

AV-PEA: Parameter-Efficient Adapter for Audio-Visual Multimodal Learning

Abduljalil Radman^a and Jorma Laaksonen^b
Department of Computer Science, Aalto University, Finland

Keywords: Parameter-Efficient, Audio-Visual Adapter, Audio-Visual Fusion, Multimodal Learning.

Abstract: Fine-tuning has emerged as a widely used transfer learning technique for leveraging pre-trained vision transformers in various downstream tasks. However, its success relies on tuning a significant number of trainable parameters, which could lead to significant costs in terms of both model training and storage. When it comes to audio-visual multimodal learning, the challenge also lies in effectively incorporating both audio and visual cues into the transfer learning process, especially when the original model has been trained with unimodal samples only. This paper introduces a novel audio-visual parameter-efficient adapter (AV-PEA) designed to improve multimodal transfer learning for audio-visual tasks. Through the integration of AV-PEA into a frozen vision transformer, like the visual transformer (ViT), the transformer becomes adept at processing audio inputs without prior knowledge of audio pre-training. This also facilitates the exchange of essential audio-visual cues between audio and visual modalities, all while introducing a limited set of trainable parameters into each block of the frozen transformer. The experimental results demonstrate that our AV-PEA consistently achieves superior or comparable performance to state-of-the-art methods in a range of audio-visual tasks, including audio-visual event localization (AVEL), audio-visual question answering (AVQA), audio-visual retrieval (AVR), and audio-visual captioning (AVC). Furthermore, it distinguishes itself from competitors by enabling seamless integration into these tasks while maintaining a consistent number of trainable parameters, typically accounting for less than 3.7% of the total parameters per task.


1 INTRODUCTION


Fine-tuning large-scale pre-trained transformers (e.g. CLIP (Radford et al., 2021), BERT (Bugliarelli et al., 2021), ViT (Dosovitskiy et al., 2021)) has proven its high efficacy in achieving remarkable performance across various downstream tasks. However, fine-tuning such large-scale models for downstream tasks using relatively small datasets can potentially lead to overfitting (Lin et al., 2023). The mismatch in scale between the model’s capacity and the available downstream data may also impede the effective generalization of large-scale pre-trained models to new downstream tasks.

In contrast to unimodal models that depend on samples from a single modality tailored for a specific unimodal task, such as audio (Gong et al., 2021b), visual (Dosovitskiy et al., 2021), or text (Brown et al., 2020), multimodal models aim to leverage correlations between different modalities. This en-

ables a more comprehensive understanding of complex tasks that involve multiple sources of information, such as audio-visual event localization (AVEL) (Xia and Zhao, 2022), audio-visual question answering (AVQA) (Li et al., 2022; Yun et al., 2021), audio-visual retrieval (AVR) (Lin et al., 2022), and audio-visual captioning (AVC) (Chen et al., 2023). These models have gained significant attention due to their ability to handle real-world scenarios where data come from diverse sources and often carry complementary information. However, the requirement for separate curation of audio and visual datasets during pre-training imposes substantial demands on memory and GPU resources. Additionally, the continuous growth in the size of transformers makes full fine-tuning increasingly challenging.

To tackle these challenges, solutions such as parameter-efficient fine-tuning approaches, exemplified by adapter modules (Houlsby et al., 2019; Lin et al., 2023; Sung et al., 2022; Pan et al., 2022), have emerged. Adapter modules have demonstrated excellent performance by introducing a limited set of train-

^a  <https://orcid.org/0000-0002-6317-9752>

^b  <https://orcid.org/0000-0001-7218-3131>

able parameters while keeping the pre-trained model parameters frozen. Freezing the pre-trained model’s parameters allows effective transfer of knowledge gained from a large-scale pre-training dataset to downstream tasks. Moreover, the frozen parameters can be readily shared among different modalities (e.g. audio and visual). This approach not only optimizes resource utilization, but also encourages seamless transfer of knowledge between distinct modalities (Houlsby et al., 2019; Lin et al., 2023).

