浙江大学
ZHEJIANG UNIVERSITY

Cross-modal Prompts

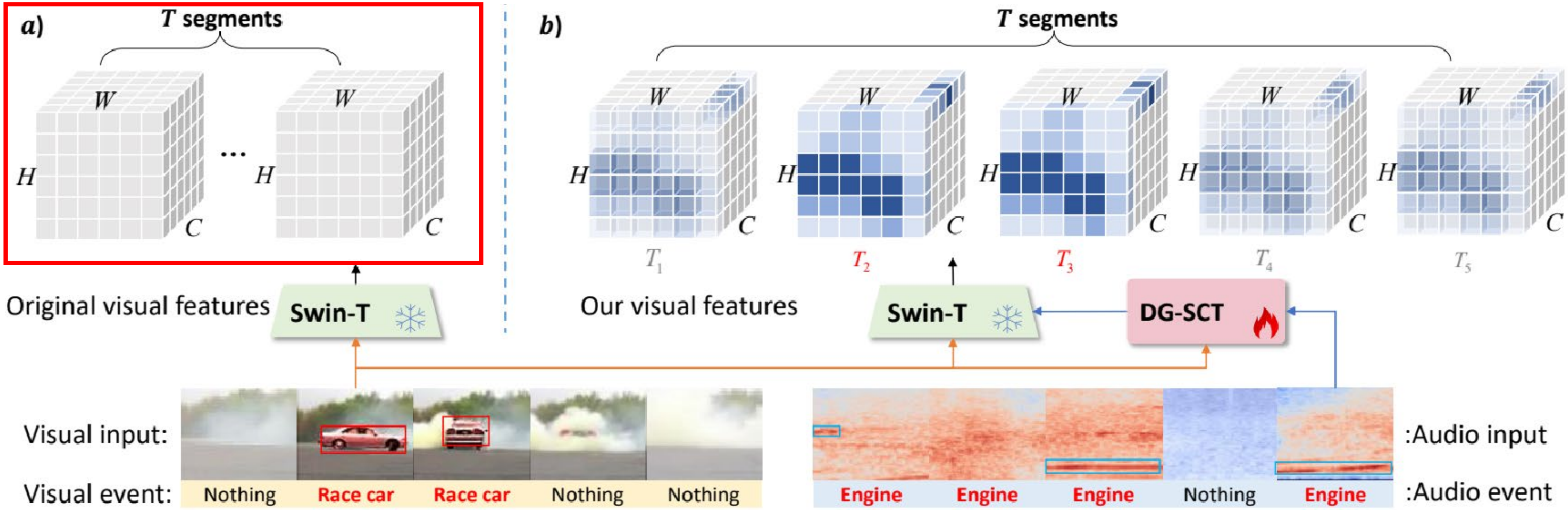
Adapting Large Pre-trained Models for Audio-Visual Downstream Tasks

