Holistic Features are almost Sufficient for Text-to-Video Retrieval

Kaibin Tian*, Ruixiang Zhao*, Zijie Xin, Bangxiang Lan, Xirong Li†

Text-to-Video Retrieval (T2VR)

CVPR
SEATTLE, WA JUNE 17-21, 2024

- aims to retrieve unlabeled videos by ad-hoc textual queries
- two objectives: 1) effective 2) efficient

Query: two girls doing a cups song

Top 1:

Top 2:

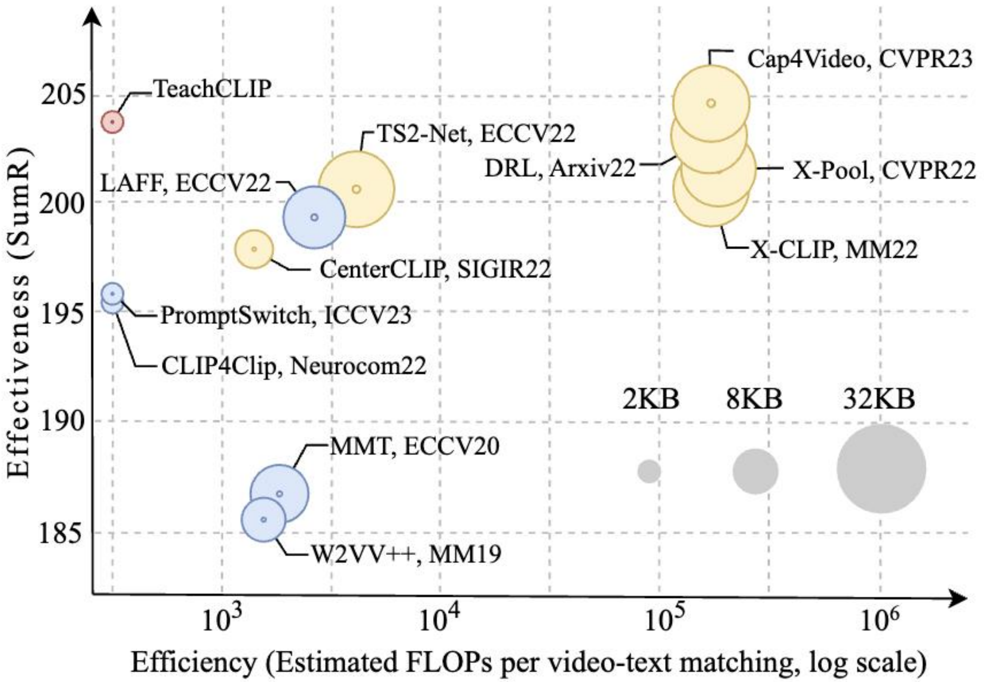
Top 3:

Motivation

- CLIP4Clip: efficient but not effective enough
- Recent methods: effective but inefficient

