Prompt Image to Watch and Hear: Multimodal Prompting for Parameter-Efficient Audio-Visual Learning

Kai Wang[1]
kaikai.wang@mail.utoronto.ca
Shentong Mo[2]
shentonm@andrew.cmu.edu
Yapeng Tian[3]
yapeng.tian@utdallas.edu
Dimitrios Hatzinakos[1]
dimitris@comm.utoronto.ca

[1] University of Toronto
Toronto, Canada
[2] Carnegie Mellon University
Pittsburgh, USA
[3] The University of Texas at Dallas
Richardson, USA

Abstract

We explore how to endow the static and pre-trained image models to watch and hear from the audio-video domain with a limited set of trainable parameters. To achieve this objective, we propose an Audio-Visual Spatial-Temporal-Fusion Prompting, called AV-STFP, to gradually adapt the shared and frozen image model to learn audio-visual representation by decoupling the adaptation into temporal-aware prompting, spatial-aware prompting, and multimodal fusion prompting. First, temporal-aware prompting introduces a set of temporal prompts into tokens along temporal frames to adapt the shared pre-trained image layers to capture temporal patterns. Next, spatial-aware prompting inserts a set of spatial prompts into the tokens to enable the shared image layers to learn audio contexts and further enhance the visual semantics. Finally, multimodal fusion prompting incorporates a set of fusion prompts between audio and visual tokens to enforce the pre-trained image layers to integrate two modalities. By only tuning the inserted learnable prompts and freezing the pre-trained image backbones, AV-STFP efficiently transfers the well-generalized image knowledge into the audio-visual domain with minimal costs. Extensive experiments on the various audio-visual understanding tasks indicate that AV-STFP achieves competitive or even superior performance to state-of-the-art methods while involving minimal trainable parameters (*i.e.* 0.6% of model parameters).

1 Introduction

Audio-visual learning aims to build multimodal systems capable of simultaneously processing audio and video information, enabling intelligent systems to perceive their surroundings by both watching and listening, similar to human perception [58]. Existing models typically adopt modality-specific encoders to separately encode modality-specific features and then aggregate them to handle various audio-visual understanding tasks like audio-visual

event localization (AVE) [39], audio-visual segmentation (AVS) [53], audio-visual action recognition (AVAR) [48] and audio-visual question answering (AVQA) [20]. Most works either leverage enormous audio-visual pairs to learn a good representation via self-supervised learning [11, 14, 47] or fine-tune audio-visual encoders that have been independently pre-trained on modality-specific data [20, 39, 46, 52, 53]. However, pre-training audio-visual encoders on massive paired data commonly requires expensive computing resources and an unaffordable training burden. Besides, the fine-tuning stage further increases the training budget and potentially degrades the well-generalized knowledge and overfits on the smaller downstream datasets.

To mitigate this challenge, LAVISH [27] adapted pre-trained image models to interact with audio-visual spatial features using a limited set of adapters. Instead, Duan *et al*. [8] introduced learnable modules to capture spatial, temporal, and channel features using unimodal encoders, involving more parameters and audio-specific encoders. To solve this issue, STG-CMA [45] proposed efficient adapters to adapt shared image pre-trained models for the audio-visual domain, while it lacks efficient fusion due to its parameter-free interaction module and consumes evident trainable parameters. Fortunately, prompt learning [15, 24] is proposed to adapt large foundation models in diverse tasks by updating extremely few learnable prompt tokens (*i.e*. less than 1% of model parameters). CoPL [53] proposed model-specific prompts for unimodality and model-interaction prompts for multimodal fusion. However, it inserts both local and global prompts into each modality token, impeding the pre-trained layers from efficiently learn spatial and temporal features. Besides, using compressed tokens for multimodal fusion would potentially lose modality-specific information. Hence, a natural question arises: *Is it possible to prompt frozen image-only foundation models to efficiently capture semantic and temporal patterns of audio-video modality and perform the multimodal fusion without losing modality-specific information?* Our proposed model affirmatively responds to this question.

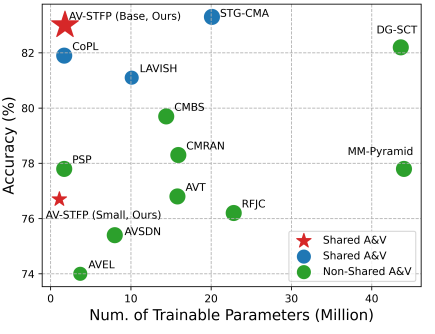

Figure 1: Comparison between AV-STFP and existing methods on AVE task. The 'Shared A&V' and 'Non-Shared A&V' mean audio and visual encoders are shared and non-shared.

