SDPT: Synchronous Dual Prompt Tuning for Fusion-based Visua Language Pre-trained Models

Yang Zhou*, Yongjian Wu*, Jiya Saiyin, Bingzheng Wei, Maode Lai, Eric Chang, Yan Xu

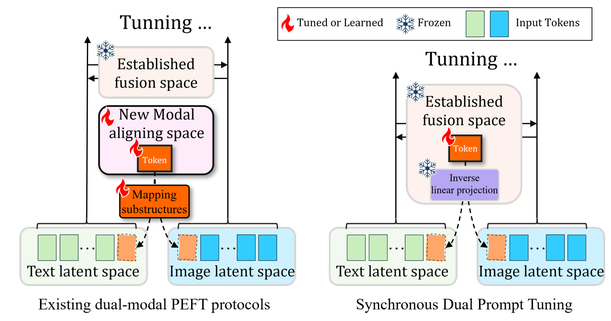
生物与医学工程学院 BUAA SCHOOL OF BIOLOGICAL SCIENCE AND MEDICAL ENGINEERING

Code

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION MILANO 2024

Motivation

- Prompt tuning methods have emerged as prevalent parameter-efficient fine-tuning (PEFT) approaches.
- **Unimodal prompt tuning methods** fail to effectively integrate new information from the other modality during downstream transfer of dual-modal VLPMs.
- **Current dual-modal prompt tuning methods** rely on additional modal mapping substructures to remodel the dual-modality alignment space. These substructures produce inaccurate distributions that reduce transfer generalization and performance and also increase training and storage costs.
- **Aim:** Fully utilize the pre-trained modal mapping inherent in deep fusion of fusion-based VLPMs to establish a dual-modal PEFT method with better accuracy and efficiency.



Our method SDPT vs. other dual-modal PEFT methods on fusion-based VLPMs. Existing dual-modal PEFT methods (left) require learning new modal mapping substructures or modality aligning spaces, whereas SDPT (right) does not.