The main goal of this work is to investigate the capacity of pre-trained vision transformers to generalize across diverse multimodal domains, with a specific emphasis on the field of audio-visual learning. In this context, the core idea revolves around the representation of audio inputs as 2D spectrogram images, which can be jointly processed alongside real visual inputs using a vision transformer. This approach eliminates the need for prior pre-training of the transformer on a separate audio dataset. To achieve this goal, we propose an innovative audio-visual parameter-efficient adapter (AV-PEA) explicitly crafted for multimodal learning. The proposed AV-PEA facilitates seamless adaptation of frozen vision transformers, initially pre-trained on images, to audio-visual tasks. It also effectively leverages the complementary nature of audio and visual modalities through a cross-attention module, all achieved with a limited set of extra trainable parameters. Specifically, within a dual-stream visual transformer, AV-PEA is employed at each layer to enhance the representations of both audio and visual inputs. This enhancement is achieved through a proficient cross-attention module, followed by a lightweight bottleneck block, wherein each stream generates a token dedicated to facilitating information exchange with the other stream. By utilizing a single token from each stream for information exchange, it significantly mitigates the quadratic costs typically associated with traditional cross-attention mechanisms, resulting in enhanced overall efficiency.

The key contributions of our work are as follows: (a) Proposing a novel adapter, AV-PEA, to adapt pre-trained vision transformers for efficient audio learning without the need for a pre-trained audio model with a large-scale dataset. (b) Introducing a simple yet effective token fusion module based on cross-attention, which operates linearly in both computation and memory usage while effectively enhancing the integration of cues from both audio and visual modalities. (c) Demonstrating that our AV-PEA outperforms contemporary audio-visual adapter modules in terms of accuracy and model parameters, achieving performance on par with or exceeding state-of-the-art (SOTA) methods in various audio-visual downstream

tasks, such as AVEL, AVQA, AVR, and AVC. (d) Offering flexibility to integrate our AV-PEA adapter and infuse visual transformers with diverse expert knowledge, eliminating the need for full parameters fine-tuning and requiring only a consistent set of additional trainable parameters within each context.

2 RELATED WORK

Audio-Visual Pre-trained Models. Vision transformer (ViT) (Dosovitskiy et al., 2021) and audio spectrogram transformer (AST) (Gong et al., 2021a) have emerged as cutting-edge solutions for image and audio classification, respectively. Beyond their original specific tasks, these models have shown significant potential as versatile foundations for transfer learning in various downstream tasks (Chen et al., 2023). Typically, they undergo training using extensive labeled datasets (such as ImageNet (Deng et al., 2009) and AudioSet (Gemmeke et al., 2017)) in a supervised manner. However, recent models (Radford et al., 2021; Wang et al., 2023; Guzhov et al., 2022) have embraced multimodal data (e.g. audio-visual and text pairs, image-text pairs, and video-text pairs) resulting in more potent representations.

Audio-Visual Learning. Audio-visual learning tasks evolve on the integration and understanding of information from both audio and visual modalities. The goal is to leverage the complementary information from both modalities to achieve improved performance in various tasks, including but not limited to AVEL (Tian et al., 2018; Xia and Zhao, 2022), AVQA (Li et al., 2022; Yun et al., 2021), AVR (Chen et al., 2023; Li et al., 2022; Yun et al., 2021), AVC (Chen et al., 2023). The AVEL task involves identifying and localizing events within a multimedia context (e.g. video) that are observable in both audio and visual data. The majority of current methods (Tian et al., 2018; Rao et al., 2022; Xia and Zhao, 2022) developed for AVEL tasks in the literature depend on pre-trained audio and visual models tailored to each modality. These models are employed to extract distinct audio and visual features, which are subsequently integrated to facilitate AVEL. AVQA is a task that combines both audio and visual modalities with natural language processing to answer human-generated questions concerning audio-visual content. Similar to the context of AVEL tasks, a significant portion of existing methods designed for the AVQA task relies on audio and vision models specialized for their respective modalities. These models are then merged through spatial and temporal grounding modules to effectively provide meaningful an-

swers (Yun et al., 2021; Li et al., 2022; Schwartz et al., 2019). However, in such contexts, irrelevant audio and visual elements processed by modality-specific models may introduce learning noise, adding complexity to the task. The AVR task involves retrieving relevant multimedia content (i.e. images, videos, or audio clips) based on a query that consists of both audio and visual input, while the AVC task involves crafting informative textual captions for multimedia content that includes both audio and visual elements. Recently, Chen et al. (2023) introduced VALOR, a novel tri-modality (vision, audio, language) pre-trained model and a dataset designed to evaluate audiovisual-language capabilities, including tasks like AVR and AVC.

