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

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), AI Division, School of Engineering, Westlake University
[2]Alibaba Group [3]Zhejiang University
{huangsiteng, wangdonglin}@westlake.edu.cn, y2li@zju.edu.cn,
a.biao.gong@gmail.com, {yanwen.pyl, jianwen.jjw, yiliang.lyl}@alibaba-inc.com

A. Discussion on Non-parameter-efficient Methods

Methods	Params (M)	R@1	R@5	R@10	MnR↓	MdR↓
X-Pool [3]	1.3 (1.1%)	40.5	64.8	75.0	18.9	2.0
VoP$^{\text{F}}$	**0.1 (0.1%)**	42.6	68.4	78.7	15.8	2.0
X-Pool [3]+**VoP**	1.4 (1.2%)	**43.1**	**69.5**	**79.5**	**14.5**	2.0

Table 1. **Comparison with non-parameter-efficient X-Pool [3] after freezing the CLIP backbone**. The $t2v$ retrieval results are obtained on the MSR-VTT-9k dataset.

Our work aims to greatly reduce the overall storage costs while achieving promising cross-modal retrieval performance. Related non-parameter-efficient methods [3,7,8] requires to fine-tune the additional parameters together with the CLIP backbone, which results in an unaffordable overhead. Despite the potential for better performance, these methods contradict our purpose. Therefore, they are not included in the fundamental comparison for fairness. To illustrate the value of studying parameter-efficient methods, in Tab. 1, we compare with the state-of-the-art X-Pool [3] by freezing the CLIP backbone. We observe that without fine-tuning the backbone, X-Pool underperforms our VoP$^{\text{F}}$ with much more parameter overhead. And equipping our simplest VoP significantly boosts its performance with negligible additional parameters. The comparison results demonstrate the superiority of our proposed methods as parameter-efficient solutions.

B. Retrieval Results with ViT-B/16

In this section, we change the visual encoder to a ViT-B/16 to examine all solutions including ours with a heavier backbone. Compared to the default ViT-B/32, ViT-B/16 splits the image into more and smaller 16×16 patches, increasing the computational effort to learn more detailed relational information while slightly reducing the number of parameters (118.1M *v.s.*119.8M). We here report the $t2v$ results obtained on MSR-VTT-9k in Tab. 2 and also compare with the method proposed by Ju *et al.* [6]. Several observations as follows: (1) Our VoP now outperforms all parameter-efficient tuning protocols including Partial, showing its ability to effectively transfer the latent knowledge with fewer trainable parameters. (2) The proposed video prompts still steadily reinforce VoP, where VoP$^{\text{F}}$ and its variants outperform Full. (3) equipping with two video prompts brings a 3.7% to 4.3% improvement to VoP, and our VoP$^{\text{F+C}}$ even yields a remarkable $t2v$ R@1 47.7%.

Methods	Params (M)	R@1	R@5	R@10	MnR↓	MdR↓
Full	118.1 (100%)	44.9	72.2	81.7	13.5	2.0
Bias [1]	0.1 (0.105%)	42.2	68.5	78.2	13.9	2.0
Proj [5]	0.7 (0.555%)	39.1	65.7	75.6	17.6	2.0
Partial [5]	7.7 (6.506%)	43.0	69.3	78.5	15.8	2.0
Adapter$^{\text{ATTN}}$ [4]	2.0 (1.680%)	41.7	66.4	76.6	15.1	2.0
Adapter$^{\text{FFN}}$ [2]	2.0 (1.680%)	41.4	66.5	77.0	15.0	2.0
Ju *et al.* [6]	4.8 (3.990%)	36.7	64.6	76.8	-	2.0
VoP	0.1 (0.104%)	43.4	69.1	80.5	14.2	2.0
VoP$^{\text{P}}$	0.5 (0.448%)	43.9	70.0	80.9	12.9	2.0
VoP$^{\text{C}}$	14.3 (12.077%)	44.6	71.8	80.2	14.6	2.0
VoP$^{\text{F}}$	0.1 (0.104%)	46.5	**73.0**	81.5	<u>12.4</u>	2.0
VoP$^{\text{F+P}}$	0.4 (0.333%)	<u>47.1</u>	<u>72.4</u>	<u>81.8</u>	12.9	2.0
VoP$^{\text{F+C}}$	14.1 (11.962%)	**47.7**	<u>72.4</u>	**82.2**	**12.0**	2.0

Table 2. $t2v$ **results on the MSR-VTT-9k dataset with ViT-B/16**.

C. Detailed Retrieval Results

We here report the detailed retrieval results on MSR-VTT-7k (Tab. 3), DiDeMo (Tab. 4), ActivityNet (Tab. 5), LSMDC (Tab. 6) for reference. Note that these results are obtained using CLIP with ViT-B/32 unless otherwise stated. The conclusions in these tables are generally consistent with those from the above experiments.