knowledge distillation

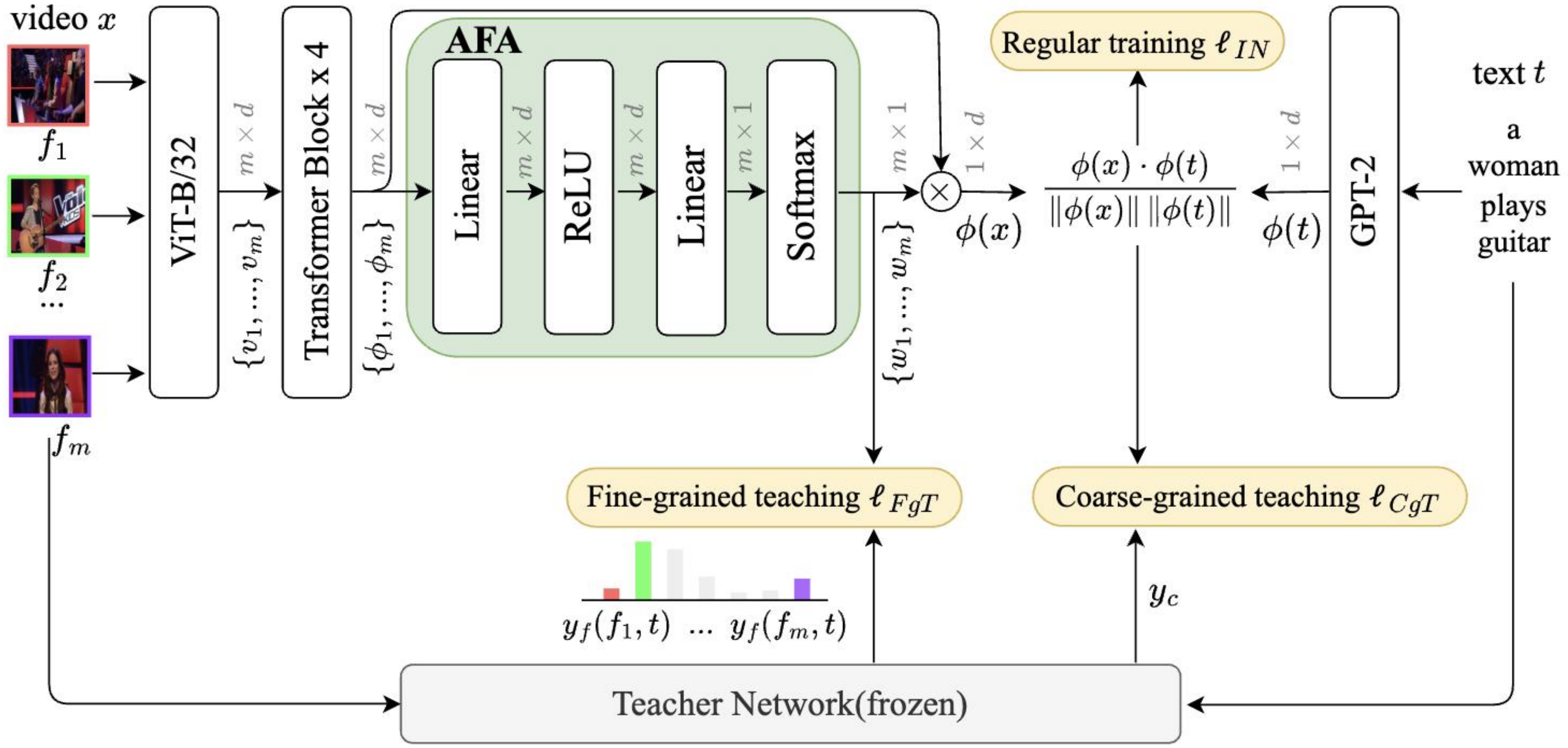
TeachCLIP: a good balance



Model	Per video-text (↓) matching (FLOPs)	Video feature (↓) storage (KB)	Video feature (↓) extraction (FLOPs)	R1 (↑)	SumR (↑)
CLIPPING[27]	0.5K	2	16.80G	40.7	–
CLIP4Clip[24]	0.5K	2	53.64G	42.8	195.5
PromptSwitch[4]	0.5K	2	59.28G	43.6	195.7
CenterCLIP[38]	1.5K	6	–	44.2	197.9
TS2-Net[22]	6.1K	24	54.27G	46.7	200.5
X-CLIP[25]	220.9K	26	53.64G	45.3	200.8
X-Pool[9]	275.0K	24	53.49G	46.0	201.5
DRL[32]	220.4K	26	53.64G	46.2	203.2
Cap4Video[34]	220.9K	28	–	**47.8**	**204.3**
TeachCLIP	0.5K	2	53.65G	46.8	203.7

