

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

Kai Wang
University of Toronto

kaikai.wang@mail.utoronto.ca

Yapeng Tian
University of Texas at Dallas

yapeng.tian@utdallas.edu

Dimitrios Hatzinakos
University of Toronto

dimitris@comm.utoronto.ca

Abstract

In this paper, we explore the cross-modal adaptation of pre-trained Vision Transformers (ViTs) for the audio-visual domain by incorporating a limited set of trainable parameters. To this end, we propose a Spatial-Temporal-Global Cross-Modal Adaptation (STG-CMA) to gradually equip the frozen ViTs with the capability for learning audio-visual representation, consisting of the modality-specific temporal adaptation for temporal reasoning of each modality, the cross-modal spatial adaptation for refining the spatial information with the cue from counterpart modality, and the cross-modal global adaptation for global interaction between audio and visual modalities. Our STG-CMA presents a meaningful finding that only leveraging the shared pre-trained image model with inserted lightweight adapters is enough for spatial-temporal modeling and feature interaction of audio-visual modality. Extensive experiments indicate that our STG-CMA achieves state-of-the-art performance on various audio-visual understanding tasks including AVE, AVS, and AVQA while containing significantly reduced tunable parameters. The code is available at <https://github.com/kaiw7/STG-CMA>.

1. Introduction

Audio-visual learning emerges as a flourishing field to simultaneously learn from both visual and auditory modalities, enabling the intelligent systems to imitate human perception for hearing and seeing the surrounding environment [50]. Generally, most audio-visual models separately encode audio and visual features and then aggregate them for various understanding tasks including audio-visual event localization (AVE) [37], audio-visual segmentation (AVS) [47], audio-visual question answering (AVQA) [19], etc. To obtain promising performance, most works either perform self-supervised learning on massive synchronized audio-visual pairs [11, 14, 41] or leverage individual audio and visual encoders pre-trained on the modality-specific data [19, 37, 40, 45, 47]. However, pre-training modality-

specific encoders commonly consumes massive paired data, expensive computing resources, and unaffordable training burden. Meanwhile, retraining such pre-trained parameters further increases the extra training budget and potentially causes the degradation of well-generalized knowledge and overfitting problems on small downstream datasets.

Recently, transformer-based foundation models pre-trained on massive data are considered as the foundation for various downstream tasks, facilitating various research fields like natural language processing (NLP) [2, 5, 31], computer vision (CV) [1, 6, 27, 38], and multimodal learning [10, 20, 21, 34]. Inspired by the remarkable generalization ability of such transformer-based architectures [39], some works [4, 9, 22, 32, 36, 43, 48] explore the knowledge transfer from pre-trained foundation models into different downstream tasks by parameter-efficient transfer learning (PETL) [12, 13, 15], where only newly-introduced lightweight parameters are tunable while maintaining the pre-trained weights frozen. However, most of them concentrate on either close-domain (*i.e.* image-to-image or image-to-video) [4, 32, 43] or vision-language domain [16, 17, 36, 49], neglecting the audio-visual multimodal scenario due to the lack of pre-trained audio-visual foundation models. To bridge this gap, LAVISH is proposed to leverage the pre-trained vision models to learn from audio and visual modalities by only keeping the newly-added lightweight parameters trainable. Differs from LAVISH only interacts with the audio-visual spatial features, Duan *et al.* [8] incorporated learnable attention modules into pre-trained vision- and audio-specific encoders for cross-modal feature interaction along spatial, temporal, and channel dimension, yielding impressive performance in downstream tasks while introducing audio-specific pre-training and more model complexity. Hence, an important question arises: *Is it possible to adapt pre-trained image-only foundation models for interacting spatial and temporal features of audio-visual modality towards efficient audio-visual representation?* Our findings yield an affirmative response to this question.

In this paper, we propose a spatial-temporal-global cross-modal adaptation (STG-CMA) to directly general-

ize the frozen vision foundation models (*i.e.* ViTs) into the audio-visual domain while only involving partial trainable parameters. To achieve this objective, our STG-CMA adopts modality-specific temporal adaptation, cross-modal spatial adaptation, and cross-modal global adaptation to gradually equip the pre-trained ViTs with the capability of temporal, spatial, and global reasoning. First, in the temporal adaptation, the pre-trained image self-attention layer is reused and combined with the T-adapter to model the relationship across different frames of audio-visual input along the temporal dimension. Then for the spatial adaptation, the pre-trained image self-attention layer followed by the proposed AV-Adapter is adopted to interact with audio and visual modalities for refining the spatial information with the cue from the counterpart modality. Finally, global adaptation further adds the global cross-modal interaction of audio-visual modality by equipping the pre-trained image feed-forward layer with the proposed AV-Adapter. To summarize, this work makes the following contributions:

1. We propose an Audio-Visual Adapter to interact with audio and visual modalities by implementing parameter-free cross-modal attention on compressed hidden features.
2. We 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, overcoming the challenges of spatial-temporal reasoning and feature interaction by exclusively relying on the vision backbone.
3. Experimental results on three audio-visual understanding tasks including AVE, AVS, and AVQA demonstrate that our proposed STG-CMA outperforms the state-of-the-art methods while involving the reduced trainable parameters.

2. Related Works

2.1. Audio-visual Understanding

Audio-visual understanding tasks aim to explore both audio and visual modalities for perceiving the audio-visual scenarios. For instance, audio-visual event localization (AVE) requires models to localize joint audio-visual events [37]. Previous works [7, 23, 40, 46] mainly rely on pre-trained modality-specific models to extract visual and audio features which are then aggregated by a fusion module for prediction. Audio-visual segmentation (AVS) is a new task to predict masks corresponding to sounding objects in visual scenes [26, 29, 47]. Authors of [47] introduce the audio semantics into the visual branch via an interaction module for guiding the visual segmentation. Moreover, the task of audio-visual question answering (AVQA) has recently been proposed to answer human-generated questions about audio-visual events by learning both audio and visual modalities [19, 35, 45]. Most existing methods leverage individual pre-trained audio and visual encoders to ex-

tract modality-specific features, which are then aggregated by spatial and temporal grounding modules [19, 35, 45]. Differing from prior methods relying on modality-specific audio and visual encoders, we study how to leverage the frozen pre-trained vision models for audio-visual data without audio-specific encoders.