Haoyi Duan

01

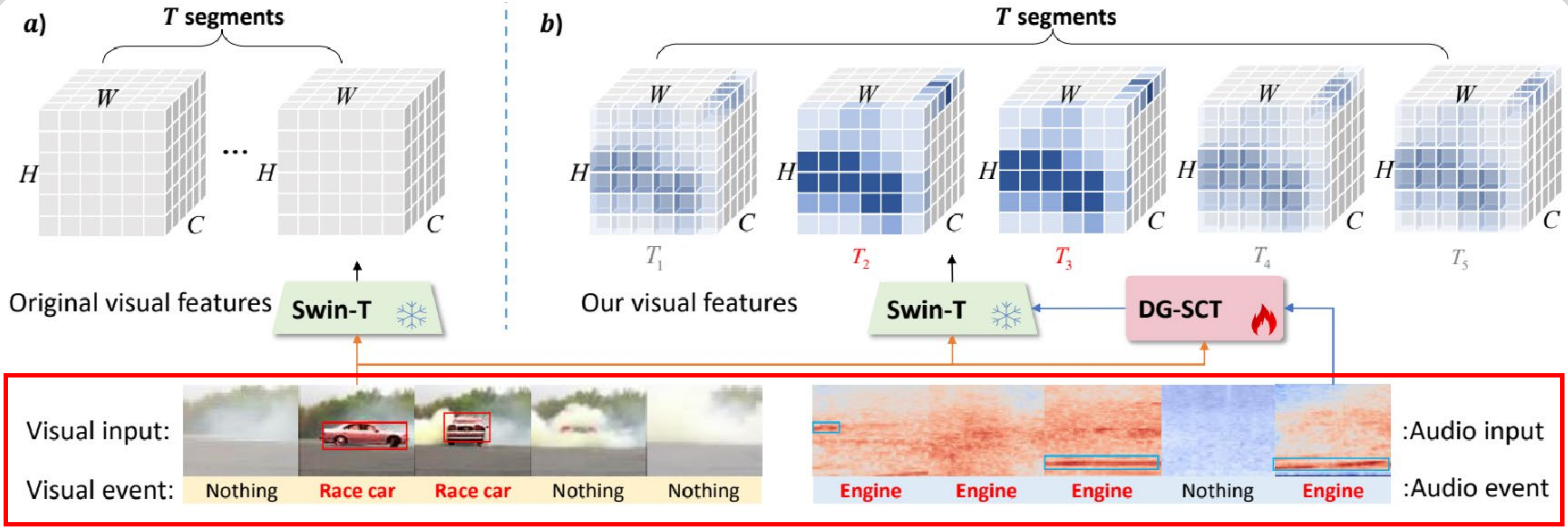
Introduction

01 Introduction



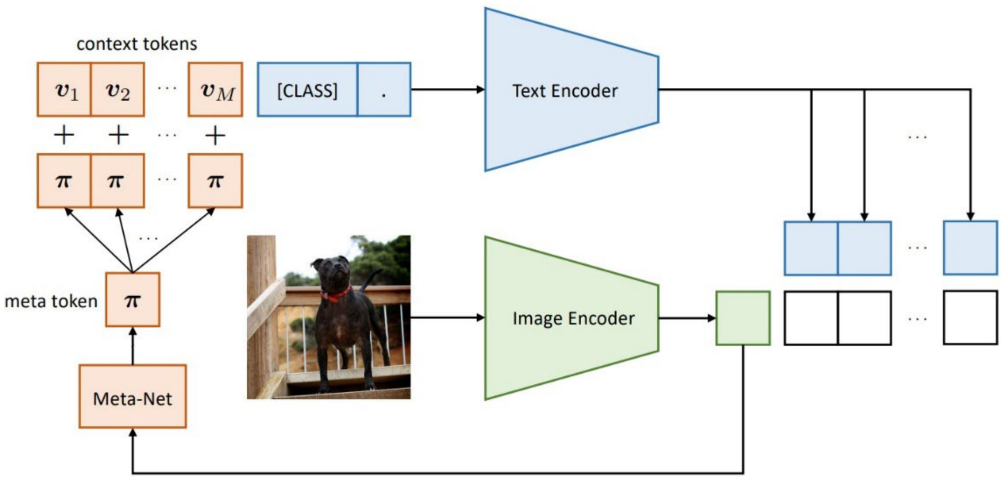
As depicted in (a), the pre-trained model equally extracts visual features and directly passes them to downstream tasks.

01 Introduction

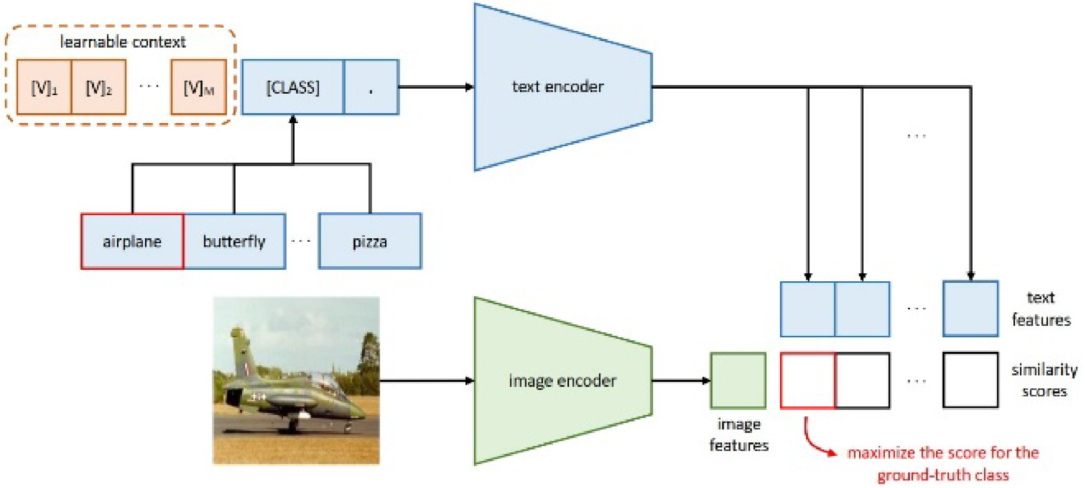


However, when perceiving the roaring sound of an engine, the visual region depicting a "car" should receive more attention than the region of "trees". Simultaneously, when observing the car, it is crucial to concentrate on the audio segments of the engine sound. Therefore, the encoder should not only equally extract modal-specific information from the current modality, but also highlight information related to other modalities to enhance feature fusion across diverse modalities in downstream tasks.

