Fine-tuned CLIP Models are Efficient Video Learners
CVPR-23

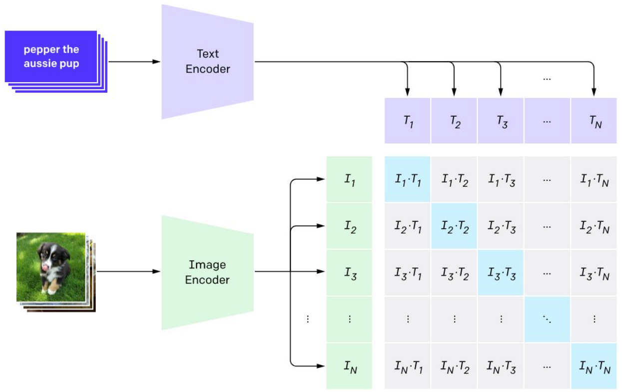
[1]Hanoona Rasheed* [1]Muhammad Uzair Khattak* [1]Muhammad Maaz [1,2]Salman Khan [1,3]Fahad Shahbaz Khan
[1]Mohamed Bin Zayed University of AI, [2]Australian National University, [3]Linköping University
*Equal Contribution

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Background

- Pretrained Vision-Language (V-L) models are open-vocabulary
- E.g: CLIP pretrained on 400 Million image-caption pairs
  - Zero-shot capability
  - Effectively transfer to downstream vision tasks
  - Generalizable



CLIP used for zero-shot classification (Radford et al., 2021)

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Effective formulation of CLIP baseline for videos

Problem statement:

Similar to CLIP for images, can we come up with VideoCLIP based V-L model for videos?

Existing solutions:

- Video-text pretraining
  - Expensive: Curating large scale video-text pairs
  - High compute requirement
- Adapt already available image-text models for videos

Common Methods to Adapt CLIP for Videos

- Introduce additional modules for temporal modeling
- e.g. Video decoders, temporal attention, inter-frame communication blocks
- Recent works e.g. XCLIP and ActionCLIP

It is Challenging

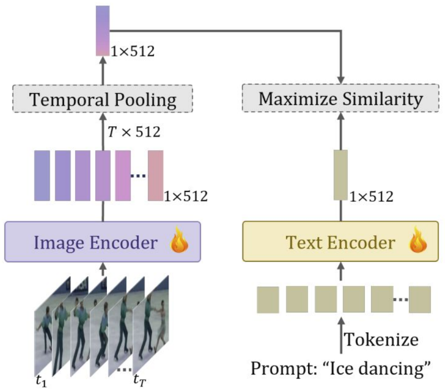
- Additional components hurts the inherent generalization ability of CLIP
- Increase compute requirements during training and inference

Effective formulation of CLIP baseline for videos

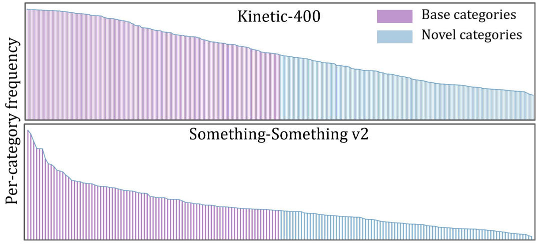
A simple Video Fine-tuned CLIP (ViFi-CLIP) is sufficient to bridge the domain gap

Our contributions:

- We formulated a simple baseline, Video Fine-tuned CLIP (ViFi-CLIP)
    - Adapts image-based CLIP for video tasks
- Introduce base-to-novel generalization benchmark for video-domain
- Propose a two-stage 'bridge and prompt' approach for adapting CLIP



ViFi-CLIP

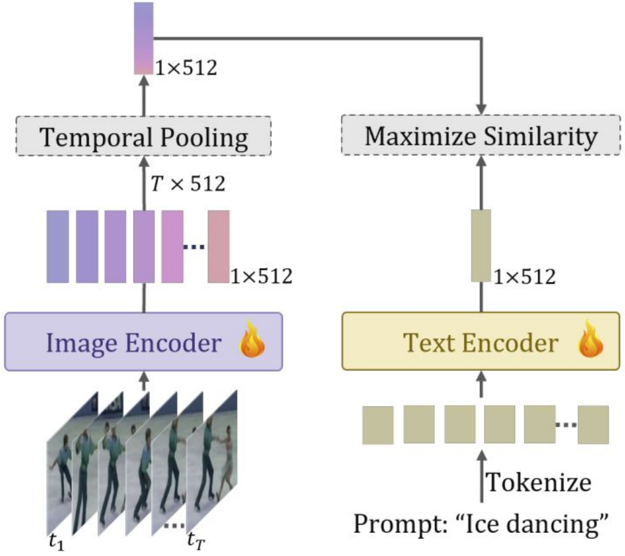


Proposed base-to-novel splits

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

ViFi-CLIP

- **Our simple baseline ViFi-CLIP for adapting CLIP to videos**
  - Fine-tune CLIP on videos with minimal design changes
  - No modality specific components that may degrade the generalization of CLIP
  - Frame-level late feature aggregation via temporal pooling allows the exchange of temporal cues

