Supplemental Material

We provide supplementary material that provides additional details and further qualitative analysis for the main paper. The contents follow the following order:

A. Other video-level tasks

We scale and evaluate our approach on video retrieval task on MSRVTT (9K) dataset with consistent settings (16 frames using ViT-B/16) as A5 model [17] and show improved performance in Table 10.

Method	R@1	R@5
Frozen	31.0	59.5
A5 [17]	36.7	64.6
ViFi-Clip	**44.8**	**72.4**

Table 10. Comparison of ViFiCLIP with methods that explicitly adapt CLIP for videos on the Video-retrieval task.

B. Implementation Details

In all our experiments on ViFi-CLIP, and its variants, individually tuned CLIP text encoder (CLIP text-FT) and image encoder (CLIP image-FT), all randomly sampled frames are pre-processed to a spatial size of 224×224. In our experiments, we use handcrafted text prompts with a template '`a photo of a <category>`'. Following CLIP [33], the maximum number of text tokens is set to 77. We use an AdamW optimizer and weight decay of 0.001. We modify the epochs, batch size and learning rate across the different experimental settings, which are detailed below.

We conduct our analysis under four experimental settings: zero-shot, base-to-novel generalization, few-shot and fully-supervised. In the *zero-shot setting*, ViFi-CLIP and its variants are trained for 10 epochs on Kinetics-400 dataset with a batch size of 256, and a learning rate of 8e-6. In the *base-to-novel generalization* and the *few-shot* setting, ViFi-CLIP is trained in a few-shot manner, with a batch size of 64, and a learning rate of 2e-6. For the *fully-supervised setting*, we train ViFi-CLIP on the kinetics-400 dataset for 30 epochs with a batch size of 512 and a learning rate of 22e-6.

We implement other baseline methods including A6 [17], ActionCLIP [40] and XCLIP [30] using their default optimal hyper-parameters as reported in their work. For Efficient prompting [17], we use their best performing A6 model for the fully-supervised setting, and use their A5 model in the zero-shot, and base-to-novel generalization and few-shot settings. In case of ActionCLIP [40], we use their best-performing variant *Transf* [40] in all our experiments.

C. Dataset details

We conduct our analysis on five established action recognition benchmarks: Kinetics-400 [19] and Kinetics-600 [6], HMDB-51 [21], UCF-101 [38] and Something-Something v2 (SSv2) [14].

Kinetics-400 and Kinetics-600: The K-400 datasets contains 400 human action classes comprising video clips taken from various YouTube videos that lasts for about 10 seconds. It contains around 240K training and 20K validation videos. The K-600 is an extension of of K-400, with around 650K video clips covering 600 action categories, consisting of around 410K training and 29K validation videos.

HMDB-51: The HMDB-51 dataset contains 71K realistic videos collected from different sources spanning 51 action categories. The standard split consist of 3570 training samples and 1530 validation samples. The training and validation are further split into three individual splits, each containing 70 and 30 clips of all action categories for training and validation, respectively.

UCF-101: UCF-101 contains 13K realistic videos collected from YouTube covering 101 action categories that includes five types of action: human-object interaction, body-motion, human-human interaction, playing instrumental music and sports. The standard split trains on 9537 videos and evaluates on 3783 videos, which are grouped into three splits.

Something-Something v2 (SSv2): The SSv2 dataset is a large collection of video clips of humans performing actions with everyday objects, spanning 174 action categories. The dataset evaluates the capacity of the model on fine-grained actions such as covering something with something or uncovering something, making the dataset more temporally biased as opposed to other datasets. The standard split consist of 168,913 training videos and 24,777 validation videos. We report the top-1 accuracy over the validation split.

D. Evaluation Protocols

We conduct our analysis on four different experimental settings: *zero-shot*, *base-to-novel generalization*, *few-shot* and *fully-supervised* setting. Across these settings, we use a sparse sampling strategy [39] to sample frames and set the number of frames to 16 or 32 (specified under each setting).

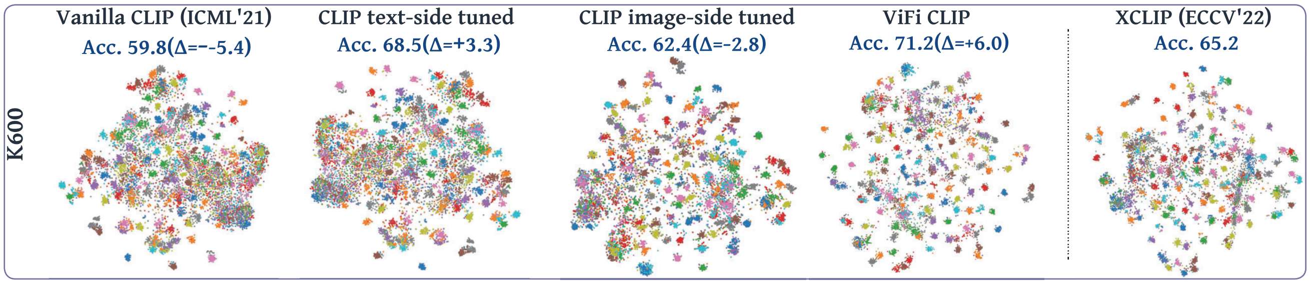

Figure 7. **t-SNE visualizations for Kinetics-600.** For K600 [6], we show the t-SNE visualizations for 160 classes that are non-overlapping with Kinetics-400. The fine-tuned models are trained on Kinetics-400 and evaluated on the non-overlapping classes of Kinetics-600.

