JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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

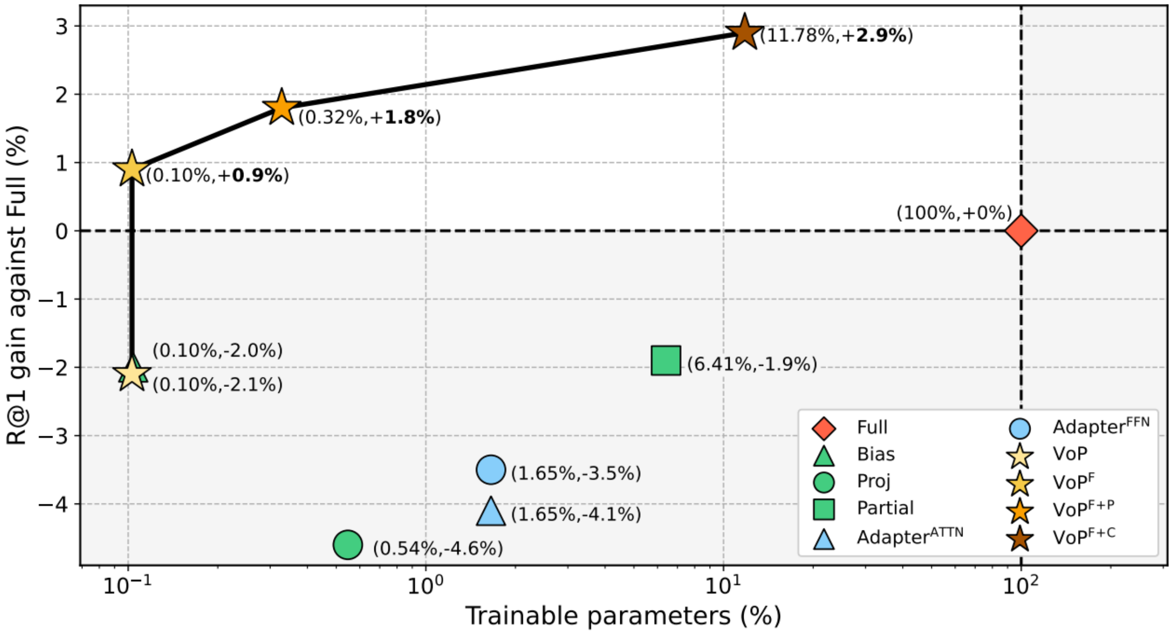
Siteng Huang[1,3], Biao Gong[2], Yulin Pan[2], Jianwen Jiang[2], Yiliang Lv[2], Yuyuan Li[3], Donglin Wang[1]✉

[1]Machine Intelligence Lab (MiLAB), Westlake University [2]Alibaba Group [3]Zhejiang University

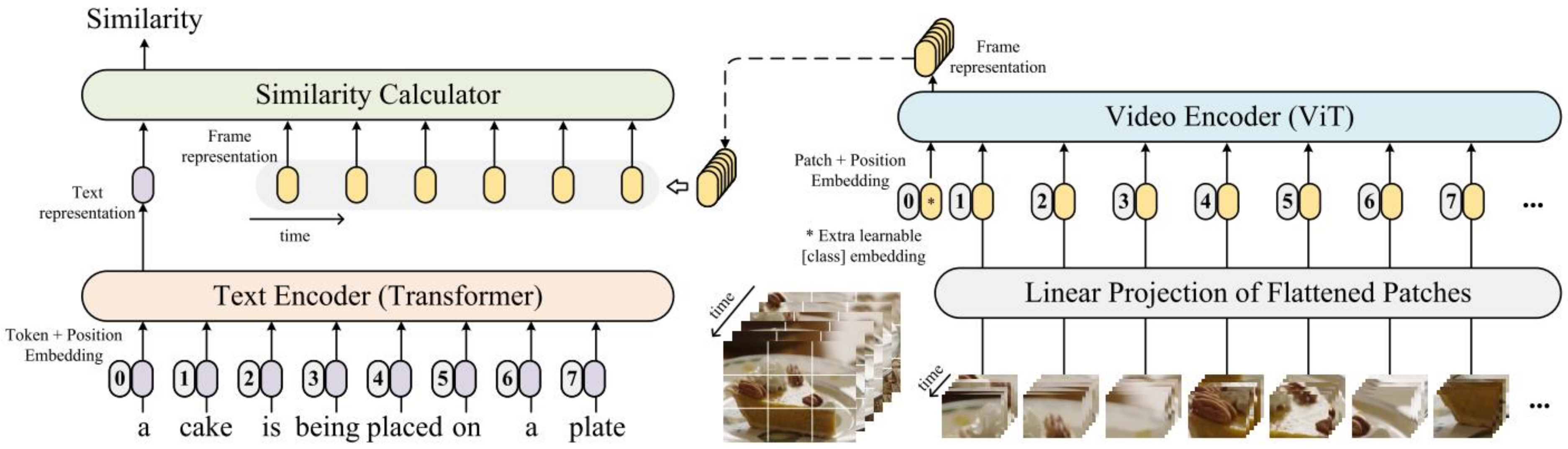
TUE-PM-233

Summary of Highlights

- **VoP**, a powerful parameter-efficient fine-tuning baseline for text-video retrieval with only **0.1%** trainable parameters.
- **Three novel video prompts**, improving VoP by excavating temporal information in a plug-and-play manner.
- Exceeding full fine-tuning by up to 2.9% with 6× less parameter overhead (R@1 on MSR-VTT-9k).