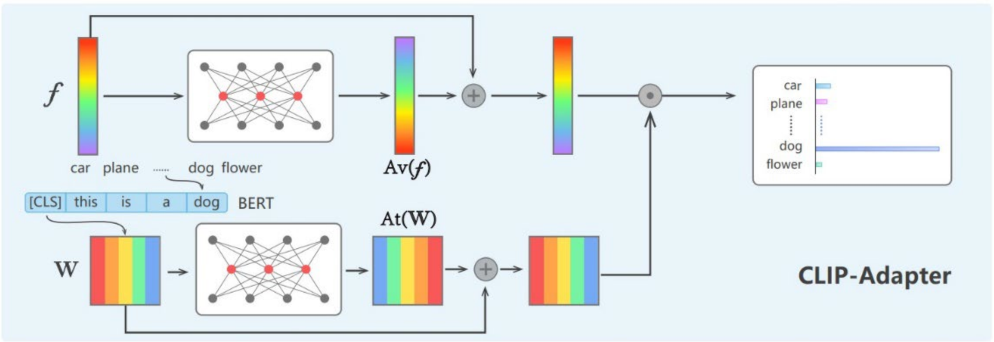
01 Introduction



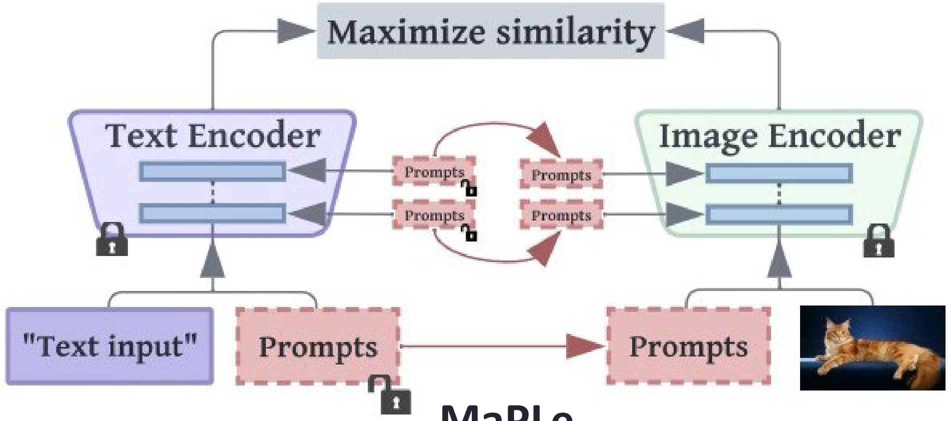
CoOp



CoCoOp

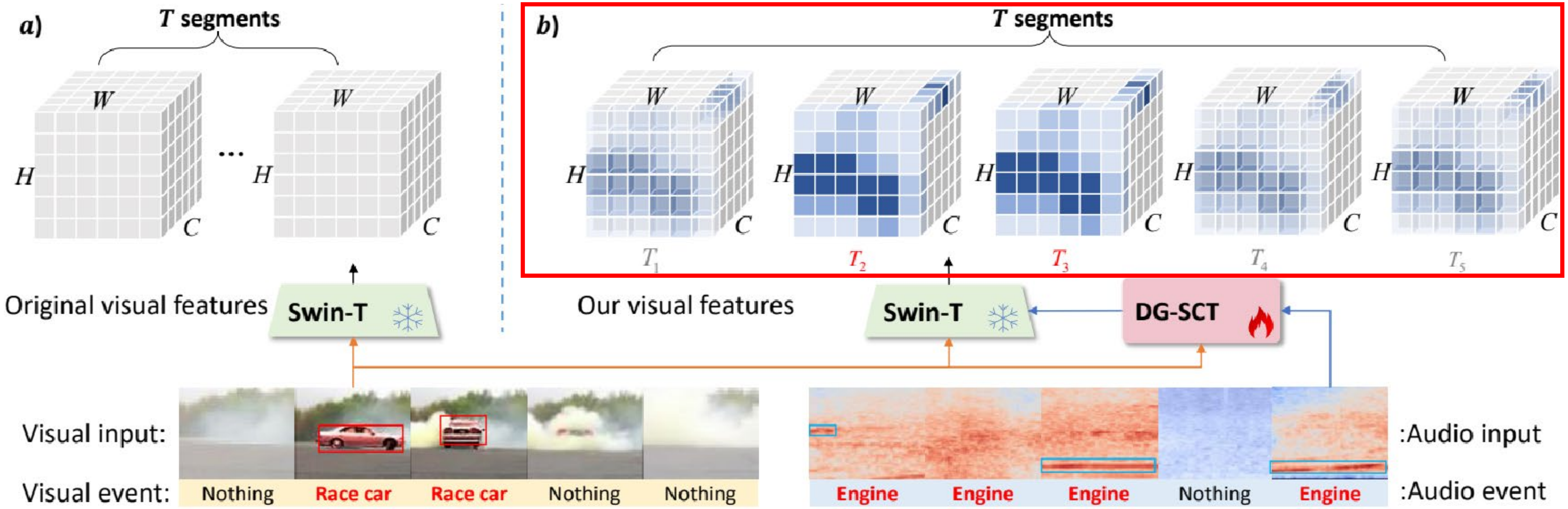


CLIP-Adapter



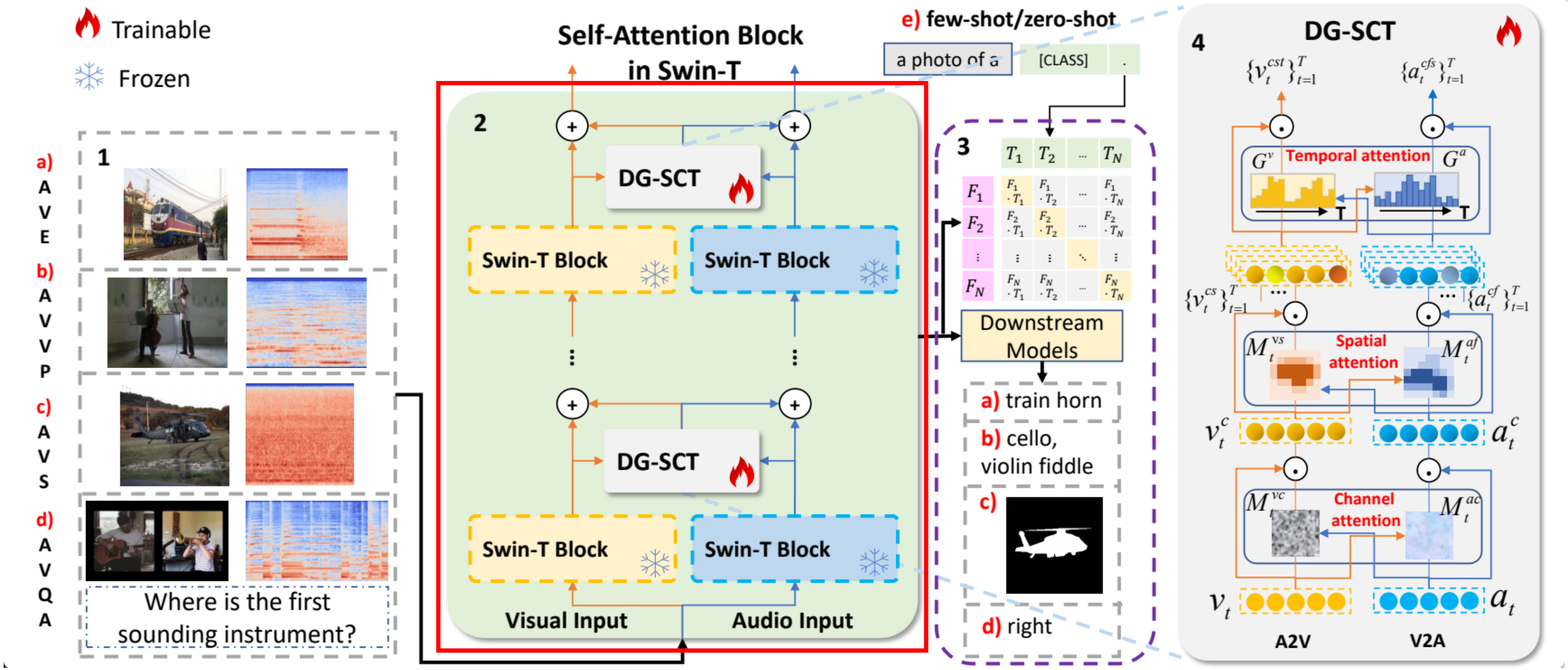
MaPLe

01 Introduction DG-SCT attention



Shown in (b), take visual modality, our visual features contain fine-grained, task-specific information under the guidance of audio prompts.