2.2. Parameter-efficient Transfer Learning

Parameter-efficient transfer learning (PETL) fine-tunes the pre-trained foundation models into various tasks by updating newly inserted parameters while keeping pre-trained models frozen. In general, PETL technologies can be categorized into adapter tuning for introducing lightweight adapter layers into pre-trained models [12], prompt tuning for injecting tunable prompt tokens at input space [15], and low-rank adaptation for learning a low-rank factorization to approximate the model weights [13]. Lin *et al.* [25] proposed LAVISH to adapt the frozen pre-trained ViTs into audio-visual tasks, where the inserted adapters only perform the interaction between audio and visual features while neglecting the spatial-temporal adaptation of modality-specific signals. Besides, one concurrent work [8] considers the semantic interaction of spatial, temporal and channel information while relying on the audio-specific pre-training and involving more model complexity. However, our STG-CMA directly adapts the frozen pre-trained ViTs to perform the spatial-temporal-global reasoning while interacting across audio and visual modalities with reduced tunable parameters, where lightweight adapters constructed from bottleneck fully connected layers are adopted due to their simplicity and efficiency.

3. Proposed Methodology

We propose a spatial-temporal-global cross-modal adaptation (STG-CMA) to adapt frozen ViTs into audio-visual data as shown in Fig. 1, consisting of the modality-specific temporal adaptation, the cross-modal spatial adaptation, and the cross-modal global adaptation.

3.1. Visual and Audio Input

For visual modality, M RGB frames are first uniformly sampled from each video clip and are then concatenated along the time dimension, yielding the video frames $V \in \mathbb{R}^{M \times H \times W \times 3}$. Then, RGB frames are projected into patch embedding as ViTs to attain the visual input $V_{in} \in \mathbb{R}^{M \times (N_v + 1) \times D}$ including a prepended class token, where N_v is the number of image patches. For audio modality, the waveform is first split into K short segments, where each one is processed by the Hanning window to obtain fbank features. Then, fbank features of all audio segments are stacked along temporal dimension to form the audio spectrogram $A \in \mathbb{R}^{K \times T \times F \times 1}$, where T and F mean the

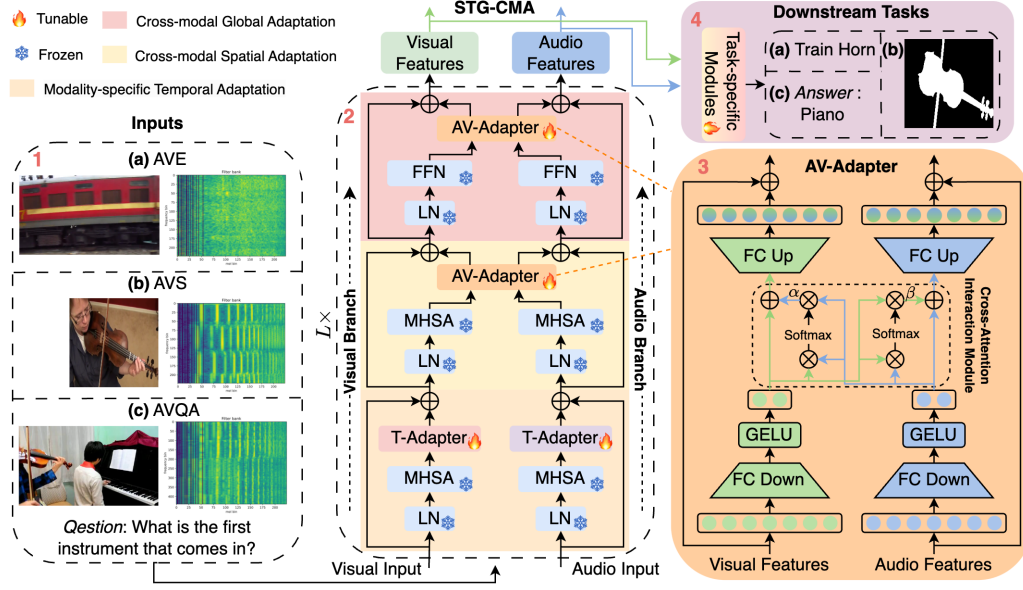


Figure 1. 1. Audio-visual inputs; 2. Overview of proposed STG-CMA; 3. Details of AV-Adapter; 4. Downstream audio-visual tasks. The LN, MHSA and FFN mean the layer normalization, multi-head self-attention, and feed-forward network in the ViT block. Note that, both audio and visual branches in each adaptation stage share the same weights of frozen pre-trained ViT. During training, only added adapters and task-specific downstream modules are trainable while all other layers are frozen.

temporal and frequency dimension. Like visual modality, the audio spectrogram is converted into audio inputs $A_{in} \in \mathbb{R}^{K \times (N_a+1) \times D}$ including an inserted class token by patching projection, where N_a is the number of audio patches. In our implementation, we directly reuse the patch embedding layer of frozen ViTs for the visual input. However, for audio input, we average the weights of the frozen patch embedding layer along the channel dimension to accommodate the layout of the audio spectrogram.

