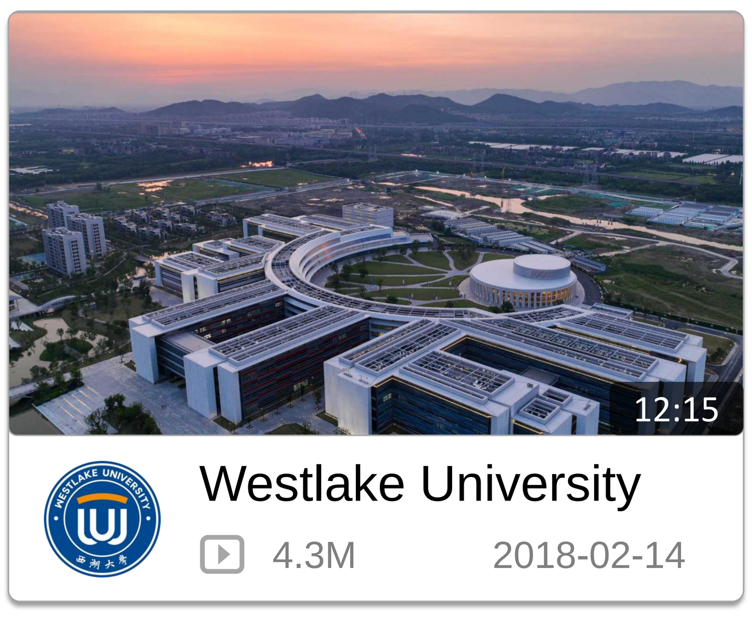
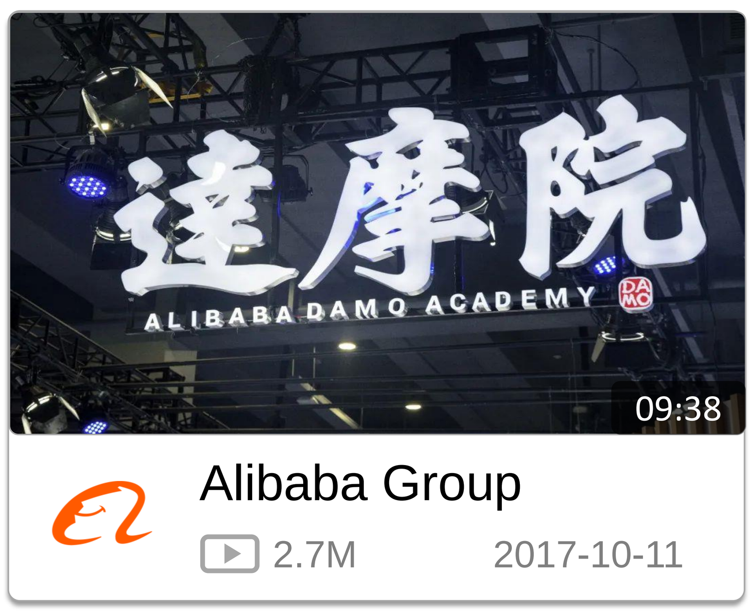
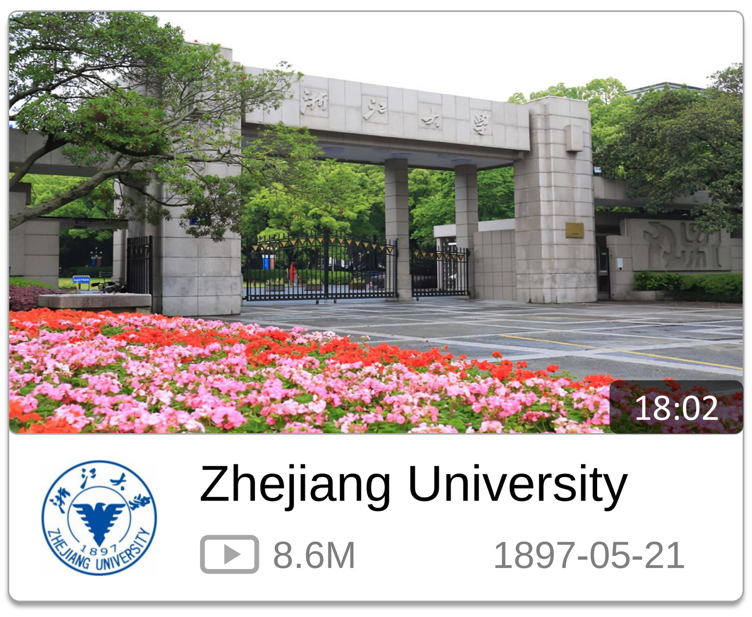
VoP: Text-Video Co-operative Prompt Tuning for Cross-Modal Retrieval





