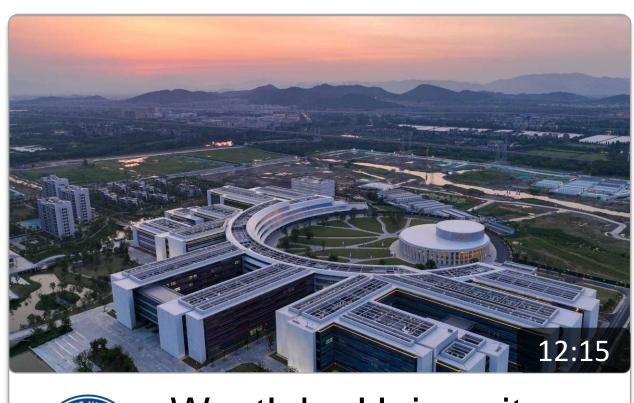
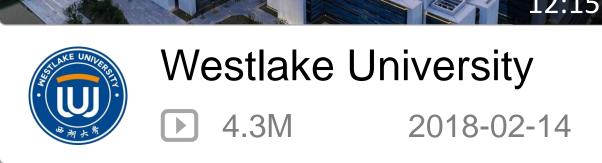
VoP: Text-Video Co-operative Prompt Tuning for Cross-Modal Retrieval





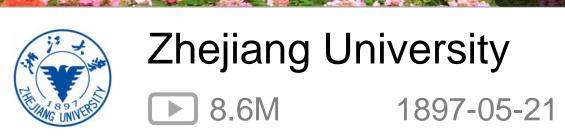






2017-10-11





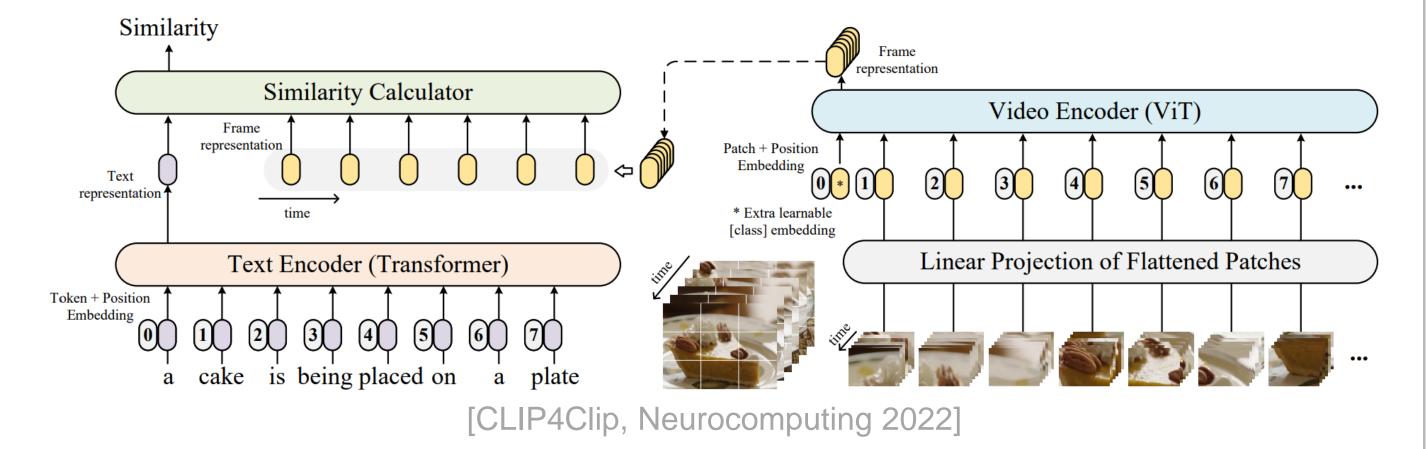


YouTube Silis Silis TikTok

Text-video retrieval is important for platforms to find relevant videos efficiently!

Introduction

Leveraging the *pre-trained CLIP* for text-video cross-modal retrieval task recently popular.



However, the dominant full fine-tuning strategy brings...

- risk of overfitting: inevitably forgetting the useful knowledge acquired in the large-scale pretraining phase.
- severe storage burdens: maintaining an independent model weight for every dataset during deployment; infeasible due to the increasing model capacity.