3.2. Audio-Visual Adapter

The audio-visual adapter (AV-Adapter) is proposed to achieve the interaction between audio and visual modalities, which consists of two parallelly connected vanilla adapters and the cross-attention interaction module (CAIM) in between as shown in Fig. 1. Different from the work [16] using adapters with weight-sharing mechanisms for multi-modal interaction, our AV-Adapter adopts the parameter-free CAIM on the compressed hidden features, involving less computation complexity and model parameters. First, AV-Adapter maps the input audio \tilde{A} and visual features \tilde{V} into compressed hidden features \tilde{A}_h and \tilde{V}_h , respectively, by using down-sampling projection layer followed by GELU non-linear activation operation. Then, the \tilde{A}_h and \tilde{V}_h undergo the CAIM for information interaction between audio and visual modalities, leading to the refined audio hidden features \tilde{A}'_h with visual information and refined visual hidden features \tilde{V}'_h with audio information. Next, \tilde{A}'_h and \tilde{V}'_h separately pass through two different up-sampling

projection layers to generate the output audio feature \tilde{A}_{out} and visual feature \tilde{V}_{out} , where the residual connection is injected into each modality branch for avoiding the overfitting issue. The whole procedure of AV-Adapter can be written as:

$$\tilde{A}_{out}, \tilde{V}_{out} = AV-Adapter(\tilde{A}, \tilde{V}) \quad (1)$$

$$\tilde{A}_h = GELU(\tilde{A}W_d^a), \quad \tilde{V}_h = GELU(\tilde{V}W_d^v) \quad (2)$$

$$\tilde{V}'_h = \tilde{V}_h + \alpha \cdot Softmax(\tilde{V}_h \tilde{A}_h^T) \tilde{A}_h \quad (3)$$

$$\tilde{A}'_h = \tilde{A}_h + \beta \cdot Softmax(\tilde{A}_h \tilde{V}_h^T) \tilde{V}_h \quad (4)$$

$$\tilde{A}_{out} = \tilde{A} + \tilde{A}'_h W_u^a, \quad \tilde{V}_{out} = \tilde{V} + \tilde{V}'_h W_u^v \quad (5)$$

where α and β are the trainable weights to control the information flow from one modality from the other one, $W_d^a, W_d^v \in \mathbb{R}^{D \times r}$ are down-sampling weights for audio and visual branches, and $W_u^a, W_u^v \in \mathbb{R}^{r \times D}$ are up-sampling weights for audio and visual branches, r is the dimension of compressed hidden features.

3.3. Modality-specific Temporal Adaptation

The modality-specific temporal adaptation is proposed to adapt the frozen ViTs to capture the temporal information of audio and visual signals as shown in Fig 1. First, the audio A_{in} and visual inputs V_{in} are first reshaped into $A_t \in \mathbb{R}^{(N_a+1) \times K \times D}$ and visual input $V_t \in \mathbb{R}^{(N_v+1) \times M \times D}$. Then, the temporal adaptation is applied in A_t and V_t to learn temporal dependency by a shared frozen LN, a shared frozen MHSA, and an individually tunable T-Adapter, operating on the dimension K and M . The whole operation is

defined as:

$$A_{t,o}^{l-1} = A_t^{l-1} + T\text{-Adapter}(MHSA(LN(A_t^{l-1}))) \quad (6)$$

$$V_{t,o}^{l-1} = V_t^{l-1} + T\text{-Adapter}(MHSA(LN(V_t^{l-1}))) \quad (7)$$

where $A_{t,o}^{l-1}$ and $V_{t,o}^{l-1}$ are output features of modality-specific temporal adaptation in $(l-1)$ th ViT blocks, and $T\text{-Adapter}(\cdot)$ denotes the operation of temporal adapter, whose structure is the same as vanilla adapter constructing from bottleneck MLP.

3.4. Cross-modal Spatial Adaptation

The cross-modal spatial adaptation is proposed to adapt the frozen ViTs to interact with audio and visual modalities for learning the refined spatial information with the cue from the counterpart modality. First, the output audio and visual features from temporal adaptation are permuted into $A_s \in \mathbb{R}^{K \times (N_a+1) \times D}$ and $V_s \in \mathbb{R}^{M \times (N_v+1) \times D}$, respectively. Then, the shared frozen LN and MHSA followed by a learnable AV-adapter are implemented on A_s and V_s to attain the refined audio spatial features \tilde{A}_s and refined visual spatial features \tilde{V}_s , where the CAIM is used to perform the interaction between audio and visual modalities on compressed feature level. The procedure can be formulated as:

$$A_{s,med}^{l-1} = MHSA(LN(A_s^{l-1})) \quad (8)$$

$$V_{s,med}^{l-1} = MHSA(LN(V_s^{l-1})) \quad (9)$$

$$\tilde{A}_s^{l-1}, \tilde{V}_s^{l-1} = AV\text{-Adapter}(A_{s,med}^{l-1}, V_{s,med}^{l-1}) \quad (10)$$

$$A_{s,o}^{l-1} = A_s^{l-1} + \tilde{A}_s^{l-1}, \quad V_{s,o}^{l-1} = V_s^{l-1} + \tilde{V}_s^{l-1} \quad (11)$$

where $A_{s,med}^{l-1}$ and $V_{s,med}^{l-1}$ are intermediate audio and visual features from frozen MHSA in $(l-1)$ th ViT block, and $A_{s,o}^{l-1}$ and $V_{s,o}^{l-1}$ are output features from spatial adaptation in $(l-1)$ th ViT block.