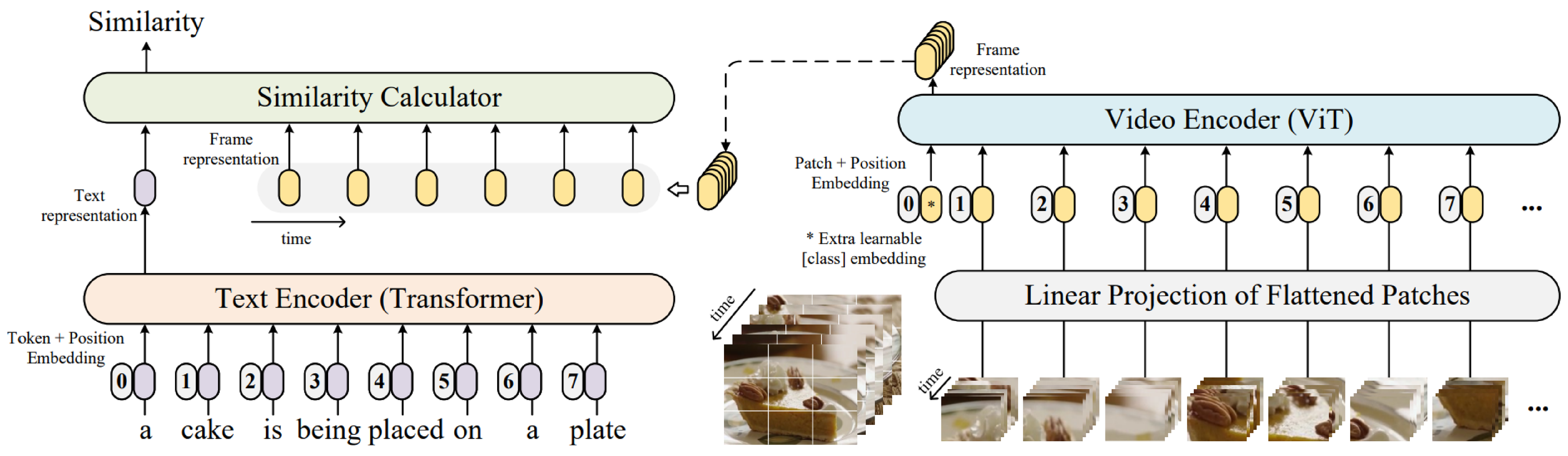


Retrieval

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.



[CLIP4Clip, Neurocomputing 2022]

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.

