Fine-tuned CLIP Models are Efficient Video Learners

Hanoona Rasheed[1*]   Muhammad Uzair Khattak[1*]   Muhammad Maaz[1]

Salman Khan[1,2]   Fahad Shahbaz Khan[1,3]

[1]Mohamed bin Zayed University of AI   [2]Australian National University   [3]Linköping University
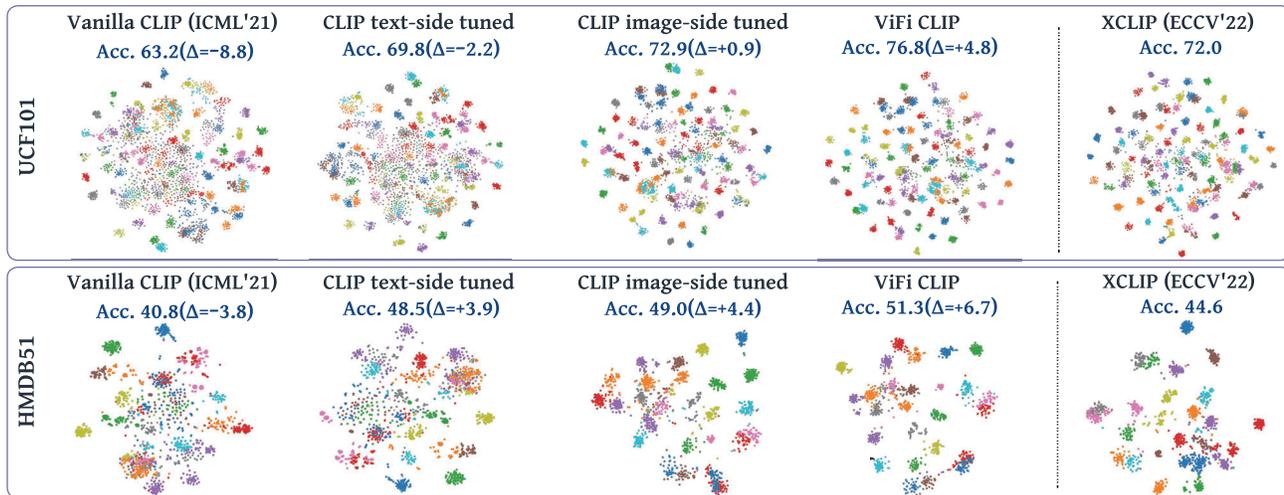
Figure 1. This work explores the capability of a simple baseline called ViFi-CLIP (Video Fintuned CLIP) for adapting image pretrained CLIP [33] to video domain. The figure compares the zero-shot performance of vanilla CLIP and several of its variants adapted for videos (trained on Kinetics-400, evaluated on UCF-101 and HMDB-51). The t-SNE visualizations of video-embeddings obtained from ViFi-CLIP (4th col.) are compared with embeddings from vanilla CLIP [33] (1st col.), individually tuned CLIP text (2nd col.) and image encoder (3rd col.) on videos, and recent state-of-the-art work, XCLIP [30] (last col.) (Δ represents difference over XCLIP). The embeddings of ViFi-CLIP are better separable, indicating that a simple fine-tuning of CLIP is sufficient to learn suitable video-specific inductive biases, and can perform competitive to more complex approaches having dedicated components designed to model temporal information in videos.

Abstract

Large-scale multi-modal training with image-text pairs imparts strong generalization to CLIP model. Since training on a similar scale for videos is infeasible, recent approaches focus on the effective transfer of image-based CLIP to the video domain. In this pursuit, new parametric modules are added to learn temporal information and inter-frame relationships which require meticulous design efforts. Furthermore, when the resulting models are learned on videos, they tend to overfit on the given task distribution and lack in generalization aspect. This begs the following question: How to effectively transfer image-level CLIP representations to videos? In this work, we show that a simple Video Fine-tuned CLIP (ViFi-CLIP) baseline is generally sufficient to bridge the domain gap from images to videos. Our qualitative analysis illustrates that the frame-level processing from CLIP image-encoder followed by feature pooling and similarity matching with corresponding

text embeddings helps in implicitly modeling the temporal cues within ViFi-CLIP. Such fine-tuning helps the model to focus on scene dynamics, moving objects and inter-object relationships. For low-data regimes where full fine-tuning is not viable, we propose a 'bridge and prompt' approach that first uses fine-tuning to bridge the domain gap and then learns prompts on language and vision side to adapt CLIP representations. We extensively evaluate this simple yet strong baseline on zero-shot, base-to-novel generalization, few-shot and fully supervised settings across five video benchmarks. Our code and pre-trained models are available at https://github.com/muzairkhattak/ViFi-CLIP.

1. Introduction

Pretrained vision-language (VL) models like CLIP [33] and ALIGN [16] have shown impressive zero-shot performance for many downstream vision applications including classification [45, 48], detection [15, 35, 49] and segmentation [22, 34]. These models are trained using millions of

*Equally contributing authors.

image-text pairs sourced from the internet and offer unique representations with strong generalization and transfer capabilities. However, such massive pretraining is laborious for videos due to the following reasons: **1)** Aligned video-text data has a limited availability and the cost of preparing such data is monumental in contrast to image-text pairs that are readily available via internet sources [16]. **2)** Videos are inherently complex and have large compute cost while the diverse appearance cues could be learned through image-text pairs with a much lower compute budget. Therefore, it is critical to devise methods effectively adapting pretrained image-language models for video-based tasks without forgetting the generic multi-modal learned representations.