3.5. Cross-modal Global Adaptation

Cross-modal global adaptation aims to add the capability of global interaction into frozen ViTs. After temporal and spatial adaptations, we attain the spatial-temporal global features from both audio and visual modalities (*i.e.*, A_{global} and V_{global}). Afterwards, the global audio and visual features are successively processed by a frozen LN, a frozen FFN and a tunable AV-Adapter for interacting with the spatial-temporal information between the audio and visual branches. The computation of cross-modal global adaptation can be written as:

$$A_{global,med}^{l-1} = FFN(LN(A_{global}^{l-1})) \quad (12)$$

$$V_{global,med}^{l-1} = FFN(LN(V_{global}^{l-1})) \quad (13)$$

$$\tilde{A}_{global}^{l-1}, \tilde{V}_{global}^{l-1} = AV\text{-Adapter}(A_{global,med}^{l-1}, V_{global,med}^{l-1}) \quad (14)$$

$$A^l = A_{global}^{l-1} + \tilde{A}_{global}^{l-1}, \quad V^l = V_{global}^{l-1} + \tilde{V}_{global}^{l-1} \quad (15)$$

where $A_{global,med}^{l-1}$ and $V_{global,med}^{l-1}$ mean the output audio and visual features from frozen FFN in $(l-1)$ th ViT block, A^l and V^l denote the final audio and visual output of $(l-1)$ th ViT block, respectively.

4. Experiments

4.1. Downstream Tasks and Datasets



The **audio-visual event localization (AVE)** task aims to predict the audio-visual events across multiple temporal segments within a video. We evaluate our STG-CMA on the AVE dataset [37] comprising 4, 143 video clips with the audio stream, where each one has a duration of 10 seconds and is labelled with events every second. The **audio-visual segmentation (AVS)** task is to predict the pixel-level segmentation map of sounding objects at the image frame. We conduct the experiments on the AVSBench-S4 dataset [47], including 4, 932 videos with the manual annotations of audible objects. The **audio-visual question answering (AVQA)** task is recently proposed to answer questions based on the objects and their associated sounds. We validate our STG-CMA on the MUSIC-AVQA dataset [19], containing 9, 288 video clips and 45, 867 question-answer pairs. For all audio-visual tasks, we report the performance score on the testing split.

































Table 1. Bottleneck ratios in ‘Tiny’ and ‘Base’ adapter configurations. The CLIP ViT blocks use the same ratio value while the Swin ViT assigns a separate value in its four blocks

Backbone	Bottleneck Ratio (α)	
	Tiny	Base
CLIP-based ViT	1/32	1/16
Swin-based ViT	[1/8, 1/8, 1/16, 1/16]	[1/2, 1/4, 1/8, 1/16]

4.2. Experimental Setups

To adapt our proposed framework to the audio-visual understanding tasks, we adopt the frozen pre-trained CLIP or Swin vision transformers as the backbone equipped with our STG-CMA to extract audio and visual features. Meanwhile, our STG-CMA has two variants (*i.e.*, Base and Tiny) by setting the different bottleneck ratios of introduced adapters as referred in Table 1. For the AVE task, both CLIP [34] and Swin transformer [27] backbone attached with our STG-CMA are used to extract audio and visual features which are concatenated to undergo a classifier for a final event prediction. To assess the performance on the AVE, we follow existing works [25, 40] to employ the classification accuracy of multi-class events over the entire video. For the AVS task, we replace the pre-trained visual and audio encoders of original baseline models with a frozen Swin vision transformer equipped with our proposed STG-CMA. We follow the evaluation metrics of baseline model [47] to calculate the mean Intersection-over-Union (mIoU) between the predicted segmentation mask and ground truth

Table 2. Comparison with SOTA methods on AVE dataset. The  and  denote the frozen and trainable parameters of visual or audio encoder, respectively. The **X** indicates that the datasets are not used for pre-training. The * represents the method re-implemented by [25].

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	X	4.7M	107.2M	75.3%
	ViT-L-16  , shared		ImageNet	X	14.5M	340.1M	78.1%
	ViT-B-16  , shared		CLIP	X	3.5M	89.6M	76.3%
STG-CMA (Tiny), ours	ViT-L-14  , shared		CLIP	X	10.7M	324.1M	82.2%
	ViT-B-16  , shared		CLIP	X	11.5M	97.5M	78.7%
STG-CMA (Base), ours	ViT-L-14  , shared		CLIP	X	20.1M	323.6M	83.3%
LAVISH [25]	Swin-B  , shared		ImageNet	X	5.0M	114.2M	78.8%
	Swin-L  , shared		ImageNet	X	10.1M	238.8M	81.1%
STG-CMA (Tiny), ours	Swin-B  , shared		ImageNet	X	5.6M	92.3M	81.1%
	Swin-L  , shared		ImageNet	X	11.7M	206.7M	82.0%
STG-CMA (Base), ours	Swin-B  , shared		ImageNet	X	10.1M	96.8M	81.4%
	Swin-L  , shared		ImageNet	X	19.0M	214.0M	82.5%

mask. For the AVQA task, we substitute the audio-visual feature extractor of baseline model [19] with a frozen Swin vision transformer augmented by our proposed STG-CMA while maintaining the original text encoder and pre-trained grounding modules. Following the baseline AVQA work, we adopt the answer prediction accuracy as the assessment criteria. For training, We utilize the Adam optimizer [18] to train our proposed framework for 20 epochs, where the weight decay is set as $5e-7$ and the momentum parameter is configured as (0.95, 0.999). Meanwhile, we use the CrossEntropy loss (for AVE and AVQA) and IoU loss (for AVS) as the objective functions and the Cosine Decay [28] as the learning rate scheduling. More implementation details are discussed in the Supplementary Material.

4.3. Experimental Results

Audio-visual Event Localization: As shown in Table 2, we compare our STG-CMA with existing state-of-the-art methods on the AVE dataset. In summary, our STG-CMA achieves superior performance than existing methods while involving significantly reduced trainable parameters. First, the ViT-B-16 backbone equipped with our Tiny STG-CMA attains a comparable accuracy performance (76.3%) while having significantly fewer trainable parameters (3.5M) than most existing methods. When adopting Base adapter configuration, our STG-CMA with ViT-B-16 backbone achieves 78.7% accuracy, outperforming most existing methods like MBT* [30] (77.8%) and LAVISH [25] (78.1%) using the same ViT structure. It is worth mentioning that the MBT* follows a full finetuning pipeline and involves more parameters than ours (172.0M vs 11.5M).

