Semantic Modulated Prompting for Few-Shot Audio-Visual Classification

Guanjie Huang†, Yawen Cui†, Danny Hin Kwok Tsang, *Fellow, IEEE*,
Wenwu Wang, *Fellow, IEEE*, Li Liu*, *Senior Member, IEEE*

***Abstract*—Few-Shot Audio-Visual Classification (FS-AVC) trains models using a limited number of labeled audio and visual sample pairs to capture the classification capability. Deep learning-based audio-visual learning methods often construct complicated frameworks with numerous parameters trained on large labeled datasets, rendering them impractical for FS-AVC. The key challenges for FS-AVC are model overfitting, multimodal fusion under temporal asynchrony, and modality imbalance. To address these challenges, we propose a novel method called Semantic Modulated Prompting (SMP) to improve the learning process of FS-AVC. This framework implants text as prompting tokens via two components: Prompt-refined Audio-Visual efficient Learner (P-AVeL) and Prompt-tuned Prototypical Regularization (P-PR). By integrating semantic prompts, adapter-based P-AVeLs conduct the prompt-guided latent attention to alleviate the overfitting and achieve effective alignment and fusion. Concurrently, P-PR, the first rebalancing method designed for few-shot scenarios, uses these semantic prompts to accurately evaluate and dynamically adjust the imbalance of two modalities. Extensive experiments demonstrate that the SMP framework consistently outperforms state-of-the-art multimodal methods by a large margin. The code is available at *https://github.com/DennisHgj/SMP_FSAVC*.**

***Index Terms*—Audio-visual learning, few-shot learning, modality rebalance.**

I. Introduction

Audio-Visual Classification (AVC) aims at classifying the event/activity based on the audio and visual inputs [1]–[4]. Previous deep learning-based AVC methods have achieved great performance, but they often construct complex frameworks with numerous parameters trained on a large amount of labeled data. In practical applications, collecting large-scale labeled data is expensive and time-consuming. It is also often hindered by privacy concerns or rare issues (e.g., abnormal action recognition in public places such as hospital wards or subway stations). In this work, facing the situation of a few available labeled samples, we conduct investigations on **F**ew-**S**hot **A**udio-**V**isual **C**lassification (FS-AVC). It requires the model to recognize new types of actions or events in videos with sound by showing it limited examples, forcing it to learn by combining what it sees and hears efficiently.

In the field of FS-AVC, we conducted an analysis of the challenges in this area, identifying three interconnected issues that remain insufficiently addressed by existing approaches: **(1) Significant Overfitting.** Multimodal models inherently contain more parameters than their unimodal counterparts [6], significantly increasing the risk of overfitting when training data is scarce (shown in Fig. 1 (b)). Existing few-shot methods are often designed for unimodal scenarios [7]–[9] and thus ignore the additional complexities of multimodal inputs. **(2) Fusion Difficulty under Temporal Asynchrony**. Effective fusion is complicated by the fact that audio and visual streams are frequently misaligned in time (shown in Fig. 1 (c)). This asynchrony demands complex cross-modal interaction to learn a proper alignment strategy. While sufficient data can enable such learning [2], [10], [11], the severe data limitation in few-shot settings cripples the model's ability to learn these alignments effectively. **(3) Severe Modality Imbalance.** The scarcity of data amplifies modality imbalance, which occurs when one modality learns at a different speed than the other [12], [13] or when dataset biases favor one modality [14]. This results in the dominant modality suppressing the other, preventing the model from fully utilizing complementary information and leading to suboptimal performance (shown in Fig. 1 (d)). Existing solutions, such as introducing additional modules [12], [15] or assessing unimodal learning status [14], [16], are ill-suited for FS-AVC; extra parameters can worsen overfitting, and accurately evaluating learning status with few samples is unreliable.

Consequently, a key unresolved problem for FS-AVC is: *how to effectively learn and aggregate limited audio-visual information while simultaneously mitigating overfitting, managing temporal asynchrony, and correcting modality imbalance?*

In this work, we focus on addressing the above three challenges of FS-AVC in a parameter-efficient manner and innovatively introduce the **S**emantic **M**odulated **P**rompting (**SMP**) framework. Prompting is an efficient technique that traditionally uses additional learnable tokens to learn task-related representations in a frozen network [17]. A similar idea also appears in audio-visual learning works that enable parameter-efficient training [18], [19]. However, randomly initialized learnable tokens still require a large amount of data to fully unlock their potential in learning multimodal representations. In contrast, existing works [20], [21] show that additional semantic information can help few-shot learning. Therefore, for the first time, we innovatively leverage semantic prompts to modulate the multimodal learning process of the model in FS-AVC. Specifically, SMP consists of two components: the **P**rompt-refined **A**udio-**V**isual **e**fficient **L**earner (**P-AVeL**) with prompt-guided latent attention module and **P**rompt-tuned **P**rototypical **R**egularization (**P-PR**).

† Equal Contribution.

* Corresponding Author: avrillliu@hkust-gz.edu.cn.

Guanjie Huang, Danny Hin Kwok Tsang, and Li Liu are with the Hong Kong University of Science and Technology (Guangzhou). Yawen Cui is with the Hong Kong Polytechnic University, work done at HKUST (GZ). Wenwu Wang is with the University of Surrey, UK.

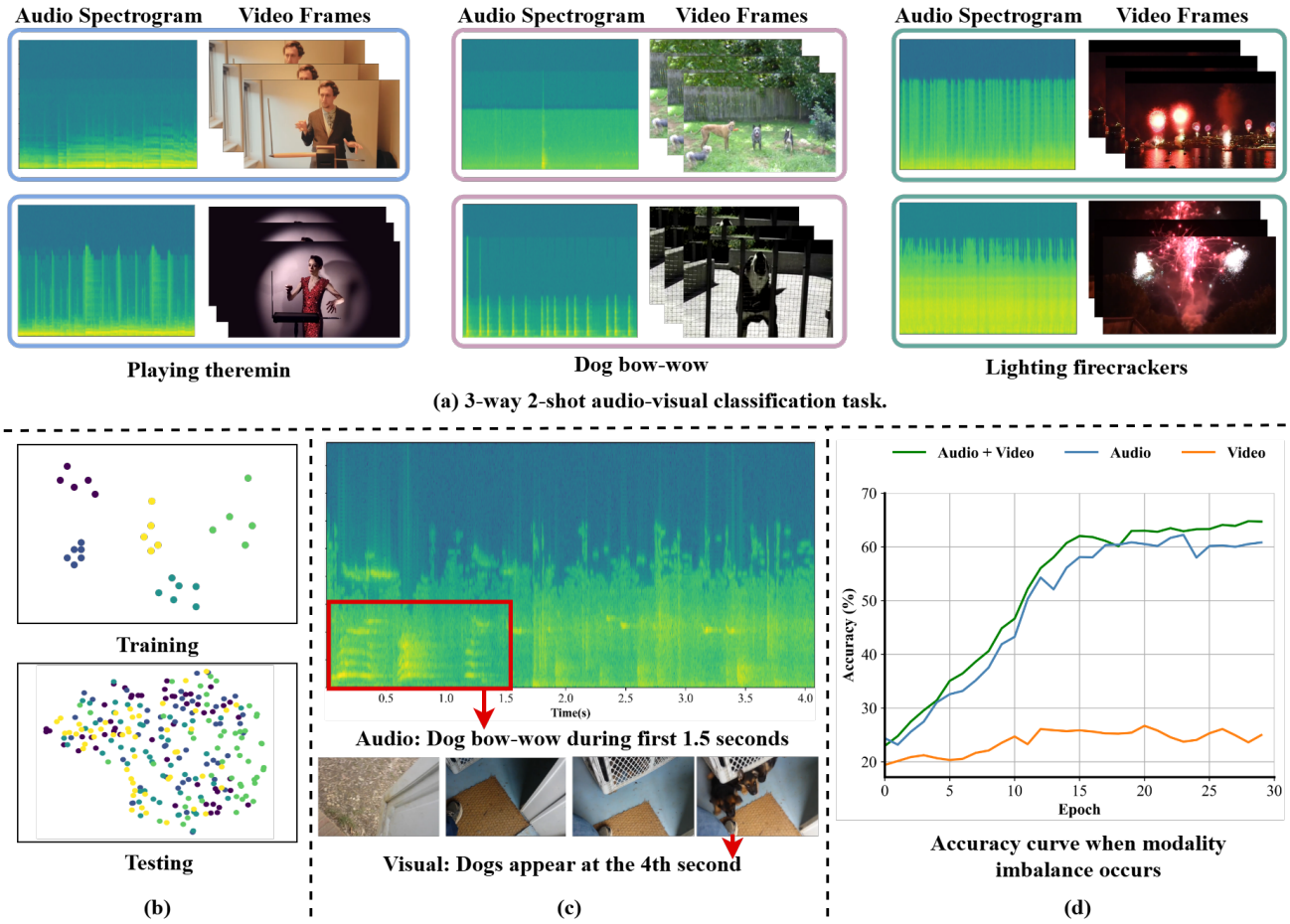

Fig. 1. **The FS-AVC configuration and three interconnected challenges. (a) The configuration.** We use the 3-way 2-shot setting as an example to indicate N-way K-shot FS-AVC problems. **(b) Significant Overfitting.** A real sample t-SNE [5] figure of an audio-visual model's training and testing performance in the FS-AVC task suffers from overfitting. Increasing the number of parameters in multimodal models makes them more prone to overfitting in limited data scenarios. **(c) Fusion Difficulty under Temporal Asynchrony.** The figure shows a real audio-visual data pair with temporal asynchrony. When the model handles modalities with temporal characteristics, learning an effective fusion mechanism is more difficult when the data is limited. **(d) Severe Modality Imbalance.** The figure shows the experimental results of a real FS-AVC test. The conventional audio-visual models exhibit varying learning speeds and annotation preference issues across different modalities, exacerbated by limited data scenarios, resulting in significant modality imbalance. The dominant modality (*e.g.*, the audio modality in the figure) causes negative effects on the learning of the other modality, which leads to convergence difficulties and suboptimal performance.

More precisely, **firstly**, to address the significant overfitting challenge, P-AVeL is built on adapters [22], [23], one of the parameter-efficient fine-tuning methods that only fine-tunes the injected, small, trainable modules while keeping the original backbone weights frozen to prevent overfitting. **Secondly**, we introduce the prompt-guided latent attention within the P-AVeL to solve the temporal asynchrony issue and enhance fusion effectiveness. This module leverages semantic prompts to guide the model's focus toward critical tokens within verbose and potentially misaligned audio and visual sequences. This semantic guide mechanism enables efficient and robust modality fusion, proving particularly effective in data-limited scenarios. **Thirdly**, P-PR is designed to overcome severe modality imbalance in FS-AVC. In data-limited scenarios, prototypes for individual modalities (e.g., audio and visual) often deviate significantly from the actual class centroids when applying the previous prototypical related rebalance algorithm [24]. P-PR mitigates this through dynamic tuning guided by semantic prompts. During each training batch, these prompts actively pull the audio and visual prototypes closer to the semantic centroid. This process allows the algorithm to continuously assess the learning state of each modality, thereby achieving an effective modal rebalance.