Recent video-based approaches adopt CLIP representations using additional learnable components for spatio-temporal modeling. These components include self-attention layers for cross-frame communication [17], textual or visual prompts [40] or dedicated video decoder modules [30] that are learned while keeping the CLIP backbone frozen or adapting the CLIP encoders as well. However, these designs require modality-specific inductive biases to be modeled in the developed architectural modules and need careful design efforts to adapt CLIP suitably for videos. Additionally, while adapting CLIP for downstream video tasks, such approaches generally do not remain a winner across all settings. For example, zero-shot adapted approaches perform lower in supervised settings, and supervised models score lower on zero-shot generalization tasks.

To address the above challenges, we frame the following two questions: **1)** Does the adaptation of CLIP for videos using additional tunable parameters tamper its generalization capacity? **2)** Is a simple video-specific fine-tuning sufficient to bridge the modality gap between images and videos? In our empirical analysis, we observe that fine-tuning pretrained CLIP encoders along with the newly introduced temporal modeling components can hinder the generalization capability of CLIP. Interestingly, a simple CLIP model when fine-tuned on a video dataset can instill suitable video-specific adaptations within the regular CLIP model and perform competitively to more complex approaches having video-specific components inbuilt.

Although existing works explore fine-tuning of CLIP encoders as a baseline, they undermine the potential of full fine-tuning of CLIP. However we note that, full fine-tuning to achieve better visual-language alignment on videos improves synergy between temporal and language cues, and perform competitive to much sophisticated approaches developed for videos (see Fig. 1). Towards understanding how this capacity is achieved by the regular CLIP model, we show that a simple frame-level late representation aggregation before loss calculation allows the exchange of temporal cues within the video fine-tuned CLIP.

While simple CLIP fine-tuning performs competitively

to more sophisticated approaches, it is not always feasible, especially on low-data regimes. Based on the finding that simple fine-tuning can efficiently adapt CLIP for videos, we propose a two-stage '*bridge and prompt*' approach for adapting CLIP for low-data regimes that first fine-tunes vanilla CLIP on videos to bridge the modality gap, followed by a vision-language prompt learning approach keeping the tuned CLIP frozen. The contributions of this work are,

- We formulate a simple but strong baseline, ViFi-CLIP (Video Finetuned CLIP), for adapting image-based CLIP to video-specific tasks. We show that simple fine-tuning of CLIP is sufficient to learn video-specific inductive biases, resulting in impressive performance on downstream tasks (Sec. 4).
- We conduct experiments on four different experimental settings including zero-shot, base-to-novel generalization, few-shot and fully-supervised tasks. We show better or competitive performance as compared to the state-of-the-art approaches (Secs. 3, 4).
- We show the effectiveness of our proposed 'bridge and prompt' approach to first bridge the modality gap through fine-tuning followed by prompt learning in both visual and language branches of the CLIP model for low-data regimes (Sec. 5).

2. Related Work

Vision Language models: Learning multi-modal representations using large-scale image-text pretraining has proved to be effective for a wide range of uni-modal and multi-modal applications [9, 18, 24, 25, 28, 29]. Foundational VL models like CLIP [33] and ALIGN [16] follow such a pretraining paradigm and are trained on large-scale image-caption pairs with contrastive self-supervised objectives. These models are open vocabulary and effectively transfer on downstream vision applications including few-shot and zero-shot recognition [45, 47, 48], object detection [15, 35, 49], and image segmentation [11, 22, 46]. However, adapting pretrained VL models to videos is a challenging task due to the lack of video-specific temporal cues in the image-level pretraining. Therefore, recent works [17, 30, 40] adapt CLIP for videos by incorporating additional learnable components such as self-attention layers, textual or vision prompts, or dedicated visual decoder, and have demonstrated improvements on video applications. However, it is still unclear how much benefit these relatively complex domain-specific components provide compared to simple alternatives such as fine-tuning on videos.

Video Action Recognition: Designing accurate video understanding models inherently requires encoding both spatial and motion cues. Recently, vision transformers based networks [2, 26, 43] proposes to effectively model the long range spatio-temporal relationships and have shown consistent improvements over the 3D CNNs [7, 10, 12, 42].

While these approaches follow independent uni-modal solutions, works like ActionCLIP [40], XCLIP [30] and Ju *et al.* [17] adopt a multi-modal approach by utilizing CLIP and steer it for video-understanding tasks. These methods utilize the rich generalized VL representations of CLIP and fuse them with additional components for temporal modeling. However, we note that such design choices can affect the generalization ability of CLIP and a simple fine-tuning approach performs competitively.

Prompt Learning: Prompting is a recently adapted paradigm to efficiently transfer a model to downstream tasks without re-learning the trained model parameters. By utilizing a few additional learnable tokens at the inputs, this approach typically aims to retain the model generalization ability in addition to better transferring the model to downstream applications. Inherited from the NLP domain, prompting has been widely used for many vision and V-L models. CoOp [47] and CoCoOp [48] propose to use continuous learnable text prompts to transfer CLIP for image recognition tasks. Bahng *et al.* [3] introduce visual prompts to probe CLIP at its vision branch. MaPLe [20] propose multi-modal prompting to effectively adapt CLIP. By keeping the original model parameters frozen, prompting is considered efficient and requires less computing and training time as compared to conventional full fine-tuning.