Each sampled frame is spatially scaled on the shorter side to 256, with a center crop of 224.

Zero-shot setting: Under the zero-shot setting, models trained on Kinetics-400 are evaluated on three cross datasets, HMBD-51, UCF-101 and Kinetic-600. For HMBD-51 and UCF-101, the methods are evaluated on their corresponding three validation splits and we report the top-1 average accuracy over them. In case of Kinetics-600, we follow [8] and evaluate the methods on 220 categories that are non-overlapping with Kinetics-400. We report top-1 and top-5 average accuracy over three randomly sampled splits, each containing 160 categories. In this setting, we use a single-view inference with 32 frames.

Base-to-novel setting: For extensive analysis on the generalization ability of various approaches, we introduce a *base-to-novel generalization* setting for video action recognition tasks, where a model is first trained on a set of *base* (seen) classes in a few-shot manner and evaluated on a set of *novel* (unseen) classes. We present comprehensive generalization analysis on four datasets, Kinetics-400, HMBD-51, UCF-101 and SSv2. For each dataset, we create three training splits, each containing randomly sampled 16-shots of every action category. The split categorizes the total categories into two equal halves, where the the most frequently occurring classes are considered as the base classes, and the rarely occurring categories are taken as the novel classes. Figure 2 shows the frequency distribution of the Kinetics-400 and depicts the resulting base-novel split. The model is evaluated on the corresponding validation splits. In case of HMDB-51 and UCF-101, the training and validation considers only their first split, while for Kinetics and SSv2, the models are evaluated on their full validation split. The setting uses 32 frames and follows a single-view inference.

Few-shot setting: The few-shot setting creates a general K-shot split, where K-samples are randomly sampled from each category for training. Specifically, we use 2, 4, 8 and 16 shots for three datasets, HMBD-51, UCF-101 and SSv2. The models are evaluated on the first validation split for HMDB-51 and UCF-101 and the full validation split in case of SSv2. In this setting, we use 32 sparsely sampled frames and evaluate with single-view inference.

Fully-supervised setting: In the fully-supervised setting, the methods are trained on Kinetics-400 are evaluated on its complete validation set. We use 16 frames and report multi-view inference with three different spatial crops and four temporal clips.

E. Additional qualitative results

The t-SNE visualizations of video-embeddings in Fig. 1 are computed for the UCF101 (1st col.) and HMDB51 (2nd col.), whereas for K-600 in Fig. 7. Here, each color represents a category. We observe the embeddings of ViFi-CLIP within a category are better separable from others, indicating the effectiveness of the proposed approach to learn suitable video-specific inductive biases. We also evaluate the quality of the clusters in the visualizations. The homogeneity (H), completeness (C), and V-measure (V) computed for vanilla CLIP, XCLIP and ViFi-CLIIP for HMDB51 in table 11, show trends consistent to our t-SNE visualizations.

Method	H (↑)	C (↑)	V (↑)
Vanilla CLIP	0.84	0.86	0.85
XCLIP	0.92	0.93	0.93
ViFi-CLIP	**0.94**	**0.95**	**0.95**

Table 11. Metrics evaluating the quality of clusters in the t-SNE visualizations.

In Fig. 7, we show additional t-SNE visualizations for the zero-shot evaluation of Kinetics-600 [6]. We compare video embedding of vanilla CLIP and its fine-tuned variants with XCLIP [30]. The models are trained on Kinetics-400 [19] and then evaluated directly on non-overlapping classes of K600. For ViFi-CLIP, the video embeddings of different classes show better separability among all other approaches. ViFi-CLIP finetunes both the text and vision encoder of CLIP, achieves better generalization performance and provides a gain of +6% on Kinetic-600 in comparison to the recent state of the art method XCLIP [30].