For both effectiveness and efficiency, we continue the vein of prompt learning and propose ...

- a strong baseline VoP that effectively adapts CLIP to text-video retrieval with only **0.1%** parameter storage.
- three video-specific prompts respectively conditioned on the frame position, frame context, and layer function, delivering an average R@1 improvement of up to 4.2% for VoP, and therefore exceed full fine-tuning by up to 1.4% with much fewer trainable parameters.

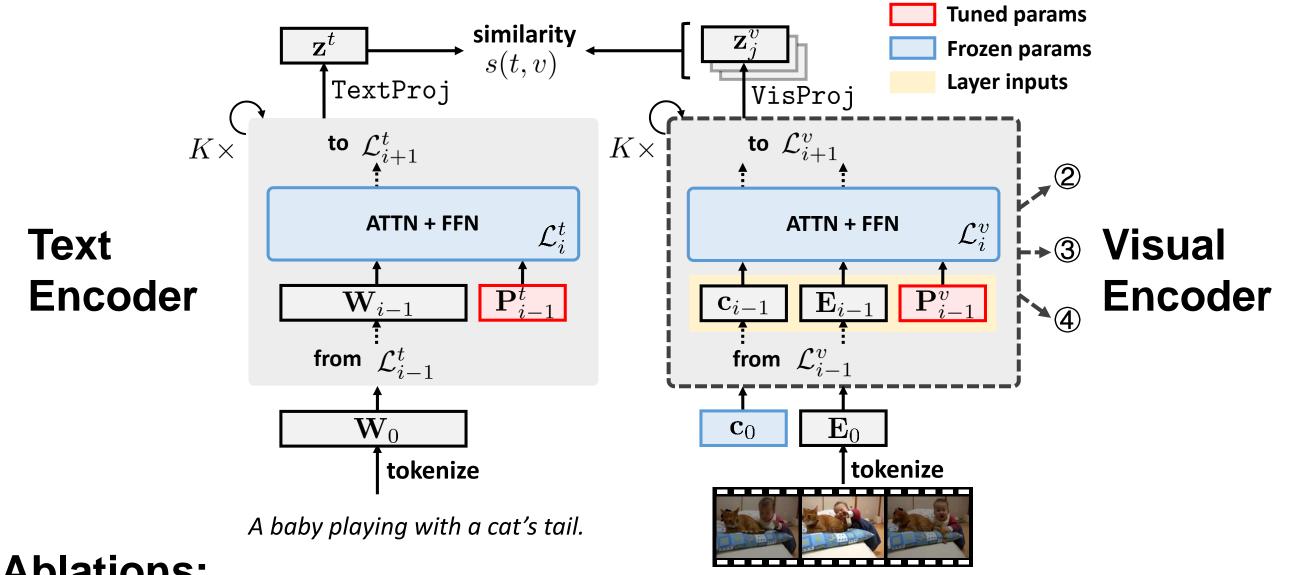
Our Proposed Framework

Baseline: (1) VoP (Text-Video Co-operative Prompt Tuning)

Motivations:

- Learning prompts only for the text branch overlooks the potential of collaboratively tuning the visual encoder.
- Prompting the mere input layer has only a relatively indirect impact on the output embeddings.

Solution: Tuning the prompts introduced in **all** layers of **both** unimodal encoders while keeping the rest of the model frozen.



Ablations:

Textual	Visual	R@1		R@10	MnR↓	MdR↓	39.5
		31.5	52.8	63.6	42.9	5.0	§ 39.0 38.8 ^{39,0}
	✓	36.5	62.7	75.1	18.3	3.0	日 38.5 38.0 37.8
✓		36.3	63.4	75.0	20.3	3.0	38.0 37.8
✓	✓	39.6	66.7	63.6 75.1 75.0 77.8	17.2	2.0	37.5
		•			•		1 1-3 1-6 1-9 1-12 Prompt depth

Equipping with Three Plug-and-Play Video Prompts

Motivation: Assisting VoP in utilizing rich temporal information.