In summary, the contributions of this work are fourfold:

- A novel prompting framework, SMP, is proposed with two components (*i.e.*, P-AVeL and P-PR) for tackling key challenges in the FS-AVC.
- With semantic prompts, the adapter-based P-AVeL prevents overfitting and achieves effective audio-visual fusion with limited labeled data under temporal asynchrony by prompt-guided latent attention.
- To address modality imbalance, the first modality rebalance method for the few-shot scenarios, P-PR, is proposed. It accurately estimates and dynamically balances the unimodal learning status by semantically tuning the audio and visual prototypes with prompts.
- Comprehensive experiments on three datasets demonstrate that the SMP significantly outperforms SOTA meth-

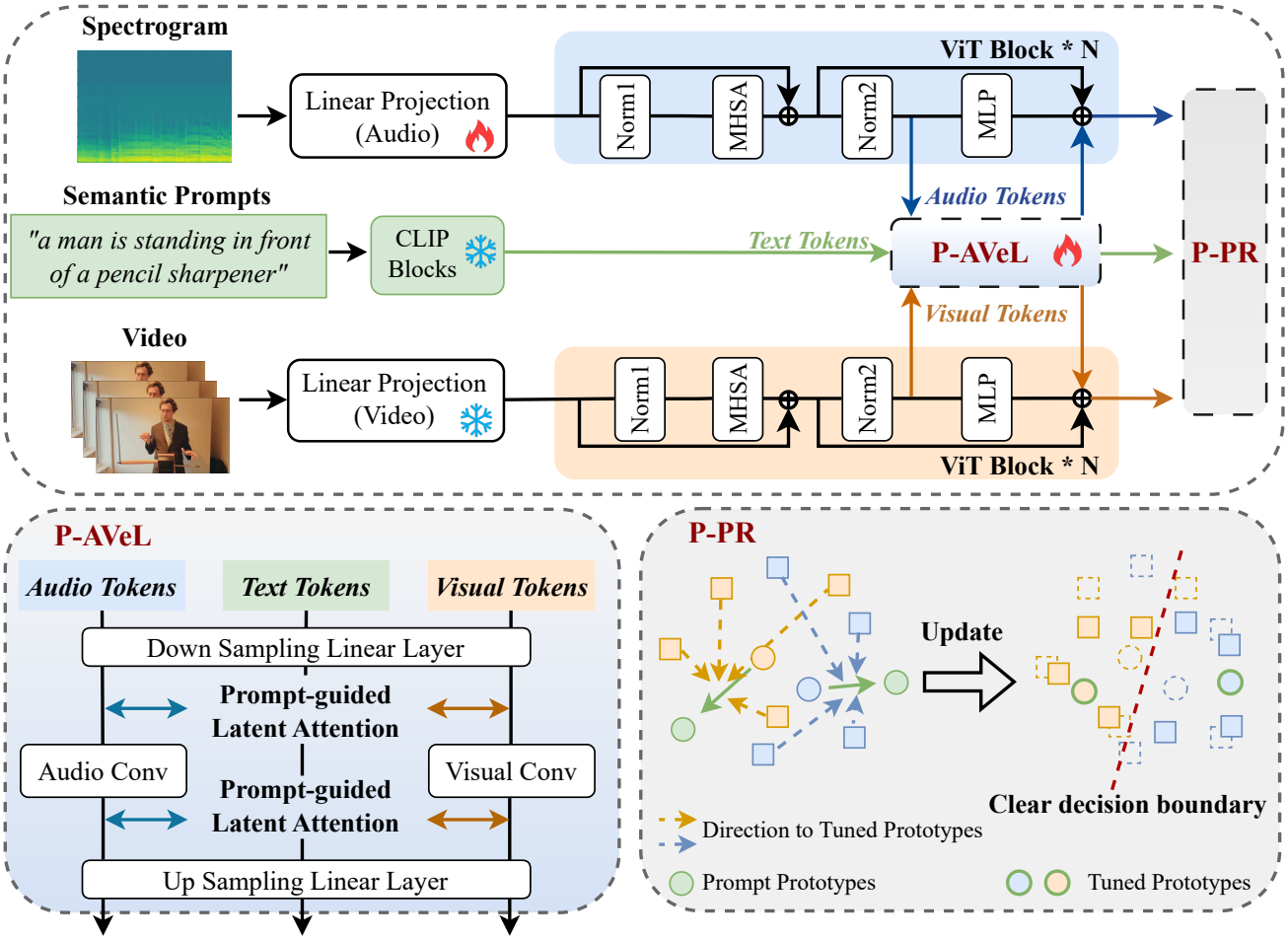

Fig. 2. **Overview of Semantic Modulated Prompting.** Two main components of SMP are **P-AVeL** and **P-PR**. P-AVeLs are parallel with MLP layers of Transformer blocks and perform modal fusion effectively via prompt-guided latent attention. P-PR takes the output of three modal encoders to calculate prototypes and perform rebalance regularization. "P-AVeL", "P-PR", "MHSA", "MLP", and "Conv" are short for Prompt-refined Audio-Visual efficient Learner, Prompt-tuned Prototypical Regularization, multi-head self-attention, multi-layer perceptron, and convolution layer. The flame and snowflake icons indicate trainable and frozen modules, respectively.

ods with fewer training parameters.

II. Related Works

A. Few-Shot Learning (FSL)

FSL aims to recognize novel classes represented with only a few available support samples. The current few-shot learning methods can be categorized into three main categories. (1) Data augmentation-based methods target enlarging the limited labeled dataset by applying data transformation [25], self-supervised learning [26], or synthesizing new data with a generative model [27]. (2) Optimization-based methods [8], [9], [28], [29] focus on designing a good initialization or optimization strategy to enable models to adapt to novel tasks quickly. MAML [9] and Reptile [29] are finding good model initializations that can be adapted to a novel few-shot classification task in a few gradient steps. (3) Metric-based methods [7], [30], [31] learn an appropriate distance function in the proper latent space and then predict results based on the similarity between support and query samples. ProtoNet [7] uses the Euclidean distance for the similarity measure. Most of the aforementioned algorithms only concentrate on unimodal learning problems. Few-shot multi-modal learning works have emerged in recent years: PROTO-CAT [32] focuses on audio-visual speech recognition; AV-Diff [33] imports a diffusion model to augment the audio-visual features in the generalized few-shot learning task. In this paper, we focus on the FS-AVC and explore a more effective methodology to tackle the challenges.

B. Audio-Visual Understanding

Audio-visual understanding tasks target the audio-visual perception with both visual and audio modalities, including the tasks of event/activity classification [1]–[4], [12], event/activity localization [18], [34]–[37], video parsing [38]–[41], audio-visual segmentation [42] and audio-visual question answering [43], [44]. In this paper, we focus on event/activity classification. This task requires the model to recognize joint audio-visual events or activities. Early work [45] focuses on fusing features in the early stage or integrating scores in the late stage. Fayek *et al.* [1] propose attention fusion models to combine audio and visual models dynamically. MBT [2] is a

representative transformer-based architecture that uses "fusion bottlenecks" for modality fusion at multiple layers.

Although transformer-based models provide satisfactory performance, training them is often expensive and time-consuming. To improve efficiency, methods with parameter-efficient fine-tuning as the core have gradually emerged since 2023. LAVISH [18] implements cross-modal adapters within frozen pre-trained transformers to process audio-visual data. At the same time, DG-SCT [46] develops dual-guided spatial-channel-temporal attention mechanisms that serve as cross-modal prompts for model augmentation. AV-MoE [47] improves this idea with the mixture of experts mechanism by involving several different adapters as experts to process multimodal and unimodal information, respectively. Although the above adapter-based approach effectively reduces training parameters and achieves good performance on conventional tasks, it is still constrained by the need for massive data to obtain an effective fusion mechanism. In this paper, we focus on event/activity classification and propose solutions to modality fusion and rebalancing in a few-shot scenario.

C. Prompt Learning

In natural language processing, the concept of prompts has demonstrated remarkable effectiveness across diverse applications since its emergence [48]–[50], as exemplified by advancements like the GPT series [51]. Researchers have subsequently explored integrating prompt-based strategies with multimodal frameworks to enhance model performance. For instance, CoOp [52] introduced trainable continuous prompts in CLIP's text encoder to refine language-vision alignment, while CoCoOp [53] extended this approach by incorporating adaptive prompt generation in the visual processing pipeline. Early methodologies primarily focused on optimizing individual modality branches, but recent innovations emphasize cross-modal interaction. CLIP-adapter [54] pioneered bidirectional feature guidance by establishing post-encoder connections between visual and textual embeddings. Building on this foundation, MaPLe [55] introduced a novel architecture with interleaved adapters within the encoder layers, enabling simultaneous semantic fusion of visual and linguistic representations throughout the feature extraction process. In audio-visual understanding, adapter-based methods LAVISH [18], DG-SCT [46], and AV-MOE [47] all borrow the mechanism of prompt learning and input randomly initialized learnable tokens or all audio/visual tokens as prompts into multimodal fusion. However, such a design is sensitive to the reduction of data volume and cannot effectively stimulate the potential of the prompt mechanism in the FS-AVC scenario. Our work innovatively involved easily obtained text modalities as semantic prompts with a special parameter-free attention fusion module to improve the efficiency of audio-visual learning, especially in data-limited scenarios.

III. Methods

A. Problem Formulation

FS-AVC takes audio and video as inputs and formulates an N-way K-shot classification task, *i.e.,* each task includes N classes with K samples for each class. We denote each audio-visual sample as $\mathcal{X} = \{\mathcal{F},\ \mathcal{A},\ y\}$. These N-way-K-shot samples are termed the support set as $\mathcal{S} = \{\mathcal{X}_n\}_{n=1}^{N\times K}$. For an input of visual modality, we uniformly sample a temporal sequence of RGB frames from a video clip as $\mathcal{F} = \{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_F\}$, where F is the temporal length. For the j-th frame in $\mathcal{F}$, $\mathbf{V}_j \in \mathbb{R}^{H\times W\times 3}$ with spatial dimensions $H \times W$. Audio is transformed into a spectrogram $\mathcal{A} \in \mathbb{R}^{M\times C}$, where M is the frequency and C is the time. y denotes the ground-truth label.

FS-AVC aims to conduct inference on the query set $\mathcal{Q}$ after perceiving limited labeled samples in the supporting set $\mathcal{S}$. In each forward pass, the visual encoder $f(\cdot)$ takes the temporal sequence of RGB frames $\mathcal{F}$ as input, and the audio encoder $g(\cdot)$ takes the spectrogram $\mathcal{A}$ as input. The fusion operation can be executed in different stages, such as before, during, or after the encoders. After that, the fusion results are sent to the decision layer for the final prediction.

For the learning paradigm of FS-AVC, we follow the hypothesis of traditional FSL [56], *i.e.*, models can access a large-scale source set and a few-shot target set. We first pretrain our framework on the source dataset, then finetune it on the FS-AVC tasks.

B. Overview of Semantic Modulated Prompting

SMP for FS-AVC is a dual-stream structure built on ViT [57]. Video frames and the audio spectrogram are first fed into corresponding linear projections, followed by N Transformer blocks. As illustrated in Fig. 2, towards the challenges of FS-AVC, a P-AVeL is incorporated into each Transformer block as a parallel module of Multi-Layer Perception (MLP). Textual knowledge from semantic prompts is introduced into the P-AVeL by the designed latent attention and goes through the forward pass along with the knowledge evolution. After the last Transformer block, P-PR is executed for modality rebalance. Finally, we put all the tokens into the classification head for the final prediction.

C. Prompt-Refined Audio-Visual Efficient Learner

Semantic prompts are introduced into P-AVeLs in SMP. With multimodality tokens, prompt-guided latent attention is executed for fusion purposes.

1) Semantic Prompts: Instead of pure learnable tokens or all audio/visual tokens used in previous audio-visual prompting methods, task-related free-accessible contexts are more suitable for few-shot scenarios. The most basic prompt can be constant text, such as "*A video of [label]*".

Moreover, using prompts with better prior knowledge can greatly stimulate the potential of SMP. In this work, we take the video captions provided by the pretrained Vision-Language Model (VLM) as semantic prompts. Specifically, a VLM takes video frames as inputs and then generates prompts based on these frames.