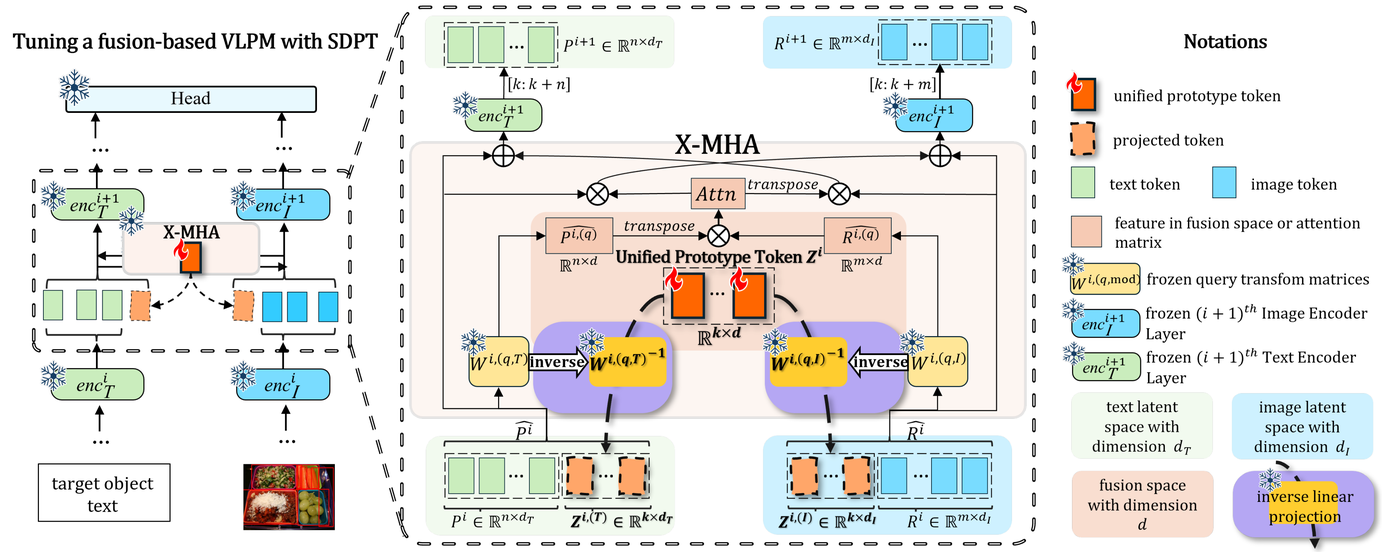
Contribution

- We propose Synchronous Dual Prompt Tuning (SDPT), the first dual-modal PEFT approach specially designed for fusion-based VLPMs. SDPT achieves superior transfer performance with only 0.04% model training parameters.
- SDPT creates learnable prototype tokens in the deep fusion space to align text and image semantics and uses accurate inverse linear projections that require no additional training.
- SDPT fully leveraging the pre-trained knowledge of fusion-based VLPMs. Extensive experiments have demonstrated the generality and flexibility of SDPT.

Method: dual-modal prompt tuning with inherent modal mapping structure

1. **Construct unified prototype tokens** within the pre-established cross-attention space, i.e. the fusion space.
2. **Establish two inverse linear projections** which require no training, enabling synchronous mapping of the unified prototype tokens back to the text and image latent spaces.

- **Basis:** Pre-trained mappings from different modalities to the fusion space already exist in the X-MHA operators, enabling the direct construction of **accurate inverse linear projections.**

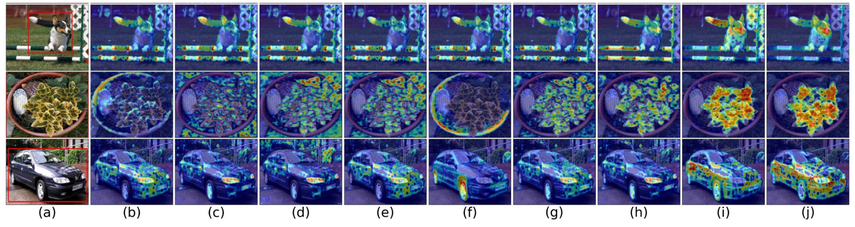


Experiment

Comparison with other PEFT methods. "#Par" denotes the number of trainable parameters, values in "(·)" representing the proportion of trainable parameters relative to the total model parameters.

Method		FLOPs(G)	#Par(M)	COCO				LVIS				ODinW13				
				mAP	AP50	AP75	AR	AP	APr	APc	APf	1-shot	3-shot	5-shot	10-shot	full-shot
Linear probing		**724.30**	1.96(0.49%)	52.7	69.9	58.1	69.4	33.8	25.3	29.8	40.6	54.1	54.7	55.0	55.8	59.2
Image	Adaptformer [4]	780.87	56.60(14.24%)	55.8	73.4	60.9	71.4	39.4	29.8	34.7	43.8	47.3	55.0	56.1	60.1	68.0
	VPT [17]	1317.41	4.26(1.07%)	57.4	75.0	62.5	72.5	40.7	29.7	32.3	44.2	58.9	61.3	63.6	65.2	68.8
Text	Adapter [13]	724.31	10.62(2.67%)	57.5	75.3	62.9	72.4	40.4	29.8	34.3	43.9	59.8	61.6	62.5	65.0	69.2
	CoOp [41]	1306.24	0.52(0.13%)	55.4	73.2	60.7	71.1	39.5	29.8	33.4	43.1	59.6	61.6	63.1	65.5	68.8
Dual	Full FT	**724.30**	397.59(100%)	**60.8**	\	\	\	41.2	31.2	34.2	45.0	59.9	62.1	64.2	64.9	68.9
	LoRA [14]	**724.30**	2.03(0.51%)	56.5	74.1	61.8	71.0	40.8	30.4	33.7	43.5	60.9	61.7	63.1	64.8	66.9
	BitFit [37]	**724.30**	0.42(0.11%)	54.5	73.0	60.1	70.6	39.1	29.7	34.2	43.1	61.2	61.7	64.1	65.6	68.3
	LoRA$_{XMHA}$ [14]	**724.30**	1.58(0.40%)	55.4	72.8	60.5	71.2	38.6	29.8	35.1	43.4	60.0	61.3	63.2	63.3	68.6
	BitFit$_{XMHA}$ [37]	**724.30**	0.21(0.05%)	52.6	69.5	57.3	70.3	36.1	26.7	29.7	41.6	61.5	62.1	62.9	63.5	65.2
	DPT [34]	925.17	80.71(20.30%)	51.1	67.9	57.1	67.3	35.3	26.1	30.3	41.5	59.0	59.7	60.6	61.6	65.2
	Apollo [5]	1150.92	3.18(0.80%)	54.3	73.1	58.5	69.5	37.4	31.5	34.0	41.4	55.8	60.2	60.4	62.5	66.7
	PMF [24]	730.59	1.53(0.38%)	56.2	74.2	60.2	70.4	39.6	30.8	33.2	42.2	58.6	61.1	61.8	64.0	67.4
	UPT [38]	785.32	2.75(0.69%)	56.8	74.5	61.3	71.7	40.2	29.8	33.4	43.3	59.6	61.8	63.3	63.9	67.7
	MaPLe [18]	801.41	2.96(0.74%)	57.2	75.1	62.2	72.3	40.8	30.6	33.8	43.7	60.8	61.9	62.8	64.1	68.7
	SDPT:k=10	724.77	**0.16**(0.04%)	57.6	75.4	63.0	72.8	41.2	31.5	34.9	45.0	61.5	62.2	64.3	65.6	69.5
	SDPT:k=120	738.36	1.97(0.50%)	58.0	**75.8**	**63.2**	**73.1**	**41.4**	**31.8**	**35.2**	**45.1**	**61.6**	**64.4**	**64.4**	**66.4**	**71.2**