01 Introduction



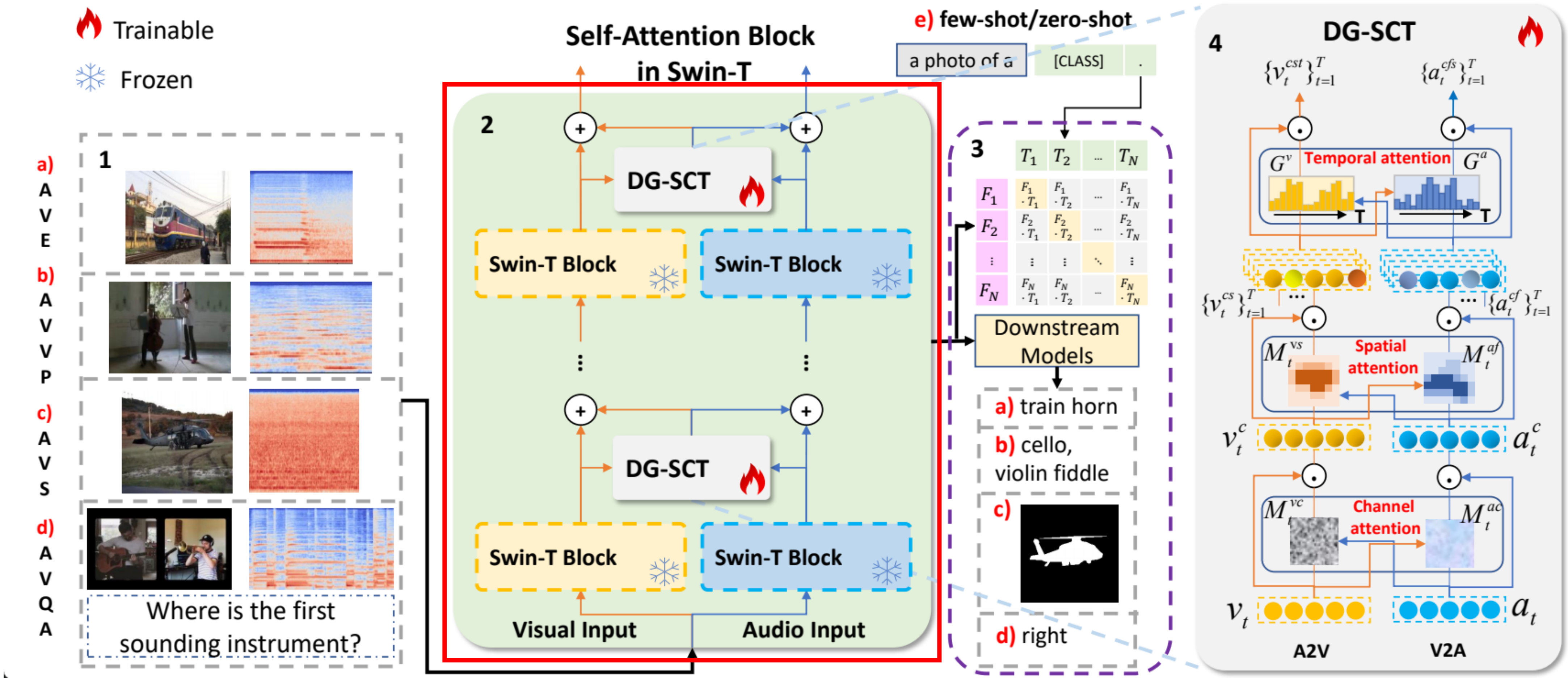
Specifically, DG-SCT is injected into every layer of frozen pre-trained audio and visual encoders. This mechanism utilizes audio and visual modalities to guide the feature extraction of their respective counterpart modalities across spatial, channel, and temporal dimensions.

