Prompting Visual-Language Models for Efficient Video Understanding

Author: Chen Ju et al.
Presenters: Myles, Wei, Taixi, Jeff

Motivation

- Current CV is task-specific
- Goal: Multi-task visual representation with minimal tuning
- Challenges with Video Understanding:
  - Resource-intensive
  - Image-text misalignment
  - Composed of frame sequences
- IVL models like CLIP, ALIGN, FILIP excel in general-purpose learning
- IVL models learn from image-caption pairs similar to how video task involve pairing sequences with relevant descriptions
- Solution:
  - Prompt-based learning for efficient video understanding
  - Adapt pre-trained CLIP for video tasks

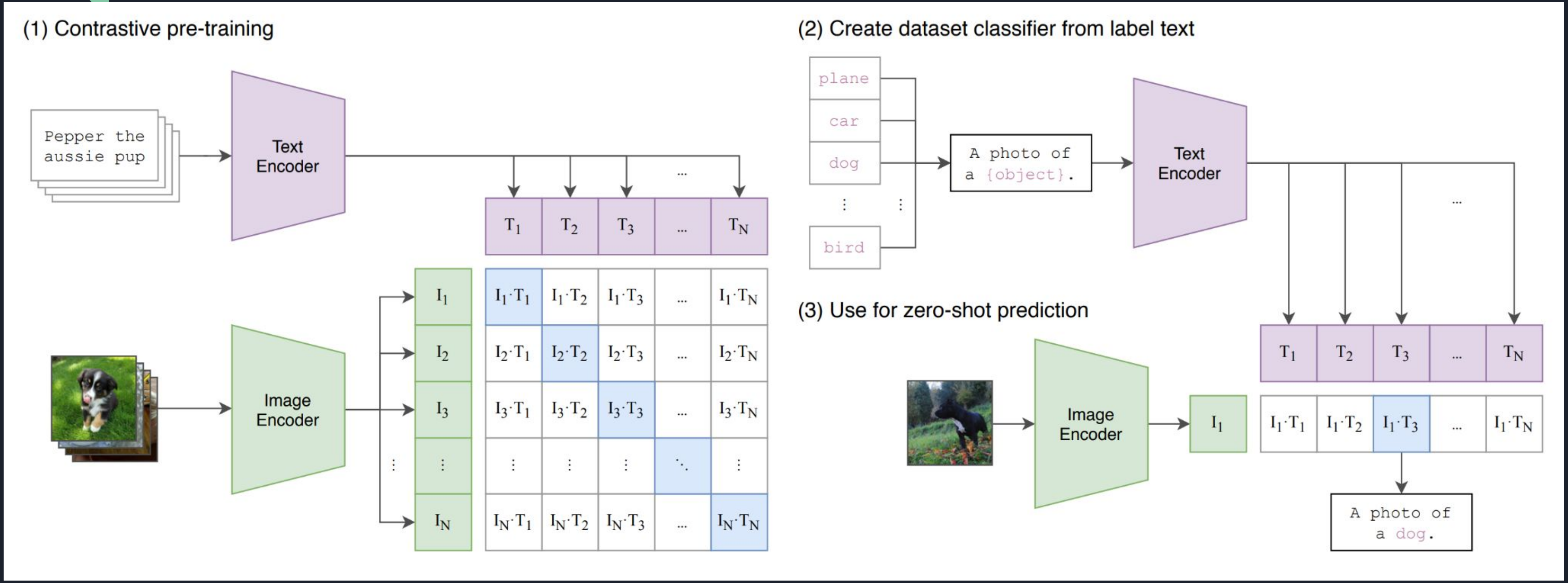
Prompt-based learning for Efficient Video Modeling

- Current Challenge:
  - Fine-tuning for each task is costly; don't want 100s of models.
- Our Solution:
  - Use CLIP-like general prompts for various tasks.
  - Optimize prompt vectors to match pre-training objectives, helping model generalize.
  - Employ lightweight transformers to encode temporal info from frame-wise features.
- Adaptation Tasks Covered:
  - Action recognition
  - Action localization
  - Text-to-video retrieval.