Method

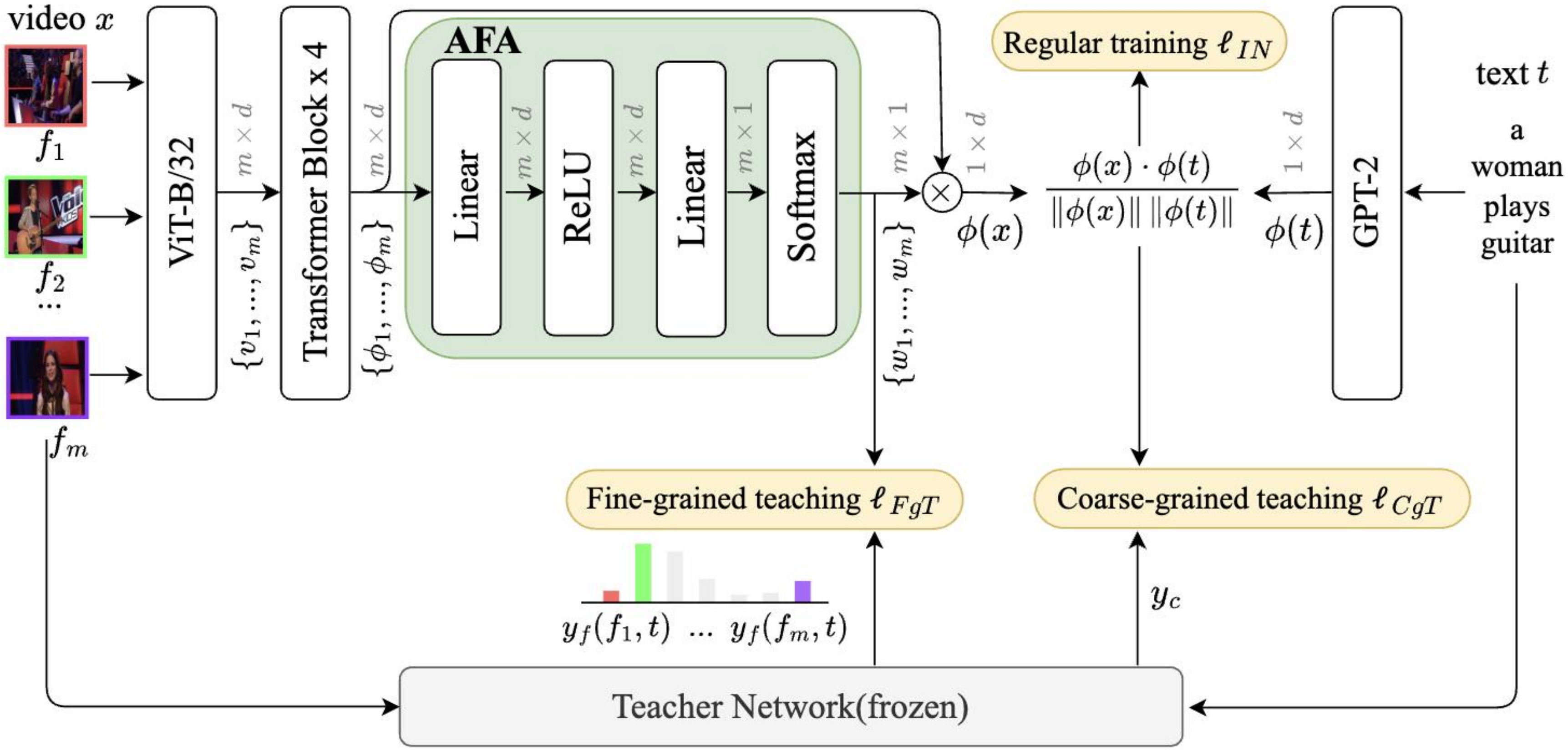
- ➢ Attentional frame-feature aggregatiog block
- ➢ Multi-grained teaching



Method

- Multi-grained teaching
  - Coarse-grained teaching

$$\ell_{CgT} := \frac{1}{b}\sum_{i=1}^{b} d_p(\sigma(B_{i,\cdot}), \sigma(y_c(v_i,\cdot))) + \frac{1}{b}\sum_{j=1}^{b} d_p(\sigma(B_{\cdot,j}), \sigma(y_c(\cdot,t_j)))$$

