Towards Efficient Audio-Visual Learners via Empowering Pre-trained Vision Transformers with Cross-Modal Adaptation

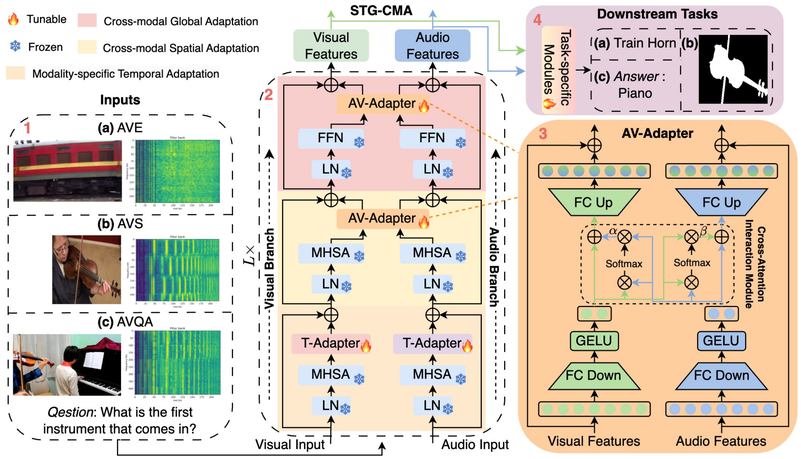
Kai Wang[1], *Yapeng Tian*[2], *Dimitrios Hatzinakos*[1]

[1]*University of Toronto* [2]*University of Texas at Dallas*

Paper

CVPR SEATTLE, WA JUNE 17-21, 2024

Overview of Proposed STG-CMA



Introduction

Motivations: **1.** Pre-training audio and visual encoders on audio-visual pairs requires massive data and expensive computing resources. **2.** Retraining pre-trained audio-visual models on downstream tasks increases the extra training budget and causes overfitting problems. **3.** Existing parameter-efficient methods either neglect the spatial-temporal adaptation of modality-specific signals (i.e. LAVISH) or rely on audio-specific pretraining (i.e. DG-SCT).

Raised Question: *Is it possible to adapt shared pre-trained image-only foundation models for interacting audio-visual modality towards efficient audio-visual representation? The answer is* ***Yes****.*

Contributions:

1. Propose an Audio-Visual Adapter to interact with audio and video.
2. Propose a spatial-temporal-global cross-modal adaptation (STG-CMA) to empower the frozen ViTs to efficiently learn spatial, temporal, and global information of audio-visual modality.
3. STG-CMA achieves state-of-the-art performance in AVE, AVS, and AVQA tasks while involving reduced trainable parameters.

Proposed Method

AV Adapter: consists of two parallelly connected vanilla adapters with a cross-attention interaction module for multimodal interaction.

Modality-specific Temporal Adaptation: adapt the frozen ViTs to capture the temporal information of audio and video signals.

Cross-modal Spatial Adaptation: adapt the frozen ViTs to interact with audio and video for learning the refined temporal information with the cue from the counterpart modality.

Cross-modal Global Adaptation: adapt the frozen ViTs for global interaction between audio and visual branches.

- ❖ Both audio and visual branches in each adaptation stage share the same weights from frozen pre-trained ViT.
- ❖ Only added adapters and task-specific modules are trainable while all other layers are frozen.

Experimental Results

Table1: Comparison with state-of-the-art AVE methods

Method	Encoder		Pretrain Dataset		Trainable Param ↓	Total Param ↓	Acc ↑
	Visual	Audio	Visual	Audio			
PSP [46]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	**1.7M**	217.4M	77.8%
AVT [23]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	15.8M	231.5M	76.8%
RFJC [7]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	22.8M	238.5M	76.2%
AVEL [37]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	3.7M	136.0M	74.0%
CMRAN [42]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	15.9M	148.2M	78.3%
MM-Pyramid [44]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	44.0M	176.3M	77.8%
CMBS [40]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	14.4M	216.7M	79.7%
AVSDN [24]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	8.0M	140.3M	75.4%
MBT* [30]	ViT-B-16 🔥	AST 🔥	ImageNet	AudioSet	172.0M	172.0M	77.8%
DG-SCT [8]	Swin-L ❄	HTS-AT ❄	ImageNet	AudioSet	43.6M	461.3M	82.2%
LAVISH [25]	ViT-B-16 ❄, shared		ImageNet	✘	4.7M	107.2M	75.3%
	ViT-L-16 ❄, shared		ImageNet	✘	14.5M	340.1M	78.1%
STG-CMA (Tiny), ours	ViT-B-16 ❄, shared		CLIP	✘	**3.5M**	**89.6M**	76.3%
	ViT-L-14 ❄, shared		CLIP	✘	10.7M	324.1M	82.2%
STG-CMA (Base), ours	ViT-B-16 ❄, shared		CLIP	✘	11.5M	97.5M	78.7%
	ViT-L-14 ❄, shared		CLIP	✘	20.1M	323.6M	**83.3%**
LAVISH [25]	Swin-B ❄, shared		ImageNet	✘	5.0M	114.2M	78.8%
	Swin-L ❄, shared		ImageNet	✘	10.1M	238.8M	81.1%
STG-CMA (Tiny), ours	Swin-B ❄, shared		ImageNet	✘	5.6M	92.3M	81.1%
	Swin-L ❄, shared		ImageNet	✘	11.7M	206.7M	82.0%
STG-CMA (Base), ours	Swin-B ❄, shared		ImageNet	✘	10.1M	96.8M	81.4%
	Swin-L ❄, shared		ImageNet	✘	19.0M	214.0M	**82.5%**