Background



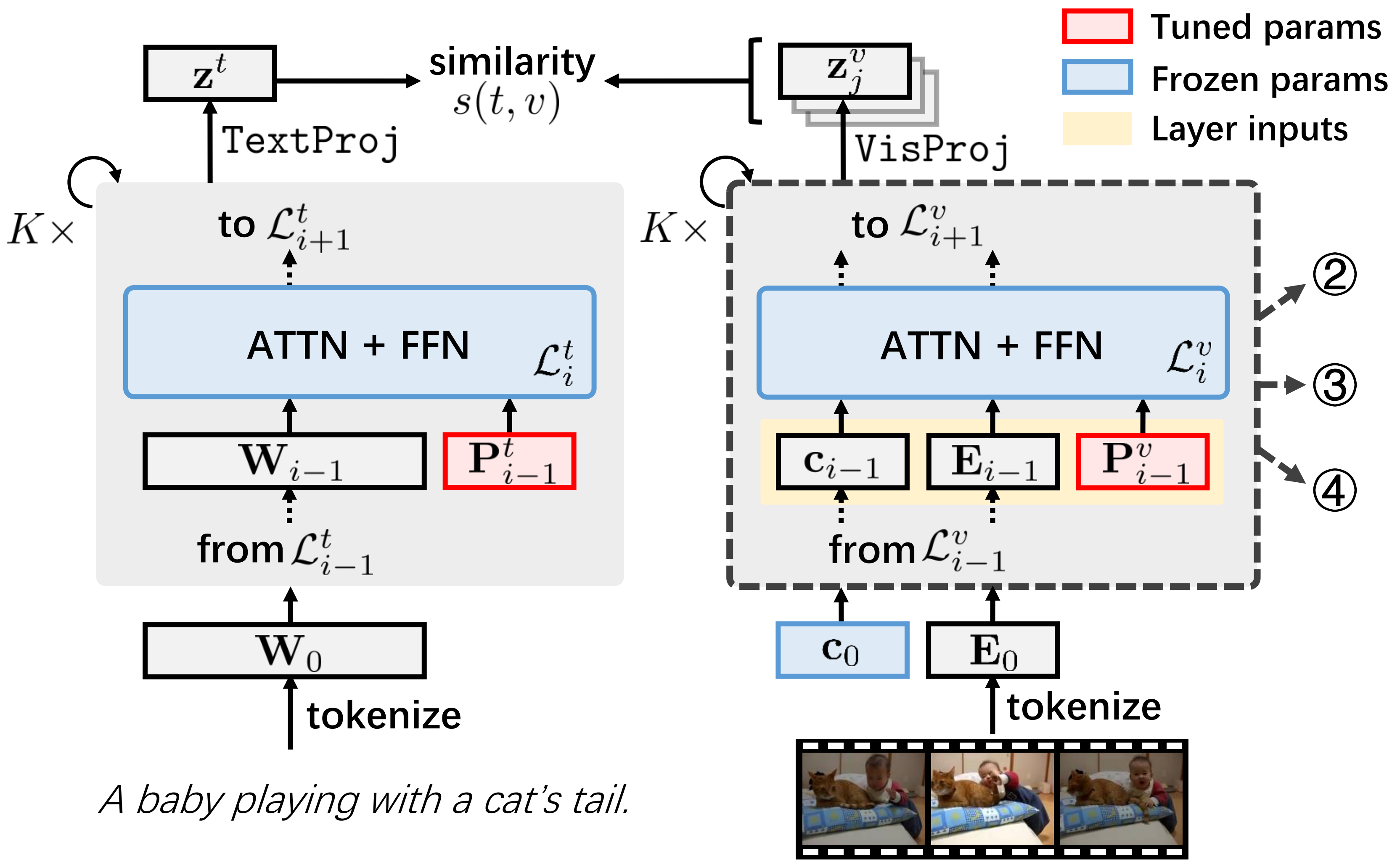
Adapting CLIP to video domains via full fine-tuning [1]

Challenges of full fine-tuning:

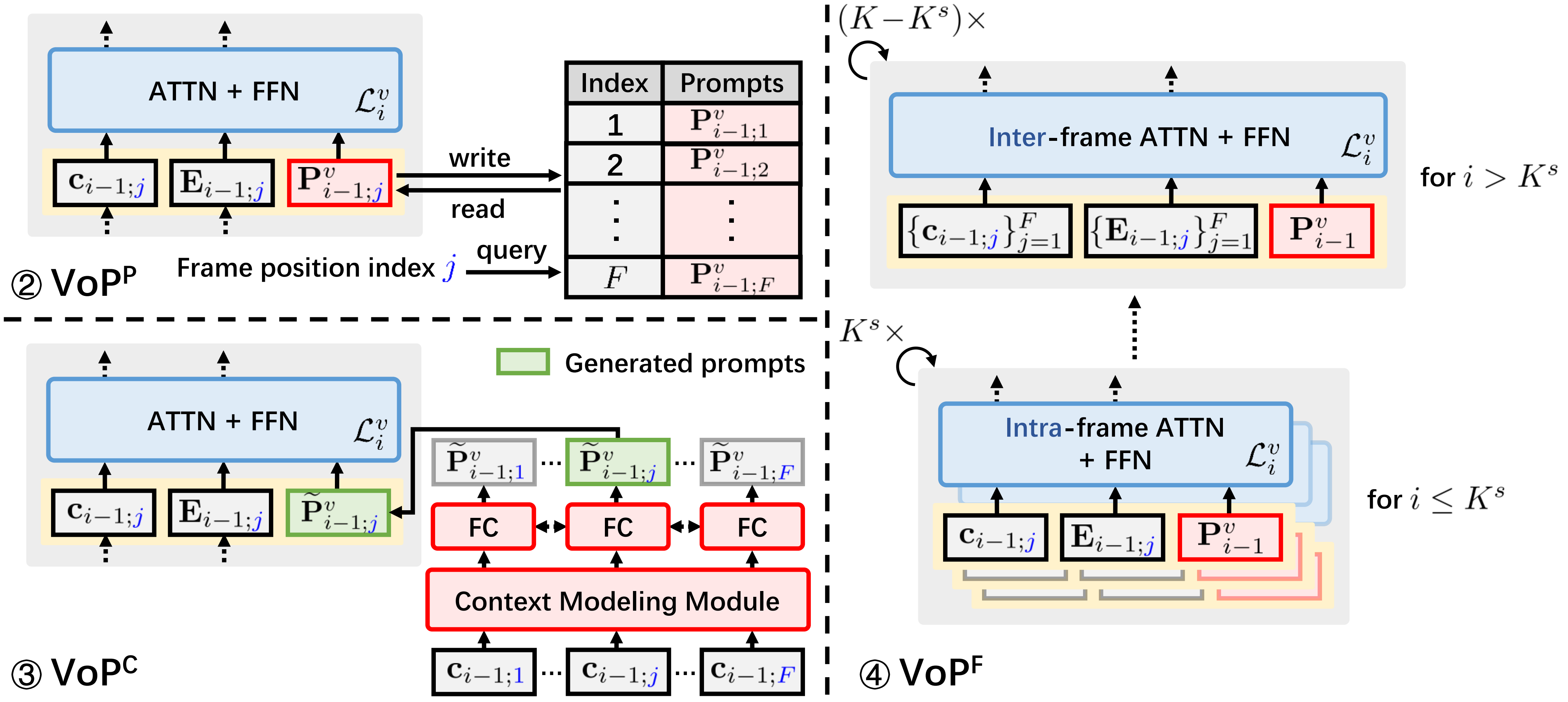
- Risk of overfitting.
- Unaffordable storage overhead.