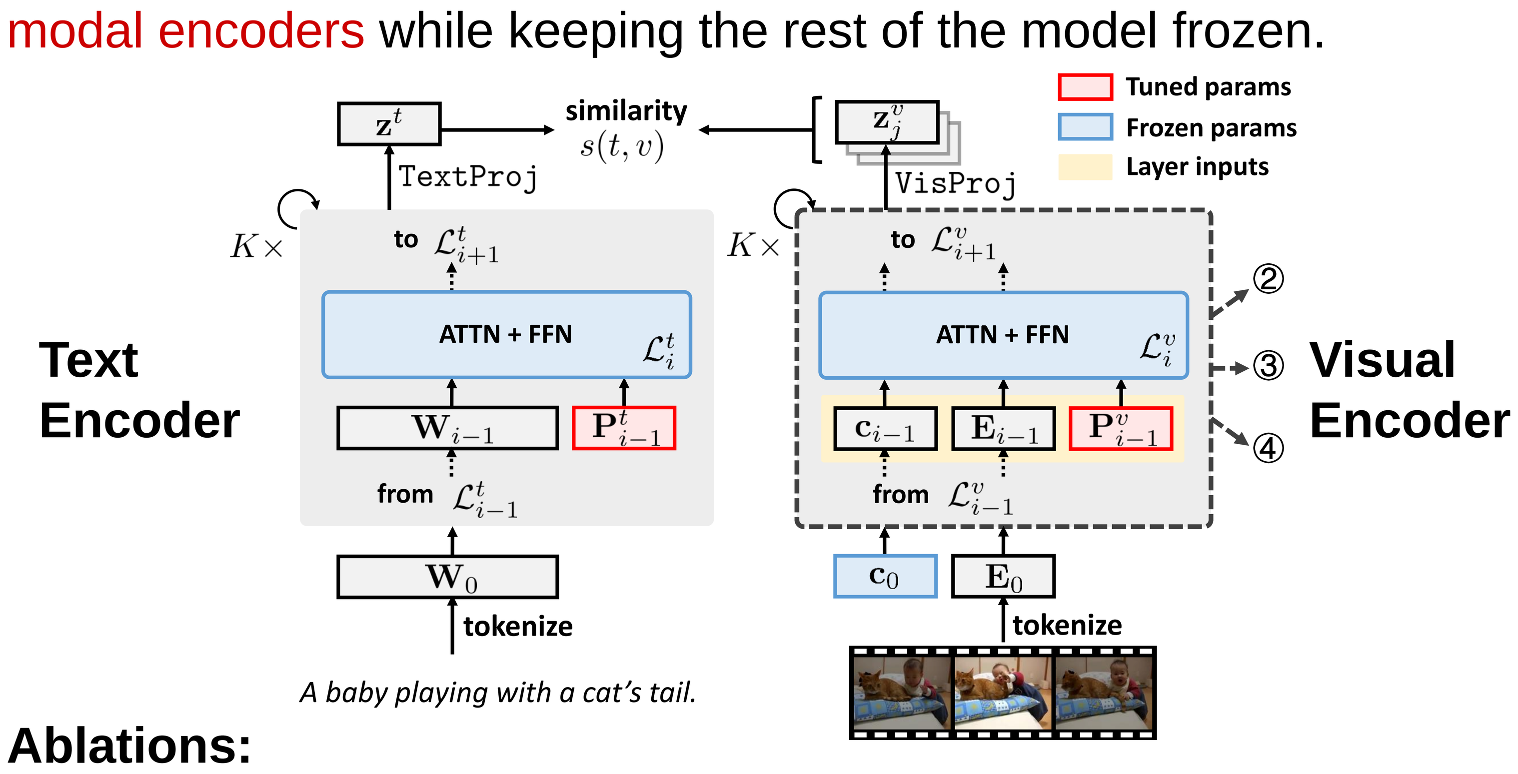
Our Proposed Framework

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

Motivations:

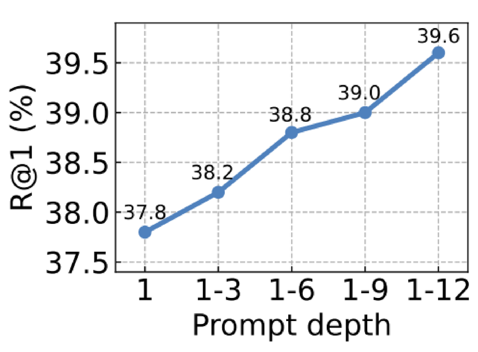
- Learning prompts only for the text branch overlooks the potential of collaboratively tuning the visual encoder.
- Prompting the mere input layer has only a relatively indirect impact on the output embeddings.