With semantic prompts, we put them into a pretrained text encoder to obtain text tokens, then introduce them to P-AVeL afterward. After being added to the first Transformer block with P-AVeL, these text tokens are passed through latent

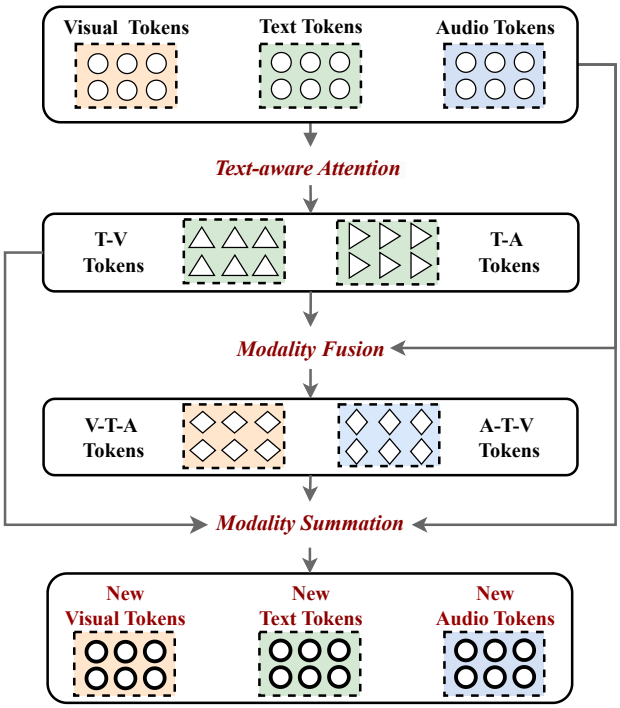

Fig. 3. **Three-phases structure of prompt-guided latent attention.** The latent attention mechanism consists of three clear phases: **1) Text-aware Attention phase** uses text tokens obtained from semantic prompts to guide the model to focus on important audio and visual tokens, thereby achieving efficient information extraction and token number compression, further reducing the computational complexity of subsequent cross-modal attention operations. **2) Modality Fusion phase** uses the key tokens obtained in the previous phase to achieve tri-modal information interaction. **3) Modality Summation phase** sums the information fusion tokens to the corresponding modality to form new modality tokens, i.e., sums tokens of the same background color.

attention with audio tokens and visual tokens, thus achieving prompts evolution.

2) Prompt-Guided Latent Attention: Fusion strategy plays an essential role in multimodal learning, which aims to effectively fuse multiple modalities for improving the performance of joint decision-making. In this paper, with tokens extracted from prompts, the specially designed latent attention is executed between adapters of the audio and video streams for effective fusion and alleviating temporal asynchrony.

In each P-AVeL, audio, video, and text tokens all go through downsampling via linear layers first. We denote the tokens after downsampling in a certain Transformer layer as $\mathcal{T} = \{\mathbf{T}^v, \mathbf{T}^a, \mathbf{T}^t\}$ with $\mathbf{T}^v$, $\mathbf{T}^a$, and $\mathbf{T}^t$ as video, audio, and text tokens, respectively. Prompt-guided latent attentions are directed before and after the convolution layers.

As illustrated in Fig. 3, the prompt-guided latent attention consists of three phases:

Text-aware Attention. With the incorporated text tokens, the purpose of this phase is to guide the model to focus on important audio and video tokens and ignore others, which also compresses visual tokens and audio tokens to a small set with the same length as text tokens. This is implemented with the following attention function:

$$\begin{aligned} \mathbf{T}^{t,v} &= \mathrm{CMA}(\mathbf{T}^t, \mathbf{T}^v, \mathbf{T}^v), \\ \mathbf{T}^{t,a} &= \mathrm{CMA}(\mathbf{T}^t, \mathbf{T}^a, \mathbf{T}^a), \end{aligned} \tag{1}$$

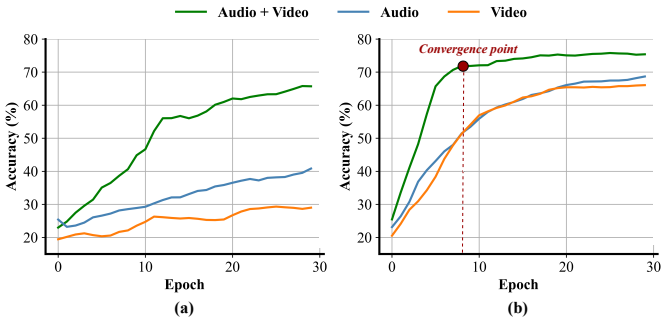

Fig. 4. **The accuracy curves of the SMP without and with P-PR in a 5-way 1-shot FS-AVC test on Kinetics-Sounds [59]**. (a) The accuracy curve of SMP without P-PR. The model converges slowly and obtains suboptimal performance because of modality imbalance. (b) The accuracy curve of SMP with the proposed P-PR. The model can converge faster and get better performance.

where CMA is the parameter-free cross-modal attention [58] and is defined as

$$\mathrm{CMA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{2}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value tokens, respectively. $\top$ is the transpose operation and $\frac{1}{\sqrt{d}}$ is the scaling factor, with d being the dimension of the keys.

Modality Fusion. This phase focuses on achieving the fusion of audio, video, and text modalities. It contains two weighted addition operations and two cross-modal attention operations:

$$\begin{aligned} \mathbf{T}^{'t,a} &= \mathbf{T}^{t,a} + \lambda^{t,a}\mathbf{T}^t, \\ \mathbf{T}^{'t,v} &= \mathbf{T}^{t,v} + \lambda^{t,v}\mathbf{T}^t, \\ \mathbf{T}^{v,t,a} &= \mathrm{CMA}(\mathbf{T}^v, \mathbf{T}^{'t,a}, \mathbf{T}^{'t,a}), \\ \mathbf{T}^{a,t,v} &= \mathrm{CMA}(\mathbf{T}^a, \mathbf{T}^{'t,v}, \mathbf{T}^{'t,v}), \end{aligned} \tag{3}$$

where $\lambda^{t,a}$ and $\lambda^{t,v}$ are learnable parameters. These four operations can take advantage of full and compressed tokens. The fusion across three modalities generates the new tokens by considering their importance from different perspectives. These generated tokens are ready for the summation phase.

Modality Summation. The summation stage sums the corresponding tokens obtained from the last two phases for each modality to form new modality tokens, *i.e.*, sums tokens of the same background color in Fig. 3. The calculations are presented as follows:

$$\begin{aligned} \overline{\mathbf{T}}^v &= \lambda^v\mathbf{T}^{v,t,a} + \mathbf{T}^v, \\ \overline{\mathbf{T}}^a &= \lambda^a\mathbf{T}^{a,t,v} + \mathbf{T}^a, \\ \overline{\mathbf{T}}^t &= \lambda^{t_1}\mathbf{T}^{t,v} + \lambda^{t_2}\mathbf{T}^{t,a} + \mathbf{T}^t, \end{aligned} \tag{4}$$

where λ^v, λ^a, λ^{t_1}, and λ^{t_2} are learnable parameters. $\mathbf{T}^{t,v}$ and $\mathbf{T}^{t,a}$ are from Text-aware Attention phase (Eq. 1), $\mathbf{T}^{v,t,a}$ and $\mathbf{T}^{a,t,v}$ are from Modality Fusion phase (Eq. 3). The new tokens gather information from different attention levels. After this phase, the summated tokens are sent back to individual modal streams through residual connections.

D. Prompt-Tuned Prototypical Regularization for Modal Rebalance

In P-AVeL, we consider audio tokens and visual tokens fairly within the latent attention. However, due to modalities learning at different speeds [12] and dataset bias [14], the modality imbalance problem appears. It manifests as the dominant modality that affects the updating direction of the slow-learning one. Moreover, when data is limited, the learning status is greatly affected by the individual sample preference. This is why the imbalance phenomenon is much more severe in the limited data scenario. It causes slow convergence, illustrated in Fig. 4, and further hinders the full realization of multimodal advantages. P-PR, executed on the generated tokens of the last Transformer block, is proposed for modal rebalancing.

The prototype usually stands for the centroid of each class in the feature space. We assume that there are C classes in total. The prototype set of these C classes is defined as $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^{C}$, where $\mathbf{P}_i = \{\mathbf{P}_i^a, \mathbf{P}_i^v, \mathbf{P}_i^t\}$ with $\mathbf{P}_i^a$ being the audio prototype, $\mathbf{P}_i^v$ being the video prototype, and $\mathbf{P}_i^t$ being the text prototype for prompts. For samples of the i-th class, we denote the output visual tokens, audio tokens, and text tokens of the last Transformer block as $\overline{\mathcal{T}}^v = \{\{\overline{\mathbf{T}}_{i_z}^v\}_{z=1}^{S_i}\}_{i=1}^{C}$, $\overline{\mathcal{T}}^a = \{\{\overline{\mathbf{T}}_{i_z}^a\}_{z=1}^{S_i}\}_{i=1}^{C}$, and $\overline{\mathcal{T}}^t = \{\{\overline{\mathbf{T}}_{i_z}^t\}_{z=1}^{S_i}\}_{i=1}^{C}$, respectively, where S_i is the number of samples for the i-th class. $\mathbf{P}_i^a$, $\mathbf{P}_i^v$, and $\mathbf{P}_i^t$ are defined as

$$\begin{aligned} \mathbf{P}_i^a &= \frac{1}{S_i}\textstyle\sum_{z=1}^{S_i} \overline{\mathbf{T}}_{i_z}^a, \\ \mathbf{P}_i^v &= \frac{1}{S_i}\textstyle\sum_{z=1}^{S_i} \overline{\mathbf{T}}_{i_z}^v, \\ \mathbf{P}_i^t &= \frac{1}{S_i}\textstyle\sum_{z=1}^{S_i} \overline{\mathbf{T}}_{i_z}^t. \end{aligned} \tag{5}$$

For each training batch, prompt-tuned prototypes of audio and video are tuned by text prototypes with dynamically calculated intensity, which form as

$$\begin{aligned} \mathbf{P}_i^{a,t} &= \mathbf{P}_i^a + \delta \mathbf{P}_i^t, \\ \mathbf{P}_i^{v,t} &= \mathbf{P}_i^v + \delta \mathbf{P}_i^t, \end{aligned} \tag{6}$$

where the tuning intensity δ of a batch B (contains B_k samples) is calculated as

$$\delta = \operatorname{norm}\left(\frac{1}{B_k} \sum_{k=1}^{B_k} \frac{\exp\left(-d\left(\overline{\mathbf{T}}_k^t, \mathbf{P}^t_{(y=i|\mathbf{T}_k^t)}\right)\right)}{\sum_{i'} \exp\left(-d\left(\overline{\mathbf{T}}_k^t, \mathbf{P}_{i'}^t\right)\right)} \right), \tag{7}$$

where y indicates the ground-truth class that the k-th sample belongs to, $d(\cdot,\cdot)$ is the distance function, which is the Euclidean Distance in this paper. $\operatorname{norm}(\cdot)$ is a normalized function, which limits δ to the interval of zero to one and suppresses the contribution of prompts when their quality is poor, ensuring the model relies less on low-quality textual information. The formulation is as follows:

$$\operatorname{norm}(x) = \frac{\sin\left(\pi \cdot x - \frac{\pi}{2}\right) + 1}{2}. \tag{8}$$

δ measures the quality of text prompts in each batch by calculating the similarity between the output text tokens and the prototype, thereby determining the prototype modulation intensity in each training step.