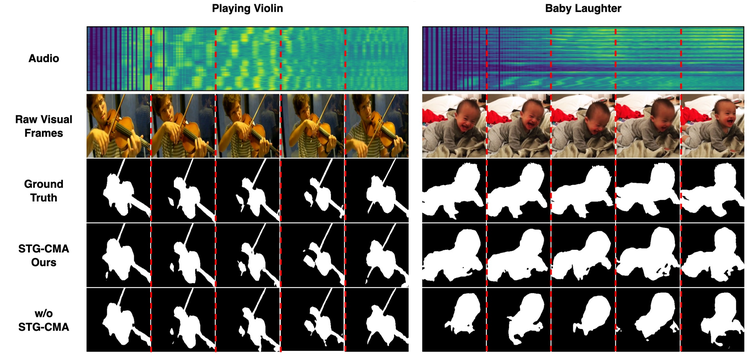
Table2: Comparison with state-of-the-art AVS methods

Method	Encoder		Pretrain Dataset		Trainable Param ↓	Total Param ↓	mIoU ↑
	Visual	Audio	Visual	Audio			
LVS [3]	ResNet-18	ResNet-18	ImageNet	AudioSet	N/A	N/A	37.8%
MMSL [33]	ResNet-18	CRNN	ImageNet	AudioSet	N/A	N/A	44.9%
AVS [47]	PVT-V2 🔥	VGGish ❄	ImageNet	AudioSet	102.4M	174.5M	78.7%
LAVISH [25]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	8.0M	140.3M	75.4%
LAVISH [25]	Swin-L ❄, shared		ImageNet	✘	37.2M	266.4M	80.1%
DG-SCT [8]	Swin-L ❄	HTS-AT ❄	ImageNet	AudioSet	61.5M	594.8M	80.9%
STG-CMA (Base), ours	Swin-B ❄, shared		ImageNet	✘	**29.7M**	**116.5M**	81.0%
	Swin-L ❄, shared		ImageNet	✘	38.6M	233.6M	**81.8%**

Table3: Comparison with state-of-the-art AVQA methods

Method	Encoder		Pretrain Dataset		Trainable Param ↓	Total Param ↓	Question ↑			
	Visual	Audio	Visual	Audio			AQ	VQ	AVQ	Avg
AVSD [35]	VGG-19	ResNet-18	ImageNet	AudioSet	N/A	N/A	68.5%	70.8%	65.5%	67.4%
Pano-AVQA [45]	Faster RCNN	VGGish	ImageNet	AudioSet	N/A	N/A	70.7%	72.6%	66.6%	68.9%
ST-AVQA [19]	ResNet-18 ❄	VGGish ❄	ImageNet	AudioSet	10.6N	94.4M	74.1%	74.0%	69.5%	71.5%
LAVISH [25]	Swin-L ❄, shared		ImageNet	✘	21.1M	249.8M	75.7%	80.4%	70.4%	74.0%
STG-CMA-B (ours)	Swin-L ❄, shared		ImageNet	✘	26.9M	221.9M	**77.1%**	**80.8%**	**70.7%**	**74.5%**
DG-SCT [8]	Swin-L ❄	HTS-AT ❄	ImageNet	AudioSet	110.4M	520.2M	77.4%	81.9%	70.7%	74.8%
STG-CMA-L (ours)	Swin-L ❄, shared		ImageNet	✘	83.1M	278.1M	**78.7%**	**83.0%**	**72.3%**	**76.2%**

Visualization examples on AVS task



Conclusion

- ➢ We proposed an STG-CMA to adapt the frozen pre-trained ViTs to efficiently learn the audio-visual representation without full fine-tuning paradigm and audio-specific pre-training.
- ➢ Our STG-CMA outperforms better than existing SOTA methods in AVE, AVS, and AVQA tasks using fewer tunable parameters.
- ➢ In future, we will explore the robust generalization ability of STG-CMA in more audio-visual scenarios.