Compared with static images, audio possesses unique acoustic characteristics and both audio and video contain rich dynamic information with temporal patterns. It is challenging for pre-trained image models to directly understand audio-visual modalities without additional parameters. Hence, we propose an Audio-Visual Spatial-Temporal-Fusion Prompting called AV-STFP to adapt pre-trained vision models into the audio-visual domain by decoupled adaptations including spatial-aware prompting, temporal-aware prompting, and multimodal fusion prompting. Differ from CoPL, our AV-STFP reuses the shared pre-trained image attention layers in each prompting stage and efficiently interacts the audio-visual modalities without losing modality-specific information. First, the temporal-aware prompting inserts a set of temporal prompts into the key-value tokens along the temporal dimension, enforcing the pre-trained image layers to capture temporal patterns across audio-visual frames. Then, the spatial-aware prompting introduces a set of spatial prompts into the tokens along the spatial dimension to adapt the pre-trained image layers to extract acoustic characteristics

and enhance the visual semantics. Finally, multimodal fusion prompting incorporates a set of fusion prompts between audio and visual tokens to enforce the pre-trained image layers for efficient integration of audio and visual modalities. This work can be summarized as:

1. We propose AV-STFP to prompt the pre-trained image-only models to efficiently learn spatial-temporal information of audio-visual modality and fuse them for feature interactions.

2. AV-STFP adopts temporal-aware prompting, spatial-aware prompting, and multimodal fusion prompting to enable cross-modal adaptation from image to video, audio, and multimodal fusion, enabling shared image models to understand dynamic video and sound.

3. Experiments on audio-visual understanding tasks demonstrate that AV-STFP achieves competitive or even better performance than state-of-the-art methods while only introducing a small amount (0.6% of model parameters) of trainable parameters, as shown in Figure 1.

2 Related Works

Audio-Visual Understanding. Audio-visual understanding tasks aim to leverage both audio and visual modalities to interpret complex audio-visual scenarios. For instance, AVE [39] requires models to identify events that involve both audio and visual components [7, 25, 46, 54]. AVS task seeks to generate masks that identify objects in visual scenes associated with specific sounds [28, 31, 55]. AVAR task aims to predict human activities by involving audio and video signals [17, 48]. AVQA task is designed to enable models to respond to human-generated questions about audio-visual events by learning representations from both modalities [20, 37, 52]. Most existing models predominantly rely on modality-specific models to extract audio-visual features, which are subsequently combined through a fusion module for specific tasks. Differ from that, we explore leveraging a shared frozen pre-trained image foundation model to perceive the audio-visual data without dedicated audio encoders.

Parameter-efficient Audio-Visual Learners. Foundation models have facilitated various fields such as NLP [2, 5, 33, 42], CV [1, 6, 29, 40], and multimodality [10, 21, 22, 35]. Re-training or fine-tuning giant foundation models for customized tasks requires unexpected computation costs. Hence, parameter-efficient transfer learning (PETL) is proposed to transfer the learned knowledge from pre-trained foundation models into various tasks [3, 9, 16, 18, 23, 34, 38, 44, 50, 56, 57] by updating newly added parameters while keeping the foundation backbone frozen. PETL methods generally fall into three categories: adapter tuning [12], prompt tuning [15], and low-rank adaptation (LoRA) [13]. Lin *et al*. [27] proposed LAVISH to adapt pre-trained image models to the audio-visual domain without considering the modality-specific spatial-temporal adaptation that is crucial for sequential modalities. Duan *et al*. [8] proposed the interaction of spatial, temporal and channel information while relying on the specific audio pre-training and involving more model complexity. Afterwards, STG-CMA [45] performed the spatial-temporal-global reasoning by pre-trained image models along with lightweight adapters, while it still requires many trainable parameters. CoPL [53] proposed collaborative prompts to fine-tune both unimodal and multimodal features with reduced parameters. However, its unimodal prompts make the pre-trained model hard to learn spatial-temporal patterns, and its multimodal fusion stage loses modality-specific information. Therefore, our AV-STFP leverages the pre-trained image models to successively learn spatial semantics and temporal dynamics of audio-video features, and aggregate them without losing unimodal information by enabling 0.6% of model parameters to be tunable.