02

Approach

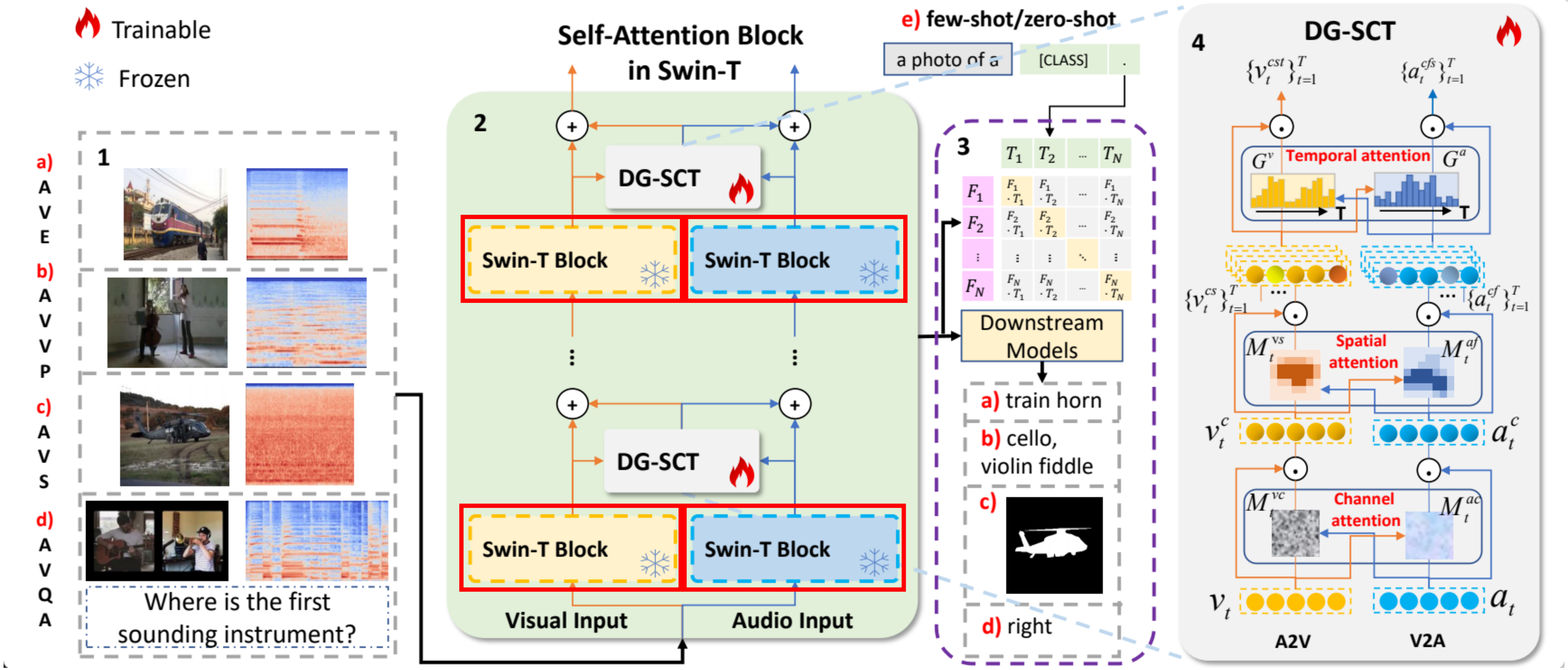
02-1 Adding DG-SCT modules to frozen encoders
02-2 DG-SCT

02-1 Adding DG-SCT modules



Now, let's have a brief introduction about how to add DG-SCT modules into Visual and Audio encoders, thus finetuning these two encoders.