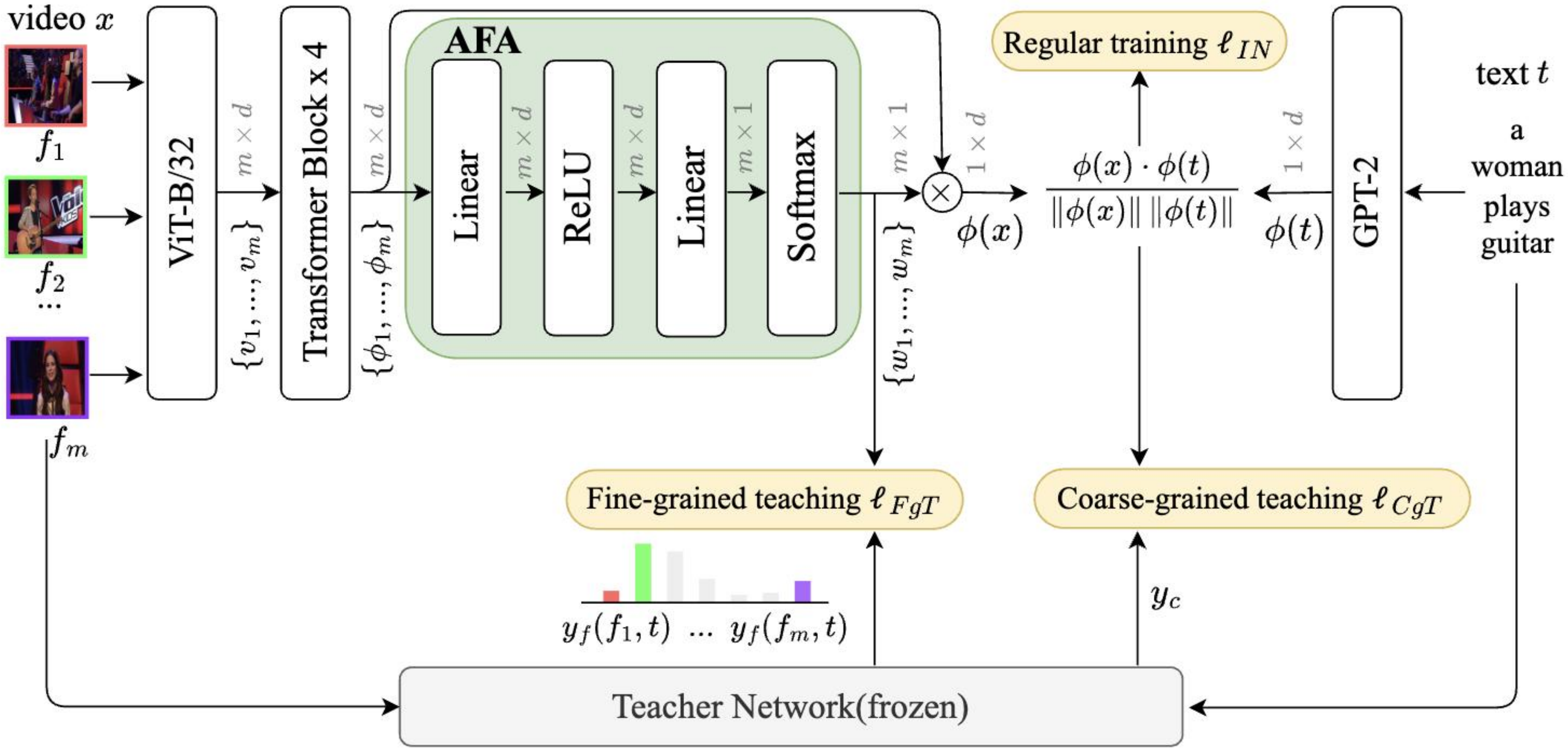


Method

➢ Multi-grained teaching

  ➢ Fine-grained teaching

$$\ell_{FgT} := -\frac{1}{b}\sum_{i=1}^{b}\sum_{k=1}^{m} y_f(f_{i,k}, t_i)\log w_{i,k}.$$