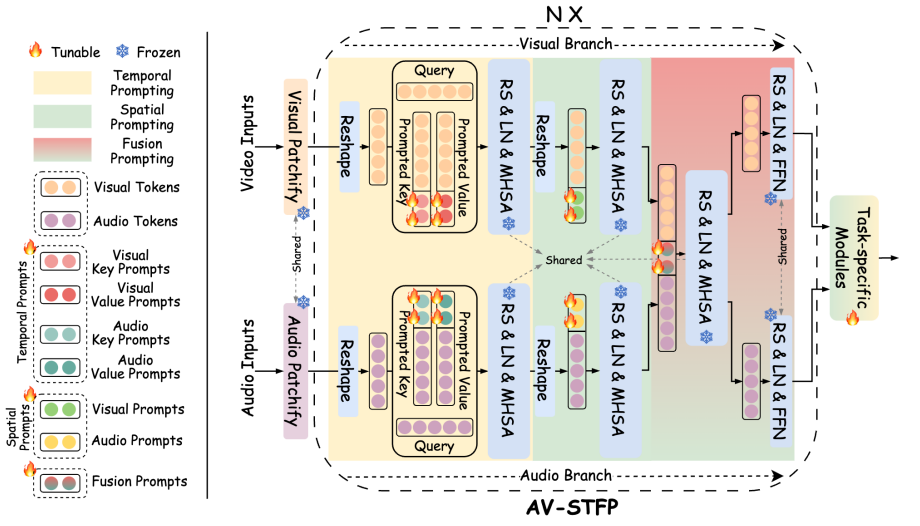

Figure 2: Illustration of our proposed AV-STFP. Our AV-STFP introduces a set of temporal-aware prompts, spatial-aware prompts, and multimodal fusion prompts to original audio-visual tokens, augmenting the shared pre-trained image layers to learn the temporal dependency, semantic information, and multimodal interaction of audio-visual modalities, respectively. During the training, only newly added prompts and task-specific modules are trainable, while the original pre-trained image backbone is frozen.

3 Proposed Methodology

We propose Audio-Visual Spatial-Temporal-Fusion Prompting (AV-STFP) to adapt frozen pre-trained image models (*e.g.*, the CLIP image encoder) to the audio-visual domain. To do so, we decouple the whole adaptation into three stages, including temporal-aware prompting, spatial-aware prompting, and multimodal fusion prompting, and reuse the multihead self-attention (MHSA) from pre-trained image models in each stage, as illustrated in Figure 2. First, the temporal-aware prompting inserts a set of temporal prompts into audio and visual tokens along the temporal dimension to capture the temporal dynamics within each modality, leveraging a shared MHSA block. Next, spatial-aware prompting injects a set of spatial prompts into tokens, enabling the pre-trained MHSA block to transfer its semantic knowledge to audio and video frames. Finally, multimodal fusion prompting positions a set of fusion prompts between audio and visual tokens, allowing the shared MHSA block to integrate information across both modalities effectively.

3.1 Audio-Visual Inputs and Patchifying

Given an input video clip with M RGB frames $V \in R^{M\times H\times W\times 3}$, we first apply a pre-trained patch embedding layer to project the frames into a visual embedding $V_{in} \in R^{M\times(N_v+1)\times D}$, including a prepended class token, where N_v is the number of image patches and D is the embedding dimension. For the audio clip, we segment the audio waveform into M short chunks, each of which is transformed into an fbank feature using the Hanning window. These audio segments are then stacked along the temporal dimension to form the input audio spectrogram $A \in R^{M\times T\times F}$, where T and F represent the temporal and frequency dimensions. Similarly,

a pre-trained patch embedding layer from the image model is applied to convert the audio input into an audio embedding $A_{in} \in R^{M\times(N_a+1)\times D}$, including an inserted class token, where N_a is the number of audio patches. To match the shape of the audio spectrogram, we sum the weights of the frozen visual patch embedding layer along the channel dimension.

3.2 Temporal-Aware Prompting

Unlike static images, audio and video contain rich dynamic information with temporal patterns across frames. Pre-trained image encoders, however, are trained on static images and thus lack the ability to capture audio and temporal information. To adapt these models for temporal dependencies, we introduce temporal-aware prompting by adding a set of key-value prompts into the key-value tokens of audio and visual embeddings, as shown in Figure 2.

First, the audio A_{in} and visual embeddings V_{in} are reshaped into $A_t \in R^{(N_a+1)\times M\times D}$ and visual input $V_t \in R^{(N_v+1)\times M\times D}$. Then, a set of audio key-value prompts $\left\{\mathbf{P}^l_{a,key}, \mathbf{P}^l_{a,val}\right\} \in R^{N^t_p\times D}$ and a set of visual key-value prompts $\left\{\mathbf{P}^l_{v,key}, \mathbf{P}^l_{v,val}\right\} \in R^{N^t_p\times D}$ are prepended into the key-value tokens of audio and visual embeddings along the temporal dimension, prompting a shared and pre-trained MHSA block to learn the temporal consistency among different frames, where N^t_p is the number of temporal-aware prompts. The specific procedure is:

$$\mathbf{K}^{l-1}_{a,t} = [A^{l-1}_t, \mathbf{P}^{l-1}_{a,key}], \quad \mathbf{V}^{l-1}_{a,t} = [A^{l-1}_t, \mathbf{P}^{l-1}_{a,val}], \quad A^l_t = A^{l-1}_t + \beta_a MHSA(LN(A^{l-1}_t, \mathbf{K}^{l-1}_{a,t}, \mathbf{V}^{l-1}_{a,t})) \quad (1)$$

$$\mathbf{K}^{l-1}_{v,t} = [V^{l-1}_t, \mathbf{P}^{l-1}_{v,key}], \quad \mathbf{V}^{l-1}_{v,t} = [V^{l-1}_t, \mathbf{P}^{l-1}_{v,val}], \quad V^l_t = V^{l-1}_t + \beta_v MHSA(LN(V^{l-1}_t, \mathbf{K}^{l-1}_{v,t}, \mathbf{V}^{l-1}_{v,t})) \quad (2)$$

where $[\cdot]$ is the concatenation operation, $\left\{\mathbf{K}^{l-1}_{a,t}, \mathbf{V}^{l-1}_{a,t}\right\}$ ($\left\{\mathbf{K}^{l-1}_{v,t}, \mathbf{V}^{l-1}_{v,t}\right\}$) are prompted audio (visual) Key-Value tokens in (l-1)th layer, (A^l_t, V^l_t) are the outputs of temporal prompting, and (β_a, β_v) denote the tunable scale factors to control the outputs from MHSA layers.

Through learning the inserted temporal prompts with self-attention, the pre-trained image layer can be enforced to capture temporal patterns in audio clips and video frames.

3.3 Spatial-Aware Prompting

Unlike images, the audio spectrogram has different semantics due to the acoustic characteristics. To enable the pre-trained image priors to learn the spatial contexts of audio and further strengthen the spatial modelling of video, we introduce spatial-aware prompting by prepending a set of spatial prompts into the tokens. First, the output audio-visual tokens from the temporal-aware prompting stage are reshaped to $A_s \in R^{M\times(N_a+1)\times D}$ and $V_s \in R^{M\times(N_v+1)\times D}$, respectively. Then, a set of audio spatial prompts $\mathbf{P}^l_{a,s} \in R^{N^s_p\times D}$ and visual spatial prompts $\mathbf{P}^l_{v,s} \in R^{N^s_p\times D}$ are separately inserted after audio and visual tokens along spatial dimension to prompt the well-learned semantic knowledge from the pre-trained MHSA to audio-video modalities, where N^s_p is the number of spatial-aware prompts. The whole operation is:

$$\mathbf{A}^{l-1}_s = [A^{l-1}_s, \mathbf{P}^{l-1}_{a,s}], \quad \mathbf{V}^{l-1}_s = [V^{l-1}_s, \mathbf{P}^{l-1}_{v,s}] \quad (3)$$

$$[A^l_s, Z^l_{a,s}] = \mathbf{A}^{l-1}_s + MHSA(LN(\mathbf{A}^{l-1}_s)), \quad [V^l_s, Z^l_{v,s}] = \mathbf{V}^{l-1}_s + MHSA(LN(\mathbf{V}^{l-1}_s)) \quad (4)$$

where A^l_s and V^l_s are the output tokens from spatial-aware prompting stage in (l-1)th layer, $Z^l_{a,s}$ and $Z^l_{v,s}$ represents the output embedding of spatial prompts.

By updating the prepended learnable spatial prompts, the pre-trained image models can mitigate the domain gap to learn the acoustic semantics and further enhance the contextual modelling of video frames.

3.4 Multimodal Fusion Prompting

We have audio and visual modalities throughout the learning task, requiring an effective integration of the two modalities to enhance the downstream prediction. However, pre-trained image models only learned the close-domain image knowledge. To enforce them to integrate audio and visual modalities without losing modality-specific information, we design a multimodal fusion prompting by incorporating a set of fusion prompts between audio and visual tokens. Instead of cross-attention-based fusion, we reuse the frozen pre-trained MHSA block to leverage its capability of connecting different sequence tokens to integrate multimodal tokens via inserted prompts. Given audio tokens $A_{s,t}$ and visual tokens $V_{s,t}$ attained from temporal- and spatial-aware prompting, we insert a set of fusion prompts $\mathbf{P}^{l}_{fusion} \in R^{N^{f}_{p} \times D}$ between them for performing the information interaction via a shared pre-trained MHSA block, where N^{f}_{p} is the number of fusion prompts. The procedure can be defined as:

$$\mathbf{AV}^{l-1} = [A^{l-1}_{s,t}, \mathbf{P}^{l-1}_{fusion}, V^{l-1}_{s,t}], \quad [A^{l}, Z^{l}_{fusion}, V^{l}] = \mathbf{AV}^{l-1} + \beta_{av} * MHSA(LN(\mathbf{AV}^{l-1})) \tag{5}$$

where Z^{l}_{fusion} represents the output embedding of fusion prompts, A^{l} and V^{l} are the output tokens from multimodal fusion prompting stage in (l-1)th layer, which then individually undergoes the shared pre-trained fully-connected network (FFN) block to obtain the final audio and visual features for the following various downstream tasks.

Through training the incorporated fusion prompts, the bridged audio and visual tokens interact with each other by self-attention operation, enforcing the pre-trained image models to effectively integrate audio and visual modalities.

4 Experiments

4.1 Downstream Tasks and Datasets

Audio-visual event localization (AVE): We assess AV-STFP on the AVE dataset [39], which comprises 4, 143 video clips with a synchronized audio stream. **Audio-visual question answering (AVQA):** We evaluate AV-STFP on the MUSIC-AVQA dataset [20], which contains 9, 288 video clips plus 45, 867 question-answer pairs. **Audio-visual action recognition (AVAR):** We evaluate AV-STFP on the Kinetics-Sound [48], including 27, 609 video clips spreading across 32 action categories. For all audio-visual tasks, the model is trained on the training split and its performance score is obtained from the testing split.

4.2 Experimental Setups

Data Pre-processing. Following existing works [8, 27, 45], we extract 10 RGB image frames as visual inputs, each one uniformly sampled from a video clip and then resized and cropped into the resolution size of 224×224. For the audio input, each audio waveform is chunked into 10 short segments, each one transformed into 128-D fbank features by the Hanning window. For AVE and AVAR tasks, we follow [50] to add stronger data augmentation for the visual signal, and follow [27] to adopt the mix-up augmentation for the audio signal. **Model Configurations.** We employ the pre-trained CLIP vision tower as the shared backbone with our spatial-temporal-fusion prompts attached to extract audio and visual features for different tasks. We employ two variants like *ViT-B/16* and *ViT-L/14* for AVE and AVAR tasks, and only utilize a large setting for the AVQA task. For all tasks, we adopt 2 temporal-aware prompts, 10 spatial-aware prompts, and 2 fusion prompts. For the AVE and AVQA

Method	Encoder		Pretrain Dataset		Trainable Param. ↓	Total Param. ↓	Acc. ↑
	Visual	Audio	Visual	Audio			
PSP [58]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	1.7M	217.4M	77.8%
AVT [25]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	15.8M	231.5M	76.8%
RFJC [2]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	22.8M	238.5M	76.2%
AVEL [43]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	3.7M	136.0M	74.0%
CMRAN [48]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	15.9M	148.2M	78.3%
MM-Pyramid [53]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	44.0M	176.3M	77.8%
CMBS [44]	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* [33]	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 [27]	Swin-L ❄, shared		ImageNet	✘	10.1M	238.8M	81.1%
CoPL [23]	Swin-L ❄, shared		ImageNet	✘	1.7M	232.3M	81.9%
LAVISH [27]	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) [15]	ViT-B-16 ❄, shared		CLIP	✘	3.5M	89.6M	76.3%
STG-CMA (BASE) [15]	ViT-L-14 ❄, shared		CLIP	✘	20.1M	323.6M	**83.3%**
AV-STFP (ours)	ViT-B-16 ❄, shared		CLIP	✘	**1.1M**	**87.1M**	76.7%
	ViT-L-14 ❄, shared		CLIP	✘	1.8M	303.4M	**83.0%**

Table 1: Comparison with SOTA methods on AVE dataset. The ❄ and 🔥 denote the frozen and trainable parameters, respectively. The ✘ indicates that the datasets are not used for pre-training. The $*$ represents the method re-implemented by [27].

tasks, the extracted audio-visual features are concatenated to pass through an MLP-based classifier to make the prediction. We follow existing works [27, 43, 46] to employ the classification accuracy as a measurement score. For the AVQA task, we substitute the feature extractor of the baseline model [20] with a frozen CLIP vision backbone with AV-STFP inserted while maintaining the original text encoder and pre-trained grounding modules. We follow the baseline work to adopt the answer prediction accuracy as the assessment criterion.