[1] Huaishao Luo, et al., CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, Neurocomputing 2022

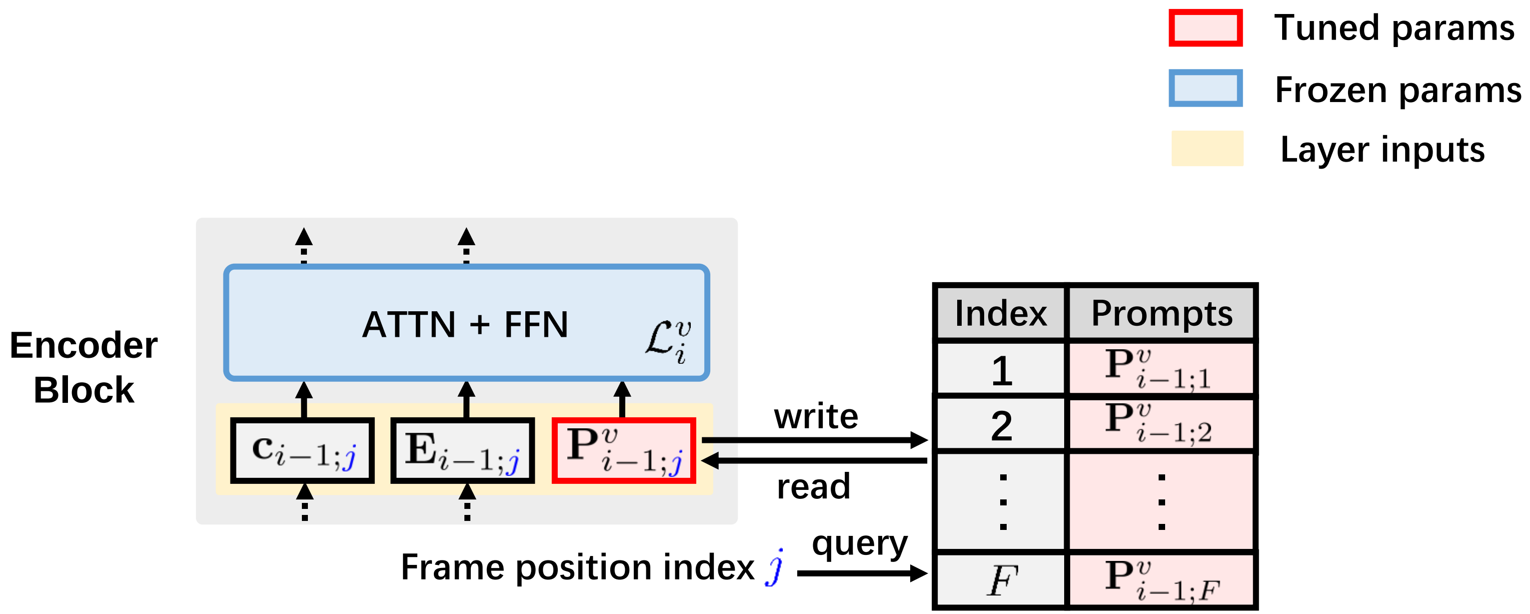
Text-Video Co-operative Prompt Tuning (VoP)