With the tuned prototypes, we apply the prototypical cross-entropy (PCE) loss and acceleration loss function [24] for the rebalance regularization. We denote the set of prompt-tuned audio prototypes as $\mathcal{P}^{a,t}$ and the set of video prototypes as $\mathcal{P}^{v,t}$. The PCE losses of audio and video are implemented as

$$\begin{aligned} \mathcal{L}^{a,PCE} &= \operatorname{CELoss}(-d(\overline{\mathcal{T}}^a, \mathcal{P}^{a,t}), \mathcal{Y}), \\ \mathcal{L}^{v,PCE} &= \operatorname{CELoss}(-d(\overline{\mathcal{T}}^v, \mathcal{P}^{v,t}), \mathcal{Y}), \end{aligned} \tag{9}$$

where $\mathcal{Y}$ is the ground-truth label set. The acceleration loss is denoted as

$$\mathcal{L} = \mathcal{L}^{CE} + \alpha \cdot \beta \mathcal{L}^{a,PCE} + \alpha \cdot \gamma \mathcal{L}^{v,PCE}, \tag{10}$$

where $\mathcal{L}^{CE}$ is cross-entropy loss, α is the hyperparameter to control the degree of modulation. β and γ are the coefficients of PCE losses which calculated as

$$\begin{cases} \beta = \operatorname{clip}\left(0, \frac{1}{\rho} - 1, 1\right), \gamma = 0 & \rho < 1 \\ \beta = 0, \gamma = \operatorname{clip}\left(0, \rho - 1, 1\right) & \rho \geq 1, \end{cases} \tag{11}$$

where the bias ratio ρ for a batch B (contains B_k samples) is calculated as

$$\rho = \frac{\sum_{k=1}^{B_k} \rho_k^a}{\sum_{k=1}^{B_k} \rho_k^v}, \tag{12}$$

and the indicators for each modality are defined as

$$\begin{aligned} \rho_k^a\left(y = i \mid \overline{\mathbf{T}}_k^a\right) &= \frac{\exp\left(-d\left(\overline{\mathbf{T}}_k^a, \mathbf{P}_i^{a,t}\right)\right)}{\sum_{i'} \exp\left(-d\left(\overline{\mathbf{T}}_k^a, \mathbf{P}_{i'}^{a,t}\right)\right)} \\ \rho_k^v\left(y = i \mid \overline{\mathbf{T}}_k^v\right) &= \frac{\exp\left(-d\left(\overline{\mathbf{T}}_k^v, \mathbf{P}_i^{v,t}\right)\right)}{\sum_{i'} \exp\left(-d\left(\overline{\mathbf{T}}_k^v, \mathbf{P}_{i'}^{v,t}\right)\right)}, \end{aligned} \tag{13}$$

where y denotes the ground-truth label that the k-th sample belongs to.

With the total loss in Eq. 10, the effect caused by the dominant modality on the updating direction of the slow-learning modality is alleviated. Therefore, the multimodal model can converge faster and achieve superior performance by fully utilizing multimodal information.

IV. Experiments

A. Experiment Setups

1) Datasets: We conducted experiments on three representative audio-visual event/activity classification datasets:

AVE [36]. AVE is an event dataset with samples that include both visible and audible content in 10-second video segments. It contains events spanning 28 categories with 4,143 videos. In our few-shot learning setting, we sample 16 categories as source classes and 12 categories as the target. We use the original testing dataset for the testing process and sample the few-shot learning tasks from the training dataset.

VGGSound100 [60]. It contains 100 classes randomly selected from the original VGGSound [61]. VGGSound is an audio-visual correspondent dataset consisting of around 200K 10-second videos from 310 classes. We sample 60 categories as the source and 40 categories as the target classes.

TABLE I

COMPARISON WITH DIVERSE CONFIGURATIONS OF THE FRAMEWORK ON AVE [36], VGGSOUND100 [60] AND KINETICS-SOUNDS [59]. THE TABLE REPORTS THE AVERAGE CLASSIFICATION ACCURACY (%) FOR 25 TEST SESSIONS UNDER DIFFERENT FRAMEWORK CONFIGURATIONS. N AND K REPRESENT N CATEGORIES AND K SAMPLES PER CATEGORY. AUDIO, VISUAL, AND AUDIO-VISUAL INDICATE THE MODEL IS FINE-TUNED ON WHICH MODALITY SETTING. THE PRESENCE OF SMP INDICATES WHETHER IT IS INTEGRATED OR IF VIT IS USED ALONE. THE "CONSTANT" AND "VLM" INDICATE WHETHER THE SEMANTIC PROMPT USED BY THE MODEL IS A SIMPLE "*A video of [label]*" OR PROVIDED BY VLM. THE RESULTS OF THE TEXT CLIP MODEL REFLECT THE QUALITY OF VLM PROMPTS ON VARIOUS DATASETS. THE AVP MODEL TAKES AUDIO-VISUAL MODALITIES AND VLM PROMPT INPUTS WITH ONLY SIMPLE FUSION. RESULTS SHOW SMP SIGNIFICANTLY IMPROVES THE PERFORMANCE OF UNIMODAL OR MULTIMODAL MODELS ON THE FS-AVC TASK, REGARDLESS OF WHETHER IT USES CONSTANT OR VLM PROMPTS.

	AVE				**VGGSound100**				**Kinetics-Sounds**			
Configurations \ (N,K)	(5,1)	(5,5)	(5,10)	(5,20)	(5,1)	(5,5)	(5,10)	(5,20)	(5,1)	(5,5)	(5,10)	(5,20)
Audio Model	35.4	56.92	61.37	66.67	37.42	62.53	71.87	77.95	34.21	48.99	54.82	60.3
Visual Model	30.51	42.66	51.5	59.75	24.08	33.99	43.32	55.61	26.95	36.61	44.31	54.17
Audio-Visual Model	33.55	47.25	59.57	66.28	31.73	55.81	67.89	76.32	27.64	43.01	51.33	60.57
Text CLIP (VLM)	48.76	78.88	87.79	91.44	30.47	48.76	60.23	71.94	38.51	53.56	64.77	74.79
AVP Model (VLM)	42.37	64.15	74.02	81.84	30.9	53.2	61.71	69.71	26.03	35.23	41.36	49.98
Audio SMP (Constant)	59.45	73.89	79.36	82.15	68.57	82.86	87.39	88.77	47.36	61.42	66.76	69.82
Visual SMP (Constant)	64.58	80.42	84.44	86.62	67.18	83.37	85.79	87.09	60.18	75.56	78.32	80.59
Audio-Visual SMP (Constant)	68.7	87.83	90.87	92.65	77.66	90.26	92.93	93.71	64.42	77.26	81.13	83.52
Audio SMP (VLM)	70.61	87.44	92.56	94.05	78.21	90.7	92.97	93.41	67.31	79.27	82.37	84.7
Visual SMP (VLM)	84.15	92.45	94.29	95.17	76.83	86.64	88.07	89.34	71.27	79.7	82.71	84.63
Audio-Visual SMP (VLM)	**85.74**	**94.82**	**95.5**	**96.46**	**80.23**	**91.48**	**93.23**	**94.29**	**74.37**	**81.98**	**84.64**	**85.92**

Kinetics-Sounds. It is a subset of Kinetics-400 [59] dataset. It contains around 24K 10-second videos from 32 human action classes. We sample 19 categories as the source classes and 13 categories as the target classes.

For all data samples used in our experiments, we evenly sample eight frames from each video to serve as visual input and to obtain prompts from VLM. Each audio clip is down-sampled to 16 kHz and converted to a mel spectrogram as the audio input.

2) Implementation Details: For the two-stream framework, to minimize pre-training bias and highlight the performance gain of different methods, we use the ViT-Base model from [57] pretrained on ImageNet [62] with 12 Transformer blocks and 768 token dimensions as the backbone. During training, we froze the parameters of ViT models except for the linear projection of the audio branch and applied Bias-terms fine-tuning [63]. A pretrained CLIP text encoder (also 12 Transformer blocks with 768 dimensions) is used to obtain the text tokens from the constant semantic prompts (*"A video of [label]"*) or VLM prompts by mPLUG-2 [64]. From the 4th Transformer block of ViTs, we incorporate P-AVeL as a parallel module of MLPs in each block, down-sampling operations in learners reduce the token dimension to 16.

After training models on the source set, we conduct 5-way 1-shot, 5-shot, 10-shot, and 20-shot comparison experiments. To reduce the uncertainty caused by random sampling of N-way K-shot tasks, we select different classes five times and randomly choose samples five times each, to test twenty-five times and report the average top-1 classification accuracy. We also set the random seed equal to 42 to ensure the training sample remains precisely the same in each session for experiments on different models. Adam [65] optimizer and learning rate 0.0001 are equipped with a batch size of 16. All experiments were conducted on NVIDIA RTX A6000 GPUs.

3) Comparison Setting of Framework Configurations: To validate the effectiveness of the framework design, we compare its performance across diverse configurations. Following the FS-AVC experiments paradigm mentioned in Sec. III, we pretrained ViTs on the source set, then finetuned them on corresponding few-shot tasks as baselines (Audio, Visual, and Audio-Visual Models in Tab. I). Also, models with SMP integrated are used to demonstrate the performance gain from SMP (Audio SMP, Visual SMP, and Audio-Visual SMP in the table) with different semantic prompts input ("Constant" for simple "*A video of [label]*" prompt and "VLM" for VLM prompts). Please note the "*[label]*" in the constant prompt is a fixed string of characters and will not change to a specific class name during the FS-AVC's training or testing. In SMP-integrated unimodal models, since there is no need to consider the modal imbalance issue and fusion difficulties of multimodal, SMP only contains a simplified version of the P-AVeL module. We also illustrate the performance of the CLIP text encoder with VLM prompts to demonstrate the quality of VLM prompts across different datasets. Additionally, we provide the performance results of the model by only concatenating the semantic prompt tokens with audio and visual tokens before the classification head (AVP model in the table).

4) Comparison Setting of Other Methods: We compare SMP with multiple representative AVL and FSL methods. For Transformer-based methods, we focus on the recent LAVISH [18], STG-CMA [19] and AV-MoE [47], which are all adapter-based. We also replace the other backbone-based models with the same ViT backbone for fair comparisons, including AV-Diff [33], which uses a diffusion model to augment features on few-shot samples; PROTO-CAT [32], which uses attention mechanism to improve the prototype of each class in the Audio-Visual Speech Recognition (AVSR) task; TBN [4], which employs temporal information; G-blend [12], which uses unimodal training process experience to prevent

TABLE II
COMPARISON WITH OTHER METHODS ON AVE [36], VGGSOUND100 [60] AND KINETICS-SOUNDS [59]. THE TABLE REPORTS THE AVERAGE CLASSIFICATION ACCURACY (%) FOR 25 TEST SESSIONS. N AND K REPRESENT FEW-SHOT SETTINGS WITH N CATEGORIES AND K SAMPLES PER CATEGORY. THE "CONSTANT" AND "VLM" INDICATE WHETHER THE SEMANTIC PROMPT USED BY SMP IS A SIMPLE "*A video of [label]*" OR PROVIDED BY VLM. THE VALUE IN THE UPPER RIGHT CORNER IS THE STANDARD DEVIATION OF ALL TEST SESSIONS. SMP ACHIEVES CLEAR PERFORMANCE LEADS IN ALL TEST SETTINGS.