Model Training. Our AV-STFP is trained for 30 epochs by Adam optimizer [19], where the weight decay is configured as 0.2. Meanwhile, we use the CrossEntropy loss functions for all tasks and the Cosine Decay [30] to dynamically adjust the learning rate during the training procedure. More specifically, the learning rate is first linearly warmed up into $3e$-4 within the first 2 epochs and then decayed into $2e$-6 with a cosine function. To better train the model, we assign different learning rates for newly introduced prompts and task-specific layers. More implementation details can be found in the Supplementary Material.

4.3 Experimental Results

Audio-Visual Event Localization. As shown in Table 1, our AV-STFP achieves an impressive performance compared with existing methods while only using 0.06% of total parameters. First, our proposed AV-STFP (ViT-B-16) obtains a competitive prediction accuracy (76.7%) using the fewest trainable parameters. When switching to the ViT-L-14 backbone, our AV-STFP boosts the accuracy prediction from 76.7% to 83.0% by only increasing the small amount (0.7M) of trainable parameters, outperforming most existing methods such as LAVISH, DG-SCT, and CoPL. Although AV-STFP has slightly inferior accuracy than STG-CMA (BASE) (83.0% vs 83.3%), it only adopts 10 times fewer tunable parameters (1.8M vs 20.1M). Compared with DG-SCT using audio-specific pre-training, AV-STFP adopts a shared image backbone to efficiently learn audio-visual features.

Audio-Visual Question Answering. We also compare AV-STFP with AVQA methods on audio questions (AQ), visual questions (VQ), and audio-visual questions (AVQ). As shown in Table 2, AV-STFP achieves a state-of-the-art average question score (77.0%). Besides, AV-STFP outperforms better than LAVISH and DG-SCT in all types of questions. Compared with CoPL [23], our method achieves inferior performance in AVQ (73.2% vs 76.3%) but

Method	Encoder		Trainable Param. ↓	Total Param. ↓	Question ↑			
	Visual	Audio			AQ	VQ	AVQ	Avg.
AVSD [21]	Non-shared 🔥		N/A	N/A	68.5%	70.8%	65.5%	67.4%
Pano-AVQA [52]	Non-shared 🔥		N/A	N/A	70.7%	72.6%	66.6%	68.9%
ST-AVQA [20]	Non-shared ❄		10.6M	94.4M	74.1%	74.0%	69.5%	71.5%
LAVISH [24]	Shared ❄		21.1M	249.8M	75.7%	80.4%	70.4%	74.0%
DG-SCT [8]	Non-shared ❄		110.4M	520.2M	77.4%	81.9%	70.7%	74.8%
STG-CMA (BASE) [15]	Shared ❄		26.9M	221.9M	77.1%	80.8 %	70.7%	74.5%
STG-CMA (LARGE) [15]	Shared ❄		83.0M	278.1M	**78.7%**	83.0%	72.3%	76.2%
CoPL [33]	Shared ❄		1.8M	240.2M	77.3%	77.6%	**76.3%**	76.7%
AV-STFP (ours)	Shared ❄		9.3M	312.7M	77.3%	**84.8%**	73.2%	**77.0%**

Table 2: Comparison results of different methods on the MUSIC-AVQA dataset, including audio, visual, and audio-visual questions.

Method	Pretrain	Backbone	Param. ↓	Acc. ↑
AVSlowFast [43]	Kinetics-400 [3]	ResNet-50 🔥	38.5M	83.7%
CrissCross [37]	Kinetics-400 [3]	R(2+1)D-18 🔥	15.4M	66.7%
MoMA [43]	ImageNet [9]	ViT-B/16 ❄	16.5M	90.6%
	ImageNet [9]	ViT-L/14 ❄	57.7M	92.2%
	CLIP	ViT-B/16 ❄	16.5M	91.1%
	CLIP	ViT-L/14 ❄	57.7M	**94.2%**
AV-STFP (ours)	CLIP	ViT-B/16 ❄	**1.10M**	85.6%
	CLIP	ViT-L/14 ❄	1.82M	92.6%

Table 3: Comparison results of our AV-STFP and existing methods on the Kinetics-Sound32 dataset for AVAR task.

Prompting			Param.	Acc.
Temporal	Spatial	Fusion		
✘	✘	✘	1.09M	78.2%
✔	✘	✘	1.28M	80.7%
✘	✔	✘	1.58M	81.4%
✘	✘	✔	1.14M	79.5%
✔	✔	✘	1.78M	82.4%
✘	✔	✔	1.63M	81.2%
✔	✘	✔	1.33M	81.5%
✔	✔	✔	1.82M	**83.0%**

Table 4: Comparison results of our model with different types of prompting stages on AVE task.

significantly better performance in VQ (84.8% vs 77.6%). Although STG-CMA (LARGE) has a slightly better accuracy score (78.7%) than ours (77.3%) in AQ, it presents a remarkably inferior performance than ours in VQ (83.0% vs 84.8%) and AVQ (72.3% vs 73.2%). Moreover, our proposed AV-STFP contains significantly fewer tunable parameters than STG-CMA (9.3M vs 83.1 M) and avoids the specific audio pre-training, revealing that the proposed learnable prompts can efficiently enable the frozen pre-trained image models to watch the video and hear the sound for better responding to the questions.