02-1 Adding DG-SCT modules

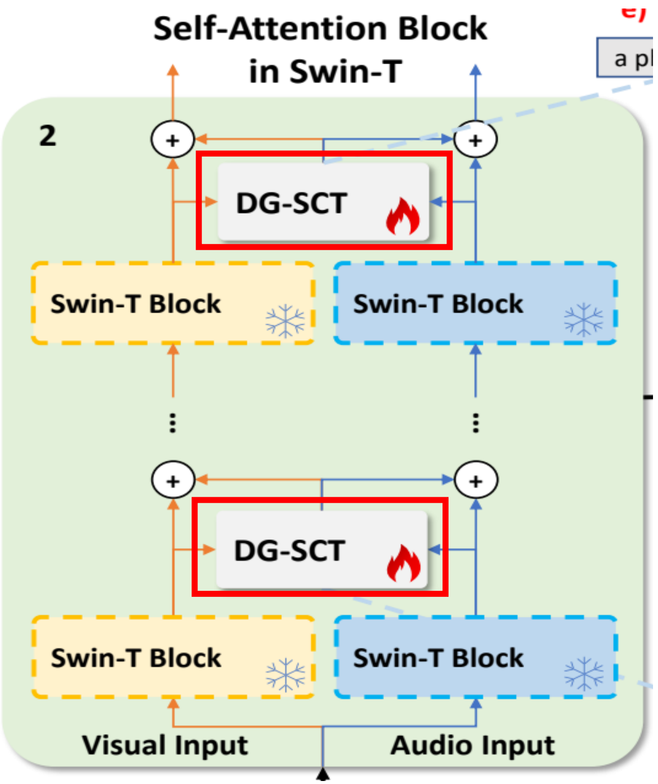


1) multi-head attention (MHA)
2) multi-layer perceptron (MLP)
3) dual-guided spatial-channel-temporal attention (DG-SCT)

02-1 Adding DG-SCT modules

How to add DG-SCT modules into Visual and Audio encoders?

$$v_y^{(l)} = v^{(\ell)} + \mathrm{MHA}(v^{(\ell)}) + \Omega^{\mathrm{a2v}}(a^{(\ell)}, v^{(\ell)}), \quad a_y^{(\ell)} = a^{(\ell)} + \mathrm{MHA}(a^{(\ell)}) + \Omega^{\mathrm{v2a}}(v^{(\ell)}, a^{(\ell)}),$$
$$v^{(\ell+1)} = v_y^{(\ell)} + \mathrm{MLP}(v_y^{(\ell)}) + \Omega^{\mathrm{a2v}}(a_y^{(\ell)}, v_y^{(\ell)}), \quad a^{(\ell+1)} = a_y^{(\ell)} + \mathrm{MLP}(a_y^{(\ell)}) + \Omega^{\mathrm{v2a}}(v_y^{(\ell)}, a_y^{(\ell)}).$$



DG-SCT
Audio as guidance (A2V)

DG-SCT
Video as guidance (V2A)

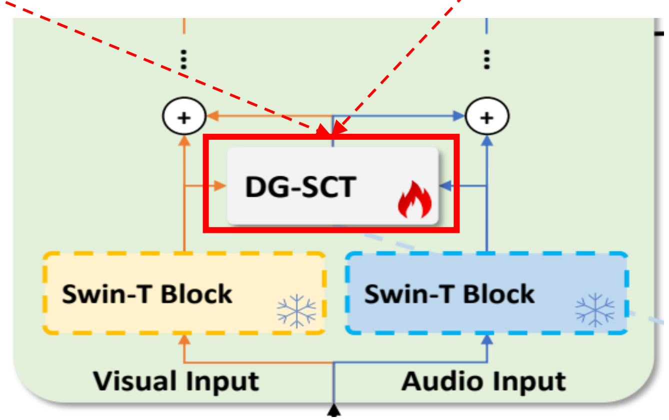
02-2 DG-SCT

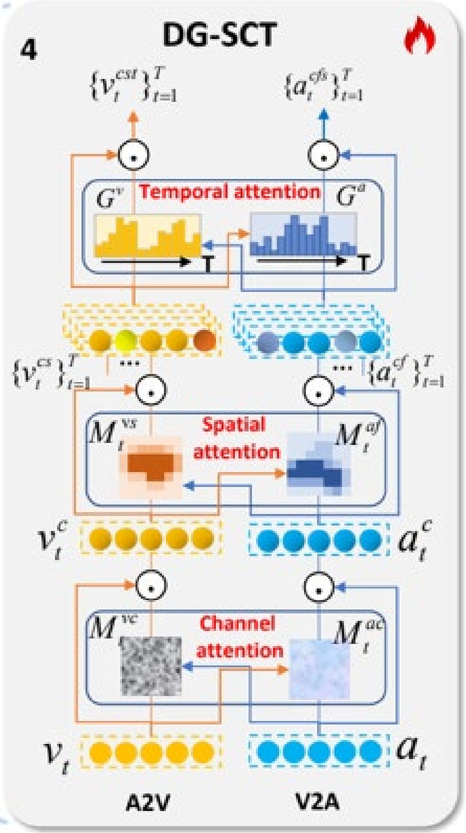
浙江大学 ZHEJIANG UNIVERSITY

$$v_t^{cs} = (\alpha \cdot M_t^{vc} + \beta \cdot M_t^{vs} + 1) \odot v_t, \quad a_t^{cf} = (\alpha \cdot M_t^{ac} + \beta \cdot M_t^{af} + 1) \odot a_t,$$