	AVE				VGGSound100				Kinetics-Sounds			
Methods \ (N,K)	(5,1)	(5,5)	(5,10)	(5,20)	(5,1)	(5,5)	(5,10)	(5,20)	(5,1)	(5,5)	(5,10)	(5,20)
Training from scratch	$35.38^{\pm 3.84}$	$51.38^{\pm 5.81}$	$58.98^{\pm 5.94}$	$66.72^{\pm 5.48}$	$32.08^{\pm 3.94}$	$47.18^{\pm 3.2}$	$59.82^{\pm 2.06}$	$69.94^{\pm 3.74}$	$26.7^{\pm 1.29}$	$38.16^{\pm 2.56}$	$44.84^{\pm 3.97}$	$52.88^{\pm 4.16}$
Finetuning	$33.55^{\pm 2.92}$	$47.25^{\pm 5.36}$	$59.57^{\pm 6.52}$	$66.28^{\pm 7.03}$	$31.73^{\pm 1.05}$	$55.81^{\pm 4.49}$	$67.89^{\pm 4.25}$	$76.32^{\pm 5.31}$	$27.64^{\pm 2.48}$	$43.01^{\pm 3.96}$	$51.33^{\pm 5.13}$	$60.57^{\pm 4.79}$
G-Blend [12]	$39.26^{\pm 2.11}$	$58.88^{\pm 6.74}$	$68.43^{\pm 6.81}$	$76.82^{\pm 5.91}$	$40.84^{\pm 3.1}$	$69.1^{\pm 4.91}$	$77.81^{\pm 4.74}$	$83.74^{\pm 4.45}$	$34.78^{\pm 1.56}$	$52.62^{\pm 4.02}$	$60.69^{\pm 3.79}$	$68.08^{\pm 5.3}$
Att. Fusion [1]	$41.36^{\pm 5.39}$	$56.72^{\pm 8.88}$	$67.5^{\pm 7.09}$	$75.63^{\pm 8.71}$	$37.12^{\pm 4.37}$	$64.01^{\pm 6.61}$	$74.61^{\pm 6.78}$	$81.68^{\pm 6.9}$	$34.01^{\pm 3.57}$	$49.78^{\pm 4.78}$	$59.31^{\pm 5.52}$	$68.52^{\pm 5.64}$
MBT [2]	$35.35^{\pm 3.9}$	$56.44^{\pm 6.37}$	$64.39^{\pm 4.54}$	$73.6^{\pm 5.08}$	$36.59^{\pm 2.71}$	$61.76^{\pm 4.32}$	$71.82^{\pm 5.06}$	$79.43^{\pm 4.73}$	$29.91^{\pm 2.95}$	$45.37^{\pm 3.34}$	$56.52^{\pm 4.48}$	$65.19^{\pm 4.18}$
CM-FSL [21]	$36.9^{\pm 3.9}$	$57.5^{\pm 3.9}$	$68.34^{\pm 3.9}$	$76.18^{\pm 3.9}$	$33.54^{\pm 3.9}$	$57.4^{\pm 3.9}$	$70.38^{\pm 3.9}$	$80.11^{\pm 3.9}$	$28.92^{\pm 3.9}$	$46.85^{\pm 3.9}$	$57.08^{\pm 3.9}$	$67.38^{\pm 3.9}$
Zorro [66]	$45.97^{\pm 2.03}$	$59.27^{\pm 4.23}$	$65.66^{\pm 4.04}$	$69.96^{\pm 4.65}$	$49.55^{\pm 2.8}$	$70.74^{\pm 3.83}$	$77.22^{\pm 4.27}$	$82.27^{\pm 4.65}$	$33.83^{\pm 3.07}$	$48.98^{\pm 4.37}$	$54.76^{\pm 5.57}$	$60.73^{\pm 5.23}$
TBN [4]	$68.09^{\pm 6.18}$	$84.40^{\pm 6.46}$	$88.03^{\pm 7.26}$	$90.58^{\pm 6.26}$	$76.32^{\pm 6.37}$	$89.65^{\pm 5.08}$	$\underline{92.02}^{\pm 4.31}$	$\underline{93.24}^{\pm 3.73}$	$\underline{63.15}^{\pm 9.92}$	$\underline{76.42}^{\pm 8.11}$	$79.42^{\pm 7.35}$	$82.45^{\pm 6.03}$
LAVISH [18]	$55.63^{\pm 7}$	$75.03^{\pm 6.26}$	$84.01^{\pm 5.8}$	$88.81^{\pm 4.65}$	$57.61^{\pm 6.2}$	$80.25^{\pm 8.24}$	$87.14^{\pm 4.63}$	$89.63^{\pm 4.61}$	$43.97^{\pm 3.03}$	$66.14^{\pm 5.44}$	$73.7^{\pm 5.25}$	$78.26^{\pm 5.01}$
AV-Diff [33]	$42.58^{\pm 2.52}$	$63.9^{\pm 8.47}$	$71.69^{\pm 7.68}$	$80.88^{\pm 8.04}$	$48.31^{\pm 3.43}$	$75.93^{\pm 5.49}$	$82.82^{\pm 5.23}$	$87.29^{\pm 5.13}$	$41.75^{\pm 3.4}$	$60.96^{\pm 9.42}$	$64.76^{\pm 9.55}$	$71.5^{\pm 8.56}$
PROTO-CAT [32]	$\underline{68.52}^{\pm 5.95}$	$84.51^{\pm 5.49}$	$89.36^{\pm 6.73}$	$\underline{92.55}^{\pm 5.79}$	$\underline{77.27}^{\pm 6.96}$	$\underline{90.14}^{\pm 4.97}$	$91.45^{\pm 4.55}$	$92.73^{\pm 4.07}$	$58.54^{\pm 7.19}$	$74.54^{\pm 9.11}$	$79.03^{\pm 6.99}$	$\underline{82.48}^{\pm 5.92}$
STG-CMA [19]	$65.43^{\pm 7.97}$	$83.77^{\pm 6.32}$	$88.47^{\pm 5.13}$	$91.03^{\pm 4.5}$	$64.46^{\pm 7.01}$	$84.47^{\pm 4.2}$	$88.45^{\pm 3.48}$	$91.07^{\pm 3.84}$	$55.08^{\pm 9.2}$	$72.55^{\pm 7.99}$	$78.71^{\pm 6.5}$	$82.1^{\pm 4.72}$
AV-MoE [47]	$66.23^{\pm 8.43}$	$\underline{84.53}^{\pm 5.06}$	$\underline{89.59}^{\pm 6.19}$	$91.77^{\pm 5.96}$	$65.19^{\pm 7.33}$	$85.22^{\pm 7.22}$	$88.67^{\pm 6.24}$	$91.84^{\pm 5.47}$	$57.92^{\pm 8.63}$	$73.37^{\pm 8.3}$	$\underline{79.81}^{\pm 7.41}$	$82.29^{\pm 7.03}$
SMP (Constant)	$\underline{68.7}^{\pm 7.51}$	$\underline{87.83}^{\pm 6.27}$	$\underline{90.87}^{\pm 5.69}$	$\underline{92.65}^{\pm 4.29}$	$\underline{77.66}^{\pm 2.81}$	$\underline{90.26}^{\pm 3.98}$	$\underline{92.93}^{\pm 3.67}$	$\underline{93.71}^{\pm 2.89}$	$\underline{64.42}^{\pm 7.48}$	$\underline{77.26}^{\pm 7.55}$	$\underline{81.13}^{\pm 6.44}$	$\underline{83.52}^{\pm 5.11}$
SMP (VLM)	$\mathbf{85.74}^{\pm 6.99}$	$\mathbf{94.82}^{\pm 5.24}$	$\mathbf{95.5}^{\pm 5.69}$	$\mathbf{96.46}^{\pm 4.83}$	$\mathbf{80.23}^{\pm 5.15}$	$\mathbf{91.48}^{\pm 5.16}$	$\mathbf{93.23}^{\pm 4.83}$	$\mathbf{94.29}^{\pm 4.52}$	$\mathbf{74.37}^{\pm 13.1}$	$\mathbf{81.98}^{\pm 10.74}$	$\mathbf{84.64}^{\pm 8.9}$	$\mathbf{85.92}^{\pm 7.56}$

overfitting; CM-FSL [21], which leverage another modality to enhance image classification; Zorro [66], which separating unimodal and fused representation flows by applying specific masking strategy; and Att. Fusion [1], which uses a simple neural network to modulate the weight of modalities. We additionally use the late fusion model as the baseline, in which training from scratch means the model is directly trained on target classes, and finetuning means the model is pretrained on source data first and then adapts to the target dataset.

B. Comparative Analysis

1) Analysis of Framework Design: The results of the comparison with diverse configurations are shown in Tab. I. Models integrated with SMP, whether using constant or VLM prompts, all illustrate significant performance improvements compared to corresponding baseline models. It shows that SMP can cope with the challenges in FS-AVC tasks under various prompt qualities. Additionally, by comparing the performance of Text CLIP, AVP, and Audio-Visual SMP models, it is evident that **models relying solely on VLM prompts or simply fusing prompts cannot effectively address the FS-AVC task**. Particularly on the VGGSound100 dataset, the performance of the AVP and Text CLIP model is even lower than that of the baseline audio model. This highlights the limitations of VLM in the FS-AVC task. However, this level of VLM prompts quality is good enough to stimulate the potential of SMP. When models integrated with SMP use VLM prompts, the performance on FS-AVC tasks is further improved compared to models with constant prompts. Similarly, by comparing the performance of the Audio-Visual SMP model with unimodal SMP models, it can be demonstrated that the complete SMP effectively addresses the fusion difficulties and modality imbalance problems in FS-AVC, leveraging complementary information from multiple modalities.

2) Performance Analysis of SMP and Other Methods: The results of comparison with other audio-visual learning methods are shown in Tab. II, SMP with different semantic prompts achieves SOTA performance in all experiments. It shows remarkable benefits when the number of training samples is further reduced, proving that with the support of P-AVeL and P-PR, the model can excellently cope with various challenges in FS-AVC. Other methods that followed closely in performance include PROTO-CAT, AV-MoE, and TBN. PROTO-CAT's performance shows the consistency of the few-shot AVSR task with FS-AVC. However, it still contains a noticeable performance gap between the post-fusion method and the SMP. AV-MoE, which equips different adapter experts to process unimodal and multimodal information, has achieved second place in several settings, but it has brought additional parameters. TBN's performance proves the importance of temporal information, but extracting it requires more complex calculations. STG-CMA and LAVISH achieve fourth and fifth place in all experiments, demonstrating that traditionally learnable prompt tokens have limited learning ability when the data is insufficient. AV-Diff, designed for the generalized FSL task, does not perform as expected in all settings. It reveals the limitations of using additional generative models for feature augmentation in few-shot settings.

3) Comparison with Few-shot Unimodal Learning Methods: Tab. III shows the experiment results of our proposed framework and representative unimodal FSL methods on the AVE dataset. Few-shot video models include the representative OTAM [67], which uses temporal alignment on frames, and the recent MASTAF [68], which uses spatial and temporal attention. Audio models include HalluAudio [69], which uses

TABLE III

COMPARATIVE WITH UNIMODAL FSL METHODS ON AVE [36]. THE TABLE REPORTS THE AVERAGE CLASSIFICATION ACCURACY (%) FOR 25 TEST SESSIONS. N AND K REPRESENT FEW-SHOT SETTINGS WITH N CATEGORIES AND K SAMPLES PER CATEGORY. THE "CONSTANT" AND "VLM" INDICATE WHETHER THE SEMANTIC PROMPT USED BY SMP IS A SIMPLE "*A video of [label]*" OR PROVIDED BY VLM. THE VALUE IN THE UPPER RIGHT CORNER IS THE STANDARD DEVIATION OF ALL TEST SESSIONS. RESULTS DEMONSTRATE THAT SMP EFFECTIVELY UTILIZES COMPLEMENTARY MULTIMODAL INFORMATION AND ACHIEVES GOOD PERFORMANCE.