In video tasks, Ju *et al.* [17] adapt CLIP via text prompts and transformer layers for temporal modeling. However, this temporal modeling hinders the CLIP generalization, and struggles to perform well in zero-shot setting.

3. Problem Settings

In this section, we introduce four problem settings for video recognition by varying the level of supervision available. This allows us to analyse the performance of our baseline and its comparison with state-of-the-art approaches across a spectrum of tasks with different degrees of generalization required. Below, we discuss the studied settings, including a newly proposed base-to-novel generalization setting for videos, in the increasing order of supervision available.

Zero-shot setting: The model is trained on a source dataset and transferred directly on downstream cross-datasets. The source dataset D_S contains samples belonging to source classes, $Y_S = \{y_i\}_{i=0}^k$. The model is evaluated on the target dataset D_T with classes Y_T such that $Y_S \cap Y_T = \phi$.

Base-to-novel generalization: To test the generalization ability of various approaches to novel classes, we introduce a *base-to-novel generalization* setting for video action recognition. A dataset D_S with labels $Y_S = \{y_i\}_{i=0}^k$ is split into base and novel classes, Y_B and Y_N such that $Y_B \cup Y_N = Y_S$ and $Y_B \cap Y_N = \phi$. The model is learned on base classes and evaluated both on base and novel classes. The proposed base and novel split categorizes the total categories into two equal halves, where the most frequently oc-
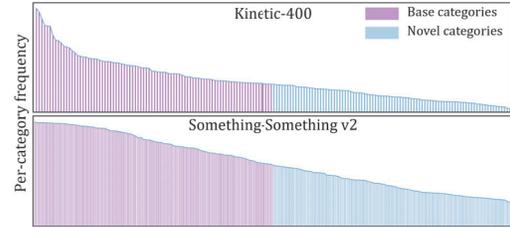

Figure 2. Frequency plot of K400 [19] and SSv2 [14].

curring classes are grouped as the base classes. Fig. 2 shows the base-novel splits of Kinetics-400 [19] and SSv2 [14].

Few-shot setting: We use this setting to test the learning capacity of the model under limited supervision. For a dataset D_S with labels $Y_S = \{y_i\}_{i=0}^k$, a general K-shot data is created, where K-samples are randomly sampled from each category $y_i \in Y_S$ for training. We use $K = 2, 4, 8$ and 16 shots. Validation set of D_S is used for evaluation.

Fully-supervised setting: This is the conventional setting for supervised approaches where for a dataset D_S with labels $Y_S = \{y_i\}_{i=0}^k$, model is trained on all training examples and evaluated on the respective test set.

4. Video Finetuned CLIP

As training vision-language (VL) models on video-caption pairs is expensive, the availability of large-scale pretrained video VL models is limited. A reliable alternative explored in literature is adaptation of large-scale pretrained image-based VL models, such as CLIP [33], for video downstream tasks. Considering the modality gap, prior methods have explored the use of various specialized attention-based components that instill communication across frames and modules to integrate the information from multiple frames [17,30,40]. On the contrary, we explore the capability of a simple baseline called ViFi-CLIP (Video Fine-tuned CLIP) for adapting CLIP [33] to video domain. Fig. 3 illus-
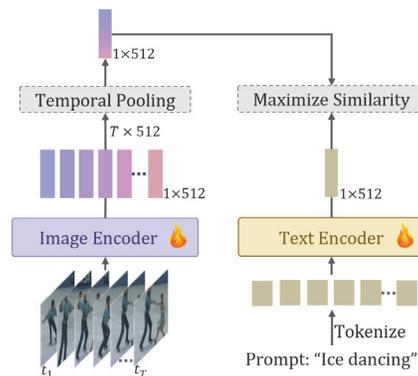

Figure 3. Overview of our simple baseline ViFi-CLIP for adapting CLIP [33] to videos. We fine-tune CLIP on videos with minimal design changes that do not include modality specific components which we find to degrade the generalization ability of CLIP 4.1. Simple frame-level late feature aggregation via temporal pooling allows the exchange of temporal cues in the CLIP representation.

Method	HMDB-51	UCF-101
Uni-modal zero-shot action recognition models		
ASR [41]	21.8 ± 0.9	24.4 ± 1.0
ZSECOC [32]	22.6 ± 1.2	15.1 ± 1.7
UR [50]	24.4 ± 1.6	17.5 ± 1.6
E2E [5]	32.7	48
ER-ZSAR [8]	35.3 ± 4.6	51.8 ± 2.9
Adapting pre-trained image VL models		
Vanilla CLIP [33]	40.8 ± 0.3	63.2 ± 0.2
ActionCLIP [40]	40.8 ± 5.4	58.3 ± 3.4
XCLIP [30]	44.6 ± 5.2	72.0 ± 2.3
A5 [17]	44.3 ± 2.2	69.3 ± 4.2
Tuning pre-trained image VL models		
CLIP image-FT	49.0 ± 0.3	72.9 ± 0.8
CLIP text-FT	48.5 ± 0.1	69.8 ± 1.1
ViFi-CLIP	**51.3** ± 0.6	**76.8** ± 0.7
	+6.7	+4.8