Parameter-Efficient Transfer Learning (PETL). The PETL principle addresses rising computational demands in natural language processing by inserting lightweight adapter modules between the layers of a pre-trained model (Houlsby et al., 2019). In the same context, PETL has gained significant traction in the computer vision domain, as evidenced by recent works (Sung et al., 2022; Pan et al., 2022; Yang et al., 2023; Lin et al., 2023). Sung et al. (2022) developed a vision-language adapter module that targets the text encoder of the CLIP model. Recently, Pan et al. (2022) and Yang et al. (2023) proposed adapter modules to adapt pre-trained image transformer models for video understanding, concentrating on the video action recognition research.

However, most existing adapter modules in the literature are designed for specific tasks and often lack the ability to effectively facilitate cross-modal information exchange. To the best of our knowledge, the latent audio-visual hybrid (LAVIS) adapter (Lin et al., 2023) stands as a singular instance of PETL modules developed for audio-visual learning. The LAVIS adapter utilizes a compact collection of latent tokens to first compress information from all modality-specific tokens (i.e., audio and video). It subsequently applies cross-attention between these latent tokens and all tokens from the other modality. This enables a two-way flow of information between the audio and video modalities, leading to an enhanced audio-visual representation.

Nonetheless, significant distinctions exist between LAVIS and our AV-PEA. First, LAVIS requires the adjustment of its hyper-parameters for each new audio-visual downstream task. In contrast, our AV-PEA seamlessly integrates into novel audio-visual tasks with a consistent design and invariant parameters, while enjoying better performance and less trainable parameters. Second, LAVIS relies on latent tokens, which are heavily influenced by the downstream

dataset size, for facilitating information exchange between audio and visual modalities. Conversely, our AV-PEA relies exclusively on the classification (*CLS*) token from each modality for cross-modal information exchange, regardless of the downstream dataset size.

3 METHOD

In this section, we introduce AV-PEA, a novel audio-visual adapter designed to fine-tune frozen pre-trained large-scale vision transformers (e.g. ViT (Dosovitskiy et al., 2021)) for various audio-visual downstream tasks (such as AVEL, AVQA, AVR, and AVC), while introducing only a limited set of new trainable parameters. We will begin with a concise overview of ViT as an example of a transformer capable of accommodating the proposed AV-PEA adapter, and then present the AV-PEA approach. Finally, we will delve into the technical details of seamlessly integrating AV-PEA into the ViT transformer.

3.1 ViT Transformer

ViT has attracted attention in the computer vision field for its ability to capture complex relationships among visual components through self-attention mechanisms, consistently achieving exceptional classification performance. In ViT, the input image is transformed into fixed-size patches (tokens) through the patch embedding layer (Figure 1a), with an added *CLS* token for global context representation. Position embeddings are incorporated into each token to capture spatial relationships. These tokens traverse stacked transformer blocks with multiheaded self-attention (MSA) and feed-forward network (FFN) layers, facilitating the integration of crucial visual information across the token sequence. The *CLS* token aggregates the information for the final classification task (Dosovitskiy et al., 2021; Chen et al., 2021).

3.2 The Proposed AV-PEA

Our AV-PEA is founded on a parameter-efficient bottleneck block, as introduced by Houlsby et al. (2019). This bottleneck block is applied on top of a simple cross-attention (CA) module as shown in Figure 1b. Particularly, our AV-PEA capitalizes on the ability of the *CLS* token in ViT to capture abstract information among patch tokens, thus enhancing audio-visual representation through the CA module. To achieve this, we propose a dual-stream ViT transformer seen in