Modality	Methods \ (N, K)	(5, 1)	(5, 5)	(5, 10)	(5, 20)
Video	Base Visual model	$30.51^{\pm 3.65}$	$42.66^{\pm 6.01}$	$51.5^{\pm 7.65}$	$59.75^{\pm 7.49}$
	OTAM [67]	$68.52^{\pm 1.54}$	$81.17^{\pm 3.78}$	$85.61^{\pm 5.64}$	$88.78^{\pm 4.95}$
	MASTAF [68]	$55.98^{\pm 2.32}$	$71.77^{\pm 4.16}$	$75.4^{\pm 5.4}$	$79.11^{\pm 6.51}$
Audio	Base Audio model	$35.4^{\pm 6.03}$	$56.92^{\pm 7.06}$	$61.37^{\pm 6.16}$	$66.67^{\pm 5.33}$
	HalluAudio [69]	$40.85^{\pm 3.61}$	$58.37^{\pm 2.97}$	$63.33^{\pm 5.71}$	$66.74^{\pm 5.85}$
	HA-ProtoNets [70]	$44.67^{\pm 5.59}$	$63.76^{\pm 4.61}$	$68.87^{\pm 5.55}$	$71.55^{\pm 4.17}$
Video & Audio	Base Audio-Visual model	$33.55^{\pm 2.92}$	$47.25^{\pm 5.36}$	$59.57^{\pm 6.52}$	$66.28^{\pm 7.03}$
	SMP (Constant)	$\underline{68.7}^{\pm 7.51}$	$\underline{87.83}^{\pm 6.27}$	$\underline{90.87}^{\pm 5.69}$	$\underline{92.65}^{\pm 4.29}$
	SMP (VLM)	$\mathbf{85.74}^{\pm 6.99}$	$\mathbf{94.82}^{\pm 5.24}$	$\mathbf{95.5}^{\pm 5.69}$	$\mathbf{96.46}^{\pm 4.83}$

features from different frequency ranges, and recent HA-ProtoNets [70], which applied hybrid attention on query and support samples. For a fair comparison, we use the same ViT backbone and list the base model results as references. The results show that even the SOTA unimodal method still has a significant performance gap from SMP, which proves our framework's ability to use complementary multimodal information. Interestingly, visual models gain greater improvements from base models than audio models. Aside from the modality preference of the dataset, video methods, such as OTAM, individually compare sampled frames with the class prototype to obtain a combined prediction, rather than treating frames as a batch. Although this operation gains good performance, it increases computational costs.

4) Analysis on the Parameter Efficiency: Compared with other adapter-based methods, SMP uses fewer trainable parameters (2.56M for SMP *vs.* 4.64M for LAVISH, 11.5M for STG-CMA, and 47.7M for AV-MoE) and achieves better performance in all settings. SMP has higher trainable parameter efficiency than other few-shot methods like AV-Diff (8.31M) and PROTO-CAT (8.22M). This is due to the designed P-AVeL, which uses a small number of text tokens from prompts to align and fuse multimodal information in the latent space with an efficient parameter structure. Coupled with the rebalance capability brought by P-PR, our model can converge fast in several epochs, which leads to an efficient training process.

5) Visualization Results: We provide t-SNE [5] and attention map visualization results to illustrate the framework's effectiveness. As shown in the t-SNE results within Fig. 5, we compared the convergence effect on test sets of our SMP and LAVISH on the two datasets under the randomly selected 5-way 1-shot tasks. Lines (a) and (b) correspond to results on VGGSound100 and Kinetics-Sounds datasets, respectively; the first and second columns correspond to LAVISH and our visualization results, respectively. It shows that SMP can obtain better decision boundaries in the 1-shot tests and proves better abilities in addressing overfitting.

In Fig. 6, we compare the attention maps between SMP with LAVISH in a randomly selected 5-way, 1-shot experiment on the Kinetics-Sounds dataset. Following [2], we use Attention Rollout [71] to compute attention maps from the CLS tokens to visual space. Due to the characteristics of the dataset and the FS-AVC tasks, the attention should be focused on sound source regions in the video. Results show that our SMP pays more accurate attention to the sound source area in the frames.

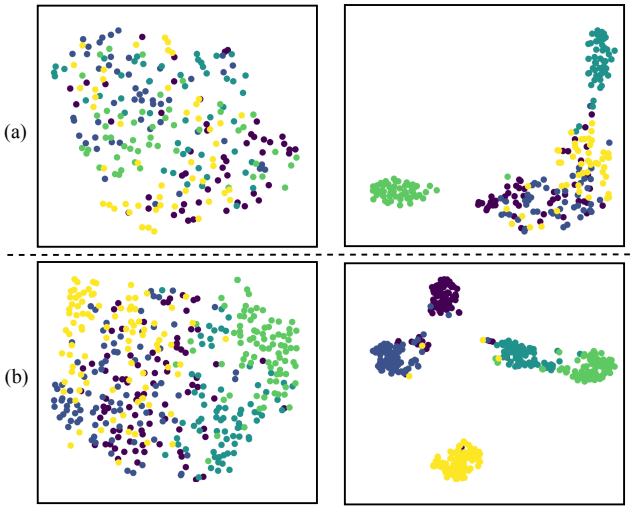

Fig. 5. **The t-SNE visualization.** (a) t-SNE visualization figure of a randomly chosen 5-way 1-shot test sample on VGGSound100. (b) t-SNE visualization figure of a randomly chosen 5-way 1-shot test sample on Kinetics-Sounds. *Left* figures are obtained with LAVISH [18], and *right* figures with clearer decision boundaries are achieved by our SMP.

C. Ablation Study

1) Effect of P-AVeL: We remove all P-AVeLs from SMP to validate their efficacy. Comparing the result in the third row of Tab. IV with the best model, the performance drops over 18% in the 1-shot setting. This fully proves the effectiveness of P-AVeL, which leverages several parameters to effectively complete information interaction between modalities and rapid learning through prompting refinement.

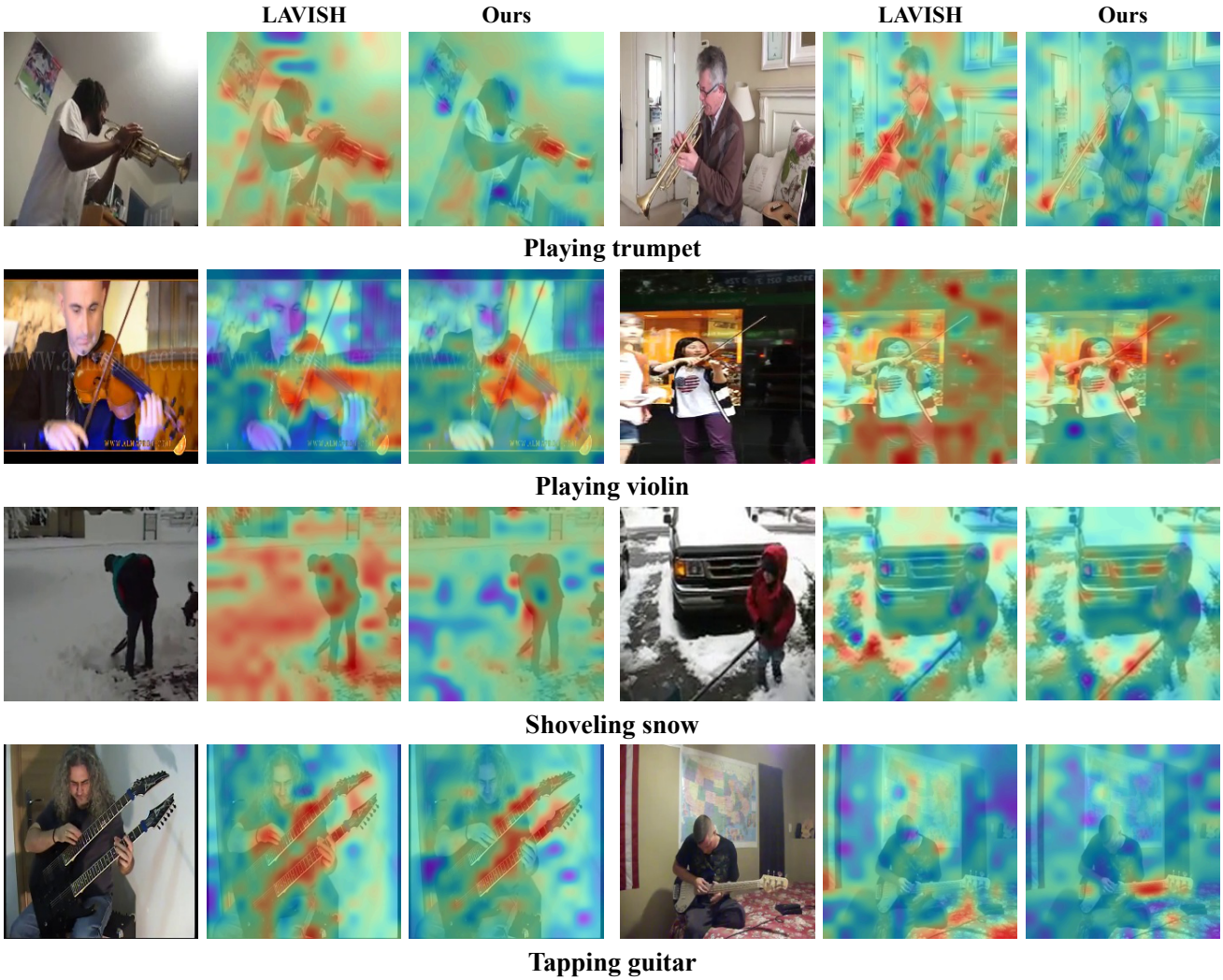

Fig. 6. **Attention maps.** We compute attention maps from the output CLS tokens of the visual space for the LAVISH [18] model and SMP in a randomly selected 5-way, 1-shot experiment on the Kinetics-Sounds test set. For each video, we show the original frame on the left with the ground truth labels at the bottom. SMP's better performance allows the model to focus on more precise sound source regions on each frame.

TABLE IV
ABLATION STUDY OF P-AVEL AND P-PR ON AVE [36] WITH VLM PROMPTS. NOTE THAT ✓ AND ✗ MEAN WHETHER THE COMPONENT IS APPLIED. EXPERIMENTS SHOW THAT P-AVEL AND P-PR EFFECTIVELY IMPROVE THE MODEL'S PERFORMANCE.

P-AVeL	P-PR	(5, 1)	(5, 5)	(5, 10)	(5, 20)
✗	✗	33.55	47.25	59.57	66.28
✓	✗	73.59	87.26	92.21	94.28
✗	✓	67.47	84.96	88.76	91.25
✓	✓	**85.74**	**94.82**	**95.5**	**96.46**

TABLE V
ABLATION STUDY FOR THE EFFECT OF SEMANTIC PROMPTING ON AVE [36]. NOTE THAT THE ✓ AND ✗ MEAN WHETHER PROMPTING IS APPLIED IN THE CORRESPONDING MODULE. WE USE PROMPTING IN BOTH P-AVEL AND P-PR BY CONSIDERING THE BEST PERFORMANCE.

P-AVeL	P-PR	(5, 1)	(5, 5)	(5, 10)	(5, 20)
✗	✗	64.3	83.19	87.93	90.31
✗	✓	68.51	86.9	91.03	92.96
✓	✗	82.71	92.41	95.09	95.92
✓	✓	**85.74**	**94.82**	**95.5**	**96.46**

2) Effect of P-PR: In the second row of Tab. IV, we only remove P-PR from SMP. Under the 1-shot setting, the P-PR affects the accuracy by over 12%. As training samples increase, the performance gap gradually becomes smaller. This phenomenon occurs because models are more easily influenced by modality imbalance in extreme data-lacking cases, which results in the inability of the model to converge within a reasonable training period.

3) Effect of Semantic Prompting: We conduct ablation studies of semantic prompting by removing it from P-AVeL and P-PR. In the second row of Tab. V, the performance significantly dropped on the model without prompt refinement

TABLE VI
ABLATION STUDY OF THE EFFICACY OF DIFFERENT PHASES OF PROMPT-GUIDED LATENT ATTENTION ON AVE [36]. NOTE THAT THE ✓ AND ✗ MEAN WHETHER THE PHASE IS CORRECTLY USED IN THE LATENT ATTENTION MECHANISM. WE UTILIZE ALL THREE PHASES OF LATENT ATTENTION TO ACHIEVE OPTIMAL PERFORMANCE.