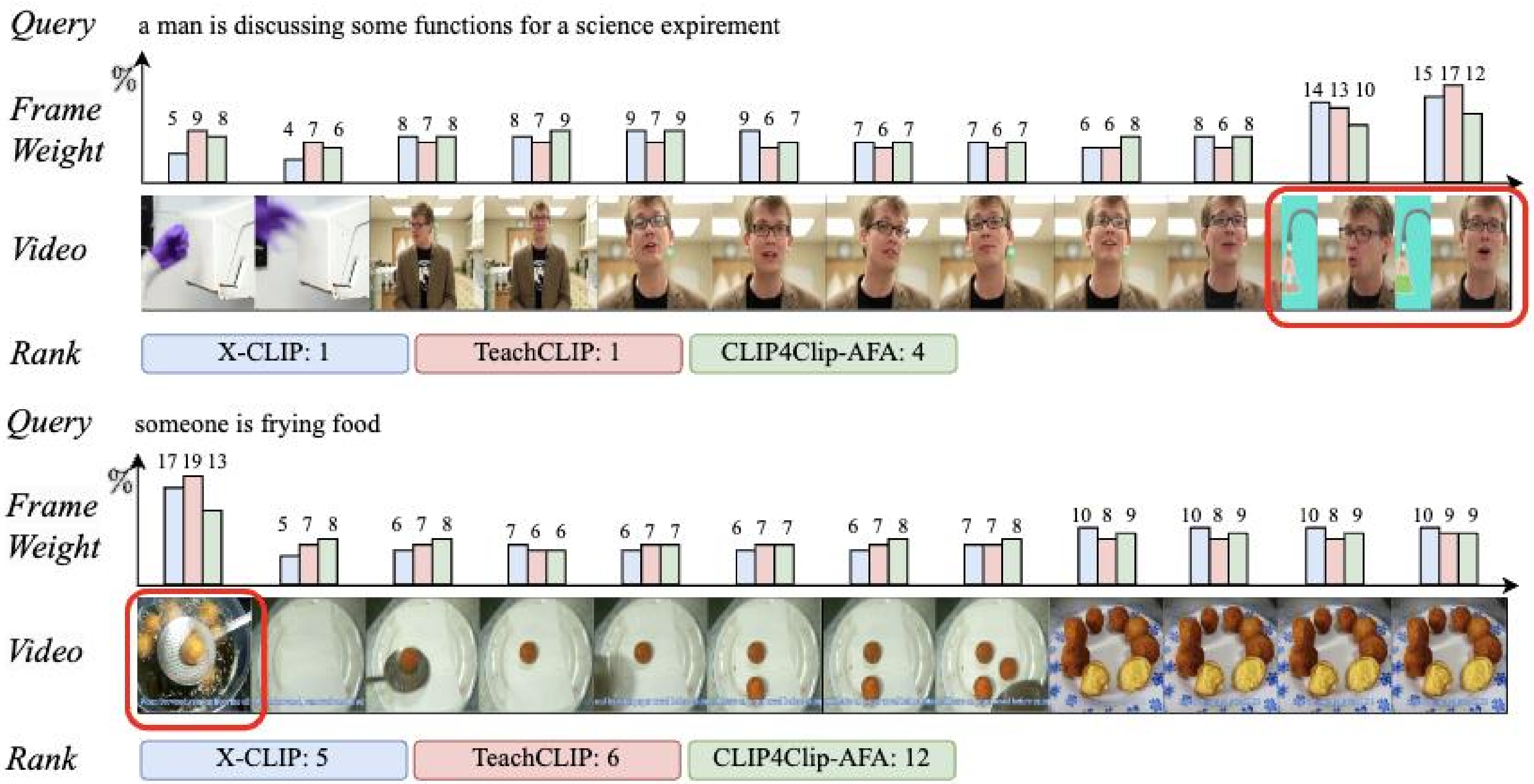
Results

➤ TeachCLIP has the same efficiency as CLIP4Clip, yet has near-SOTA effectiveness.