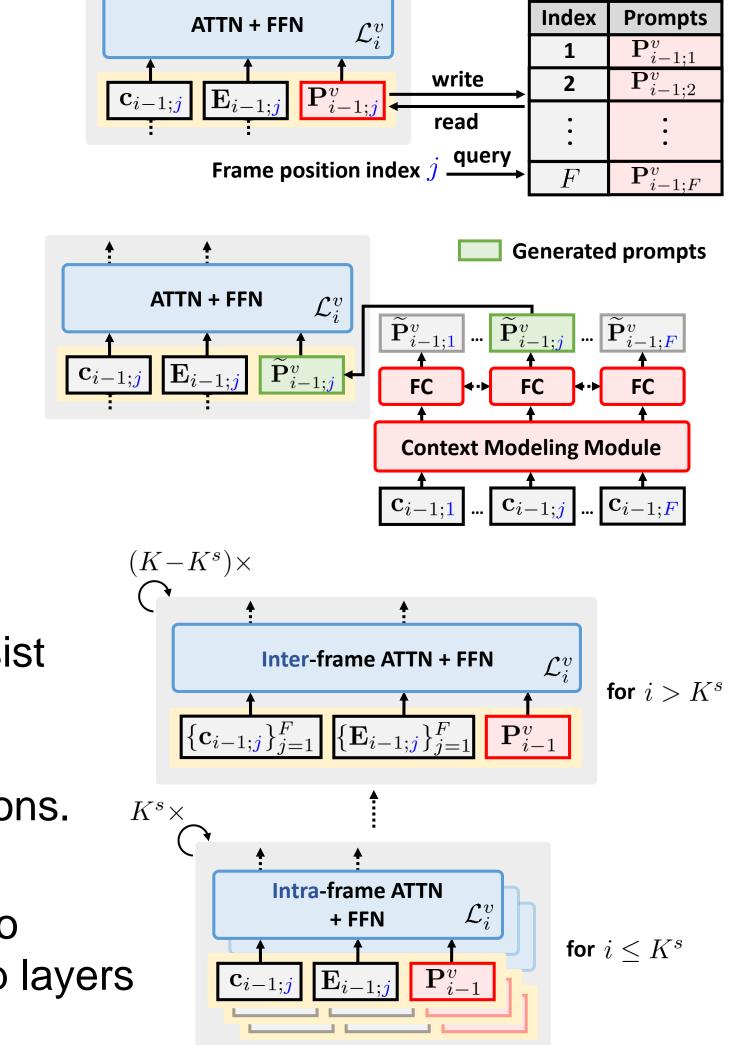
2 VoPP: position-specific video prompts model the information shared between frames at the same relative position.

3 VoPc: generated contextspecific video prompts integrate injected contextual message from the frame sequence into the intraframe modeling.

4 VoP^F: function-specific video prompts adaptively assist

to learn intra- or inter-frame affinities by sensing the transformation of layer functions.

VoP^F can be combined with position/context-specific video prompts by deploying in deep layers (VoPF+P / VoPF+C).



Experiments

Main Results (CLIP ViT-R/32)