Method	K600 (Top-1)	K600 (Top-5)
Uni-modal zero-shot action recognition models		
SJE [1]	22.3 ± 0.6	48.2 ± 0.4
ESZSL [36]	22.9 ± 1.2	48.3 ± 0.8
DEM [44]	23.6 ± 0.7	49.5 ± 0.4
GCN [13]	22.3 ± 0.6	49.7 ± 0.6
ERZSAR [8]	42.1 ± 1.4	73.1 ± 0.3
Adapting pre-trained image VL models		
Vanilla CLIP [33]	59.8 ± 0.3	83.5 ± 0.2
ActionCLIP [40]	66.7 ± 1.1	91.6 ± 0.3
XCLIP [30]	65.2 ± 0.4	86.1 ± 0.8
A5 [17]	55.8 ± 0.7	81.4 ± 0.3
Tuning pre-trained image VL models		
CLIP image-FT	62.4 ± 1.0	85.8 ± 0.5
CLIP text-FT	68.5 ± 1.2	89.6 ± 0.3
ViFi-CLIP	**71.2** ± 1.0	**92.2** ± 0.3
	+4.5	+0.6

Table 1. **Zero-shot setting:** We compare ViFi-CLIP with uni-modal methods specifically designed for zero-shot action recognition and methods that explicitly adapt CLIP for videos. Models are trained on Kinetics-400 and evaluated directly on HMDB-51, UCF-101 (left) and Kinetics-600 (right). ViFi-CLIP acheives strong generalization. Accuracy gains over prior best are indicated in blue. We underline the second best numbers.

	K-400			HMDB-51			UCF-101			SSv2		
Method	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
Adapting pre-trained image VL models												
Vanilla CLIP [33]	62.3	53.4	57.5	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
ActionCLIP [40]	61.0	46.2	52.6	69.1	37.3	48.5	90.1	58.1	70.7	13.3	10.1	11.5
XCLIP [30]	74.1	56.4	64.0	69.4	45.5	55.0	89.9	58.9	71.2	8.5	6.6	7.4
A5 [17]	69.7	37.6	48.8	46.2	16.0	23.8	90.5	40.4	55.8	8.3	5.3	6.4
Tuning pre-trained image VL models												
CLIP image-FT	72.9	58.0	64.6	62.6	47.5	54.0	86.4	65.3	74.4	9.2	8.5	8.8
CLIP text-FT	73.4	59.7	65.8	70.0	51.2	59.1	90.9	67.4	77.4	12.4	9.5	10.8
ViFi-CLIP	**76.4**	**61.1**	**67.9**	**73.8**	**53.3**	**61.9**	**92.9**	**67.7**	**78.3**	**16.2**	**12.1**	**13.9**
	+2.3	+4.7	+3.9	+4.4	+6.5	+6.9	+2.4	+4.1	+7.1	+2.9	+2.0	+2.4

Table 2. **Base-to-novel generalization:** We compare the generalization ability of ViFi-CLIP with models that adapt CLIP [33] for video tasks on Kinetics-400, HMDB-51, UCF-101 and SSv2. Here, HM refers to harmonic mean which measures the trade-off between base and novel accuracy. Gains over prior best are shown in blue.

trates an overview of the proposed baseline ViFi-CLIP.

With the additional temporal information in videos, the important question is how to leverage this information into image-based CLIP model. We explore the capability of *full fine-tuning of CLIP* to bridge the modality gap in video domain. ViFi-CLIP fine-tunes both image and text encoder.

Given a video sample $V_i \in \mathbb{R}^{T \times H \times W \times C}$ with T frames, and corresponding text label Y, the CLIP image encoder encodes the T frames independently as a batch of images and produce frame level embeddings $x_i \in \mathbb{R}^{T \times D}$. These frame-level embeddings are average-pooled to obtain a video-level representation $v_i \in \mathbb{R}^D$. We refer to this as temporal pooling as this operation implicitly incorporates temporal learning via aggregation of multiple frames.

The CLIP text encoder encodes the class Y, wrapped in a prompt template such as 'a photo of a <category>' to produce text embedding $t \in \mathbb{R}^D$. For a batch of videos, the cosine similarity $sim(.)$, between all the video-level embeddings v_i and the corresponding text embeddings t_i is maximized to fine-tune the CLIP model via cross-entropy

(CE) objective with a temperature parameter τ,

$$\mathcal{L} = -\sum_i \log \frac{\exp(sim(v_i, t_i)/\tau)}{\sum_j \exp(sim(v_i, t_j)/\tau)}.$$

Experimental setup: We use ViT-B/16 based CLIP model for our experiments. For zero-shot, base-to-novel and few-shot settings, we use 32 sparsely sampled frames with single view evaluation. In fully supervised setting, we use 16 frames and multi-view inference with 4 spatial crops and 3 temporal views. We conduct our analysis on five action recognition benchmarks: Kinetics-400 and 600 [6, 19], HMDB-51 [21], UCF-101 [38] and Something Something V2 (SSv2) [14]. See Appendix B and D for more details.