Next, when switching to a larger ViT-L-14 backbone, our STG-CMA (Tiny) yields better accuracy than LAVISH using Swin-L (82.2% vs 81.1%). Although DG-SCT achieves the same performance as ours, it adopts audio-specific pre-training and contains much more tunable parameters than ours (43.6M vs 10.7M). Furthermore, when using the Base adapter configuration, our STG-CMA improves the accuracy from 82.2% to 83.3%, achieving a new SOTA on the AVE dataset.

Second, our proposed STG-CMA also presents an impressive performance on the AVE dataset when switching to adopt the Swin ViT as the frozen backbone as indicated in Table 2. For instance, our STG-CMA (Tiny) with Swin-B achieves 81.1% accuracy on AVE showing a competitive or even better performance than existing methods. Furthermore, LAVISH with Swin-L setting obtains the same performance, but it possesses about 2 times more tunable parameters (10.1M vs 5.6M). When replacing the tiny configuration with base one for proposed adapters, our Swin-B based model further achieves a 0.03% absolute accuracy improvement, indicating that the capacity of adaptation has a crucial influence on performance. Once a large and powerful Swin-L backbone is applied to our method, our Swin-B with Tiny and Base adapter configuration can attain 82.0% and 82.5% accuracy, respectively, showing that our parameter-efficient adaptation is benefited by powerful foundation ViT models. Compared with STG-CMA (Base) using ViT-L-14, our STG-CMA (Base) employing Swin-L presents a slight performance drop (from 83.3% to 82.5%) while possessing fewer trainable parameters (19.0M) and fewer total parameter (214.0M), achieving a good trade-off

Table 3. 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%
LAVIS _H [25]	ResNet-152 🌸	VGGish 🌸	ImageNet	AudioSet	8.0M	140.3M	75.4%
LAVIS _H [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%

Table 4. 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%
LAVIS _H [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%

between performance and model complexity.

Audio-visual Segmentation: We also compare our STG-CMA with existing AVS methods on the AVSBench-S4 dataset, where the Swin transformers adapted by the proposed adaptation are adopted for extracting audio-visual features. As indicated in Table 3, we observe that our proposed framework outperforms all existing state-of-the-art methods. First, our STG-CMA (Base) using Swin-B or Swin-L backbones achieves an impressive performance on AVS (81.0% vs 81.8%), showing our proposed adaptation can efficiently adapt the frozen ViTs into complex dense prediction tasks like segmentation. Second, our proposed STG-CMA (Base) with Swin-L backbone attains a better performance than the previous best method (81.8% vs 80.9%) while involving fewer trainable parameters (38.6M vs 61.5M). In addition, compared with DG-SCT using the extra audio-specific pre-training, our STG-CMA adopts the same vision transformer backbone in both visual and audio encoders, achieving efficient knowledge transfer from pre-trained image models into the audio-visual domain.

Audio-visual Question Answering: Finally, we compare our STG-CMA with existing AVQA methods on the MUSIC-AVQA dataset, including audio questions (AQ), visual questions (VQ), and audio-visual questions (AVQ). We adopt a frozen Swin-L backbone equipped with our STG-CMA (Base) as the audio-visual feature extractor and utilize the same grounding modules from LAVISH and DG-SCT, yielding two different model variants (*i.e.* STG-CMA-B and STG-CMA-L) for fair comparisons. As shown in Table 4, our STG-CMA-B achieves better performance than LAVISH on all question types including AQ (77.1% vs 75.7%), VQ (80.8% vs 80.4%), and AVQ (70.7% vs 70.4%). Besides, our proposed STG-CMA-L outperforms

Table 5. Comparison with reference models using various configurations on AVE dataset. ‘A2V’ and ‘V2A’ mean ‘Audio-to-Visual’ and ‘Visual-to-Audio’, respectively.

Model	Audio Adapt	Visual Adapt	A2V Fusion	V2A Fusion	Trainable Param	Acc
Single-Modality Branch						
Audio-only	✓	✗	✗	✗	9.6M	63.7%
Visual-only	✗	✓	✗	✗	9.6M	80.8%
Audio-Visual Branch with Signal-Modality Adaptation						
Audio-only	✓	✗	✗	✗	10.6M	64.3%
Visual-only	✗	✓	✗	✗	10.6M	82.1%
Interaction Variants in AV-Adapter						
Audio-to-Visual	✓	✓	✓	✗	20.1M	82.9%
Visual-to-Audio	✓	✓	✗	✓	20.1M	82.2%
Without Interaction	✓	✓	✗	✗	20.1M	82.4%
Finetuning Scheme						
Full Finetuning	✗	✗	✗	✗	606.8M	57.6%
STG-CMA (ours)	✓	✓	✓	✓	20.1M	83.3%

the leading DG-SCT baseline on AQ (78.7% vs 77.4%), VQ (83.0% vs 81.9%), and AVQ (72.3% vs 70.7%) question types, respectively, achieving a new state-of-the-art performance on AVQA task. Moreover, our proposed STG-CMA contains significantly fewer tunable parameters than DG-SCT (83.1M vs 110.4 M) and avoids the specific audio pre-training, revealing the proposed adaptation can efficiently equip the frozen vision transformers with the capability of audio-visual learning.

4.4. Ablation Studies

In this section, we conduct some ablation studies to investigate how the performance of our proposed STG-CMA is affected by various configurations.