Model (I-VL)

- Review I-VL
- Describe proposed method

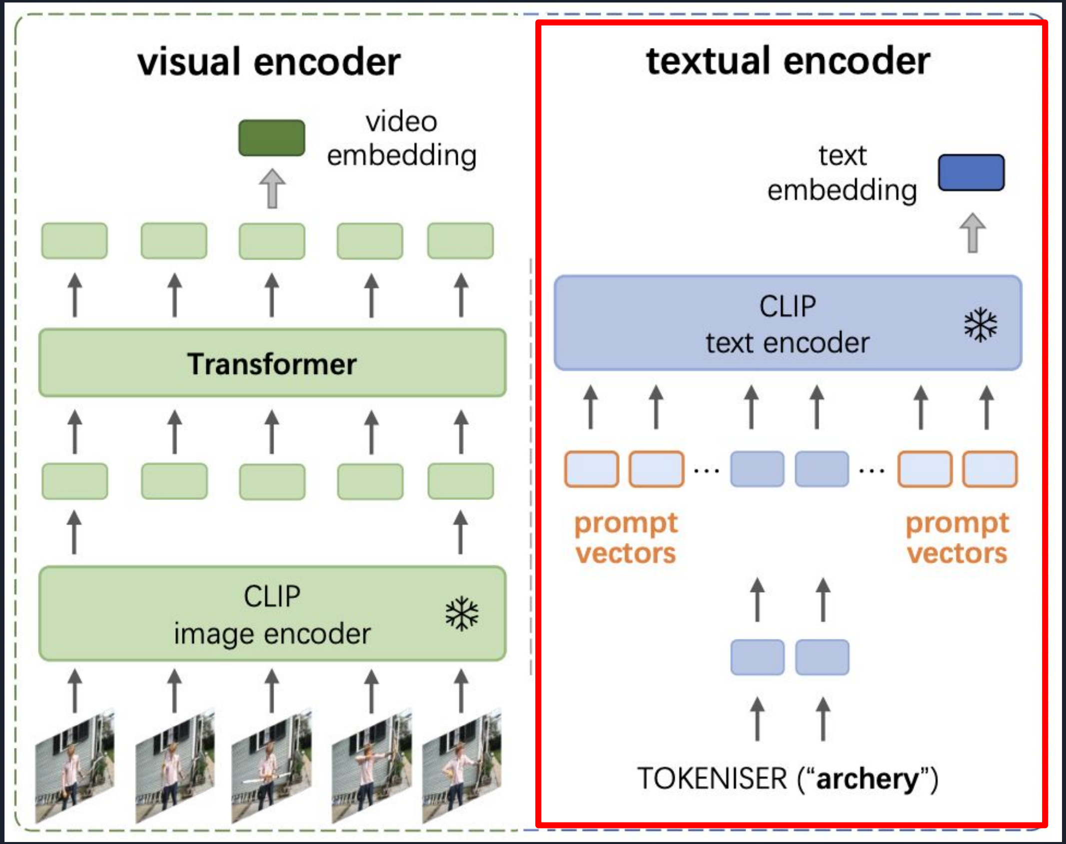
Model (I-VL -> CLIP)