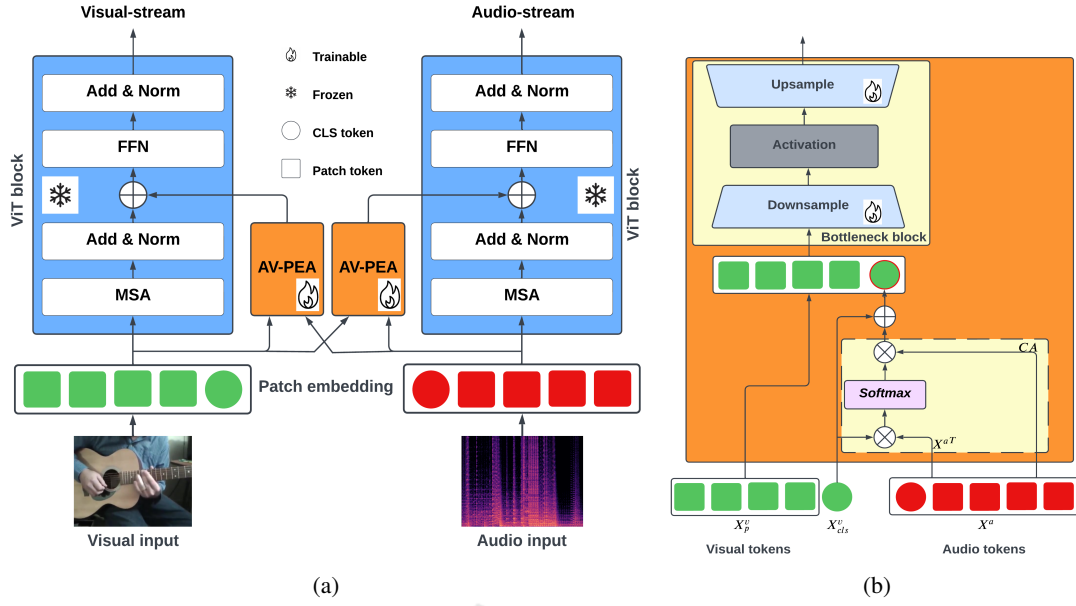


Figure 1: (a) Integration of the proposed AV-PEA into the ViT transformer. (b) Details of the proposed AV-PEA, highlighting the cross-attention (CA) module enclosed by a dashed rectangle.

Figure 1a: the *visual-stream* for processing visual input and the *audio-stream* for processing audio input. Within each block of both streams, we integrate our AV-PEA to efficiently adapt the ViT transformer to audio input (which is unseen during the training phase of ViT) while also enabling seamless information exchange between the audio and visual streams.

In the CA module, the *CLS* token of each stream serves as an intermediary to facilitate information exchange with the token sequence from the other stream. The *CLS* token is then back-projected to its respective stream, allowing it to interact with its own patch tokens once again in the bottleneck block. This enables the learned information from the other stream to be effectively conveyed to each patch token, thereby enriching the representation of individual patch tokens and ensuring comprehensive integration of multimodal representations.

3.3 Technical Integration of AV-PEA into the ViT Transformer

Within our proposed dual-stream ViT transformer (Figure 1a), consider the visual tokens $X^v \in \mathbb{R}^{(n+1) \times D}$, comprising both the patch tokens $X_p^v \in \mathbb{R}^{n \times D}$ and the *CLS* token $X_{cls}^v \in \mathbb{R}^{1 \times D}$ directed to the visual stream. Similarly, the audio tokens $X^a \in \mathbb{R}^{(n+1) \times D}$ consist of the patch tokens $X_p^a \in \mathbb{R}^{n \times D}$ and the *CLS* token $X_{cls}^a \in \mathbb{R}^{1 \times D}$ directed to the audio stream, where n and D represent the number of patch tokens and the embedding dimension, respectively.

Before we integrate our AV-PEA into the ViT block of each stream, let's first outline the standard operations of a ViT block ℓ within the visual stream v . The block ℓ begins by applying the multiheaded self-attention layer (MSA) as:

$$Y_\ell^v = X_\ell^v + \text{MSA}(X_\ell^v). \quad (1)$$

Subsequently, the intermediate representation Y_ℓ^v from MSA is passed through the feed-forward network (FFN) of the block ℓ , resulting in:

$$X_{\ell+1}^v = Y_\ell^v + \text{FFN}(Y_\ell^v). \quad (2)$$

These MSA and FFN operations are iteratively applied to the visual tokens X^v in each block of v . The same procedure is applied to the audio stream a , with the only difference being the interchange of the indices v and a .

The integration of AV-PEA into each block ℓ of the dual-stream ViT transformer proceeds as follows:

$$\begin{aligned} X_{\ell+1}^v &= Y_\ell^v + \text{FFN}(Y_\ell^v) \\ Y_\ell^v &= X_\ell^v + \text{MSA}(X_\ell^v) + B_\ell^v \end{aligned} \quad (3)$$

$$\begin{aligned} X_{\ell+1}^a &= Y_\ell^a + \text{FFN}(Y_\ell^a) \\ Y_\ell^a &= X_\ell^a + \text{MSA}(X_\ell^a) + B_\ell^a, \end{aligned} \quad (4)$$