Configuration Variants. In Table 5, we present some

Table 6. Comparison results with different AVQA methods on the MUSIC-AVQA dataset, where different types of questions are presented, like audio-only, visual-only, and audio-visual questions

Method	Audio Question (%)			Visual Question (%)			Audio-Visual Question (%)						Overall Avg. (%)
	Counting	Comparative	Avg.	Counting	Location	Avg.	Existential	Location	Counting	Comparative	Temporal	Avg.	
AVSD [35]	72.4	61.9	68.5	67.4	74.2	70.8	81.6	58.8	63.9	61.5	61.4	65.5	67.4
Pano-AVQA [45]	74.4	64.6	70.7	69.4	75.7	72.6	81.2	59.3	64.9	64.2	63.2	66.6	68.9
ST-AVQA [19]	78.2	67.1	74.1	71.6	76.4	74.0	81.8	64.5	70.8	66.0	63.2	69.5	71.5
STG-CMA-B (ours)	83.1	67.1	77.1	80.4	81.1	80.8	81.6	65.9	74.4	64.0	66.4	70.7	74.5
STG-CMA-L (ours)	84.8	68.2	78.7	81.5	84.5	83.0	81.3	68.5	76.8	65.8	67.4	72.3	76.2

findings to verify the effectiveness of our STG-CMA with various configurations, where all experiments are conducted on the AVE dataset using our STG-CMA (Base) with CLIP ViT-L-14 backbone. First, to investigate the importance of audio-visual multimodal features on performance, we design two reference models with audio-only or visual-only branches, respectively. From Table 5, we observe that removing either the audio or visual branch may lead to an apparent drop in performance, showing that audio-visual features are crucial for promising performance. Meanwhile, the model with a visual-only branch (80.8%) obtains a better performance than the one with an audio-only branch (63.7%), implying that the visual modality dominates the performance on the event localization. Second, we exploit the effect of our proposed adaptations on the performance by only removing all adapters from the visual or audio branch. As shown in Table 5, we can find that either visual-only or audio-only adaptation based models degrade the performance of our STG-CMA using both audio and visual adaptation from 83.3% into 82.1% and 64.3%, respectively, revealing that visual adapters can efficiently adapt the frozen ViTs to learn the spatial-temporal visual features due to the modality consistency.

Third, to explore the benefits of audio-visual interaction in our AV-Adapter, we compare our STG-CMA with three reference models using audio-to-visual fusion, visual-to-audio fusion, and fusion-free choices as shown in Table 5. We can observe that our STG-CMA with cross-modality fusion outperforms other comparison models by efficiently refining the audio and visual features. After omitting the cross-modal attention from all AV-Adapters, the performance will be dropped from 83.3% to 82.4%, presenting that the cross-modal interaction is crucial for capturing the audio-visual representation. In addition, the reference models with unidirectional fusion present a slight accuracy drop when compared with our STG-CMA with bidirectional fusion. Especially, the model with visual-to-audio fusion is slightly inferior to the one with audio-to-visual fusion (82.2% vs 82.9%), exhibiting that the audio guidance benefits the visual branch more than visual guidance helps the audio branch. In the end, we compare our adapta-

Table 7. Ablation study of our different adapters in modality-specific temporal adaptation, cross-modal spatial adaptation, and cross-modal global adaptation.

Adaptation			Trainable Param	mIoU
Temporal	Spatial	Global		
x	x	x	21.2M	55.9%
✓	x	x	24.7M	80.8%
x	✓	x	28.1M	76.6%
x	x	✓	28.1M	76.1%
✓	✓	x	31.6M	81.4%
x	✓	✓	35.1M	77.0%
✓	x	✓	31.6M	81.5%
✓	✓	✓	38.6M	81.8%

tion scheme with a full finetuning paradigm. From table 5, we find that our proposed STG-CMA performs much better than the full finetuning scheme (83.3% vs 57.6%) while involving significantly fewer tunable parameters (20.1M vs 606.8M). It demonstrates that our spatial-temporal-global cross-modal adaptation can efficiently generalize the frozen ViTs to enhance audio-visual learning with reduced trainable parameters while tackling the knowledge degradation brought by full fine-tuning.

Performance in AVQA with Different Questions. In Table 6, we present more comparison results between our proposed method and existing AVQA methods on the different types of questions in the MUSIC-AVQA dataset. We can find that our proposed STG-CMA-L achieves the best performance in most types of questions, yielding the remarkable efficiency of our proposed adaptation scheme in empowering the frozen image models to solve the question-answering task. Meanwhile, it is interesting that our STG-CMA-B with a smaller configuration still attains competitive performance in answering different sorts of questions in the MUSIC-AVQA dataset.

Adaptation Design. In Table 7, we further conduct more ablation studies to investigate the effect of different components of our proposed adaptation on performance. More specifically, we design different reference models by removing the inserted adapters (*i.e.* T-Adapter in temporal adaptation, AV-Adapters in spatial or global adaptations) to explore the effectiveness of our proposed adaptation modules as shown in Table 7. All experiments are