$$\{v_t^{cst}\}_{t=1}^{T} = (\gamma \cdot G^v + 1) \odot \{v_t^{cs}\}_{t=1}^{T}, \quad \{a_t^{cft}\}_{t=1}^{T} = (\gamma \cdot G^a + 1) \odot \{a_t^{cf}\}_{t=1}^{T},$$

$$\Omega^{\mathrm{a2v}}(\{a_t\}_{t=1}^{T}, \{v_t\}_{t=1}^{T}) \qquad \Omega^{\mathrm{v2a}}(\{v_t\}_{t=1}^{T}, \{a_t\}_{t=1}^{T})$$





03

Experiments

03-2 Audio-visual downstream tasks

03-3 Few-shot/zero-shot tasks

03-1 Audio-visual downstream tasks

浙江大学 ZHEJIANG UNIVERSITY

- **AVE**

Method	Visual Encoder	Audio Encoder	Acc
AVEL(Audio-Visual) [30]	VGG-19	VGG-like	71.4
AVEL(Audio-Visual+Att) [30]	VGG-19	VGG-like	72.7
AVSDN [15]	VGG-19	VGG-like	72.6
CMAN [35]	VGG-19	VGG-like	73.3
DAM [32]	VGG-19	VGG-like	74.5
CMRAN [34]	VGG-19	VGG-like	77.4
CMBS [33]	VGG-19	VGG-like	79.3
LAVisH [16]	Swin-V2-L (shared)		81.1
LAVisH [16]	Swin-V2-L (shared)		79.7
LAVisH*[2]	Swin-V2-L	HTS-AT	78.6
Ours	Swin-V2-L	HTS-AT	**82.2**

- **AVVP**

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
AVE [30]	49.9	37.3	37.0	41.4	43.6	43.6	32.4	32.6	36.2	37.4
AVSDN [15]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [29]	60.1	52.9	48.9	54.0	55.4	**51.3**	48.9	43.0	47.7	48.0
MGN [22]	**60.7**	55.5	50.6	55.6	**57.2**	51.0	52.4	44.4	49.3	**49.2**
Ours	59.0	**59.4**	**52.8**	**57.1**	57.0	49.2	**56.1**	**46.1**	**50.5**	49.1

03-2 Audio-visual downstream tasks

浙江大学 ZHEJIANG UNIVERSITY

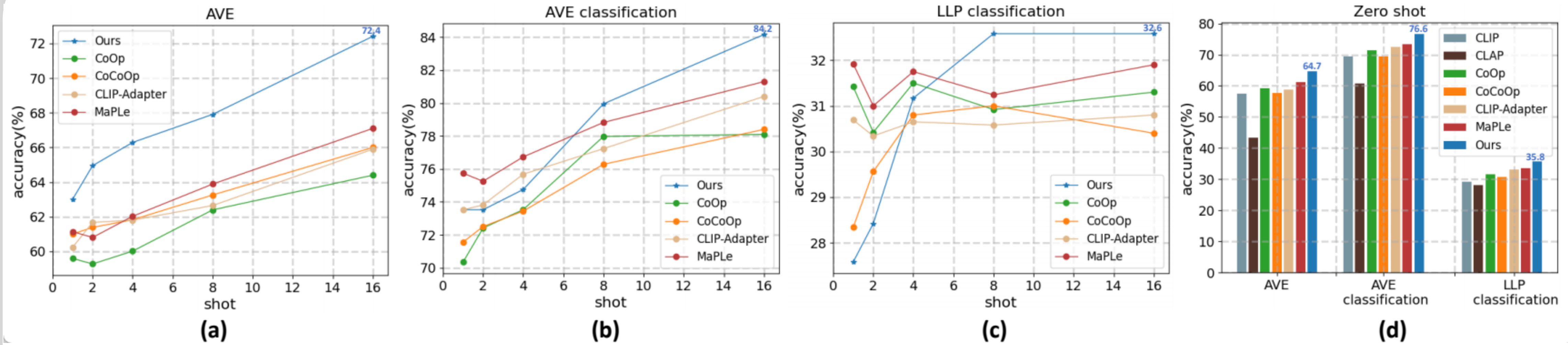
- **AVS**

Method	Visual Encoder	Audio Encoder	Setting			
			S4		MS3	
			$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
AVS [39]	PVT-v2	VGG-like	78.7	87.9	**54.0**	**64.5**
LAVisH [16]	Swin-V2-L (shared)		80.1	88.0	49.8	60.3
LAVisH*	Swin-V2-L	HTS-AT	78.0	87.0	49.1	59.9
Ours	Swin-V2-L	HTS-AT	**80.9**	**89.2**	53.5	64.2