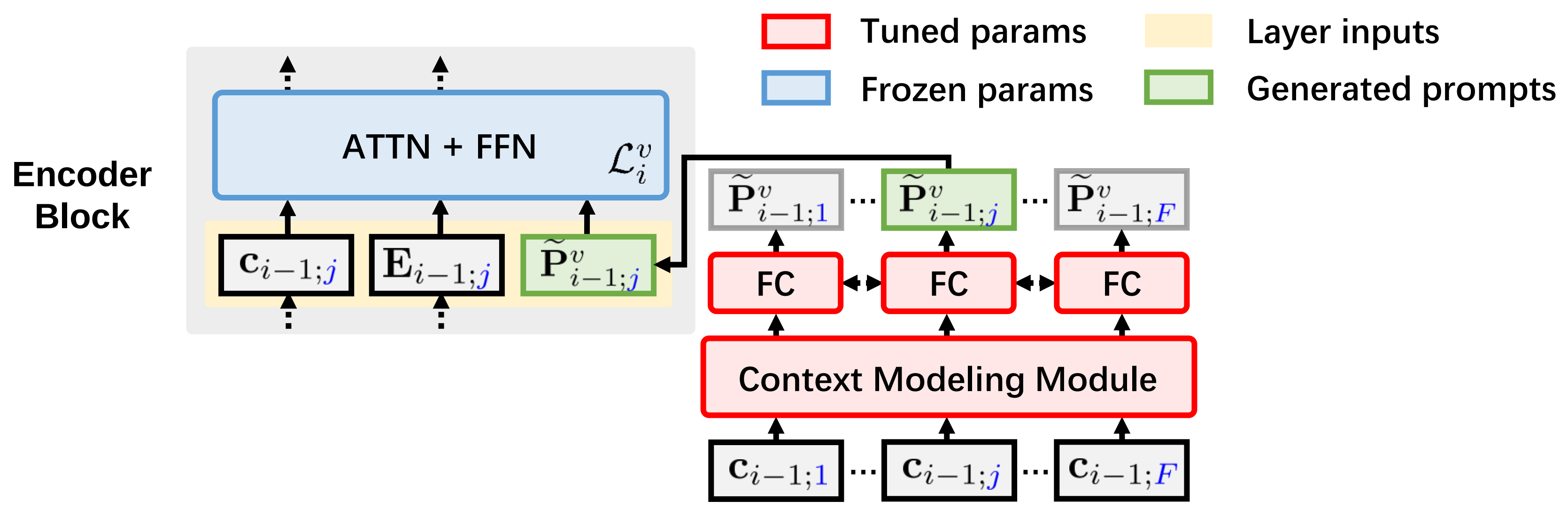
VoP with Video-specific Prompts



Position-specific Video Prompts

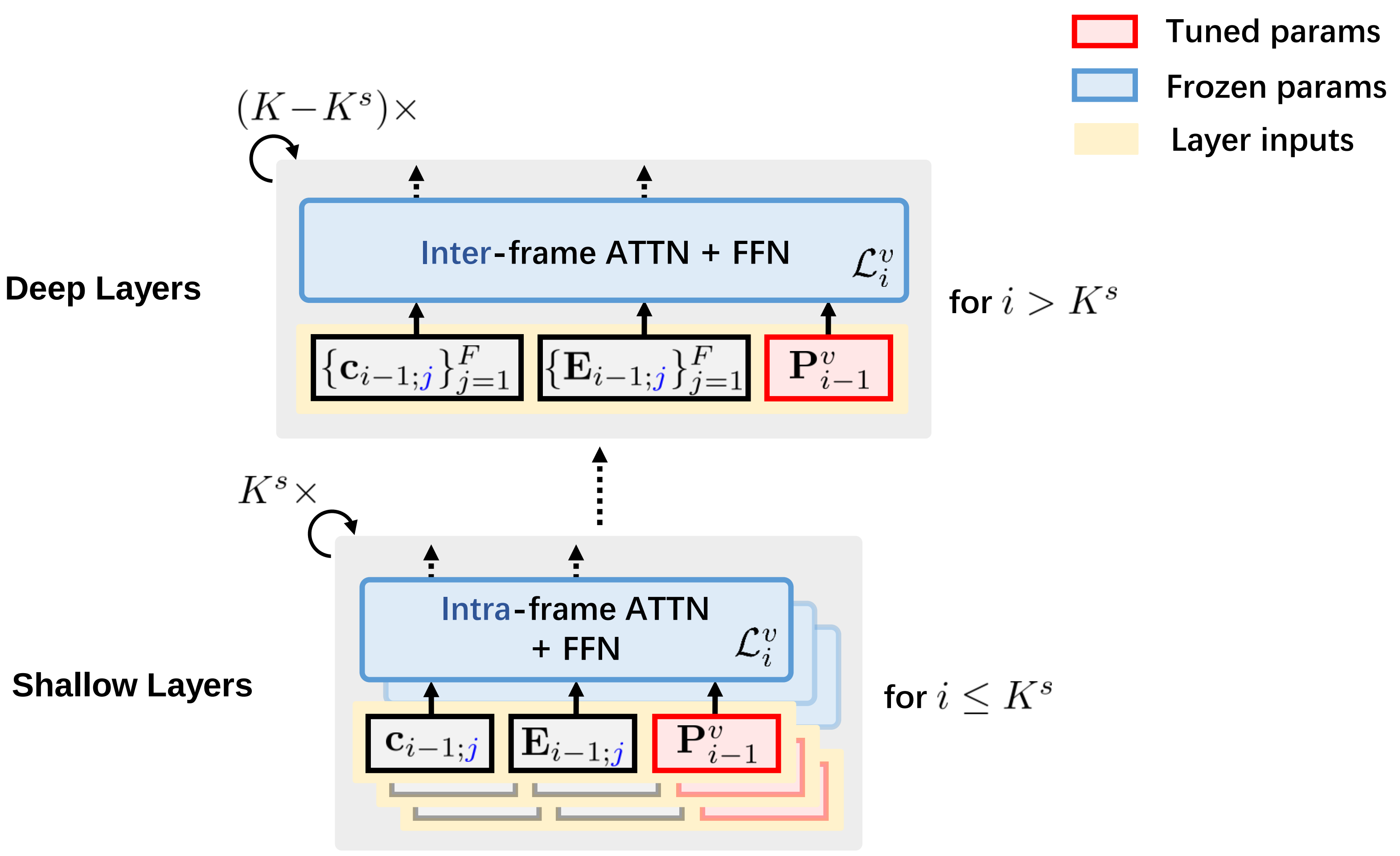


Context-specific Video Prompts

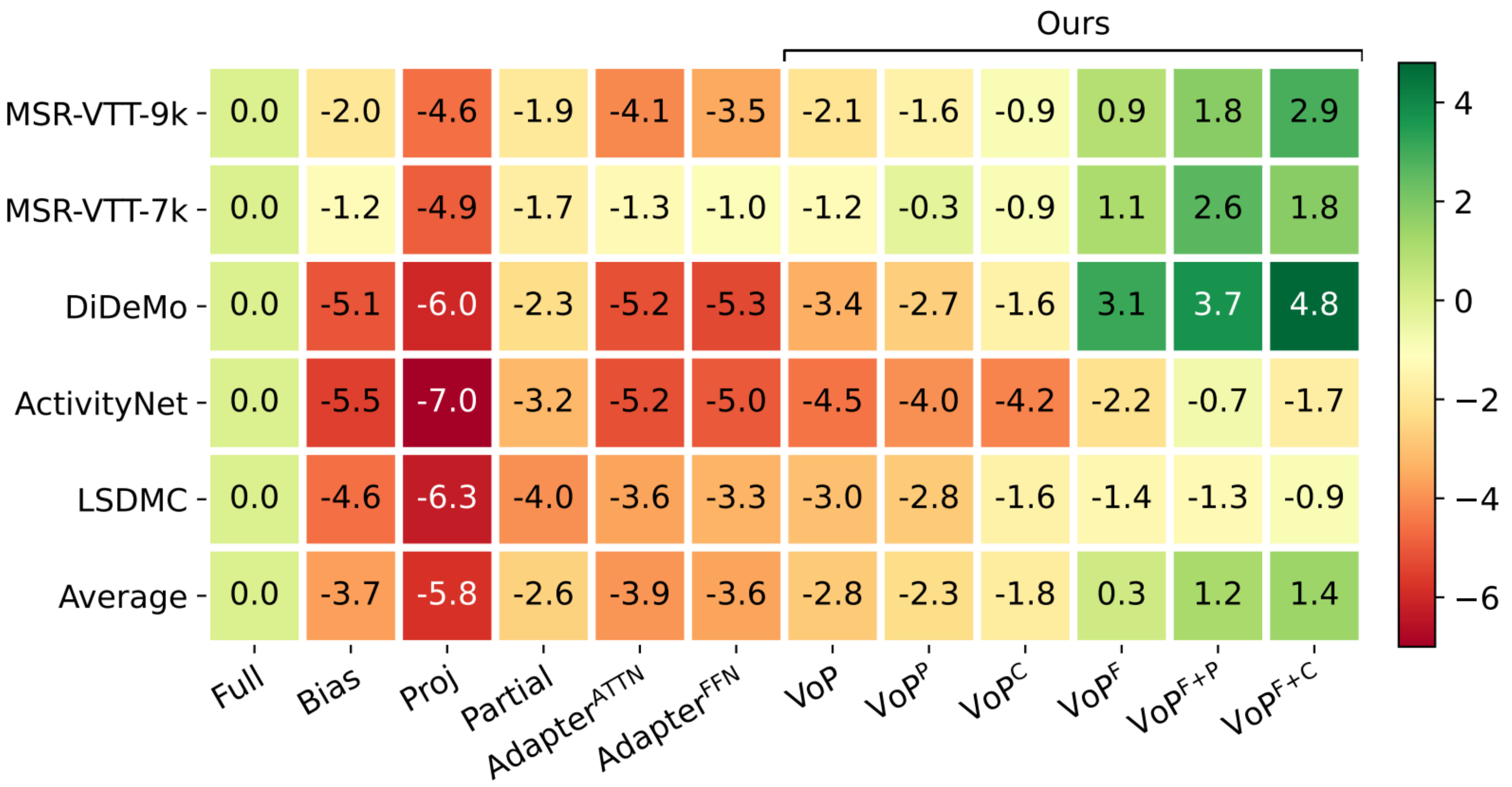


Choice of CMM	MSR-VTT-9k			MSR-VTT-7k			DiDeMo			ActivityNet			LSMDC		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Transformer	40.1	68.2	78.8	39.5	68.2	78.1	**40.4**	67.3	77.3	32.0	61.5	74.9	20.3	39.5	47.8
LSTM	40.6	**69.5**	**79.7**	39.5	**69.3**	78.0	38.6	66.7	77.0	32.4	62.0	75.4	19.6	38.2	47.7
BiLSTM	**40.8**	68.1	79.0	**40.0**	67.3	**78.2**	40.0	**68.0**	**78.5**	**32.6**	**62.5**	**76.5**	**20.4**	**40.0**	**48.1**