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



Ablations:

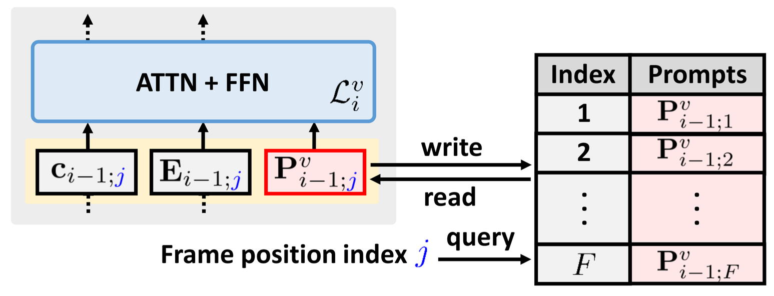
Textual	Visual	R@1	R@5	R@10	MnR↓	MdR↓
		31.5	52.8	63.6	42.9	5.0
	✓	36.5	62.7	75.1	18.3	3.0
✓		36.3	63.4	75.0	20.3	3.0
✓	✓	**39.6**	**66.7**	**77.8**	**17.2**	**2.0**



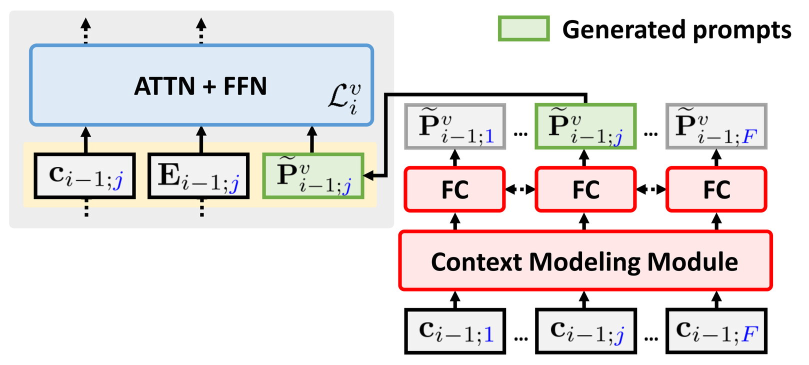
Equipping with Three Plug-and-Play Video Prompts

Motivation: Assisting VoP in utilizing rich *temporal* information.

② **VoP^{P}**: **position-specific** video prompts model the information shared between frames at the *same relative position*.

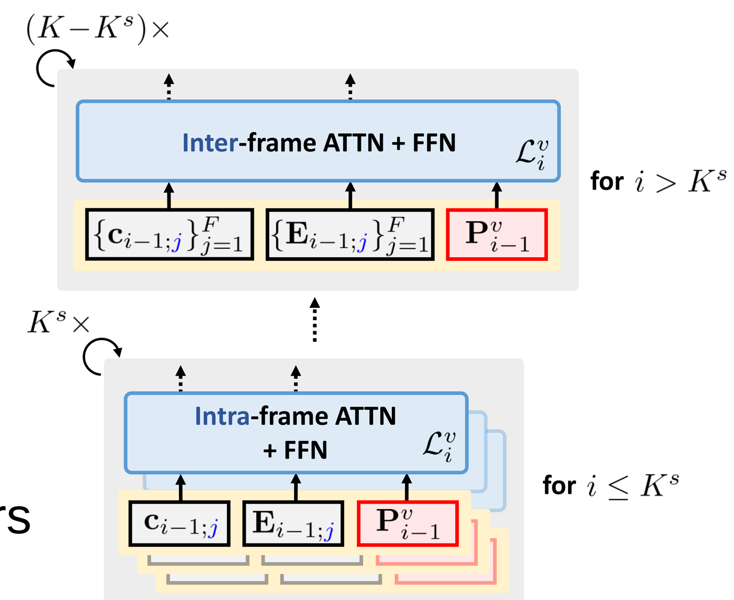


③ **VoP^{C}**: generated **context-specific** video prompts integrate injected *contextual* message from the frame sequence into the intra-frame modeling.