- **AVQA**

Method	Visual Encoder	Audio Encoder	AQ	VQ	AVQ	Avg
AVSD [26]	VGG-19	VGG-like	68.5	70.8	65.5	67.4
Pano-AVQA [36]	Faster RCNN	VGG-like	70.7	72.6	66.6	68.9
ST-AVQA [13]	ResNet-18	VGG-like	74.1	74.0	69.5	71.5
LAVisH [16]	Swin-V2-L(shared)		75.7	80.4	70.4	74.0
LAVisH*	Swin-V2-L	HTS-AT	75.4	79.6	70.1	73.6
Ours	Swin-V2-L	HTS-AT	**77.4**	**81.9**	**70.7**	**74.8**

03-3 Few-shot/zero-shot tasks

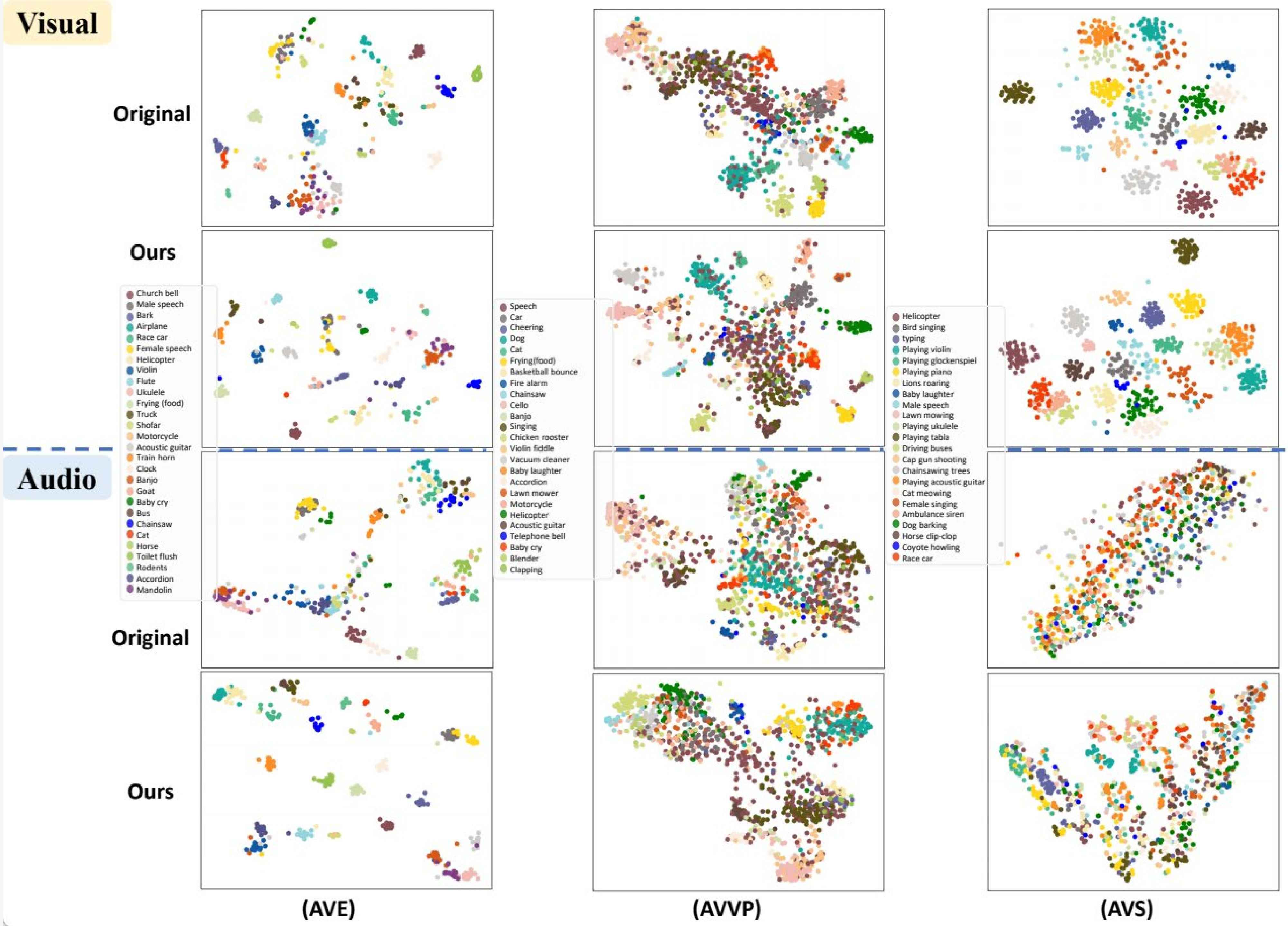


our approach demonstrates robust generalizability and holds potential for application in more audio-visual scenarios in the future.

04

Qualitative analysis

04 Qualitative analysis



浙江大学
ZHEJIANG UNIVERSITY

Cross-modal Prompts

Adapting Large Pre-trained Models for Audio-Visual Downstream Tasks

Thank you

Haoyi Duan