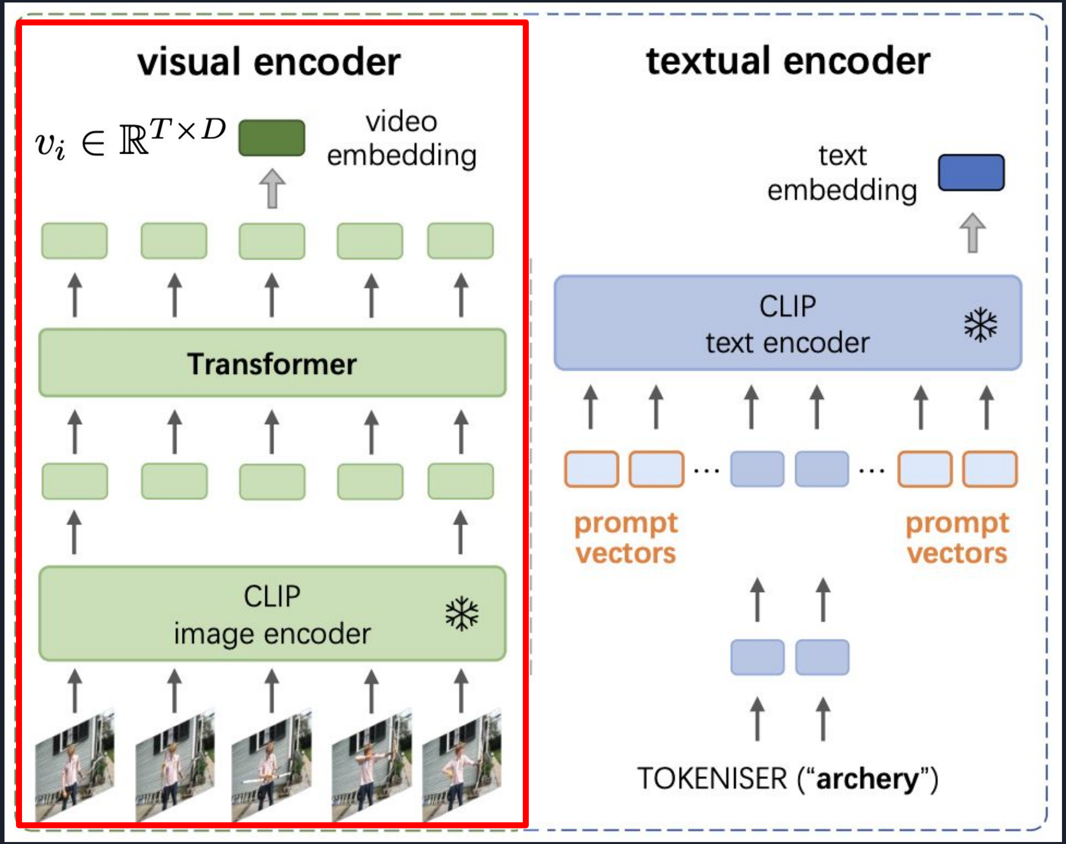
Model(CLIP -> Video)

- Why?

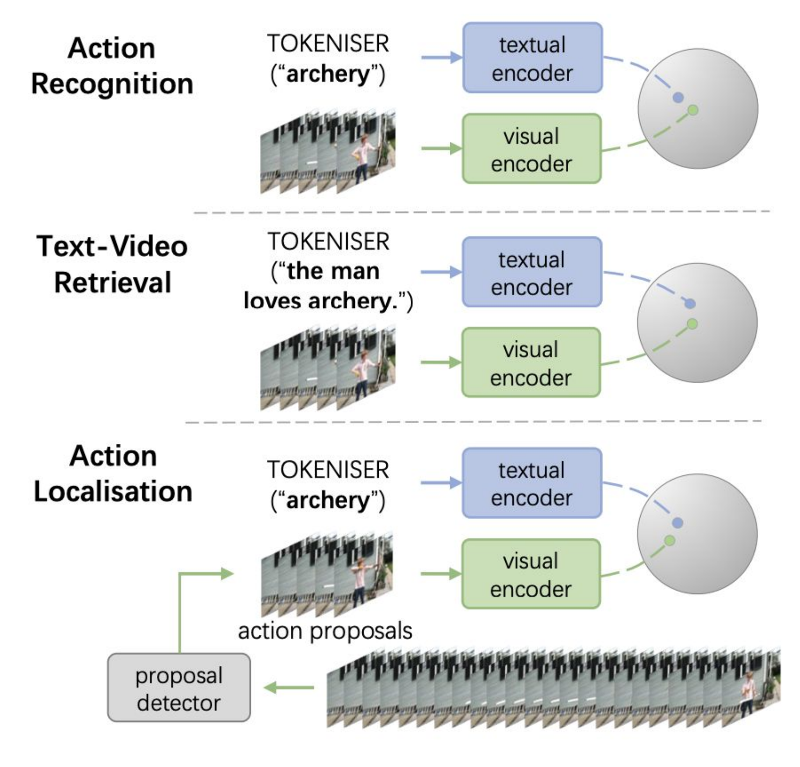
Model (Model Adaption)



Model (Temporal Modeling)



Model(Downstream Tasks)



Model (Loss)