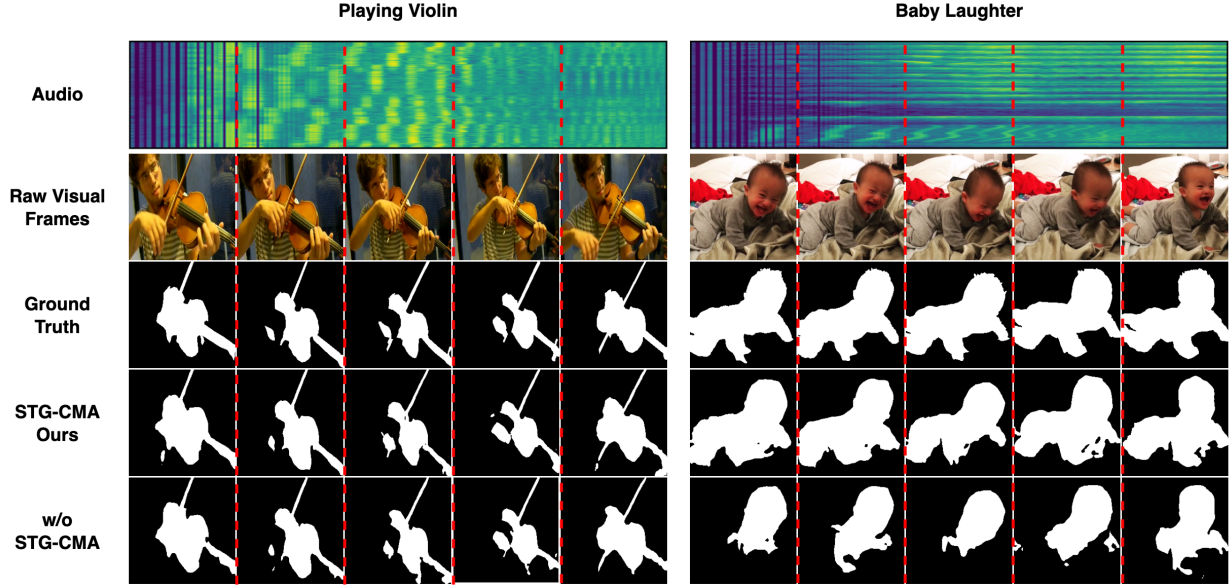


Figure 2. Qualitative examples of our STG-CMA approach and the baseline model (without our STG-CMA) evaluated on the AVS task

implemented in the AVS task by using the Swin-L backbone and Base adapter configuration. First, we observe that the STG-CMA adopting all adaptation modules attains the best mIoU score (81.8%) in the AVE task. Instead, omitting all adaptation modules will lead to the worst performance (66.9%) of STG-CMA on segmentation. It shows that all the proposed adaptation stages are indispensable for adapting frozen image models into the audio-visual domain. Next, when introducing any proposed adaptation module, the STG-CMA will present varying degrees of performance improvement compared with the adaptation-free one. Especially, the STG-CMA with temporal-only adaptation (80.8%) obtains better performance than the one with spatial-only (76.6%) or global-only (76.1%) adaptations, showing that it is crucial to adapt the frozen image models to learn the temporal dependency of video and audio signals. Moreover, the performance will be further improved when inserting any two adaptation modules into the frozen ViT backbone. For instance, the STG-CMA equipped with temporal-spatial or temporal-global adaptations performs better than the one with spatial-global adaptation, indicating that modality-specific temporal adaptation plays an important role in achieving the knowledge transfer from image models into the audio-visual domain.

4.5. Qualitative Results

In this section, we provide some visualization examples of our proposed STG-CMA implemented on AVS task, where the frozen pre-trained Swin-L vision transformer augmented with our proposed adaptations is used for extracting audio-visual features. To evaluate the effectiveness of the proposed STG-CMA, we also design a base-

line model by removing the temporal adaptation stage and inserted adapters in spatial-global adaptation. It means that the baseline model only adopts the frozen pre-trained Swin vision transformer and downstream task-specific layers without our proposed STG-CMA. As shown in Fig. 2, we observe that our STG-CMA can segment better shapes of sounding objects than the baseline model. For example, in the right part of Fig. 2, our proposed STG-CMA correctly locates a laughing baby across whole video frames. However, the baseline model cannot completely outline the region representing baby laughter in each video frame. It indicates that our STG-CMA efficiently empowers the frozen image models to extract the audio-visual features for better segmentation performance.

5. Conclusion

In this paper, we studied how to leverage frozen ViTs pre-trained on image-only data to generalize their learned knowledge into the audio-visual domain. Thereby, we propose a Spatial-Temporal-Global Cross-Modal Adaptation (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. Extensive experiments on audio-visual understanding tasks, including AVE, AVS, and AVQA, indicate that our proposed STG-CMA outperforms the state-of-the-art methods while involving significantly reduced tunable parameters. In the future, we will explore the robust generalization ability of our model in more audio-visual scenarios.

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. **1**
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. **1**
- [3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. **6**
- [4] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. **1**
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. **1**
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2010. **1**
- [7] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4013–4022, 2021. **2, 5**
- [8] Haoyi Duan, Yan Xia, Mingze Zhou, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *arXiv preprint arXiv:2311.05152*, 2023. **1, 2, 5, 6**
- [9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. **1**
- [10] Rohit Girdhar, Alaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. **1**
- [11] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. **1**
- [12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. **1, 2**
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. **1, 2**
- [14] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024. **1**
- [15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. **1, 2**
- [16] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanolin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022. **1, 3**
- [17] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. **1**
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. **5, 1**
- [19] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Jirong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. **1, 2, 4, 5, 6, 7**
- [20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. **1**
- [21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. **1**
- [22] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613, 2023. **1**
- [23] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020. **2, 5**
- [24] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. **5**
- [25] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference*

- on *Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 2, 4, 5, 6, 1
- [26] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5604–5614, 2024. 2
- [27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 4
- [28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5, 1
- [29] Shentong Mo and Bhiksha Raj. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [30] Arsha Nagrai, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 5
- [31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1
- [32] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 1
- [33] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer, 2020. 6
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 4
- [35] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019. 2, 6, 7
- [36] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 1
- [37] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 1, 2, 4, 5
- [38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1
- [40] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19989–19998, 2022. 1, 2, 4, 5
- [41] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1
- [42] Haoming Xu, Runhao Zeng, Qingyao Wu, Minghui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3893–3901, 2020. 5
- [43] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. 1
- [44] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6241–6249, 2022. 5
- [45] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021. 1, 2, 6, 7
- [46] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 2, 5
- [47] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 1, 2, 4, 6
- [48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1
- [49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1
- [50] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18:351–376, 2021. 1

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

Supplementary Material

6. Dataset Details

We evaluate our proposed STG-CMA on three different audio-visual understanding tasks including audio-visual event localization (AVE), audio-visual segmentation (AVS), and audio-visual question answering (AVQA). For all downstream tasks, we summarize the details of the used dataset separately in the following and present some descriptions in Table 8. In addition, we train our model on the training set and report the evaluation score on the testing set.

