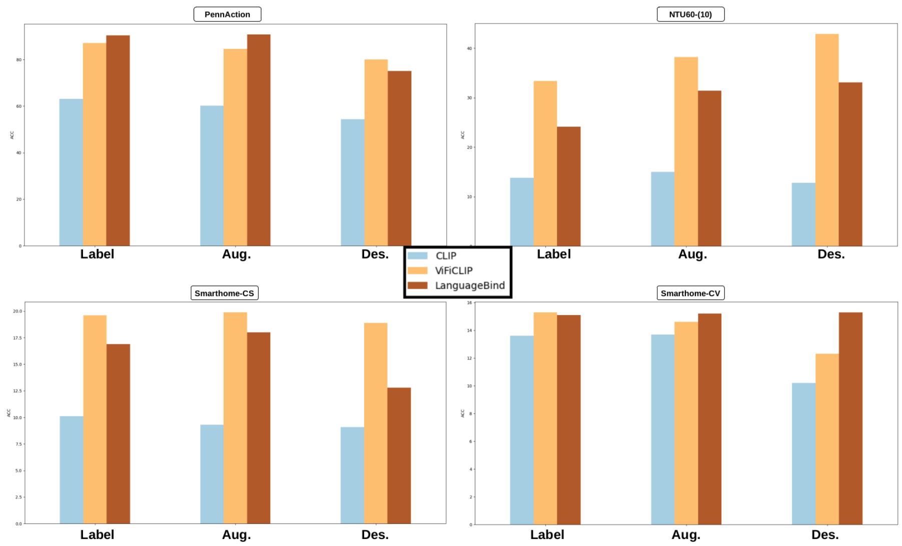

Figure 4: Comparisons of text information using raw action labels, augmented action labels (Aug.) and full action description (Des.) on PennAction, NTU10 and Smarthome.

Methods	TSU		Charades
	CS(%)	CV(%)	mAP(%)
TimeChat [25]	2.5	3.4	14.7
UniVTG [16]	2.4	3.2	**17.7**
VTimeLLM [12]	1.9	2.8	12.3
PDAN w/ ViFi-CLIP [23]	**28.6**	**15.9**	16.4

Table 7: 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 [25]	42.4	23.1	9.4	28.0
VTimeLLM [12]	40.2	20.3	7.8	22.8
UniVTG [16]	**55.8**	**29.2**	**10.6**	**31.8**

Table 8: R1@IOU on Charades for comparison of VQA methods with zero-shot action segmentation.

predicted by the VQA models into frame-level predictions and evaluate using mean Average Precision (mAP). Results in Tab. 7 demonstrate that UniVTG delivers better performance than ViFi-CLIP on the Charades dataset, even without additional re-training. However, when dealing with more complex scenarios like the TSU dataset, where multiple actions often overlap within the same video, all VLLMs struggles to manage these challenges effectively. In such cases, two-stage approaches utilizing VLMs features remain more robust, underscoring the limitations of VQA-based models in highly dynamic environments.

3.6 Further Study

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

Can Augmenting Action Labels Improve Zero-shot Results? Raw action labels often lack the depth necessary to fully capture video content, which affects vision-language alignment. To address this, we enhance the labels in two ways: by creating augmented labels and by providing action descriptions, as shown in Fig.3 in the supplementary. We then re-evaluated zero-shot action classification using the Smarthome dataset , PennAction dataset, and a subset of NTU-RGB+D60, referred to as NTU-10[17], which consists of the 10 evaluated actions. The results, detailed in Tab.9 and Fig.4, show that VLMs respond well to text embedding on NTU-10 and PennAction, with action descriptions improving text features for zero-shot classification. In contrast, for datasets like Smarthome, where the original labels are already detailed (e.g., "person eat at the table"), augmenting the labels does not significantly enhance performance.

Are sampling frames strategies effective on zero-shot Results? Due to the high computation to process the videos, all the methods resort to sampling only a few frames from each video to reduce complexity while maintaining relevant visual information. In this section, we evaluate the impact of different frame sampling methods on the visual encoder performance of Vision-Language

Methods	Smarthome						NTU-10			PennAction		
	CS(%)			CV(%)			Top-1(%)			Top-1(%)		
	Label	*Aug.*	*Des.*	*Label*	*Aug.*	*Des.*	*Label*	*Aug.*	*Des.*	*Label*	*Aug.*	*Des.*
CLIP [22]	10.1	9.3	9.1	13.6	13.7	10.2	13.8	15.0	12.8	63.1	60.2	54.4
ViFi-CLIP [23]	19.6	**19.9**	18.9	**15.3**	14.6	12.3	33.4	38.2	**42.9**	87.1	84.5	80.0
LanguageBind [43]	16.9	18.0	12.8	15.1	15.2	15.3	24.1	31.4	33.1	90.4	**90.8**	75.1

Table 9: Study of zero-shot action classification on Smarthome and NTU-10 with different text embeddings: original label (Label), augmented label (Aug.), and action description (Des.).