Function-specific Video Prompts

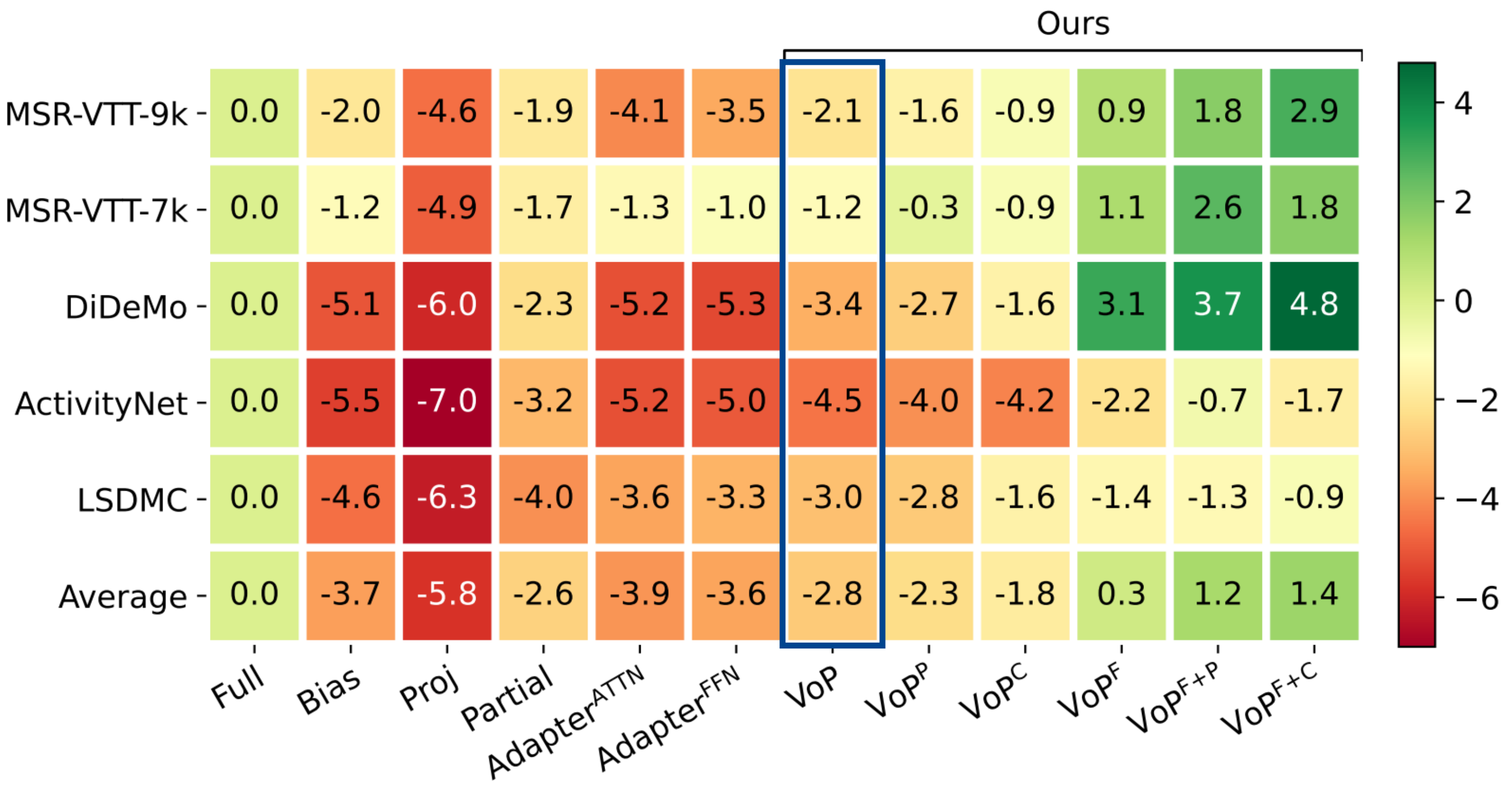


Main Results



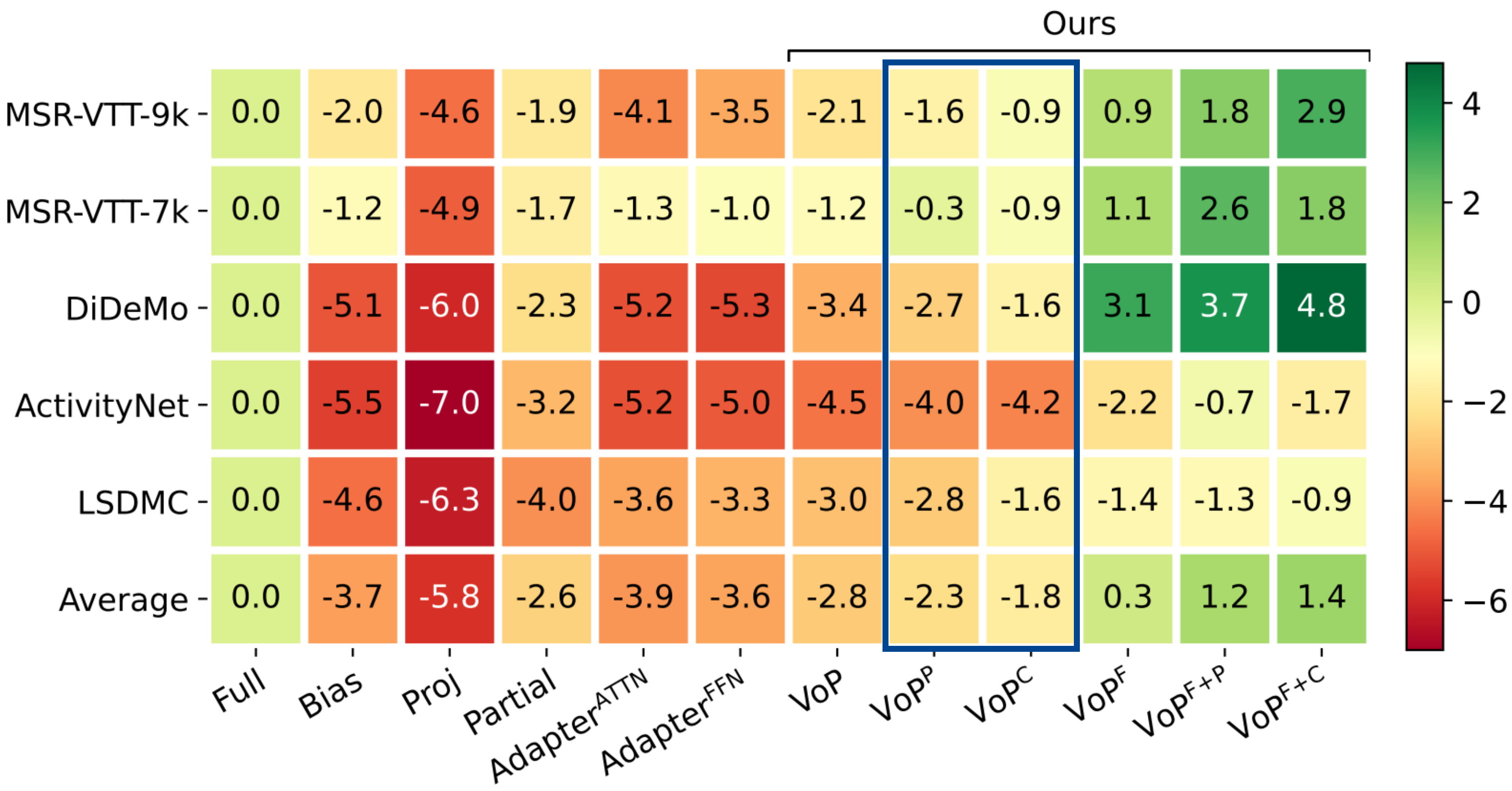
Text-to-video R@1 gains of all methods in comparison against full fine-tuning