AVE dataset: We evaluate the performance of our STG-CMA on the AVE task using the AVE dataset [37] that is extracted from AudioSet. The AVE dataset consists of 4, 143 video clips, each one having a duration of 10 seconds and is labeled with corresponding events every 1 second. Meanwhile, the dataset includes 28 different event categories (*i.e.*, Dog Barking, Church Bell, Bus, Violin) and one background category without involving event information. Moreover, the AVE dataset is split into 3,339 video clips for training, 402 video clips for validation, and 402 video clips for testing.

AVSBench-S4 dataset: We validate our proposed STG-CMA on the AVSBench-S4 dataset [47] for the AVS task. The dataset contains 4, 932 video clips, where each one has a duration of 5 seconds. Meanwhile, the AVSBench-S4 dataset includes 23 categories including playing violin, baby laughter, lions roaring, etc. In addition, the pixel-level mask images are provided as the ground truth to represent the objects that generate the sound in the given RGB image frame. Moreover, the dataset is split into 3,452 videos for training, 740 videos for validation, and 740 videos for testing.

MUSIC-AVQA dataset: We conduct the experiments on the MUSIC-AVQA dataset [19] to evaluate the performance of our STG-CMA on the AVQA task. The dataset consists of 9, 288 video clips representing 45, 867 question-answer pairs including 33 question templates spanning over different modality scenarios (*i.e.* *Audio-Visual*, *Audio-only* and *Visual-only*) and question types (*i.e.* *Counting*, *Location*, *Comparative*, *Existential*, *Temporal*). In addition, the MUSIC-AVQA dataset has 42 types of answers for different question contents, such as ‘Yes’ for existential questions, ‘One’ for counting questions, ‘Violin’ for temporal questions, etc.

Table 8. The description of used audio-visual datasets. Each dataset consists of raw video clips, which are then extracted into visual frames and audio waveforms as the audio-visual inputs. The ‘Annotation Type’ row presents whether the frames are annotated by category, pixel-level mask, or answer.

	AVE[37]	AVSBench-S4 [47]	MUSIC-AVQA [19]
Video Clips	4, 143	4, 932	9, 288
Visual Frames	41, 430	24, 660	45, 867
Classes	29	23	42
Annotation Types	Event Category	Pixel-level Mask	Answer
Evaluation Score	Accuracy	mIoU	Accuracy

7. Implementation Details

In this section, we provide more implementation details about data pre-processing, model training and pre-trained vision backbone. We conduct all experiments with one NVIDIA GeForce RTX 3090 GPU by using the PyTorch framework.

7.1. Data Pre-processing

For visual input, following existing works [8, 25], our proposed STG-CMA receives 10 RGB image frames for AVE and AVQA tasks ($M=10$) and 5 RGB image frames for AVS ($M=5$), where the image frames are uniformly sampled from each video clip and each frame is then resized and cropped into the resolution of 224×224 . For audio input, each audio waveform is chunked into 10 short segments for AVE and AVQA tasks ($K=10$) and 5 short segments for AVS task ($M=5$), where each one is converted into either 128-D (for CLIP-based backbone) or 224-D (for Swin-based backbone) fbank features by using Hanning window. In addition, for the AVE task, we follow [43] to adopt stronger data augmentation for visual signal (*i.e.*, random augmentation and random erasing), and follow [25] to use the mix-up augmentation for the audio signal.

7.2. Model Training

We train our proposed model in all audio-visual downstream tasks for 20 epochs by using the Adam [18] optimizer. Besides, we adopt the Cosine Decay [28] to dy-

Table 9. Training configurations and hyperparameters used for different audio-visual understanding tasks

Config	AVE		AVS	AVQA	
Model	STG-CMA (Tiny or Base)		STG-CMA (Base)	STG-CMA-B	STG-CMA-L
Backbone	CLIP-ViT	Swin-ViT	Swin-ViT	Swin-ViT	
Optimizer	Adam				
Adapter LR	5e-5		3e-4	5e-5	2.5e-5
Task-specific LR	5e-6		3e-4	5e-5	2.5e-5
Minimal LR	2e-6		2e-5	5e-6	2e-6
Weight Decay	5e-7				
Optimizer Momentum	(0.95, 0.999)				
Batch Size	1		2	2	
LR Schedule	Cosine Decay				
Warmup Epochs	2		5	2	
Loss Function	CE		IoU	CE	
Mixup	Yes		No	No	
Stronger Augmentation	Yes		No	No	

namically adjust the learning rate during the training procedure. More specifically, the learning rate is first linearly warmed up into initial value within the first several epochs and then decayed into the minimal one with a cosine function. To better train the model, we also assign different learning rates for updating parameters of newly introduced adapter layers and downstream layers for various audio-visual datasets. All training configurations or hyperparameters are summarized in Table 9.

7.3. Pre-trained Vision Transformers

We adopt the off-the-shelf pre-trained vision transformers (*i.e.* CLIP and Swin transformers) as the backbone for both visual and audio encoders. The CLIP consists of vision and text transformer encoders, which are pre-trained on massive image-text pairs by using contrastive learning. We just employ the vision transformer encoder from CLIP as the frozen backbone in our method. In addition, the Swin transformer can efficiently produce the hierarchical feature representation while reducing the computational complexity using shift windows and cross-window attention, which is very useful for dense prediction tasks like audio-visual segmentation.