where B_ℓ^v and B_ℓ^a denote the bottleneck blocks of AV-PEA on the v and a streams, respectively. Mathematically, the expressions for the B_ℓ^v and B_ℓ^a bottleneck blocks are:

$$B_\ell^v = h_v \cdot f^v(\text{CA}_v \parallel X_p^v) \quad (5)$$

$$B_{\ell}^a = h_a \cdot f^a(CA_a \parallel X_p^a), \quad (6)$$

where f is the projection function of the bottleneck block, \parallel denotes concatenation, and h is a trainable scalar parameter that acts as a learnable gate to regulate the flow of information through the model. The CA_v and CA_a denote the cross-attention process within the AV-PEA of the v and a streams, respectively, and can be mathematically expressed as:

$$CA_v(X_{cls}^v, X^a) = g_v \cdot \Theta_v X^a, \text{ where} \quad (7)$$

$$\Theta_v = \text{Softmax}(X_{cls}^v X^{aT})$$

$$CA_a(X_{cls}^a, X^v) = g_a \cdot \Theta_a X^v, \text{ where} \quad (8)$$

$$\Theta_a = \text{Softmax}(X_{cls}^a X^{vT}),$$

where g is a trainable scalar parameter utilized to control the flow of information between the two streams. Equations 7 and 8 reveal that only the CLS token is used as the query, ensuring that the generation of the attention maps Θ maintain linear computation and memory complexity. In addition to the CA process, the bottleneck block in AV-PEA involves projecting the original D -dimensional tokens into a lower-dimensional space with dimensionality d . Subsequently, a non-linear activation function $ReLU$ is applied before projecting the tokens back into their original D -dimensional space. This dimensionality reduction, achieved by setting $d \ll D$, substantially decreases the number of additional parameters.

4 EXPERIMENTS

4.1 Experimental Settings

For the AVEL and AVQA experiments: we employed the conventional ViT (Dosovitskiy et al., 2021) model, which underwent supervised pre-training on annotated data sourced from ImageNet-21K (Deng et al., 2009) as our base pre-trained model. The ViT-B/16 and ViT-L/16 variants, optimized for processing patches of size 16×16 , were used in most of our experiments.

In the context of the AVR and AVC experiments, we integrated our AV-PEA into the VALOR pre-trained model (Chen et al., 2023). While this model shares foundational principles with the ViT transformer, it has undergone supervised pre-training on the VALOR-1M audio-visual-language dataset (Chen et al., 2023).

To conduct a comprehensive comparison with the SOTA models, we just replaced the visual and audio encoders of the SOTA models with the frozen ViT

(except explicitly stated otherwise) transformer augmented by our AV-PEA, as explained in Section 3. We followed the evaluation procedures of the SOTA approaches, including the extraction of audio and visual features, to ensure methodological alignment. Throughout the training process, the parameters of the pre-trained transformer remained frozen, while the parameters of the AV-PEA were randomly initialized to meet the specific requirements of the audio-visual downstream task. Across all our experiments, we maintained a consistent learning rate of $3 \cdot 10^{-4}$, set $D = 8 \cdot d$, and initialized g , h_a , and h_v from zero.

4.2 Downstream Tasks and Results

AVEL: The audio-visual event (AVE) dataset (Tian et al., 2018) was used to assess the performance of our AV-PEA within the audio-visual event localization task. To this end, AV-PEA was incorporated into the cross-modal background suppression (CMBS) model (Xia and Zhao, 2022) with replacing its pre-trained visual and audio encoders by the frozen ViT transformer. Following the procedure outlined in the CMBS work, the event category label for each second within the videos was predicted, and the model's performance was evaluated using the overall accuracy metric for predicting event categories.

The comparison results with SOTA models on the AVE dataset were presented in Table 1. Our primary emphasis was on the CMBS model, well-known for its attainment of SOTA results on the AVE benchmark dataset. Furthermore, we conducted comparative analyses with the published outcomes derived from the multimodal bottleneck transformer (MBT) (Nagrani et al., 2021), the recent LAVISH adapter, and the dual perspective network (DPNet) (Rao et al., 2022) on the AVE dataset. Importantly, the LAVISH adapter employed the same pre-trained ViT models as those integrated with our AV-PEA.