Comparison of attention map visualization on PascalVOC. (a) Original image and ground truths, (b) LoRA, (c) BitFit, (d) DPT, (e) Apollo, (f) PMF, (g) UPT, (h) MaPLe, (i) SDPT (k=10), (j) SDPT (k=200).



Compatibility with other fine-tuned PEFT modules, with the COCO as an old task and the Aquarium as a new task. PEFT components with subscript "COCO" refer to those fine-tuned on the old task.

Method	Setting	COCO mAP	Aquarium mAP
GLIP_L + Adapter$_{coco}$	Adapter$_{coco}$ has been fine-tuned on COCO. Zero-shot reference on Aquarium	56.8	30.7
GLIP_L + Adapter$_{coco}$ + Adapter	Keep other parameters frozen and only tune a new Adapter on Aquarium	35.4	55.6
GLIP_L + Adapter$_{coco}$ + SDPT	Keep other parameters frozen and only tune SDPT on Aquarium	53.2	58.7
GLIP_L + VPT$_{coco}$	VPT$_{coco}$ has been fine-tuned on COCO. Zero-shot reference on Aquarium	55.4	27.6
GLIP_L + VPT$_{coco}$ + VPT	Keep other parameters frozen and only tune a new VPT on Aquarium	30.8	53.1
GLIP_L + VPT$_{coco}$ + SDPT	Keep other parameters frozen and only tune SDPT on Aquarium	51.9	58.3

Effectiveness of synchronous dual-modal knowledge incorporation. "Sync." indicates whether the knowledge incorporating strategy is synchronous spatially and temporally.

Setting	Sync.	#Par(M)	Aquarium mAP	AP50	AP75	AR
Unshared prototype tokens	×	6.56	53.3	80.2	56.3	66.1
Asynchronous training	×	**3.28**	54.8	81.7	57.4	67.8
Standard setting	✓	**3.28**	**57.7**	**85.0**	**60.6**	**70.5**

Effectiveness of inverse linear projections. We compared our inverse linear projection with the learnable linear projection. "Modal" indicates the modal branch receiving the projected tokens.

Method	Modal	#Par(M)	Aquarium mAP	AP50	AP75	AR
Learnable linear projection	Image	13.82	51.7	81.0	55.2	65.5
	Text	3.58	52.8	81.8	55.6	65.8
	Dual	17.40	54.3	83.1	57.4	67.2
Inverse linear projection	Image	**3.28**	53.1	82.9	56.8	66.1
	Text	**3.28**	55.4	83.6	58.7	68.1
	Dual	**3.28**	**57.7**	**85.0**	**60.6**	**70.5**