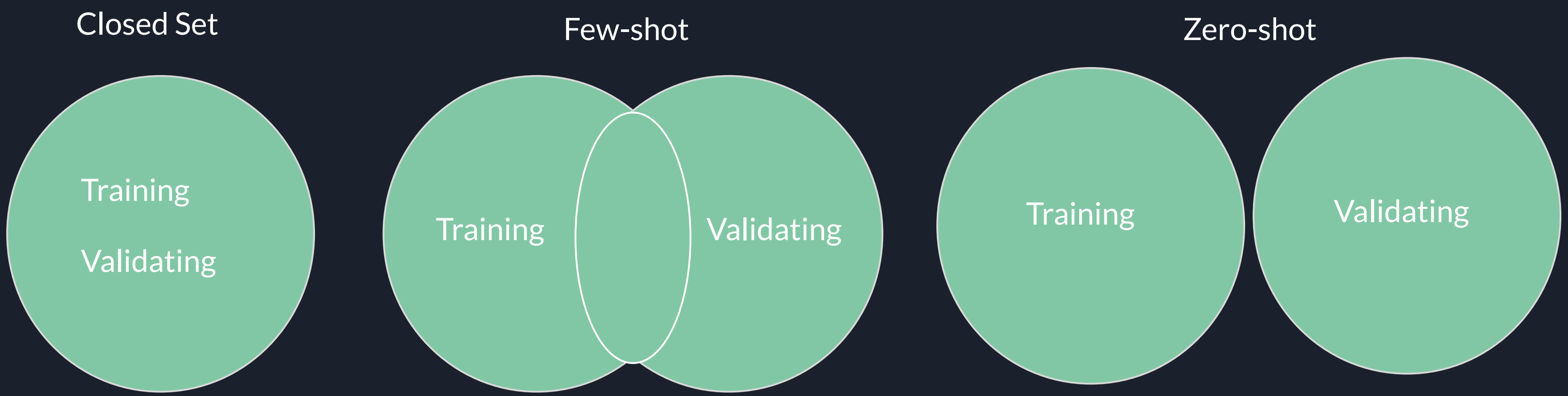
- action recognition & text-video retrieval
- action localisation

$$\overline{v}_i = \Phi_{\text{POOL}}(v_i) \in \mathbb{R}^{1\times D}$$

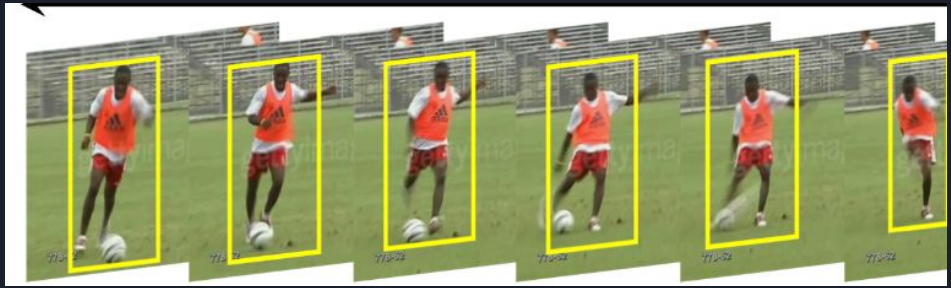
- Overall NCE loss

$$\mathcal{L} = -\sum_i \Big(\log \frac{\exp(\overline{v}_i \cdot c_i/\tau)}{\sum_j \exp(\overline{v}_i \cdot c_j/\tau)} \Big)$$

Experiment Definition



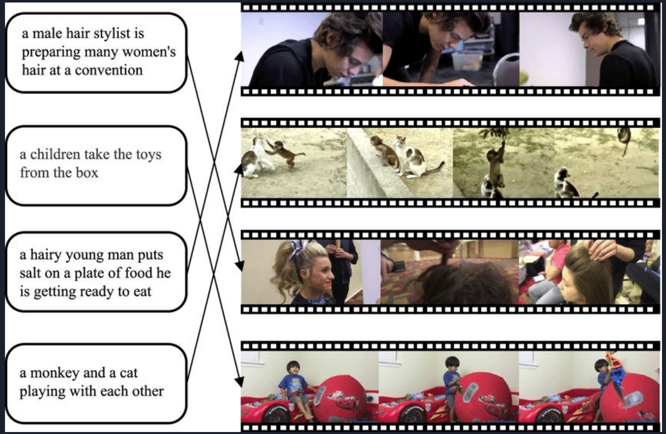
Experiment Tasks



Action Recognition

Action Localization

Text-video Retrieval



Experiment: Closed-set ablation

Table 1: **Ablation study for closed-set action recognition.**

			K-400			K-700		
Model	Prompt	Temporal	TOP1	TOP5	AVG	TOP1	TOP5	AVG
Baseline-I [69]	hand-craft	✗	–	–	–	–	–	52.4
Baseline-II [69]	✗	✗	–	–	–	–	–	66.1