Main Results



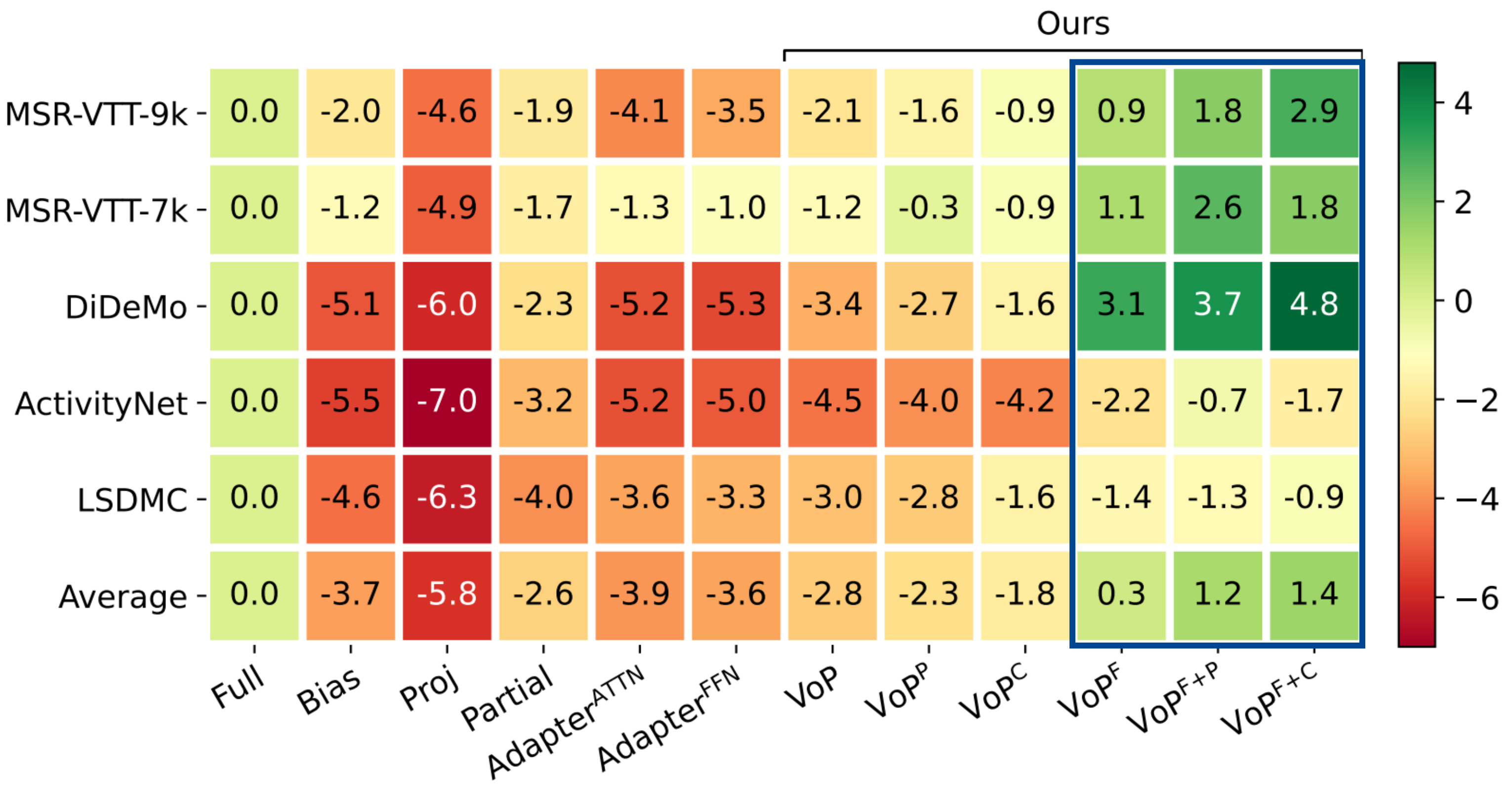
Text-to-video R@1 gains of all methods in comparison against full fine-tuning

Main Results



Text-to-video R@1 gains of all methods in comparison against full fine-tuning

Main Results



Text-to-video R@1 gains of all methods in comparison against full fine-tuning

Main Results: MSR-VTT-9k

Methods	Params (M)	t2v R@1	t2v R@5	t2v R@10	t2v MnR↓	t2v MdR↓	v2t R@1	v2t R@5	v2t R@10	v2t MnR↓	v2t MdR↓
Full	119.8 (100%)	41.7	69.2	79.0	16.5	2.0	42.5	<u>70.9</u>	**81.4**	**11.0**	2.0
Bias [6]	0.1 (0.104%)	39.7	66.5	77.3	17.3	2.0	41.1	68.4	79.2	13.6	2.0
Proj [17]	0.7 (0.547%)	37.1	63.0	76.1	20.5	3.0	37.2	64.6	75.9	16.7	3.0
Partial [17]	7.7 (6.410%)	39.8	65.3	75.9	19.3	2.0	37.9	66.1	77.4	15.5	3.0
Adapter^{ATTN} [12]	2.0 (1.655%)	37.6	63.2	75.8	18.7	3.0	39.6	66.5	76.8	14.7	2.0
Adapter^{FFN} [7]	2.0 (1.655%)	38.2	63.5	76.4	17.9	3.0	39.9	66.8	77.7	14.2	2.0
VoP	0.1 (0.103%)	39.6	66.7	77.8	17.2	2.0	42.1	68.8	80.7	12.4	2.0
VoP^{P}	0.5 (0.441%)	40.1	65.7	77.7	16.9	2.0	42.5	70.0	79.9	12.4	2.0
VoP^{C}	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
VoP^{F}	0.1 (0.103%)	42.6	68.4	78.7	<u>15.8</u>	2.0	42.4	70.5	81.0	**11.0**	2.0
VoP^{F+P}	0.4 (0.328%)	<u>43.5</u>	<u>69.3</u>	<u>79.3</u>	**14.8**	2.0	<u>43.6</u>	**71.2**	<u>81.2</u>	**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