4.1. ViFi-CLIP Generalizes Well!

When adapting CLIP to video tasks that demand high generalization ability, two key elements must be satisfied: i) modality gap should be bridged by adapting image-based CLIP for video domain ii) modality adaptation must happen without hurting the in-build generalization. To analyze the generalization ability of the simple CLIP fine-tuning approach, we evaluate two problem settings: 1) zero-shot

Model	HMDB-51				UCF-101				SSv2			
	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
Adapting pre-trained image VL models												
Vanilla CLIP [33]	41.9	41.9	41.9	41.9	63.6	63.6	63.6	63.6	2.7	2.7	2.7	2.7
ActionCLIP [40]	47.5	_57.9_	57.3	59.1	70.6	71.5	73.0	**91.4**	4.1	_5.8_	_8.4_	_11.1_
XCLIP [30]	_53.0_	57.3	_62.8_	_64.0_	48.5	75.6	83.7	**91.4**	3.9	4.5	6.8	10.0
A5 [17]	39.7	50.7	56.0	62.4	_71.4_	_79.9_	_85.7_	89.9	_4.4_	5.1	6.1	9.7
Tuning pre-trained image VL models												
CLIP image-FT	49.6	54.9	57.8	62.0	74.4	79.1	85.3	90.5	4.9	6.0	7.2	10.4
CLIP text-FT	54.5	61.6	63.1	65.0	80.1	82.8	85.8	88.1	6.2	6.1	6.3	9.1
ViFi-CLIP	**57.2**	**62.7**	**64.5**	**66.8**	**80.7**	**85.1**	**90.0**	**92.7**	**6.2**	**7.4**	**8.5**	**12.4**
	+4.2	+4.8	+1.7	+2.8	+9.3	+5.2	+4.3	+1.3	+1.8	+1.6	+0.1	+1.3

Table 3. **Few-shot setting:** We compare ViFi-CLIP with approaches that explicitly adapt CLIP for video action-recognition on HMDB-51, UCF-101 and SSv2. Gains over the best previous methods that adapt CLIP are indicated in blue and underlined the second best.

setting to evaluate the cross-dataset generalization, and 2) base-to-novel generalization to test performance on novel categories. The later setting has not been studied before for videos and we introduce new base-to-novel splits for videos. Further details of the splits are given in Appendix C.

(i) Zero-shot Setting: We investigate the cross-dataset generalization ability of the simple baseline, ViFi-CLIP, in a zero-shot setting. We train the model on a large video action recognition dataset, Kinetics-400 and evaluate across different datasets, HMDB-51, UCF-101 and Kinetics-600. In Table 1, we compare ViFi-CLIP with: 1) uni-modal methods that are specifically designed for zero-shot action recognition, and 2) models that adapt image-based multi-modal VL models for video action recognition. The direct zero-shot evaluation of vanilla CLIP shows impressive generalization performance as compared to uni-modal methods. Further, adapting CLIP with video-specific components helps in improving the generalization in most of the scenarios, indicating the importance of bridging the modality gap. However, the simple fine-tuning approach shows better capability to bridge the domain gap, without disrupting the generalization learned in the pretraining stage of CLIP. Note that we also fine-tune image and text encoders (denoted with CLIP image-FT and CLIP text-FT respectively) and compare with fully fine-tuned CLIP (ViFi-CLIP) where the latter gives stronger generalization due to better alignment of visual and text representations on video tasks. ViFi-CLIP achieves consistent gains of +6.7%, +4.8% and +4.5% in HMDB-51, UCF-101 and K-600 respectively.

(ii) Base-to-Novel Generalization Setting: In Table 2, we evaluate the generalization from base to novel classes on four datasets, K-400, HMDB-51, UCF-101 and SSv2. In comparison to XCLIP [30] and ActionCLIP [40] which use additional components to model video-specific inductive biases, ViFi-CLIP with minimal design modifications provides better base accuracy, and shows noticeable gains in novel accuracy. It provides a better base-to-novel trade-off with an overall best harmonic mean on all datasets. Further, ViFi-CLIP shows better understanding of scene dynamics

even on temporally-challenging datasets like SSv2.

4.2. CLIP directly adapts to Video tasks

We explore the capability of a simple fine-tuning approach in bridging the domain gap on supervised video action recognition tasks under different experimental settings: 1) few-shot learning, 2) fully-supervised setting.

(i) Few-shot Setting: In Table 3, we show the effect of ViFi-CLIP in the few-shot setting along with methods that adapt CLIP for videos. We note that ViFi-CLIP consistently improves performance with increasing shots. Across the three datasets HMDB-51, UCF-101 and SSv2 in all the shots ($K = 2, 4, 6, 8$), it provides better performance against all the compared methods. Interestingly, it achieves relatively larger gains in extremely limited data scenarios demonstrating robustness towards overfitting. For example, it achieves gains of +9.3% and 4.2% in UCF-101 and HMDB-51 respectively compared to prior best methods.

(ii) Fully-supervised Setting: We compare the performance of ViFi-CLIP trained on Kinetics-400 with uni-

Method	Frames	Top-1	Top-5	Views	GFLOPs	TP
Uni-modal architectures						
Uniformer-B [23]	32	83.0	95.4	4×3	259	-
TimeSformer-L [4]	96	80.7	94.7	1×3	2380	-
Mformer-HR [31]	16	81.1	95.2	10×3	959	-
Swin-L [27]	32	83.1	95.9	4×3	604	-
Adapting pre-trained image VL models						
ActionCLIP [40]	32	83.8	96.2	10×3	563	67.7
X-CLIP [30]	16	**84.7**	**96.8**	4×3	287	58.5
A6 [17]	16	76.9	93.5	-	-	-
Tuning pre-trained image VL models						
CLIP image-FT	16	82.8	96.2	4×3	281	71.1
CLIP text-FT	16	73.1	91.2	4×3	281	71.1
ViFi-CLIP	16	_83.9_	_96.3_	4×3	281	71.1