Among the models of Table 1 that employ AudioSet pre-training and demand modality-specific dual encoders (visual and audio), the MBT model demonstrated the lowest accuracy (77.80%), lagging behind both DPNet and CMBS (79.68% and 79.70%, respectively). This is a significant observation, especially considering that the MBT model underwent full parameter tuning. Without the need for extensive audio pre-training on AudioSet, the LAVISH and our AV-PEA approaches, based on ViT-B and utilizing a shared pre-trained encoder for both visual and audio inputs, achieved comparable results ranging from 75.30% to 75.65%. However, our AV-PEA achieved this while utilizing fewer adapter parameters than LAVISH (3.7M vs. 3.9M), and amount-

Table 1: Audio-Visual Event Localization (AVEL): comparison with SOTA on the AVE dataset. Within this context, "PD" stands for pre-trained dataset, "N/A" abbreviates not available, * indicates the absence of official code, ✗ denotes a non-relevance criterion, * signifies frozen, 🔥 means full fine-tuning, and "Acc" abbreviates accuracy.

Method	Visual Encoder	Audio Encoder	Visual PD	Audio PD	Parameters (M) ↓			Acc% ↑
					Adapter	Total		
DPNet* (Rao et al., 2022)	VGG-19	VGGish	ImageNet	AudioSet	✗	N/A	N/A	79.68
CMBS (Xia and Zhao, 2022)	ResNet-152 *	VGGish *	ImageNet	AudioSet	✗	14.4	202.3	79.70
MBT (Nagrani et al., 2021)	ViT-B/16 🔥	AST 🔥	ImageNet	AudioSet	✗	172	✗	77.80
LAVISH (Lin et al., 2023)	ViT-B/16 * (shared)		ImageNet	✗	3.9	4.7	102.5	75.30
LAVISH (Lin et al., 2023)	ViT-L/16 * (shared)		ImageNet	✗	13.4	14.5	325.6	78.10
CMBS+AV-PEA (Ours)	ViT-B/16 * (shared)		ImageNet	✗	3.7	17.8	102.5	75.65
CMBS+AV-PEA (Ours)	ViT-L/16 * (shared)		ImageNet	✗	12.9	27.2	325.6	79.90

Table 2: Audio-Visual Question Answering (AVQA) using the Music-AVQA dataset. We reported accuracy spans three question categories: audio, visual, and audio-visual. "Avg" denotes the average accuracy.

Method	Visual Encoder	Audio Encoder	Visual PD	Audio PD	Parameters (M) ↓			Question% ↑			
					Adapter	Total		Audio	Visual	Audio-visual	Avg
AVSD* (Schwartz et al., 2019)	VGG-19	VGGish	ImageNet	AudioSet	✗	N/A	N/A	68.52	70.83	65.49	68.28
Pano-AVQA* (Yun et al., 2021)	Faster RCNN	VGGish	ImageNet	AudioSet	✗	N/A	N/A	70.73	72.56	66.64	69.98
AVQA (Li et al., 2022)	ResNet-18 *	VGGish *	ImageNet	AudioSet	✗	10.6	94.4	74.06	74.00	69.54	72.53
AVQA (Li et al., 2022)	Swin-V2-L 🔥	VGGish *	ImageNet	AudioSet	✗	240	312.1	73.16	73.80	73.16	73.37
AVQA+LAVISH	ViT-B/16 * (shared)		ImageNet	✗	4.4	13.1	102.5	73.14	68.73	64.93	68.93
AVQA+LAVISH	ViT-L/16 * (shared)		ImageNet	✗	14.8	23.8	325.6	75.05	79.44	70.34	74.94
AVQA+AV-PEA (Ours)	ViT-B/16 * (shared)		ImageNet	✗	3.7	12.4	102.5	76.16	78.82	69.72	74.90
AVQA+AV-PEA (Ours)	ViT-L/16 * (shared)		ImageNet	✗	12.9	21.9	325.6	74.49	80.06	71.26	75.27

ing to just 3.1% of the total parameters (3.7M vs. (17.8+102.5)M). Significantly, our AV-PEA with ViT-L outperformed all other methods, attaining an accuracy of 79.90%, even surpassing the analogous LAVISH adapter with ViT-L (78.10%). Worth noting is that the performance of LAVISH degraded on larger models like ViT-L due to its substantial reliance on latent tokens. On the contrary, our AV-PEA model demonstrated continuous improvement, all while utilizing fewer adapter parameters than LAVISH (12.9M vs. 13.4M), accounting for only 3.7% of the total parameters (12.9M vs. (27.2+325.6)M), all the while capitalizing on its seamless plug-and-play functionality.