Phase 1	Phase 2	Phase 3	(5, 1)	(5, 5)	(5, 10)	(5, 20)
✗	✗	✗	73.07	85.91	90.27	93.58
✗	✓	✓	80.22	92.59	95.15	96.32
✓	✗	✗	77.11	90.23	92.77	94.32
✓	✓	✗	80.66	91.35	93.03	93.55
✓	✗	✓	84.61	94.32	95.36	96.07
✓	✓	✓	**85.74**	**94.82**	**95.5**	**96.46**

TABLE VII
ABLATION STUDY OF P-AVEL LOCATION ON AVE [36]. LOC. 1 REPRESENTS THAT P-AVELS ARE IN PARALLEL WITH THE MULTI-HEAD SELF-ATTENTION LAYER IN TRANSFORMER BLOCKS, AND LOC. 2 REPRESENTS THAT THEY ARE IN PARALLEL WITH THE MLP LAYERS. NOTE THAT THE ✓ AND ✗ MEAN WHETHER THE P-AVEL IS APPLIED AT THE CORRESPONDING LOCATION OF THE VIT BLOCKS. WE APPLY P-AVEL ONLY AT LOC. 2 OF VIT BLOCKS, CONSIDERING PERFORMANCE AND PARAMETER EFFICIENCY.

Loc. 1	Loc. 2	(5, 1)	(5, 5)	(5, 10)	(5, 20)
✗	✗	75.23	88.2	93.08	94.94
✓	✗	84.28	93.75	94.88	96.12
✓	✓	85.42	94.26	95.35	96.28
✗	✓	**85.74**	**94.82**	**95.5**	**96.46**

in P-AVeL, in which latent attention only involves audio and visual tokens. It reveals that semantic prompting in P-AVeL can effectively refine the learning process of audio and visual modalities, bringing more than a 17% performance increase in 1-shot scenarios. Additionally, the third row of Tab. V shows the accuracy of the model without prompt-tuned operation in P-PR. In this case, modality prototypes obtained with Eq. 6 only contain $\mathbf{P}_i^a$ and $\mathbf{P}_i^v$. Results show that prompt modulation can bring more than 2% performance improvements in 1- and 5-shot settings. It proves that prompt-tuned prototypes can better tackle the challenges of modality imbalance in data-limited scenarios.

4) Effect of Different Phases of Prompt-Guided Latent Attention: In Tab VI, we compare the impact of the three phases of prompt-guided latent attention on performance in detail. In Sec. III and Fig. 3, we introduced that the three phases correspond to Text-aware Attention, Modality Fusion, and Modality Summation. We compared the impact of correctly applying the three phases on model performance.

Precisely, latent attention will not execute the Text-aware Attention phase when phase 1 is not applied correctly (in the settings with a cross at "phase 1" in the table). At this point, prompt-guided latent attention will be performed as follows:

In phase 2, Modality Fusion:

$$\begin{aligned} \mathbf{T}^{v,a} &= \mathrm{CMA}(\mathbf{T}^v, \mathbf{T}^a, \mathbf{T}^a), \quad \mathbf{T}^{v,t} = \mathrm{CMA}(\mathbf{T}^v, \mathbf{T}^t, \mathbf{T}^t), \\ \mathbf{T}^{a,v} &= \mathrm{CMA}(\mathbf{T}^a, \mathbf{T}^v, \mathbf{T}^v), \quad \mathbf{T}^{a,t} = \mathrm{CMA}(\mathbf{T}^a, \mathbf{T}^t, \mathbf{T}^t), \\ \mathbf{T}^{t,v} &= \mathrm{CMA}(\mathbf{T}^t, \mathbf{T}^v, \mathbf{T}^v), \quad \mathbf{T}^{t,a} = \mathrm{CMA}(\mathbf{T}^t, \mathbf{T}^a, \mathbf{T}^a). \end{aligned} \tag{14}$$

In phase 3, Modality Summation:

$$\begin{aligned} \overline{\mathbf{T}}^v &= \lambda^{v,a}\mathbf{T}^{v,a} + \lambda^{v,t}\mathbf{T}^{v,t} + \mathbf{T}^v, \\ \overline{\mathbf{T}}^a &= \lambda^{a,v}\mathbf{T}^{a,v} + \lambda^{a,t}\mathbf{T}^{a,t} + \mathbf{T}^a, \\ \overline{\mathbf{T}}^t &= \lambda^{t_1}\mathbf{T}^{t,v} + \lambda^{t_2}\mathbf{T}^{t,a} + \mathbf{T}^t. \end{aligned} \tag{15}$$

Since phases 2 and 3 implement information fusion and summation, if these two modules are removed as a whole, latent attention will not be able to interact with information as designed. Thus, we focus on whether to apply the learnable weight parameter λ to tokens in phase 2 and phase 3. In Tab. VI, settings with crosses at "phase 2" or "phase 3" means we remove all learnable weights λ in corresponding equations (Eq. 3, Eq. 4 and Eq. 15).

Comparing the first row of results in the table with the final results, it shows that when all phases are not working correctly, the model's performance is greatly affected. Although the performance has been improved to a certain extent with the addition of learnable parameters of tokens (second line in Tab. VI), there is still a performance gap of more than 5% in the 1-shot experiment from the fully functional version. Moreover, in the model that uses phase 1 correctly, we only need four cross-modal attentions to achieve information interaction between the three modalities. Compared with a model without phase 1 that requires six cross-modal attentions, the computational complexity is reduced by one-third, and more efficient modality fusion is achieved.

The effectiveness of the learnable weight parameter λ in different phases can be demonstrated by comparing the results of the third to sixth rows in Tab. VI. It demonstrates that letting the model learn how to balance tokens of different modalities in phases 2 and 3 can bring practical performance improvements. Among them, the learnable weights in phase 3 have a more significant gain in model performance.

In the above experiments, we prove the rationality and effectiveness of the design of prompt-guided latent attention in P-AVeL through ablation experiments.

5) Robustness Study on Temporal Asynchrony: To quantitatively evaluate the robustness of SMP against temporal asynchrony, we conducted a stress test on the AVE dataset by manually introducing fixed time delays (1, 3, and 5 seconds) to the audio track. As reported in Tab. VIII, SMP demonstrates superior stability compared to existing methods. While the performance of LAVISH [18] significantly deteriorates as the delay increases (e.g., a drop of up to 13.34% at a 5s delay), SMP maintains a remarkably stable performance with only marginal fluctuations (e.g., less than 2% drop in most cases). Notably, comparing SMP with the "AVeL + PR" baseline (our model without semantic prompt modulation) reveals that the inclusion of semantic prompts significantly buffers the impact of temporal shifts. This suggests that the high-level semantic guidance provided by our VLM prompts allows the model to focus more on categorical and contextual correspondences rather than being strictly constrained by low-level temporal synchronization. These results confirm that SMP's structural design effectively handles extreme temporal misalignment,

TABLE VIII
ABLATION STUDY OF ROBUSTNESS TO TEMPORAL ASYNCHRONY. WE MANUALLY INTRODUCE TIME DELAYS (1, 3 AND, 5 SECONDS) TO THE AUDIO SIGNALS WITHIN THE TEST SAMPLES OF THE AVE DATASET TO EVALUATE THE MODEL'S STABILITY AGAINST EXTREME TEMPORAL MISALIGNMENT. THE RESULTS DEMONSTRATE SMP'S SUPERIOR ROBUSTNESS TO AUDIO-VISUAL ASYNCHRONY COMPARED TO BASELINE METHODS, OWING TO ITS STRUCTURAL DESIGN AND SEMANTIC PROMPT GUIDANCE. NOTE THE AVEL+PR INDICATES THE SMP WITHOUT PROMPT MODULATION. THE NUMBER IN THE UPPER RIGHT CORNER REPRESENTS THE PERFORMANCE DEGRADATION COMPARED TO THE NO-DELAY TEST.

	Delay 1 second				**Delay 3 seconds**				**Delay 5 seconds**			
Models	(5,1)	(5,5)	(5,10)	(5,20)	(5,1)	(5,5)	(5,10)	(5,20)	(5,1)	(5,5)	(5,10)	(5,20)
LAVISH [18]	$49.86^{-5.77}$	$70.6^{-4.43}$	$77.77^{-6.24}$	$84.75^{-4.06}$	$47.3^{-8.33}$	$66.86^{-8.17}$	$73.57^{-10.44}$	$81.88^{-6.93}$	$45.6^{-10.03}$	$64.63^{-10.4}$	$70.67^{-13.34}$	$78.2^{-10.61}$
AVeL + PR	$62.91^{-1.39}$	$81.25^{-1.94}$	$86.8^{-1.13}$	$89.79^{-0.52}$	$62.1^{-2.2}$	$80.58^{-2.61}$	$86.27^{-1.66}$	$89.3^{-1.01}$	$61.36^{-2.94}$	$79.98^{-3.21}$	$85.89^{-2.04}$	$89.11^{-1.21}$
SMP	$85.04^{\mathbf{-0.7}}$	$93.33^{\mathbf{-1.49}}$	$94.53^{\mathbf{-0.97}}$	$95.97^{\mathbf{-0.49}}$	$84.82^{\mathbf{-0.92}}$	$92.94^{\mathbf{-1.88}}$	$94.85^{\mathbf{-0.65}}$	$95.68^{\mathbf{-0.78}}$	$83.83^{\mathbf{-1.91}}$	$92.89^{\mathbf{-1.93}}$	$94.34^{\mathbf{-1.16}}$	$95.6^{\mathbf{-0.86}}$

TABLE IX
ABLATION STUDY OF DIFFERENT CONFIGURATIONS OF SEMANTIC PROMPTS. WE REPORT THE PERFORMANCE ON THE AVE DATASET WITH 50% IN-BATCH PROMPT SHUFFLING AND COMPARE THE EFFECTIVENESS OF SEMANTIC PROMPTS DERIVED FROM DIFFERENT MODALITIES (AUDIO-BASED (ALM) VS. VISION-BASED (VLM) PROMPTS).

Configurations	**(5, 1)**	**(5, 5)**	**(5, 10)**	**(5, 20)**
SMP (50% shuffle)	79.18	90.87	92.37	93.68
Text CLIP (ALM)	32.92	48.29	62.43	78.75
SMP (ALM)	80.65	91.07	94.09	95.61
Text CLIP (VLM)	48.76	78.88	87.79	91.44
SMP (VLM)	85.74	94.82	95.5	96.46

TABLE X
ABLATION STUDY OF THE LOCATION OF PROMPT-GUIDED LATENT ATTENTION IN P-AVEL ON AVE [36]. REFER TO THE P-AVEL STRUCTURE ILLUSTRATED AT THE BOTTOM LEFT OF FIG. 2, LA. 1 MEANS THE LATENT ATTENTION PROCESSES BEFORE THE MODALITY CONVOLUTIONS, AND LA. 2 MEANS IT PROCESSES AFTER THE MODALITY CONVOLUTIONS IN P-AVEL. NOTE THAT THE ✓AND ✗MEAN WHETHER THE LATENT ATTENTION IS APPLIED AT THE CORRESPONDING LOCATION IN P-AVEL. WE APPLY LATENT ATTENTION AT BOTH LOCATIONS, CONSIDERING THE BEST PERFORMANCE.

LA. 1	**LA. 2**	**(5, 1)**	**(5, 5)**	**(5, 10)**	**(5, 20)**
✗	✗	80.13	92.16	93.89	94.92
✓	✗	81.97	94.5	95.32	95.75
✗	✓	84.87	94.63	95.38	96
✓	✓	**85.74**	**94.82**	**95.5**	**96.46**

ensuring its reliability in complex real-world environments.

6) Robustness and Modality-Agnostic Study of Semantic Prompts: To further investigate the robustness of the proposed SMP framework on different semantic prompt settings, we conducted a series of experiments, with results summarized in Tab. IX.

Firstly, to examine the robustness of SMP to semantic prompt quality, we introduced a "50% in-batch shuffling" mechanism during the test, where half of the samples were intentionally paired with semantic prompts belonging to incorrect categories from the same batch. As observed in Tab. IX, while the performance of the SMP (50% shuffle) configuration naturally exhibits a slight decrease compared to the clean VLM-based prompts, it maintains a remarkably high accuracy. This resilience suggests that our SMP architecture does not merely "copy" information from the text prompt; instead, it utilizes the prompt as a flexible semantic guide, effectively cross-referencing it with the actual audio-visual evidence to maintain robust performance.