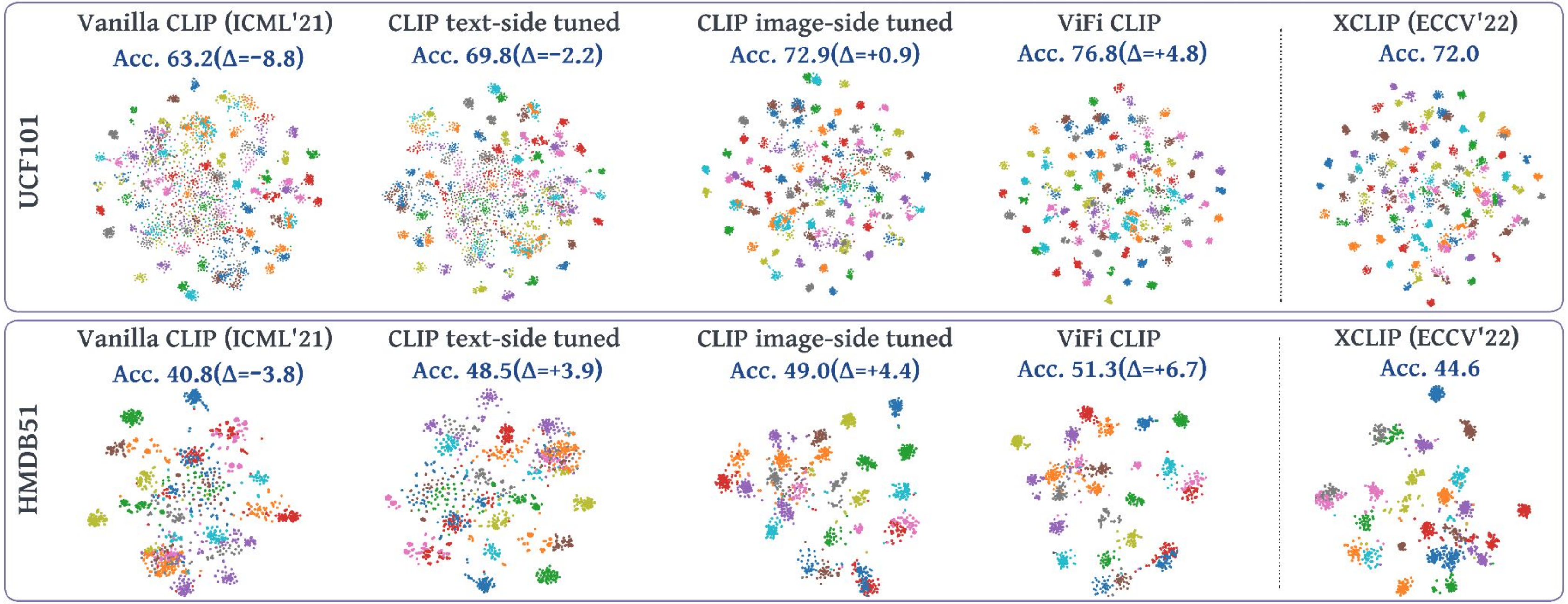


ViFi-CLIP

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

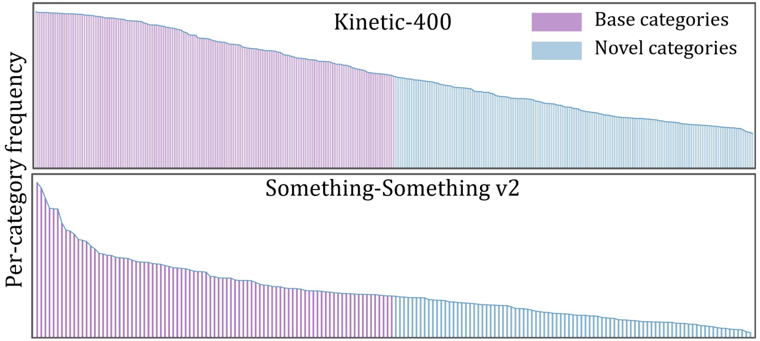
ViFi-CLIP

- ViFi-CLIP can learn suitable video representations with minimal design changes



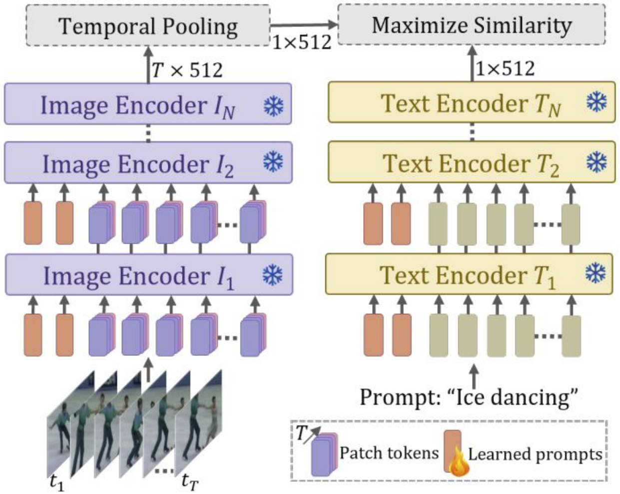
Base to Novel generalization benchmark

- Introduce a base-to-novel generalization benchmark
  - Evaluating model's generalization ability within a dataset
  - First open-vocabulary video recognition protocol
  - Splits datasets into base and novel classes



Bridge and Prompt approach in low-data regimes

- We explore a two-stage approach, **‘bridge and prompt’**
  - Fine-tuning on a video dataset to bridge the modality gap.
  - Model is then adapted to downstream tasks for better generalization via prompting.



Experiments

We conduct experiments on four different benchmark settings

Generalization benchmarks:

- Zero-shot
  - Pretrain models on K-400
  - Evaluate models on: UCF101, HMDB-51, K-600
- Base-to-novel generalization
  - Train models on base classes
  - Evaluate models on base and novel classes

Supervised learning benchmarks:

- Few-shot
- Fully-supervised tasks

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

ViFi-CLIP Generalizes Well

Zero-shot setting

- Modality gap bridged by adapting CLIP for video domain (K-400 pretraining)
- Without loss in generalization ability towards cross datasets

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]	<u>44.6</u> ± 5.2	<u>72.0</u> ± 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]	<u>66.7</u> ± 1.1	<u>91.6</u> ± 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

ViFi-CLIP Generalizes Well

Base-to-novel generalization

- We compare ViFi-CLIP with