Table 4. **Fully-supervised setting:** We compare ViFi-CLIP with uni-modal methods and models specifically designed to adapt CLIP for video tasks on Kinetics-400. In addition to accuracy, we report FLOPs and throughput (TP).
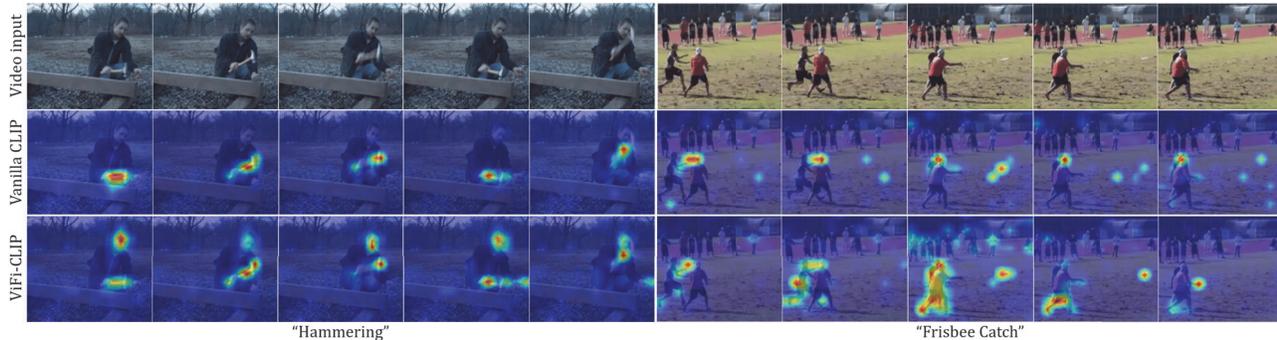
Figure 4. **Attention map visualizations** of ViFi-CLIP in comparison with vanilla CLIP on two examples from UCF-101 validation set. ViFi-CLIP learns inter-object relationships and scene-dynamics from temporal cues and focuses on fast-moving parts and objects, thereby demonstrating the ability to encode video specific information. **(Left):** An example on action class 'hammering'. While vanilla CLIP focuses only on the object (hammer), ViFi-CLIP attends to the interaction between the person and the object. **(Right):** Example on 'frisbee catch' category. Vanilla CLIP uses only appearance cues and confuses a hat with the frisbee, while ViFi-CLIP focuses on the players and pays more attention on fast-moving parts like the hands and legs of the players, and correctly locates the frisbee.

modal video-specific models and other methods that tailor CLIP for videos in Table 4. The simple approach of *fully fine-tuning* CLIP provides competitive performance in comparison to methods that use additional carefully designed learnable components for video-specific temporal modeling. Further, the ablation of fine-tuning image-encoder and text-encoder indicates the effectiveness of fine-tuning the full CLIP model to address the domain gap.

4.3. How simple fine-tuning bridges domain gap?

Having shown the effectiveness of ViFi-CLIP for adapting CLIP for video action recognition, we explore how this approach encodes video specific information that enables bridging the modality gap. We conduct our experiments by ablating on the fusion mechanism that is used to combine frame-level information. In the proposed baseline, we adopt an embedding-level fusion, where individual frames are encoded by the image encoder, and the resulting image embeddings are then fused together to obtain a video-level visual representation. We explore two alternate fusion mechanisms; 1) *Decision-level fusion:* image embeddings from individual frames are used separately to compute similarity (logits) with the corresponding text embeddings. The frame-level logits are then averaged to obtain video-level logits. 2) *Image-level fusion:* the frames of a video are considered as individual images, and losses are computed across each frame, thus removing all temporal information.

In Table 5, we compare the two fusion mechanisms with our simple embedding level fusion across four datasets, Kinetics-400 tiny (a smaller split of full K-400), few-shot ($K = 16$) splits of SSv2, HMDB-51 and UCF-101. The analysis shows that fusing the frame-level embeddings helps the model learn the temporal relations between different frames, thus implicitly establishing inter-frame communication. We note a significant gain on SSv2 with

Method	K400-tiny	SSv2	HMDB-51	UCF-101
Vanilla CLIP [33]	51.6	2.7	41.9	63.6
Decision-level fusion	74.6	10.8	62.2	90.5
Image-level fusion	74.0	11.2	64.4	91.7
Embedding fusion	**76.9**	**12.4**	**66.8**	**92.7**

Table 5. Analysis on different fusion mechanisms that integrate temporal information in the model. ViFi-CLIP uses an embedding-level fusion, where representations of multiple frames are combined together to integrate the temporal information.

embedding-level fusion that further supports the intuition, as SSv2 demands rich temporal modeling due to its fine-grained actions as compared to Kinetics-400.

4.4. How effective are the video-specific representations learned during simple fine-tuning?

We conduct qualitative analysis on the generalization performance of ViFi-CLIP for the zero-shot setting, as shown in Fig. 1. The t-SNE visualisation of video-embeddings from ViFi-CLIP are compared with vanilla CLIP, other alternatives (CLIP image-FT and CLIP text-FT) and prior state-of-the art method XCLIP [30]. The feature representations improve with the fine-tuning of either text or image-encoder over vanilla CLIP. When both text and image encoder are tuned, the learned representations further improve and show better separability in the latent space. Additionally, we note that ViFi-CLIP achieves competitive performance when compared with XCLIP, that requires dedicated components to model temporal information.