Secondly, to mitigate potential vision-centric bias inherent in video-derived prompts, we explored the integration of audio-derived semantic guidance. By utilizing the Qwen3-Omni [72] model to generate acoustic captions (denoted as ALM in Tab. IX), we evaluated the framework's ability to ground semantic information from an alternative modality. The results indicate that SMP (ALM) achieves competitive performance levels that are nearly on par with the vision-based counterpart. Notably, both ALM and VLM configurations of SMP significantly outperform the baseline Text CLIP method across all settings. This once again proves that SMP can effectively utilize imperfect semantic prompts to achieve significant performance gains.

7) Impact of the Location of P-AVeL: In Tab. VII, we conduct an ablation study of the impact of different location settings of P-AVeL. In the table, "Loc. 1" represents the case where P-AVeLs are in parallel with the self-attention layers in Transformer blocks, and "Loc. 2" represents the case where P-AVeLs are in parallel with the MLP layers. Results show that deploying P-AVeLs at both locations simultaneously does not yield further performance variation compared to using Loc. 2 alone. The design of P-AVeL is similar to imitating the functions of MLP layers, and the sufficient self-attention features are more conducive to the interaction of modal information. Therefore, we only incorporated P-AVeL at "Loc. 2" by considering performance and parameter efficiency.

8) Impact of the Location of Prompt-Guided Latent Attention: In Tab. X, we compare the location impact of the prompt-guided latent attention in the proposed P-AVeL. In the table, LA. 1 represents the latent attention process before modality convolution, and LA. 2 represents the latent attention process after modality convolutions. Compared to the model without applying latent attention (i.e., the first line in Tab. X), it is clear that latent attentions play key roles in modality fusion. Since latent attention is an essential channel for information interaction between modalities, we chose to apply latent attention in two places simultaneously for better results.

9) Impacts of Fusion Beginning Layer and Downsampling Dimension: We also conducted experiments on the hyperpa-

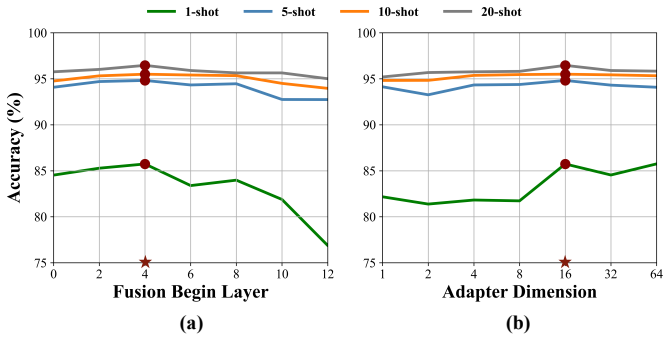

Fig. 7. **The impact of the fusion begin layer (a) and adapter downsampling dimension (b) on AVE [36].** Red markers indicate the best performance points.

rameters that affect model performance, namely the fusion beginning layer and the downsampling dimension. The fusion beginning layer indicates from which Transformer block we equip P-AVeL to conduct multi-modal fusion. The downsampling dimension means that when tokens enter the P-AVeL, they will be downsampled into the dimension needed for prompt-guided latent attentions. As shown in Fig. 7, the model achieves the best results when the fusion beginning layer is set to 4, and the downsampling dimension is set to 16. This setting enables the model to achieve a balance between unimodal feature extraction and multimodal feature interaction.

V. Conclusion

In this work, we have presented a novel prompting framework, SMP. It innovatively introduces semantic prompts to address overfitting, temporal asynchrony, and modality imbalance issues in FS-AVC. SMP equips P-AVeLs, which leverage semantic prompts to modulate the multimodal learning process and effectively achieve multimodal alignment. SMP also proposes P-PR, the first modal rebalance method designed for few-shot scenarios, which evaluates the learning status of every batch to rebalance the contributions of modalities and mitigate the impact of prototype deviation by prompt modulation.

Experiments show SMP achieved SOTA performance across all FS-AVC test settings. SMP achieves good performance even with the most straightforward constant prompts, revealing the framework's high adaptability to prompt quality. The performance improvement brought by the VLM prompts reveals the potential to achieve higher performance when better semantic prompts are provided. Extensive ablation studies validate the contributions of key components like P-AVeL and P-PR. In addition, the steady performance improvement in FS-AVC reaffirms the effectiveness of semantic prompting in data-efficient scenarios.

This work marks a new pivotal advancement in data—and parameter-efficient audio-visual learning. In combination with efficient multimodal fusion and dynamic modality rebalancing, SMP addresses the three main challenges in FS-AVC. It opens new horizons for developing data-limited audio-visual learning and modality rebalancing. Future work could try to apply the semantic prompting mechanism to a broader range of data-limited audio-visual learning scenarios.

VI. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62471420), in part by Guangdong Basic and Applied Basic Research Foundation (2025A1515012296), in part by 2025 Tencent AI Lab Rhino-Bird Program, in part by Guangdong Provincial Project under Grant 2021JC02X149, in part by Guangzhou Municipal Science and Technology Project under Grant 2023A03J0011, in part by Guangzhou Municipal Key Laboratory on Future Networked Systems (024A03J0623), and in part by Guangdong Provincial Key Laboratory of Integrated Communications, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

References

[1] H. Fayek and A. Kumar, "Large scale audiovisual learning of sounds with weakly labeled data," in *IJCAI*, 2021.
[2] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in NeurIPS*, vol. 34, pp. 14 200–14 213, 2021.
[3] J. B. Li, K. Ma, S. Qu, P.-Y. Huang, and F. Metze, "Audio-visual event recognition through the lens of adversary," in *ICASSP*, 2021.
[4] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019.
[5] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[6] A. T. de Pablos, "Complementary-contradictory feature regularization against multimodal overfitting," *WACV*, 2024.
[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in NeurIPS*, vol. 30, 2017.
[8] S. Baik, J. Choi, H. Kim, D. Cho, J. Min, and K. M. Lee, "Meta-learning with task-adaptive loss function for few-shot learning," in *CVPR*, 2021.
[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.
[10] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *Advances in NeurIPS*, vol. 35, pp. 28 708–28 720, 2022.
[11] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, 2021.
[12] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*, 2020.
[13] Y. Sun, S. Mai, and H. Hu, "Learning to balance the learning rates between various modalities via adaptive tracking factor," *IEEE Signal Processing Letters*, vol. 28, pp. 1650–1654, 01 2021.
[14] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022.
[15] C. Du, T. Li, Y. Liu, Z. Wen, T. Hua, Y. Wang, and H. Zhao, "Improving multi-modal learning with uni-modal teachers," *arXiv*, 2021.
[16] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv*, 2020.
[17] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.
[18] Y.-B. Lin, Y.-L. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in *CVPR*, 2023.
[19] K. Wang, Y. Tian, and D. Hatzinakos, "Towards efficient audio-visual learners via empowering pre-trained vision transformers with cross-modal adaptation," in *CVPR*, 2024.
[20] F. Yang, R. Wang, and X. Chen, "Semantic guided latent parts embedding for few-shot learning," *WACV*, 2023.
[21] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models," in *CVPR*, 2023.
[22] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*, 2019.
[23] S. Jie and Z.-H. Deng, "Convolutional bypasses are better vision transformer adapters," *arXiv*, 2022.

[24] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in *CVPR*, 2023.

[25] J. Li, Z. Wang, and X. Hu, "Learning intact features by erasing-inpainting for few-shot classification," in *AAAI*, 2021.

[26] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervised learning for few-shot medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1837–1848, 2022.

[27] Y. Guo and N.-M. Cheung, "Attentive weights generation for few-shot learning via information maximization," in *CVPR*, 2020.

[28] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2016.

[29] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv*, 2018.

[30] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018.

[31] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, and C. Gagné, "Matching feature sets for few-shot image classification," in *CVPR*, 2022.

[32] Y.-K. Zhang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Audio-visual generalized few-shot learning with prototype-based co-adaptation." in *Interspeech*, 2022.

[33] O.-B. Mercea, T. Hummel, A. S. Koepke, and Z. Akata, "Text-to-feature diffusion for audio-visual few-shot learning," in *DAGM GCPR*, 2023.

[34] T. Mahmud and D. Marculescu, "Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization," in *WACV*, 2023.

[35] V. Rao, M. I. Khalil, H. Li, P. Dai, and J. Lu, "Dual perspective network for audio-visual event localization," in *ECCV*, 2022.

[36] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.

[37] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *CVPR*, 2019.

[38] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *ICCV*, 2019.

[39] H. Cheng, Z. Liu, H. Zhou, C. Qian, W. Wu, and L. Wang, "Joint-modal label denoising for weakly-supervised audio-visual video parsing," in *ECCV*, 2022.

[40] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang, "Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing," *Advances in NeurIPS*, vol. 34, pp. 11 449–11 461, 2021.

[41] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," *Advances in NeurIPS*, vol. 35, pp. 34 722–34 733, 2022.

[42] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, "Audio-visual segmentation," in *ECCV*, 2022.

[43] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *CVPR*, 2022.

[44] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in *ICCV*, 2021.

[45] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

[46] H. Duan, Y. Xia, Z. Mingze, L. Tang, J. Zhu, and Z. Zhao, "Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks," *Advances in NeurIPS*, vol. 36, pp. 56 075–56 094, 2023.

[47] Y. Cheng, Y. Li, J. He, and R. Feng, "Mixtures of experts for audio-visual learning," in *Advances in NeurIPS*, vol. 37, 2024, pp. 219–243.

[48] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv*, 2021.

[49] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv*, 2021.

[50] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," *arXiv*, 2021.

[51] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[52] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[53] K. Zhou, J. Yang, and C. C. Loy, "Conditional prompt learning for vision-language models," in *CVPR*, 2022.

[54] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[55] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *CVPR*, 2023.

[56] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *ICLR*, 2020.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in NeurIPS*, vol. 30, 2017.

[59] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv*, 2017.

[60] W. Pian, S. Mo, Y. Guo, and Y. Tian, "Audio-visual class-incremental learning," in *ICCV*, 2023.

[61] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[63] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv*, 2021.

[64] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou, "mplug-2: A modularized multi-modal foundation model across text, image and video," *ArXiv*, 2023.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, 2017.

[66] A. Recasens, J. Lin, J. Carreira, D. Jaegle, L. Wang, J.-B. Alayrac, P. Luc, A. Miech, L. Smaira, R. Hemsley, and A. Zisserman, "Zorro: the masked multimodal transformer," 2023.

[67] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *CVPR*, 2020.

[68] X. Liu, H. Zhang, and H. Pirsiavash, "Mastaf: A model-agnostic spatio-temporal attention fusion network for few-shot video classification," in *WACV*, 2023.

[69] Z. Yu, S. Wang, L. Chen, and Z. Cheng, "Halluaudio: Hallucinate frequency as concepts for few-shot audio classification," in *ICASSP*, 2023.

[70] Y. Wang and D. V. Anderson, "Hybrid attention-based prototypical networks for few-shot sound classification," in *ICASSP*, 2022.

[71] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *ACL*, 2020.

[72] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, Y. Lv, Y. Wang, D. Guo, H. Wang, L. Ma, P. Zhang, X. Zhang, H. Hao, Z. Guo, B. Yang, B. Zhang, Z. Ma, X. Wei, S. Bai, K. Chen, X. Liu, P. Wang, M. Yang, D. Liu, X. Ren, B. Zheng, R. Men, F. Zhou, B. Yu, J. Yang, L. Yu, J. Zhou, and J. Lin, "Qwen3-omni technical report," *arXiv*, 2025.