Audio-Visual Action Recognition. We further compare our AV-STFP with existing AVAR methods such as AVSlowFast [43], CrissCross [37] and MoMA [43]. From Table 3, AV-STFP (CLIP ViT-B/16) (85.6%) achieves better performance than AVSlowFast (83.7%) and CrissCross (66.7%) while using only 1.10 M trainable parameters. It is worth mentioning that AVSlowFast and CrissCross pre-train their models on the action recognition dataset (Kinetics-400) and then fine-tune them on the Kinetics-Sound32, enabling the overlapped action categories. However, AV-STFP adapts the models pre-trained on the image-only dataset into Kinetics-Sound32, which is unseen for pre-trained backbones. When using the CLIP ViT-L/14 backbone, AV-STFP improves the performance from 85.6% to 92.6% with an increase of only 0.72M parameters. Although the MoMA (CLIP ViT-L/14) slightly outperforms AV-STFP (94.2% vs 92.6%), it utilizes around 30 times more tunable parameters (57.7M vs 1.82M). It presents that AV-STFP can efficiently transfer the knowledge from pre-trained image models to benefit the AVAR task with minimal model parameters.

4.4 Ablation Studies

In this section, we conduct extensive ablation studies on the AVE dataset to validate the effectiveness of different components of our approach.

Spatial-Temporal-Fusion Prompts Design. In Table 4, we conduct studies to explore the

Model	A.P.	V.P.	F.P.	Param.	Acc.
Audio Branch					
w/o Prompt	✘	✘	✘	0.06M	44.0%
w/ Prompt	✔	✘	✘	0.4M	65.5%
Visual Branch					
w/o Prompt	✘	✘	✘	0.06M	75.8%
w/ Prompt	✘	✔	✘	0.4M	79.4%
Audio-Visual Branch Without Fusion					
w/ Audio Prompt	✔	✘	✘	1.43M	80.8%
w/ Visual Prompt	✘	✔	✘	1.43M	79.0%
w/ AV Prompt	✔	✔	✘	1.8M	81.6%
Ours vs Finetuning					
Full Finetuning	✘	✘	✘	606.8M	57.6%
AV-STFP(ours)	✔	✔	✔	1.8M	**83.0%**

Table 5: Comparison results with different prompting configurations. '**A.P.**', '**V.P.**', and '**F.P.**' mean the 'Audio Prompting', 'Video Prompting', and 'Fusion Prompting', respectively.

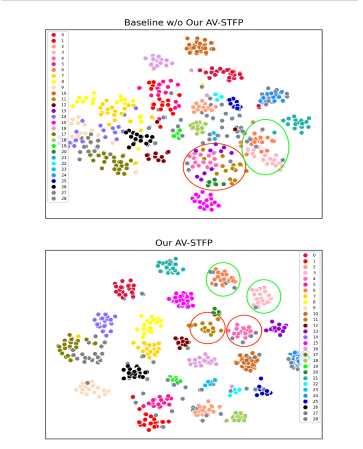

Figure 3: The t-SNE visualization of baseline model w/o AV-STFP (left) and with AV-STFP (right). The number $0-28$ means the indices for event categories in the AVE dataset. Each color corresponds to a specific event.

effectiveness of different types of proposed prompts in our AV-STFP on the AVE dataset. First, we observe that the reference model omitting all learnable prompts obtains the worst performance, showing the proposed temporal-aware, spatial-aware, and multimodal fusion prompts are indispensable for adapting the frozen image models into the audio-visual domain. Then, we find that introducing any type of proposed prompts into the pre-trained image model will enhance the performance to a different degree. Especially, spatial-aware prompts play a more important role in the cross-modal adaptation from image to audio and video modalities. Afterwards, the performance will be further improved when injecting any two types of learnable prompts into the pre-trained image backbone. For instance, inserting spatial-temporal prompts performs better than adding spatial-fusion prompts, indicating that temporal-aware prompts are crucial to adapt the static image models to capture the temporal dependency of audio and video frames. Finally, removing the fusion prompts from AV-STFP would degrade the performance from the best 83.0% into 82.4%, presenting the efficiency of our prompt-based fusion mechanism on interacting with audio and visual modalities.