Model	MSRVTT-1k			MSRVTT-3k			MSVD			VATEX			ActNetCap			DiDeMo			Mean
	R1	R5	SumR	R1	R5	SumR	R1	R5	SumR	R1	R5	SumR	R1	R5	SumR	R1	R5	SumR	
Feature re-learning w/o CLIP feature:																			
W2VV++ [18]	18.9	45.3	121.7	11.1	29.6	81.2	22.4	51.6	138.8	–	–	–	–	–	–	–	–	–	–
DualE [6]	21.1	48.7	130.0	11.6	30.3	83.2	–	–	–	36.8	73.6	194.1	–	–	–	–	–	–	–
CE [21]	20.9	48.8	132.1	10.0	29.0	80.2	19.8	49.0	132.6	–	–	–	17.7	46.6	–	–	–	–	–
SEA [19]	23.8	50.3	137.9	13.1	33.4	91.5	24.6	55.0	147.5	–	–	–	–	–	–	–	–	–	–
MMT [8]	24.6	54.0	145.7	–	–	–	–	–	–	–	–	–	22.7	54.2	–	–	–	–	–
TeachText [3]	29.6	61.6	165.4	15.0	38.5	105.2	25.4	56.9	153.6	53.2	87.4	233.9	23.5	57.2	–	–	–	–	–
Feature re-learning with CLIP feature:																			
SEA	37.2	67.1	182.6	19.9	44.3	120.7	34.5	68.8	183.8	52.4	90.2	238.5	–	–	–	–	–	–	–
W2VV++	39.4	68.1	185.6	23.0	49.0	132.7	37.8	71.0	190.4	55.8	91.2	243.0	–	–	–	–	–	–	–
MMT	39.5	68.3	186.1	24.9	50.5	137.4	40.6	72.0	194.3	54.4	89.2	238.6	–	–	–	–	–	–	–
LAFF [13]	45.8	71.5	199.3	29.1	54.9	149.8	45.4	70.6	200.6	59.1	91.7	247.1	–	–	–	–	–	–	–
CLIP-based end-to-end (visual backbone: ViT-B/32):																			
CenterCLIP [38]	44.2	71.6	197.9	–	–	–	47.3	76.8	209.7	–	–	–	43.9	**74.6**	**204.3**	–	–	–	–
CLIP4Clip [24]	42.8	71.6	195.5	29.4	54.9	150.1	45.6	76.1	206.6	61.6	91.1	248.5	39.7	71.0	194.1	42.0	69.0	189.2	197.3
TS2-Net [22]	46.7	72.6	200.5	29.9	56.4	153.6	44.6	75.8	204.9	61.1	91.5	248.6	37.3	69.9	190.4	40.2	69.4	188.4	197.7
X-CLIP [25]	45.3	73.7	200.8	**31.2**	**57.4**	**156.7**	47.2	77.0	210.1	62.2	90.9	248.5	**44.4**	**74.6**	204.1	45.0	73.1	200.2	**203.4**
X-Pool [9]	46.0	72.8	201.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
DRL [32]	46.2	74.0	203.2	–	–	–	–	–	–	–	–	–	–	–	–	**47.9**	**73.8**	**204.4**	–
PromptSwitch [4]	43.6	71.5	195.7	–	–	–	46.3	75.8	206.6	–	–	–	–	–	–	–	–	–	-
CLIP-ViP [36]	46.5	72.1	201.1	–	–	–	–	–	–	–	–	–	–	–	–	40.6	70.4	190.3	–
STAN [20]	**46.9**	72.8	202.5	–	–	–	–	–	–	–	–	–	–	–	–	46.5	71.5	198.9	–
Cap4Video [34]	*47.8*	*73.8*	*204.3*	–	–	–	–	–	–	–	–	–	–	–	–	*52.0*	*79.4*	*218.9*	–
*CLIP-ViP**	*50.1*	*74.8*	*209.5*	–	–	–	–	–	–	–	–	–	–	–	–	*48.6*	*77.1*	*210.1*	–
UMT [17]	*51.0*	*76.5*	*211.7*	–	–	–	*71.9*	*94.5*	*264.2*	–	–	–	*58.3*	*83.9*	*233.7*	*61.6*	*86.8*	*239.9*	–
TeachCLIP	46.8	**74.3**	**203.7**	30.9	57.1	156.0	**47.4**	**77.3**	**210.2**	**63.6**	**91.9**	**251.6**	42.2	72.7	200.1	43.7	71.2	196.0	202.9

Results

- The weights by TeachCLIP are closer to the query-dependent weights by the teacher, especially on salient frames (manually marked out by red rectangles).



Conclusion

- Main contribution
  - We propose TeachCLIP, letting a CLIP4Clip based network learn from more advanced yet computationally heavy T2VR models.
  - We propose Attentional frame-Feature Aggregation (AFA) to convey fine-grained cross-modal knowledge ,which introduces no extra storage / computation overhead at the retrieval stage.

- Codes:
  - https://github.com/ruc-aimc-lab/TeachCLIP

- Contact:
  - ruixiangzhao@ruc.edu.cn