*Work done during internship at Alibaba DAMO Academy.
†Corresponding author.

Methods	Params (M)	*t2v*					*v2t*				
		R@1	**R@5**	**R@10**	**MnR↓**	**MdR↓**	**R@1**	**R@5**	**R@10**	**MnR↓**	**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 [1]	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 [5]	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 [5]	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$^{\text{ATTN}}$ [4]	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$^{\text{FFN}}$ [2]	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$^{\mathbf{P}}$	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$^{\mathbf{C}}$	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$^{\mathbf{F}}$	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$^{\mathbf{F+P}}$	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$^{\mathbf{F+C}}$	14.1 (11.785%)	42.7	**68.2**	**79.3**	**15.9**	2.0	**44.2**	69.6	**80.8**	11.4	2.0

Table 3. **Retrieval results on the MSR-VTT-7k dataset**.

Methods	Params (M)	*t2v*					*v2t*				
		R@1	**R@5**	**R@10**	**MnR↓**	**MdR↓**	**R@1**	**R@5**	**R@10**	**MnR↓**	**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 [1]	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 [5]	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 [5]	7.7 (6.410%)	39.3	65.5	75.7	22.3	2.0	36.9	64.2	74.5	17.0	2.0
Adapter$^{\text{ATTN}}$ [4]	2.0 (1.655%)	36.4	62.8	73.9	23.5	3.0	36.3	64.4	74.8	15.4	2.0
Adapter$^{\text{FFN}}$ [2]	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$^{\mathbf{P}}$	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$^{\mathbf{C}}$	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$^{\mathbf{F}}$	0.1 (0.103%)	44.7	70.8	79.7	15.7	2.0	43.5	70.9	81.4	9.8	2.0
VoP$^{\mathbf{F+P}}$	0.4 (0.328%)	45.3	**72.3**	80.4	13.8	2.0	**44.7**	71.2	81.1	9.9	2.0
VoP$^{\mathbf{F+C}}$	14.1 (11.785%)	**46.4**	71.9	**81.5**	**13.6**	2.0	44.4	**71.8**	**81.8**	**9.5**	2.0

Table 4. **Retrieval results on the DiDeMo dataset**.

Methods	Params (M)	*t2v*					*v2t*				
		R@1	**R@5**	**R@10**	**MnR↓**	**MdR↓**	**R@1**	**R@5**	**R@10**	**MnR↓**	**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 [1]	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 [5]	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 [5]	7.7 (6.410%)	33.6	64.0	77.8	10.6	3.0	33.4	64.6	77.8	10.2	3.0
Adapter$^{\text{ATTN}}$ [4]	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$^{\text{FFN}}$ [2]	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$^{\mathbf{P}}$	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$^{\mathbf{C}}$	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$^{\mathbf{F}}$	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$^{\mathbf{F+P}}$	0.4 (0.328%)	36.1	65.5	78.5	10.9	3.0	36.3	65.9	79.2	10.1	3.0
VoP$^{\mathbf{F+C}}$	14.1 (11.785%)	35.1	63.7	77.6	11.4	3.0	35.6	65.9	77.8	10.4	3.0

Table 5. **Retrieval results on the ActivityNet dataset**.

Methods	Params (M)	*t2v*					*v2t*				
		R@1	**R@5**	**R@10**	**MnR↓**	**MdR↓**	**R@1**	**R@5**	**R@10**	**MnR↓**	**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 [1]	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 [5]	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 [5]	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$^{\text{ATTN}}$ [4]	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$^{\text{FFN}}$ [2]	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.* [6] †	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$^{\mathbf{P}}$	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$^{\mathbf{C}}$	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$^{\mathbf{F}}$	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$^{\mathbf{F+P}}$	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$^{\mathbf{F+C}}$	14.1 (11.785%)	21.1	**40.9**	49.6	60.1	11.0	**22.3**	40.3	**50.7**	51.1	10.0

Table 6. **Retrieval results on the LSMDC dataset**. † denotes that it uses CLIP with ViT-B/16.

References

[1] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. TinyTL: Reduce memory, not parameters for efficient on-device learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 11285–11297, 2020. 1, 2

[2] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022. 1, 2

[3] Satya Krishna Gorti, Noel Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guang Wei Yu. X-Pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4996–5005, 2022. 1

[4] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Proceedings of the International Conference on Learning Representations*, 2022. 1, 2

[5] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, pages 709–727, 2022. 1, 2

[6] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 105–124, 2022. 1, 2

[7] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. TS2-Net: Token shift and selection transformer for text-video retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 319–335, 2022. 1

[8] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 638–647, 2022. 1