To better understand what temporal relations are captured by fine-tuning CLIP on a video-dataset, we present attention map visualizations in Fig. 4. Our empirical analysis stipulates two interesting findings: 1) ViFi-CLIP learns inter-object relationships from temporal cues to recognize the action. For example, in Fig. 4 (left), the model fo-

Figure 5. **Generalization to out-of-distribution examples**: Attention map visualizations from ViFi-CLIP shows impressive generalization. **(Left):** Visualization on a diffusion model generated (synthetic) video from Imagen [37] shows how ViFi-CLIP focuses on the body of the astronaut, the horse and its moving feet. **(Right):** Example of a rare scenario 'giraffe diving'. ViFi-CLIP consistently attends to the giraffe at various locations: diving board, air, water-surface and under the pool.

cuses not only on the object (hammer), but attends to the interaction between the person and the object. 2) ViFi-CLIP focuses on scene dynamics and moving objects. In Fig. 4 (right), ViFi-CLIP attends to the fast-moving parts of the scene and learns to attend to salient parts of the body important for temporal understanding such as the legs and hands of the players. Intuitively, these observations indicate that temporal relations can be implicitly modeled by simply fine-tuning CLIP on a video-dataset. Additionally, we test ViFi-CLIP on extreme out-of-distribution examples. Attention maps shown in Fig. 5 demonstrate good generalization. It supports the claim that ViFi-CLIP learns temporal relations and aids in boosting the generalization of CLIP [33] by keeping CLIP image and text encoders intact.

4.5. Is fine-tuning efficient w.r.t adapting CLIP?

Being a competitive alternative to other methods in-terms of accuracy, we further study the compute complexity of ViFi-CLIP in comparison to other methods.

Method	GFLOPs	TP	Params (M)
ActionCLIP [40]	563	67.7	168.5
XCLIP [30]	287	58.5	131.5
ViFi-CLIP	281	71.1	124.7

Table 6. Compute comparison of ViFi-CLIP with methods that adapt CLIP with additional components. Throughput per view (TP) is measured using a single A100 GPU. ViFi-CLIP enjoys efficiency in-terms of GFLOPs, throughput and parameter count.

Table 6 shows that ViFi-CLIP provides high throughput (TP) of 71.1 images/sec as compared to other methods that adapt CLIP for videos. This is attributed mainly due to its simple design that avoids using any additional video-specific components. This also leads to lower FLOPs and fewer training parameters as compared to other approaches.

5. Bridge and Prompt in low-data regimes

ViFi-CLIP shows that a simple fine-tuning approach is effective in bridging the domain gap in video. However, fine-tuning the CLIP model may not always be feasible as it requires training large number of parameters. Particularly in case of low-data regimes, where availability of training data is extremely limited, we explore an important question: How can one efficiently steer CLIP towards various downstream tasks, after bridging the modality gap, that favours both effectiveness and efficiency in-terms of performance and compute respectively? We explore a two-stage framework, '*bridge and prompt*': i) The model is first fine-tuned on a video dataset to bridge the modality gap, ii) Model is adapted to downstream tasks for better generalization through context optimization via prompting. Here the entire model is frozen, and prompts are adapted and learned for a specific task. Ju *et al.* propose a strong baseline that learns task-specific vision prompts for adapting CLIP for video tasks [17] and use lightweight transformers for temporal modeling. Although this efficient prompting technique proves to adapt CLIP for video tasks, the model struggles to generalize towards unseen classes due to the late fusion through the transformer layers.

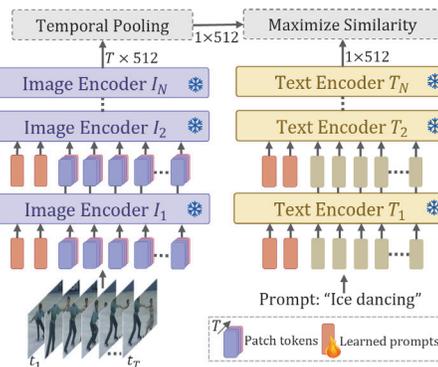To this end, we develop an extended baseline that

Figure 6. We use Vision-Language prompting approach to adapt CLIP for videos. Vision and textual prompt tokens are used in the vision and language branch of CLIP which are the model's only learnable parameters. These prompts steer CLIP towards downstream tasks in low data regime scenarios without losing the original generalization ability of CLIP. Deep contextual prompts are used in both branches across multiple transformer layers.

Model	HMDB-51				UCF-101				SSv2			
	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
Vanilla CLIP [33]	41.9	41.9	41.9	41.9	63.6	63.6	63.6	63.6	2.7	2.7	2.7	2.7
ActionCLIP [40]	54.3	56.2	59.3	66.1	76.7	80.4	87.6	91.8	4.8	6.9	9.1	12.3
XCLIP [30]	60.5	**66.8**	69.3	71.7	89.0	91.4	94.7	96.3	6.6	7.8	9.9	**13.7**
A5 [17]	46.7	50.4	61.3	65.8	76.3	84.4	90.7	93.0	4.5	6.7	7.2	9.5
VL Prompting	**63.0**	65.1	**69.6**	**72.0**	**91.0**	**93.7**	**95.0**	**96.4**	**6.7**	**7.9**	**10.2**	13.5

Table 7. VL prompting effectively improves over other methods in few-shot setting. Models are pretrained on Kinetics-400 to bridge the modality gap.