Prompting Configuration. In Table 5, we design comparison experiments to verify the effectiveness of our AV-STFP with different prompting configurations on the AVE dataset using the frozen CLIP ViT-L-14 backbone. First, to explore the influence of spatial-temporal prompts on a single modality, we design two groups of reference models only considering either the audio branch or the visual branch. We can find that inserting spatial-temporal prompts will improve the audio-only model from 44.0% to 65.5% and the visual-only model from 75.8% to 79.4%, showing that the prepended prompts are very crucial for adapting the frozen image models to understand sequential consistency of audio and video. We also observe that the visual-only model achieves better performance than the audio-only one, implying that the performance of the AVE task is dominated by the visual modality. Second, we investigate the impact of removing audio or visual prompts from the audio-visual branch. We can observe that omitting prompts from either audio or visual branch will lead to perfor-

mance degradation, showing that our dual-branch prompting benefits in correctly predicting the event in the AVE task. Finally, we compare our efficient audio-visual prompting scheme with the full fine-tuning paradigm. From Table 5, we can find that our proposed prompt-based method significantly exceeds the full-finetuning scheme while only occupying 0.3% of its trainable parameters, demonstrating the efficiency of our AV-STFP in generalizing the pre-trained image models into the audio-visual domain and overcoming the knowledge forgetting caused by full fine-tuning.

Qualitative Analysis. We adopt the t-SNE [41] to visualize the output audio-visual features from the baseline model with our AV-STFP and without our AV-STFP as depicted in Figure 3. To obtain the baseline without AV-STFP, we only remove all inserted learnable prompts and unchange other components of the architecture of AV-STFP. We adopt the CLIP ViT-L/14 as the pre-trained image backbone for both models and evaluate them on the AVE dataset. As we can observe, compared with the baseline removing all proposed prompts, our AV-STFP yields more intra-class compact and inter-class separable audio-visual features in terms of circled classes in Figure 3, indicating that our novel multimodal prompting technique can efficiently adapt the pre-trained image model to successfully predict event categories.

5 Conclusions

In this paper, we propose an Audio-Visual Spatial-Temporal-Fusion Prompting (AV-STFP), a novel prompt learning framework to transfer the learnt knowledge from the pre-trained image foundation model into the audio-visual domain. By introducing a set of temporal-aware, spatial-aware, and multimodal fusion learnable prompts into original audio-visual tokens, the shared pre-trained image layers can be efficiently adapted to learn the temporal consistency, semantic reasoning, and multimodal interaction of audio and video modalities. The extensive experiments on the audio-visual understanding tasks present that our AV-STFP attains competitive or even better performance than state-of-the-art methods while only involving a small number of tunable parameters.

Future Work. While our AV-STFP enhances the advancement of parameter-efficient audio-visual learning, it is necessary to evaluate our model in more challenging audio-visual environments. In addition, our proposed method relies on a shared pre-trained image foundation model to handle audio data, which maybe hard to sufficiently understand the acoustic characteristics due to the domain gap. In the future, we will apply our AV-STFP to the pre-trained audio-visual backbones to further improve the performance.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[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.

[3] 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.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[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.

[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 2020. *arXiv preprint arXiv:2010.11929*, 2010.

[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.

[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.

[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.

[10] Rohit Girdhar, Alaaeldin 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.

[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.

[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.

[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.

[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.

[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.

[16] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022.

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[18] 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.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong 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.

[21] 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.

[22] 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.

[23] 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.

[24] Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, and Qiang Zhu. Querying as prompt: Parameter-efficient learning for multimodal language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26855–26865, 2024.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[31] Shentong Mo and Bhiksha Raj. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Arsha Nagrani, 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.

[33] 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.

[34] 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.

[35] 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.

[36] Pritam Sarkar and Ali Etemad. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9723–9732, 2023.

[37] 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.

[38] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-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.

[39] 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.

[40] 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.

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[42] 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.

[43] Kai Wang and Dimitrios Hatzinakos. Moma: Mixture-of-modality-adaptations for transferring knowledge from image models towards efficient audio-visual action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8055–8059. IEEE, 2024.

[44] Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024.

[45] Kai Wang, Yapeng Tian, and Dimitrios Hatzinakos. Towards efficient audio-visual learners via empowering pre-trained vision transformers with cross-modal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1837–1846, 2024.

[46] 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.

[47] 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.

[48] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[49] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui 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.

[50] 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.

[51] 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.

[52] 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.

[53] Yihan Zhao, Wei Xi, Yuhang Cui, Gairui Bai, Xinhui Liu, and Jizhong Zhao. Copl: Parameter-efficient collaborative prompt learning for audio-visual tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4455–4464, 2024.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.