AVQA. In Table 2, we further evaluated the effectiveness of our AV-PEA in the context of audio-visual question answering task, utilizing the MUSIC-AVQA (Li et al., 2022) dataset. In these experiments, we implemented a more robust AVQA (Li et al., 2022) baseline using the frozen ViT augmented with our AV-PEA. The MUSIC-AVQA dataset comprises 9,288 videos and 45,867 question-answer pairs. It includes 33 question templates encompassing 9 question types, which span across audio, visual, and audio-visual domains. Each of these question templates is associated with a specific answer, resulting in a pool of 42 potential answers.

Table 2 showed that the best performance among the methods utilizing AudioSet pre-training is achieved by AVQA (Li et al., 2022) with the Swin-V2-L visual encoder. This configuration of AVQA achieved a marginal accuracy improvement of 0.84% compared to the baseline AVQA (Li et al., 2022) employing a ResNet-1 visual encoder. However, achiev-

ing this modest improvement demanded the integration of an extra 229.4M trainable parameters. These experiments also highlight the limitations of the LAVISH adapter with larger datasets such as the MUSIC-AVQA dataset. Remarkably, LAVISH with ViT-B/16 presented inferior performance compared to its own baseline AVQA model (68.93% vs. 73.37%). This is despite the introduction of additional latent tokens, as evidenced by the contrast in the number of adapter parameters of the AVEL (Table 1) and AVQA (Table 2) tasks (3.9M vs. 4.4M).

On the contrary, our AV-PEA with ViT-B/16 not only outperformed audio-visual scene-aware dialog (AVSD) (Schwartz et al., 2019) and Pano-AVQA (Yun et al., 2021), but also surpassed various AVQA baseline variants, including LAVISH with ViT-B/16. Additionally, it obtained comparable results to LAVISH with ViT-L/16 (74.90% vs. 74.94%), while utilizing only 25% of trainable parameters used by LAVISH with ViT-L/16. Finally, we noted a consistent improvement in accuracy through our AV-PEA with ViT-L/16, achieving an accuracy of 75.27%, and amounting to just 3.7% of the total parameters (12.9M vs. (21.9+325.6)M).

It is noteworthy that our AV-PEA adapter maintains parameter consistency across diverse tasks, coupled with its user-friendly design that enables effortless integration into new tasks, eliminating the need for parameter adjustments.

AVR and AVC. For audio-visual retrieval and captioning tasks, AV-PEA was integrated into the frozen VALOR model, using its visual encoder for both visual and audio inputs. To ensure a fair comparison with LAVISH, we also integrated the LAVISH

Table 3: Comparison of performance results on the VALOR-32K dataset, covering Text-to-Audio-Visual Retrieval (AVR) and Audio-Visual Captioning (AVC), along with results on the MUSIC-AVQA dataset, which focuses on the Audio-Visual Question Answering (AVQA) benchmark.

Method	AVR \uparrow				AVC \uparrow				AVQA \uparrow
	R@1	R@5	R@10	Avg	BLEU4	METEOR	ROUGE-L	Avg	Acc%
VALOR	67.90	89.70	94.40	84.00	9.60	15.40	31.80	18.93	78.90
VALOR+LAVISH	64.70	86.70	92.00	81.10	11.14	19.53	36.66	22.44	77.93
VALOR+AV-PEA (Ours)	64.10	86.60	92.40	81.00	11.37	19.09	37.06	22.51	78.63

adapter into the frozen VALOR model. Both adapters underwent evaluation on the VALOR-32K dataset (Chen et al., 2023). Just like the VALOR evaluation protocol, the recall at rank K ($R@K$, $K = 1, 5, 10$) were used as metrics for the AVR task, whereas BLEU4, METEOR, and ROUGE-L were used as metrics for the AVC task. On top of these, our evaluation extended to re-evaluating the performance of both the AV-PEA and LAVISH approach, now integrated into the VALOR model, using the MUSIC-AVQA dataset. This evaluation was conducted in line with the VALOR framework. Worth noting is that while the AVQA framework in Table 2 primarily pertains to a classification problem where answers are retrieved from a pool of 42 potential answers, the VALOR framework formulates the AVQA task as a generative problem, aiming to directly generate the answer based on the input question.