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 (**VoP^{F+P}** / **VoP^{F+C}**).

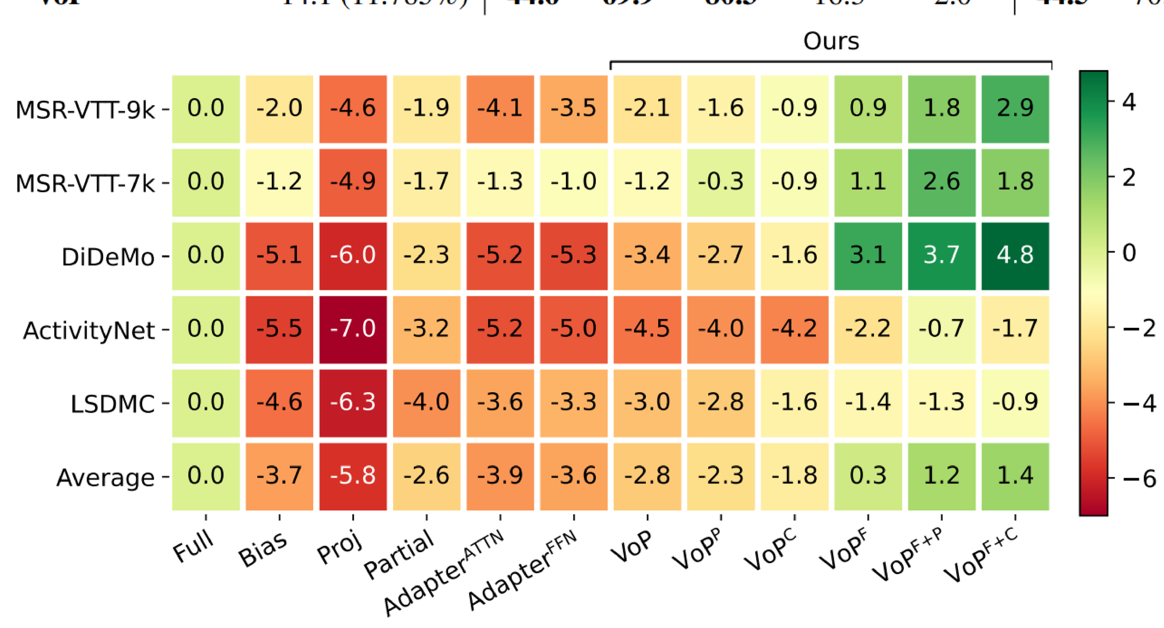


Experiments

Main Results *(CLIP ViT-B/32)*

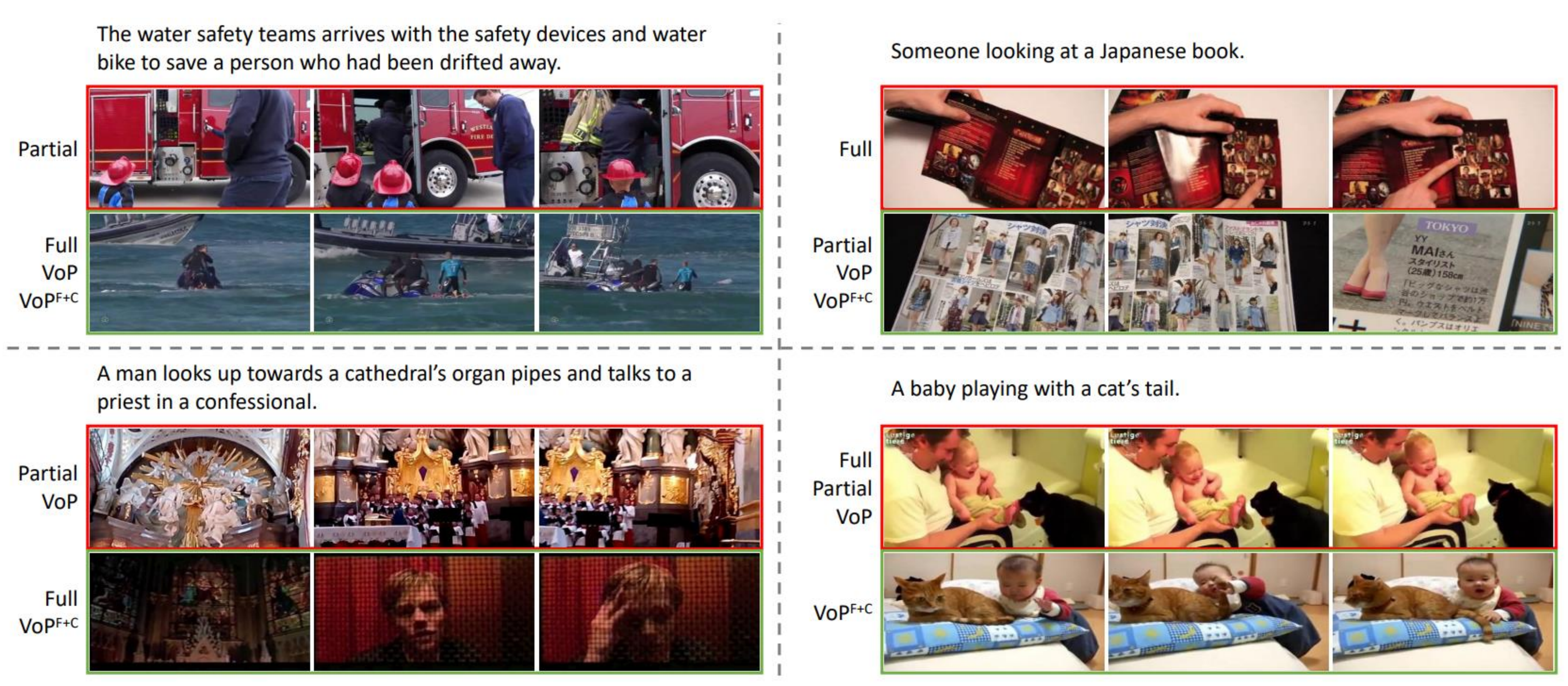
Methods	Params (M)	t2v R@1	t2v R@5	t2v R@10	t2v MnR↓	t2v MdR↓	v2t R@1	v2t R@5	v2t R@10	v2t MnR↓	v2t MdR↓
Full	119.8 (100%)	41.7	69.2	79.0	16.5	2.0	42.5	70.9	**81.4**	**11.0**	2.0
Bias [6]	0.1 (0.104%)	39.7	66.5	77.3	17.3	2.0	41.1	68.4	79.2	13.6	2.0
Proj [17]	0.7 (0.547%)	37.1	63.0	76.1	20.5	3.0	37.2	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	15.8	2.0	42.3	70.1	81.1	11.4	2.0
VoP^{F}	0.1 (0.103%)	42.6	68.4	78.7	15.8	2.0	42.4	70.5	81.0	**11.0**	2.0
VoP^{F+P}	0.4 (0.328%)	43.5	69.3	79.3	**14.8**	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

↑MSR-VTT-9k



←t2v relative results on all datasets

Qualitative Results (color: incorrect or ground truth)



Authors:

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

Organizations:

1 Machine Intelligence Laboratory (MiLAB), Westlake University

2 Alibaba Group

3 Zhejiang University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

ArXiv Homepage Youtube Github