Main Results: MSR-VTT-7k

Methods	Params (M)	*t2v* R@1	*t2v* R@5	*t2v* R@10	*t2v* MnR↓	*t2v* MdR↓	*v2t* R@1	*v2t* R@5	*v2t* R@10	*v2t* MnR↓	*v2t* MdR↓
Full	119.8 (100%)	40.9	67.9	78.4	18.3	2.0	41.7	69.6	79.7	12.7	2.0
Bias [6]	0.1 (0.104%)	39.7	65.9	76.7	17.9	2.0	41.2	66.6	78.9	14.0	2.0
Proj [17]	0.7 (0.547%)	36.0	63.6	74.6	21.4	3.0	36.9	63.6	74.6	17.8	3.0
Partial [17]	7.7 (6.410%)	39.2	64.0	74.7	20.9	3.0	37.7	63.6	74.9	16.9	3.0
Adapter$^{\text{ATTN}}$ [12]	2.0 (1.655%)	39.6	65.4	76.8	16.8	2.0	41.6	67.6	79.8	12.4	2.0
Adapter$^{\text{FFN}}$ [7]	2.0 (1.655%)	39.9	65.3	76.9	16.8	2.0	41.6	67.6	79.2	12.7	2.0
VoP	0.1 (0.103%)	39.7	66.7	77.9	16.7	2.0	41.4	68.8	**80.8**	12.5	2.0
VoP$^{\text{P}}$	0.5 (0.441%)	40.6	66.0	76.7	16.6	2.0	41.6	69.0	79.5	12.3	2.0
VoP$^{\text{C}}$	14.3 (11.898%)	40.0	67.3	78.2	17.0	2.0	41.7	69.4	79.1	12.3	2.0
VoP$^{\text{F}}$	0.1 (0.103%)	42.0	67.4	78.2	16.2	2.0	42.8	68.4	79.8	12.3	2.0
VoP$^{\text{F+P}}$	0.4 (0.328%)	**43.5**	68.1	79.2	16.0	2.0	43.4	**71.0**	80.4	**11.3**	2.0
VoP$^{\text{F+C}}$	14.1 (11.785%)	42.7	**68.2**	**79.3**	**15.9**	2.0	**44.2**	69.6	**80.8**	11.4	2.0

Main Results: DiDeMo

Methods	Params (M)	t2v R@1	t2v R@5	t2v R@10	t2v MnR↓	t2v MdR↓	v2t R@1	v2t R@5	v2t R@10	v2t MnR↓	v2t MdR↓
Full	119.8 (100%)	41.6	68.4	78.2	17.7	2.0	40.2	68.4	78.7	11.9	2.0
Bias [6]	0.1 (0.104%)	36.5	63.4	75.2	24.8	3.0	36.8	65.7	75.8	15.1	2.0
Proj [17]	0.7 (0.547%)	35.6	61.3	72.6	24.4	3.0	34.5	60.9	72.6	18.8	3.0
Partial [17]	7.7 (6.410%)	39.3	65.5	75.7	22.3	2.0	36.9	64.2	74.5	17.0	2.0
$\text{Adapter}^{\text{ATTN}}$ [12]	2.0 (1.655%)	36.4	62.8	73.9	23.5	3.0	36.3	64.4	74.8	15.4	2.0
$\text{Adapter}^{\text{FFN}}$ [7]	2.0 (1.655%)	36.3	63.4	75.4	22.9	3.0	35.6	64.3	75.6	14.8	3.0
VoP	0.1 (0.103%)	38.2	66.9	76.1	19.8	2.0	38.1	65.7	76.5	13.5	2.0
VoP^{P}	0.5 (0.441%)	38.9	67.7	78.1	17.2	2.0	40.6	68.3	78.6	11.6	2.0
VoP^{C}	14.3 (11.898%)	40.0	68.0	78.5	18.3	2.0	39.1	65.3	76.7	13.8	3.0
VoP^{F}	0.1 (0.103%)	44.7	70.8	79.7	15.7	2.0	43.5	70.9	<u>81.4</u>	<u>9.8</u>	2.0
$\text{VoP}^{\text{F+P}}$	0.4 (0.328%)	<u>45.3</u>	**72.3**	<u>80.4</u>	<u>13.8</u>	2.0	**44.7**	<u>71.2</u>	81.1	9.9	2.0
$\text{VoP}^{\text{F+C}}$	14.1 (11.785%)	**46.4**	<u>71.9</u>	**81.5**	**13.6**	2.0	<u>44.4</u>	**71.8**	**81.8**	**9.5**	2.0

Main Results: ActivityNet

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Methods	Params (M)	t2v R@1	t2v R@5	t2v R@10	t2v MnR↓	t2v MdR↓	v2t R@1	v2t R@5	v2t R@10	v2t MnR↓	v2t MdR↓
Full	119.8 (100%)	**36.8**	**66.9**	**80.1**	**9.3**	3.0	**38.9**	**70.1**	**81.9**	**8.4**	2.0
Bias [6]	0.1 (0.104%)	31.3	60.3	74.2	13.4	3.0	33.7	63.8	77.6	11.4	3.0
Proj [17]	0.7 (0.547%)	29.8	59.1	73.3	14.2	4.0	31.1	60.6	74.6	13.1	3.0
Partial [17]	7.7 (6.410%)	33.6	64.0	77.8	<u>10.6</u>	3.0	33.4	64.6	77.8	10.2	3.0
Adapter$^{\text{ATTN}}$ [12]	2.0 (1.655%)	31.6	60.5	74.4	13.1	3.0	33.3	63.6	77.1	11.3	3.0
Adapter$^{\text{FFN}}$ [7]	2.0 (1.655%)	31.8	61.0	75.0	12.8	3.0	33.6	63.9	77.3	11.1	3.0
VoP	0.1 (0.103%)	32.3	61.9	75.5	12.4	3.0	33.7	64.7	77.2	11.1	3.0
VoP$^{\text{P}}$	0.5 (0.441%)	32.8	62.3	75.4	12.3	3.0	34.8	65.0	78.2	10.7	3.0
VoP$^{\text{C}}$	14.3 (11.898%)	32.6	62.5	76.5	12.0	3.0	34.2	64.8	78.4	10.7	3.0
VoP$^{\text{F}}$	0.1 (0.103%)	34.6	62.6	76.4	11.6	3.0	35.5	65.1	77.4	10.2	3.0
VoP$^{\text{F+P}}$	0.4 (0.328%)	<u>36.1</u>	<u>65.5</u>	<u>78.5</u>	10.9	3.0	<u>36.3</u>	<u>65.9</u>	<u>79.2</u>	<u>10.1</u>	3.0
VoP$^{\text{F+C}}$	14.1 (11.785%)	35.1	63.7	77.6	11.4	3.0	35.6	<u>65.9</u>	77.8	10.4	3.0