Method	HMDB-51			UCF-101			SSv2		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
Vanilla CLIP [33]	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
ActionCLIP [40]	69.0	**57.2**	62.6	85.6	**75.3**	80.1	8.1	8.7	8.4
XCLIP [30]	75.8	52.0	61.7	95.4	74.0	83.4	14.2	11.0	12.4
A5 [17]	70.4	51.7	59.6	95.8	71.0	81.6	12.9	5.7	7.9
VL prompting	**77.1**	54.9	**64.1**	**95.9**	74.1	**83.6**	**15.8**	**11.5**	**13.3**

Table 8. Comparison of VL prompting approach in base-to-novel generalization setting. VL prompting shows consistent gains on base classes while also improving on novel classes. It performs competitive even against fine-tuning based approaches [30,40]. All models are first pretrained on Kinetics-400.

efficiently adapts CLIP in low-data regimes via vision-language (VL) prompt learning for videos. Fig. 6 shows the overall architecture of our proposed framework. In contrast to previous approaches that learn prompts only at the language branch for video adaption [17], we use a vision-language prompt learning design where prompt vectors are learnt on both vision and language branch. Moreover, we introduce prompts at deeper layers of both encoders, to capture hierarchical contextual representations.

5.1. Prompting is effective on fine-tuned CLIP

We compare our VL prompting with other methods that adapt CLIP for videos. Following the first stage of our 'bridge and prompt' approach, all methods are pretrained on Kinetics-400 and then evaluated in two problem settings: few-shot transfer and base-to-novel generalization. Note that ActionCLIP [40] and XCLIP [30] fine-tune CLIP encoders on the corresponding datasets.

(i) Few-shot Setting: Table 7 shows the results for few-shot transfer. Vanilla CLIP [33] is the lower bound and A5 [17] is most similar to our approach as it only adapts prompts and few transformer layers, keeping the CLIP model frozen. We note that VL prompting consistently provides better performance over A5 and even performs competitively against fine-tuning approaches. Particularly for extreme cases where $K = 2$, it provides the best results by providing absolute gains of 2.5% and 2% over XCLIP on HMDB-51 and UCF-101 respectively. This suggests the significance of learning prompts in low-data regimes.

(ii) Base-to-Novel Generalization Setting: In Table 8, we compare the results on base-to-novel setting among different methods. In comparison to vanilla CLIP, all fine-tuning approaches improves their generalization ability for novel classes. We note that vision-language (VL) prompting provides competitive performance against prior prompting [17] designs and highly competitive fine-tuning methods [30,40]

without any video-specific attention modules. The results suggest that VL prompting is effective in steering pretrained CLIP model towards downstream tasks without compromising on generalization.

5.2. Is prompting efficient w.r.t CLIP adaptation?

We perform an analysis on complexity of various methods in the low-data regime as detailed in Table 9. CLIP adaptation methods provide less throughput due to the use of video-specific learnable components in addition to the vanilla CLIP model. A5 [17] requires fewer FLOPs but achieves less throughput, due to the additional transformer blocks for temporal modeling. On the other hand, VL prompting shows higher efficiency in terms of throughput and comparable FLOPs as compared to prior approaches.

Method	GFLOPs	TP
ActionCLIP [40]	563	67.8
XCLIP [30]	287	58.5
A5 [17]	284	62.5
VL prompting	287	71.6

Table 9. Computational complexity comparison of VL prompting with other methods in-terms of GFLOPs and throughput (TP).

6. Conclusion

This work shows the significance of an often neglected but simple baseline for transferring image-based CLIP model to video domain. We demonstrate that simply fine-tuning both the vision and text encoders on video data performs favourably on supervised as well as generalization tasks. The results show the scalabiltiy and advantage of a simple solution with respect to sophisticated approaches developed dedicatedly for videos in majority of the settings. In cases where fine-tuning is not possible, we also propose a bridge and prompt scheme that uses the video fine-tuned representations to quickly adapt to downstream video applications.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 4

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 5

[5] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4

[6] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 4, 11, 12

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[8] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4, 12

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *The European Conference on Computer Vision*, 2020. 2

[10] R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *Advances in Neural Information Processing Systems*, 2016. 2

[11] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[13] Pallabi Ghosh, Nirat Saini, Larry S Davis, and Abhinav Shrivastava. All about knowledge graphs for actions. *arXiv preprint arXiv:2008.12432*, 2020. 4

[14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense.

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 1, 2

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 2021. 1, 2

[17] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *The European Conference on Computer Vision*, 2022. 2, 3, 4, 5, 7, 8, 11

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR– Modulated Detection for End-to-End Multi-Modal Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 4, 11, 12

[20] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv:2210.03117*, 2022. 3

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2011. 4, 11

[22] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 1, 2

[23] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. In *ICLR*, 2022. 5

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. 2020. 2

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 3, 4, 11

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViL-BERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks . In *Advances in Neural Information Processing Systems*, 2019. 2

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *The European Conference on Computer Vision*, 2022. 2

[30] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *The European Conference on Computer Vision*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 11, 12

[31] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems*, 2021. 5

[32] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, 2017. 4

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 3, 4, 5, 6, 7, 8, 11

[34] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[35] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[36] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, 2015. 4

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 7, 13

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4, 11

[39] 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 *The European Conference on Computer Vision*, 2016. 11

[40] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2, 3, 4, 5, 7, 8, 11

[41] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *ECML PKDD*, 2017. 4

[42] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[43] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[44] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 4

[45] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *The European Conference on Computer Vision*, 2022. 1, 2

[46] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. In *The European Conference on Computer Vision*, 2022. 2

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 2, 3

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3

[49] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *The European Conference on Computer Vision*, 2022. 1, 2

[50] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *CVPR*, 2018. 4