Sampling method	Smarthome				PennAction	
	CS(%)		CV(%)			
	ViFi-CLIP	*LanguageBind*	*ViFi-CLIP*	*LanguageBind*	*ViFi-CLIP*	*LanguageBind*
Random	18.2	16.5	14.5	15.0	86.6	90.2
Uniform	18.9	16.6	14.7	14.9	**87.2**	90.4
TSN [33]	**19.6**	**17.1**	**15.3**	**15.1**	**87.2**	**91.0**

Table 10: Ablation study on different methods for sampling frames from videos

Models (VLMs). As shown in Tab. 10, we compare the results of three sampling methods: random, uniform, and TSN (Temporal Segment Network) [33]. Our findings indicate that TSN is the most effective approach, as it segments the video into a predefined number of equal segments and then randomly selects a frame from each segment. This method ensures better coverage of the video content, enhancing the model's ability to capture temporal dynamics compared to random and uniform sampling methods.

Is Vision-Language Alignment Impacted by Different Viewpoints and Fine-Grained Labels? In Tab. 11, we report the results of zero-shot action classification on the EgoExo4D dataset [9], which provides fine-grained action labels across N different viewpoints (ego, exo1, exo2, exo3, exo4,..etc) for each video,see Fig 5 in the supplementary.Vision-Language Models (VLMs) [23, 43] struggle with this type of fine-grained action classification due to the difficulty in distinguishing between similar action labels, as shown in the first row Tab. 11. To address this, we grouped similar fine-grained actions into coarse-grained categories using GPT-3.5, as detailed in Fig 4 in the supplementary. This adjustment led to improved alignment between vision-language representations, particularly in the ego view, which proved more informative and aligned with the labels, as shown in the second row Tab. 11. While VLMs still face challenges with fine-grained actions, they show promise as an initial stage for zero-shot action classification when using coarse-grained action labels.

Methods	EgoExo4D							
	Action type	Ego (%)	Exo1 (%)	Exo2 (%)	Exo3 (%)	Exo4 (%)	Exos (%)	Ego+Exos (%)
ViFi-CLIP [23]	Fine-grained	**4.0**	1.8	2.1	2.1	2.0	2.3	2.7
LanguageBind [43]	Fine-grained	**3.2**	2.4	2.2	1.9	1.8	2.0	2.6
ViFi-CLIP [23]	coarse-grained	**30.2**	17.6	19.0	15.7	15.4	21.7	26.3
LanguageBind [43]	coarse-grained	**27.5**	13.2	12.3	11.4	9.0	12.7	14.4

Table 11: Zero-shot action classification on EgoExo4D using Fine-grained and Coarse-grained label from different viewpoints.

4 Conclusions and Novel Direction

In this study, we evaluat SoTA VLMs and VLLMs on their performance and generalization capabilities in fine-grained video understanding tasks. From our study, we highlight several key insights: (i) among the VLMs, ViFi-Clip [23] demonstrates superior performance in most of action classification and video retrieval tasks with strong transfer ability. (ii) For VLLMs, the LLAVIDal [3], LonVA [41] and Video-LLaMA2 [4] respectively achieve the highest accuracy on particular datasets for action forecasting and video description. They show the power of LLMs in more complex video-based tasks. (iii) Several limitations of the foundation models still remain, such as long-term temporal modeling, multi-modal learning, and the ability to handle fine-grained activities in complex, real-world scenarios, *e.g.*, Smarthome, UAV-Human, TSU and Charades, require further improvement. Based on our findings, we suggest that future research would focus on improving temporal modeling, where more advanced architectures that can better capture long-term dependencies and compositional actions in untrimmed videos. Moreover, multi-modal pre-training (*e.g.*, with audio [24], optical flow [30] and human motion [39]) will enhance the ability of models to generalize to a wider range of video understanding tasks.

References

[1] Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, June 2019.

[2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, 2019.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[19] 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.

[20] 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.

[21] Lina Mezghani, Piotr Bojanowski, Karteek Alahari, and Sainbayar Sukhbaatar. Think before you act: Unified policy for interleaving language reasoning with actions. *arXiv:2304.11063*, 2023.

[22] 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.

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

[24] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aäron van den Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. In *ICCV*, 2021.

[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] Mustafa Shukor, Corentin Dancette, and Matthieu Cord. ep-alm: Efficient perceptual augmentation of language models. In *ICCV*, 2023.

[27] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. Unified model for image, video, audio and language tasks. In *ICCVW*, 2023.

[28] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. UnIVAL: Unified model for image, video, audio and language tasks. *Transactions on Machine Learning Research*, 2023.

[29] 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.

[30] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *ICCV*, 2021.

[31] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[32] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. In *CVPR*, 2023.

[33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[34] 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.

[35] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv:2212.03191*, 2022.

[36] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021.

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

[38] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023.

[39] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Via: View-invariant skeleton action representation learning via motion retargeting. *IJCV*, 2024.

[40] 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 (ICCV)*, pages 11975–11986, October 2023.

[41] 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.

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

[43] 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.