Main Results: LSMDC

Methods	Params (M)	t2v R@1	t2v R@5	t2v R@10	t2v MnR↓	t2v MdR↓	v2t R@1	v2t R@5	v2t R@10	v2t MnR↓	v2t MdR↓
Full	119.8 (100%)	**22.0**	39.9	**49.9**	**56.8**	11.0	21.9	40.0	48.2	**50.7**	12.0
Bias [6]	0.1 (0.104%)	17.4	36.2	44.9	73.2	14.0	18.0	36.0	44.9	62.2	15.0
Proj [17]	0.7 (0.547%)	15.7	32.7	40.8	83.7	20.0	17.1	32.6	39.9	76.4	21.0
Partial [17]	7.7 (6.410%)	18.0	33.8	41.8	79.9	18.0	15.9	33.2	41.5	72.3	18.0
Adapter$^{\text{ATTN}}$ [12]	2.0 (1.655%)	18.4	38.0	46.4	68.9	13.0	19.7	37.6	46.3	55.4	13.0
Adapter$^{\text{FFN}}$ [7]	2.0 (1.655%)	18.7	38.9	47.3	63.6	13.0	19.8	38.4	47.0	57.8	12.0
Ju *et al.* [18] †	4.8 (3.990%)	18.8	38.5	47.9	-	12.3	-	-	-	-	-
VoP	0.1 (0.103%)	19.0	37.9	46.5	66.9	14.0	18.5	36.1	45.3	59.5	14.0
VoP$^{\text{P}}$	0.5 (0.441%)	19.2	38.3	47.3	64.4	12.0	19.7	38.9	48.1	55.4	12.0
VoP$^{\text{C}}$	14.3 (11.898%)	20.4	40.0	48.1	65.9	12.0	20.3	38.7	48.5	56.9	11.0
VoP$^{\text{F}}$	0.1 (0.103%)	20.6	39.5	49.1	60.3	11.0	21.2	39.4	49.2	52.3	11.0
VoP$^{\text{F+P}}$	0.4 (0.328%)	20.7	40.7	49.7	59.1	11.0	21.5	**40.6**	**50.7**	50.8	10.0
VoP$^{\text{F+C}}$	14.1 (11.785%)	21.1	**40.9**	49.6	60.1	11.0	**22.3**	40.3	**50.7**	51.1	10.0

Ablation Study

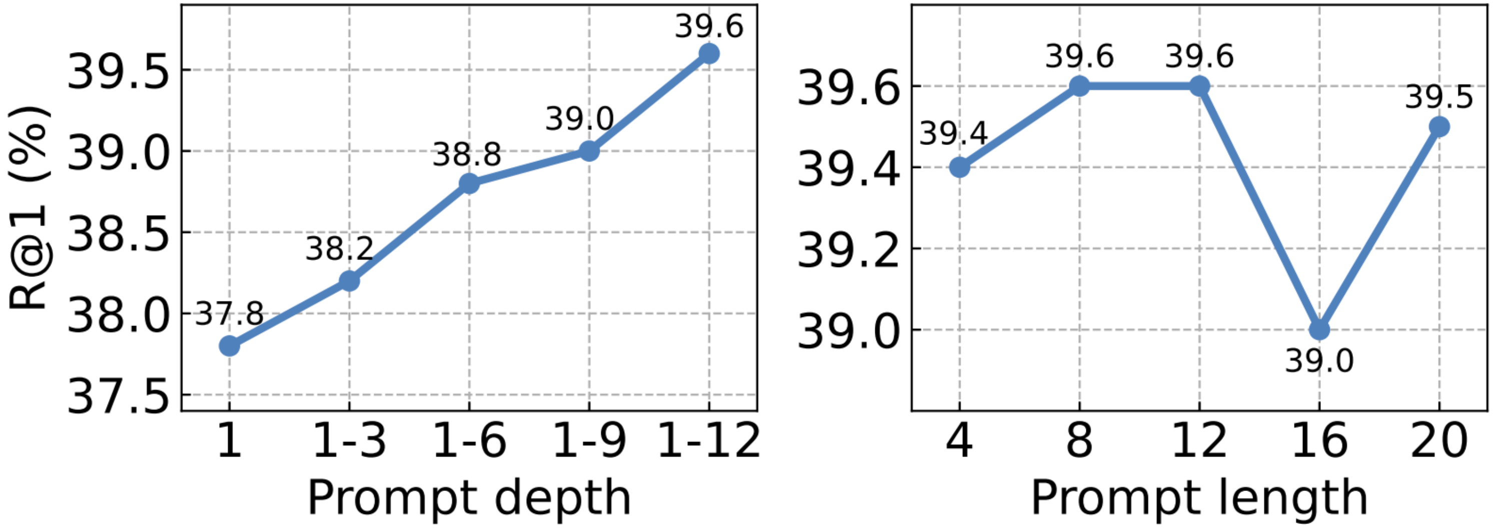
Textual	Visual	R@1	R@5	R@10	MnR↓	MdR↓
		31.5	52.8	63.6	42.9	5.0
	✓	36.5	62.7	75.1	18.3	3.0
✓		36.3	63.4	75.0	20.3	3.0
✓	✓	**39.6**	**66.7**	**77.8**	**17.2**	**2.0**

Prompting both encoders (i.e. VoP) > Uni-modal prompts > applying CLIP without tuning

Ablation Study

1. Inserting prompts into **every layer** of both encoders contributes to the best results.

2. Using **only 8 prompt tokens** remains a competitive performance with parameter efficiency.



Qualitative Results

Figure 7. **Qualitative results of four tuning methods: Full, Partial, VoP and VoP^{F+C}.** Given the query text, we represent the rank-1 retrieval result of each method, which can be incorrect (each first row) or ground truth (each second row).

Thank you for listening!

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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

Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, Donglin Wang

arXiv

Project page

Github