To analyze the type of temporal information captured by ViFi-CLIP, we present additional attention map visual-

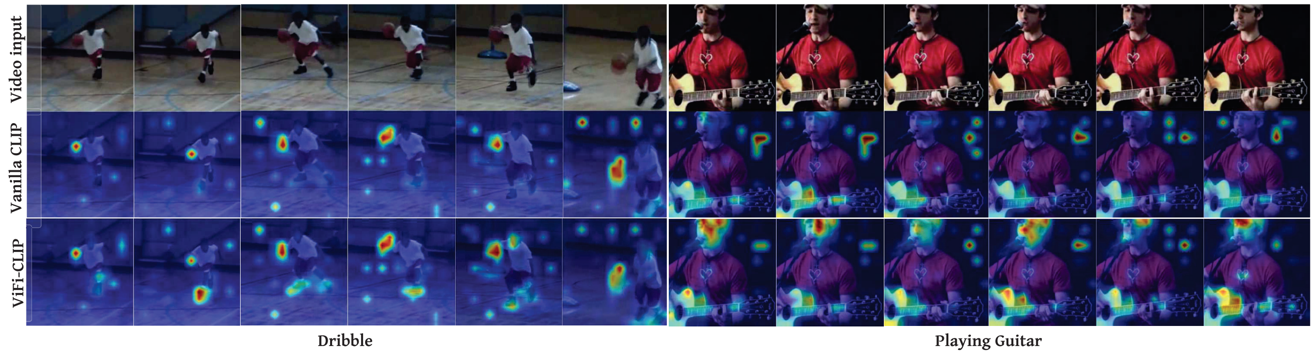

Figure 8. **Attention map visualizations** of ViFi-CLIP in comparison with vanilla CLIP on two examples from HMDB-51 (left) and UCF-101 (right) validation set. Fine-tuning CLIP on video-datasets in ViFi-CLIP helps the model learn inter-object relationships and scene-dynamics from temporal cues. The model focus on moving objects and fast-moving parts which indicates that ability of ViFi-CLIP to encode video specific information. **(Left):** An example on an action class with fast motion, 'dribble'. While vanilla CLIP focuses only on the ball, ViFi-CLIP attends to the interaction between the player and the object. Moreover, it always focuses on fast-moving parts of the player (legs), thus shows ability to focus on temporal cues. **(Right):** Example from 'playing guitar' category. While vanilla CLIP uses only appearance cues and attends to the guitar, ViFi-CLIP focuses on the interaction between the singer and the guitar, and pays attention on moving parts like lips of the players.

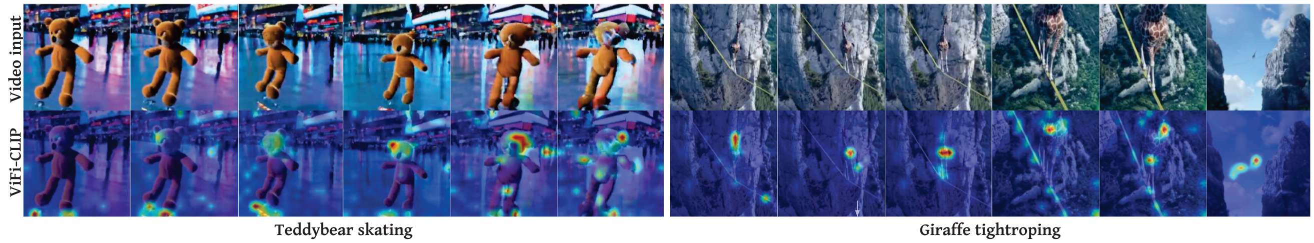

Figure 9. **Generalization to out-of-distribution examples**: Attention map visualizations from ViFi-CLIP shows good generalization. **(Left):** Visualization on a synthetically generated video from Imagen [37] shows how ViFi-CLIP focuses on inter-object relationships, like the teddy-bear and the skating shoes. **(Right):** Example of a rare scenario '`giraffe tight-roping`'. ViFi-CLIP attends to the giraffe at difference scene variations and additionally focuses on the tight-rope, showing ability to capture inter-object relationships.

izations in Fig. 8. As discussed earlier, the visualization indicates that fine-tuning CLIP on a video dataset helps in learning inter-object relationships from temporal cues, which plays a key role in recognizing the action category. Additionally, it steers the models to focus on scene dynamics, moving parts and objects in the scene. For example, in Fig. 8 (left), the model focuses on the moving ball, the child and the fast-moving body parts like the legs. Similarly in Fig. 8 (right), while vanilla CLIP only focuses on the guitar, ViFi-CLIP learns the interaction between the singer and the guitar.

In Fig. 9, we show additional attention map visualizations on extreme out-of-distribution examples. We test ViFi-CLIP on synthetically generated videos to test the generalization ability of the model. Fig. 9 (left) shows that the models successfully focus on the skating shoes, and the interaction with the teddy-bear. When tested on a rare scenario like '`giraffe tight-roping`' (shown in Fig. 9 (right)), ViFi-CLIP shows good generalization in recognizing the action using both appearance and temporal cues. These visualizations indicate that temporal relations can be implicitly modeled by simply fine-tuning CLIP on a video-dataset.