  - Methods that explicitly adapt CLIP for videos

Method	K-400 Base	K-400 Novel	K-400 HM	HMDB-51 Base	HMDB-51 Novel	HMDB-51 HM	UCF-101 Base	UCF-101 Novel	UCF-101 HM	SSv2 Base	SSv2 Novel	SSv2 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

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

ViFi-CLIP directly adapts to supervised video tasks

Few-shot learning

- ViFi-CLIP supasses other approaches that explicitly adapts CLIP for videos

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	<u>57.9</u>	57.3	59.1	70.6	71.5	73.0	<u>91.4</u>	4.1	<u>5.8</u>	<u>8.4</u>	<u>11.1</u>
XCLIP [30]	<u>53.0</u>	57.3	<u>62.8</u>	<u>64.0</u>	48.5	75.6	83.7	<u>91.4</u>	3.9	4.5	6.8	10.0
A5 [17]	39.7	50.7	56.0	62.4	<u>71.4</u>	<u>79.9</u>	<u>85.7</u>	89.9	<u>4.4</u>	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

ViFi-CLIP directly adapts to supervised video tasks

Fully supervised setting (K400)

- ViFi-CLIP performs competitive in fully supervised setting

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	<u>83.9</u>	<u>96.3</u>	4×3	281	71.1

Further analysis

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

- We compare the compute complexity of ViFi-CLIP with methods that explicitly adapt CLIP for videos

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

Visualizations

Attention maps

- ViFi-CLIP learn Inter-object relationships and scene-dynamics from temporal cues

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Conclusion

- We propose a simple and effective baseline for adapting CLIP to videos
- Performs favourably well against existing complex approaches on four benchmark in video action recognition
- Introduce base to novel generalization benchmark for videos
- Bridge and Prompt for low data regimes