Experiment: Closed-set ablation

Table 1: **Ablation study for closed-set action recognition.**

			K-400			K-700		
Model	Prompt	Temporal	TOP1	TOP5	AVG	TOP1	TOP5	AVG
Baseline-I [69]	hand-craft	✗	–	–	–	–	–	52.4
Baseline-II [69]	✗	✗	–	–	–	–	–	66.1
A0	2+X+2	✗	65.4	88.7	77.1	56.3	81.9	69.1
A1	4+X+4	✗	66.1	89.0	77.6	56.6	82.4	69.5
A2	8+X+8	✗	67.9	90.0	79.0	57.4	83.0	70.2
A3	16+X+16	✗	68.8	90.1	79.5	57.8	83.1	70.5
A4	16+X+16	1-TFM	75.8	92.9	84.4	64.2	87.3	75.8
A5	16+X+16	2-TFM	76.6	93.3	85.0	64.7	88.5	76.6
A6	16+X+16	3-TFM	76.9	93.5	85.2	64.8	88.4	76.6
A7	16+X+16	4-TFM	76.8	93.5	85.2	64.9	87.9	76.4

Experiment: Closed-set action recognition

Table 2: **Comparison on closed-set action recognition.** On all datasets, our model performs comparably to existing methods, by training far fewer parameters.

	HMDB-51		UCF-101		K-400		K-700	
Method	TOP1	TOP5	TOP1	TOP5	TOP1	TOP5	TOP1	TOP5
I3D [13]	74.3	–	95.1	–	71.6	90.0	58.7	81.7
S3D-G [85]	75.9	–	96.8	–	74.7	93.4	–	–
R(2+1)D [79]	74.5	–	96.8	–	72.0	90.0	–	–
TSM [50]	–	–	–	–	74.7	–	–	–
R3D-50 [33]	66.0	–	92.0	–	–	–	54.7	–
NL-I3D [83]	66.0	–	–	–	76.5	92.6	–	–
SlowFast [20]	–	–	–	–	77.0	92.6	–	–
X3D-XXL [18]	–	–	–	–	80.4	94.6	–	–
TimeSformer-L [4]	–	–	–	–	80.7	94.7	–	–
Ours (A5)	66.4	92.1	93.6	99.0	76.6	93.3	64.7	88.5

Experiment: Few-shot action recognition

Method	K-shot	N-way	Prompt	Temporal	UCF-101	HMDB-51	K-400
CMN [101]	5	5	–	–	–	–	78.9
TARN [5]	5	5	–	–	–	–	78.5
ARN [94]	5	5	–	–	83.1	60.6	82.4
TRX [68]	5	5	–	–	96.1	75.6	85.9
Baseline-I [69]	–	5	hand-craft	✗	91.9	68.9	95.1
Ours	5	5	✓	✗	**98.3**	**85.3**	**96.4**
	5	5	✓	✓	97.8	84.9	96.0
Baseline-I [69]	–	$\mathcal{C}_{\text{ALL}}$	hand-craft	✗	64.7	40.1	54.2
Ours	5	$\mathcal{C}_{\text{ALL}}$	✓	✗	77.6	56.0	57.1
	5	$\mathcal{C}_{\text{ALL}}$	✓	✓	**79.5**	**56.6**	**58.5**