Main R			•				-	t2i	,					v2t		
Methods		Params (M)		R@	R@1 R@5		R@10	M	nR↓	$MdR\downarrow$	R@1	R@5	R@10	MnR↓	$MdR \downarrow$	
Full		11	119.8 (100%)		41	.7	69.2	79.0	1	6.5	2.0	42.5	70.9	81.4	11.0	2.0
Bias [6]		0.	0.1 (0.104%)		39	39.7 66.5		77.3	1	7.3	2.0	41.1	68.4	79.2	13.6	2.0
Proj [17]		0.	0.7 (0.547%)		37	37.1 63.0		76.1	2	20.5 3.0		37.2	64.6	75.9	16.7	3.0
Partial [17]			7.7 (6.410%)		39	39.8 65.3		75.9	1	9.3	2.0	37.9	66.1	77.4	15.5	3.0
Adapter ^{ATTN} [12]			2.0 (1.655%)		37	37.6		75.8	1	8.7	3.0	39.6	66.5	76.8	14.7	2.0
Adapter ^{FFN} [7]		2.	2.0 (1.655%)		38	.2	63.5	76.4	1	7.9	3.0	39.9	66.8	77.7	14.2	2.0
VoP			0.1 (0.103%)		39	39.6		77.8	1	7.2	2.0	42.1	68.8	80.7	12.4	2.0
VoPP		0.5 (0.441%)			40.1 65		77.7		6.9	2.0	42.5	70.0	79.9	12.4	2.0	
VoPC		14.3 (11.898%)			40.8 68.1		79.0	_	<u>15.8</u> 2.0		42.3	70.1	81.1	<u>11.4</u>	2.0	
VoPF		0.1 (0.103%)		42	42.6 68.4		78.7		<u>15.8</u> 2.0		42.4	70.5	81.0	11.0	2.0	
VoPF+P		0.4 (0.328%)		43	43.5		<u>79.3</u>	1	4.8	2.0	43.6	71.2	81.2	11.0	2.0	
VoP^{F+C}		14.1 (11.785%)		44	.6	69.9	80.3	16.3		2.0	44.5	70.7	80.6	11.5	2.0	
									Οι	urs				+ N 1 C	ים או	TT OI
													- 4	IIVIS	K-V	TT-9l
MSR-VTT-9k -	0.0	-2.0	-4.6	-1.9	-4.1	-3.5	-2.1	-1.6	-0.9	0.9	1.8	2.9	4			
	0.0	1.0	4.0	1.7	1.0	1.0	1.0	0.2	0.0	1 1	2.6	1.0	2			
MSR-VTT-7k -	0.0	-1.2	-4.9	-1./	-1.3	-1.0	-1.2	-0.3	-0.9	1.1	2.6	1.8	2			
	0.0	- 1	C 0	2.2	- 2		2.4	2.7	1.6	2.1	2.7	4.0				
DiDeMo -	0.0	-5.1	-6.0	-2.3	-5.2	-5.3	-3.4	-2.7	-1.6	3.1	3.7	4.8	- 0			
	0.0		7.0	2.2	F 2	- O	4.5	4.0	4.2	2.2	0.7	1 7	_			
ActivityNet -	0.0	-5.5	-7.0	-3.2	-5.2	-5.0	-4.5	-4.0	-4.2	-2.2	-0.7	1./	-2			
	0.0	4.6	C 2	4.0	2.6		2.0	2.0	1.0	1 4	1.2	0.0				
LSDMC -	0.0	-4.6	-4.6 -6.3	-4.0	-3.6	-3.3	-3.0	-2.8	-1.0	-1.4	-1.3 -	J. 9	-4	←t2v relative		
A	0.0	2 7	50	2.6	2.0	2.6	2.0	-2.3	1 0	0.3	1.2	1.4		* • • • • • • •	10 010	الم
Average -		-3.7 -5.8 -2.6 -3.9 -3.6											- –6	results on all		
	أارع	135	o(o)	rial,	TIN	FFN	10P	LOPP	PC'	LOPF'	ar [*] P	<u> -</u> دٰ	_	datas	sets	
	()	0,0	k, bs	arcia pter	dapte	Š.C.	1 ° ,	10. 10	J, /	10. 10	DE+b, NOBE			aatat		
	_			_												
Qualita	tiv	e R	est	ılts	(co	lor:	inco	orrec	t or	gro	ound	truth)				
												,				
The water sa bike to save	Contraction Contract				Section Checken	rices a	na wate	1		Soi	meone loc	king at a J	apanese bo	ook.		
	# W				- H.			415								
				GESTA FIRE DE				Ų0	_			CARDON		No. 19		Marita
tial		17		E		0			F	ull	200	S. Sec.		No.		Representation
Mail .						OI					16	Ship)	J. J.	10 12	200	THE TOTAL
Full			SARA .		1				Part	ial	A BA	POAR	A POPULAR	A COL	A M	TOKYO Alek
OP A) A	-		10 1			oP /	A SEE		Star Control		(25)	イリスト 後) 158cm
PF+C		9	2/18:2	oral to					VoP	F+C	WAL	A		1	100000	カップを約7万 よみかをベルト いを対象スト プスはオリエ
															'-A	-191
	and the second second							1								
A man looks priest in a co			cathedra	al's orgai	n pipes	and ta	lks to a	1		At	aby playir	ng with a c	at's tail.			

Authors:

Siteng Huang^{1,3}, Biao Gong², Yulin Pan², Jianwen Jiang², Yiliang Lv², Yuyuan Li³, Donglin Wang¹ **Organizations:**

- 1 Machine Intelligence Laboratory (MiLAB), Westlake University
- 2 Alibaba Group
- 3 Zhejiang University











Youtube **Github**