Experiment: Closed-set action localization

			THUMOS14						ActivityNet1.3			
Method	Date	Modality	0.3	0.4	0.5	0.6	0.7	AVG	0.5	0.75	0.95	AVG
CDC [73]	2017	RGB+Flow	40.1	29.4	23.3	13.1	7.9	22.8	45.3	26.0	0.2	23.8
TALNET [14]	2018	RGB+Flow	53.2	48.5	42.8	33.8	20.8	39.8	38.2	18.3	1.3	20.2
BSN [53]	2018	RGB+Flow	53.5	45.0	36.9	28.4	20.0	36.8	46.5	30.0	8.0	30.0
DBS [29]	2019	RGB+Flow	50.6	43.1	34.3	24.4	14.7	33.4	–	–	–	–
BUTAL [96]	2020	RGB+Flow	53.9	50.7	45.4	38.0	28.5	43.3	43.5	33.9	**9.2**	30.1
A2NET [89]	2020	RGB+Flow	58.6	54.1	45.5	32.5	17.2	41.6	43.6	28.7	3.7	27.8
GTAD [88]	2020	RGB+Flow	66.4	60.4	51.6	37.6	22.9	47.8	50.4	34.6	9.0	34.1
BSN++ [77]	2021	RGB+Flow	59.9	49.5	41.3	31.9	22.8	41.1	51.3	**35.7**	8.3	**34.9**
AFSD [49]	2021	RGB+Flow	**67.3**	**62.4**	**55.5**	**43.7**	**31.1**	**52.0**	**52.4**	35.3	6.5	34.4
TALNET [14]	2018	RGB	42.6	–	31.9	–	14.2	–	–	–	–	–
A2NET [89]	2020	RGB	45.0	40.5	31.3	19.9	10.0	29.3	39.6	25.7	2.8	24.8
Baseline-III	2022	RGB	36.3	31.9	25.4	17.8	10.4	24.3	28.2	18.3	3.7	18.2
Ours	2022	RGB	**50.8**	**44.1**	**35.8**	**25.7**	**15.7**	**34.5**	**44.0**	**27.0**	**5.1**	**27.3**

Experiment: Zero-shot action localisation

Table 7: **Results of zero-shot action localisation.** Baseline-III uses the same proposal detector as our method, but adopts the original CLIP with handcrafted prompts as the proposal classifier. Our model is trained on 75% (or 50%) action categories and tested on the remaining 25% (or 50%) action categories.

		THUMOS14						ActivityNet1.3			
Method	Train $v.s$ Test	0.3	0.4	0.5	0.6	0.7	AVG	0.5	0.75	0.95	AVG
Baseline-III	75% $v.s$ 25%	33.0	25.5	18.3	11.6	5.7	18.8	35.6	20.4	2.1	20.2
Ours	75% $v.s$ 25%	**39.7**	**31.6**	**23.0**	**14.9**	**7.5**	**23.3**	**37.6**	**22.9**	**3.8**	**23.1**
Baseline-III	50% $v.s$ 50%	27.2	21.3	15.3	9.7	4.8	15.7	28.0	16.4	1.2	16.0
Ours	50% $v.s$ 50%	**37.2**	**29.6**	**21.6**	**14.0**	**7.2**	**21.9**	**32.0**	**19.3**	**2.9**	**19.6**

Experiment: text-video retrieval

Table 8: **Results of text-video retrieval.** Baseline-IV refers to the original CLIP model with text query naïvely encoded, *i.e.* without using any prompt. E2E denotes if the model has been trained end-to-end. As these methods are pre-trained on different datasets with variable sizes, it is unlikely to make fair comparisons.

		MSRVTT (9K)		LSMDC		DiDeMo		SMIT	
Method	E2E	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CE [55]	✗	21.7	51.8	12.4	28.5	16.1	41.1	–	–
MMT [23]	✗	24.6	54.0	13.2	29.2	–	–	–	–
TT-CE+ [15]	✗	29.6	61.6	**17.2**	**36.5**	21.6	48.6	–	–
Baseline-IV	✗	31.2	53.7	11.3	22.7	28.8	54.6	39.3	62.8
Ours	✗	**36.7**	**64.6**	13.4	29.5	**36.1**	**64.8**	**66.6**	**87.8**
Frozen [3]	✓	31.0	59.5	15.0	30.8	34.6	65.0	–	–
CLIP4Clip [58]	✓	44.5	71.4	22.6	41.0	43.4	70.2	–	–

I-VL Spec

Transformer x2: **5 Million**

Prompt Vector x16 x2: **16k**