The results presented in Table 3 reveal several findings. Firstly, our AV-PEA presented superior average performance in comparison to the baseline VALOR model for the AVC task (22.51 vs. 18.93), despite not using a pre-trained audio encoder or undergoing extensive AudioSet pre-training like the VALOR model.

Secondly, our AV-PEA performed comparably to the VALOR model for the AVQA task (78.63% and 78.90%). Thirdly, our AV-PEA showcased a slight performance improvement over LAVISH for both the AVC (22.51 vs. 22.41) and AVQA (78.63% vs. 77.93%) tasks, while maintained parity on the AVR task (81.00% and 81.10%).

Finally, it is truly impressive to witness the remarkable efficacy of adapter modules, including our AV-PEA and LAVISH, when seamlessly incorporated into pre-trained models. Even with a relatively modest count of additional trainable parameters and without the need for extensive AudioSet pre-training, these adapter modules manage to attain comparable or even superior performance across a range of downstream tasks.

4.3 Ablation Studies

To validate the efficiency of our AV-PEA, we explored different design scenarios, integrating it into both visual and audio streams (Figure 1a) or omitting it from

Table 4: Effectiveness of AV-PEA on audio-visual learning.

Method	Audio stream	Visual stream	Acc% \uparrow
CMBS	\times	\times	72.01
CMBS	AV-PEA	\times	72.71
CMBS	\times	AV-PEA	74.68
CMBS	AV-PEA	AV-PEA	75.65

either, using the ViT-B/16 pre-trained model on the AVE dataset (Tian et al., 2018). We replaced the visual and audio encoders of the CMBS (Xia and Zhao, 2022) model with the frozen ViT-B/16 transformer following the methodology in Section 3.3.

As shown in Table 4, AV-PEA significantly improved audio input handling, reflected in the results when integrated into the audio stream (72.71% vs. 72.01%). It’s worth noting that the frozen ViT pre-trained model did not undergo AudioSet pre-training. A substantial enhancement in the visual stream (74.68% vs. 72.01%) was also observed, primarily attributed to the CA module (Figure 1b), which effectively facilitates information exchange between the audio and visual modalities, robustly establishing audio-visual cues in both streams. Integrating AV-PEA into both streams surpasses the highest single adapter result achieved by augmenting only the visual stream with AV-PEA (75.65% vs. 74.68%).

5 CONCLUSIONS

In this paper, we introduced AV-PEA, a novel audio-visual parameter-efficient adapter module that serves a dual purpose: (1) simplifying the integration of audio inputs into frozen vision transformers without the need for audio pre-training, and (2) enabling seamless information exchange between the audio and visual modalities, all achieved with a limited set of additional trainable parameters. Through a lightweight bottleneck block on top of a simple cross-attention module that employs only the CLS token from both modalities as an intermediary for cross-modal information exchange. AV-PEA demonstrated efficacy across various audio-visual tasks, including audio-visual event localization (AVEL), audio-visual question answering (AVQA), audio-visual retrieval (AVR), and audio-visual captioning (AVC).

Although the presented results are preliminary, the experiments strongly indicate a promising direction.

AV-PEA’s flexibility allows its adoption on any visual transformer supported by the *CLS* token.

ACKNOWLEDGEMENTS

This work is supported by the Academy of Finland in project 345791. We acknowledge the LUMI super-computer, owned by the EuroHPC Joint Undertaking, hosted by CSC and the LUMI consortium.

REFERENCES

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *NeurIPS*, 33:1877–1901.
- Bugliarello, E., Cotterell, R., Okazaki, N., and Elliott, D. (2021). Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.
- Chen, C.-F. R., Fan, Q., and Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, pages 357–366.
- Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., and Liu, J. (2023). Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houtsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
- Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780.
- Gong, Y., Chung, Y.-A., and Glass, J. (2021a). Ast: Audio spectrogram transformer. In *Interspeech*, pages 571–575.
- Gong, Y., Chung, Y.-A., and Glass, J. (2021b). Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306.
- Guzhov, A., Raue, F., Hees, J., and Dengel, A. (2022). Audioclip: Extending clip to image, text and audio. In *ICASSP*, pages 976–980.
- Houlsby, N., Giurghi, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799.
- Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.-R., and Hu, D. (2022). Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19108–19118.
- Lin, Y.-B., Lei, J., Bansal, M., and Bertasius, G. (2022). Eclipse: Efficient long-range video retrieval using sight and sound. In *ECCV*, pages 413–430.
- Lin, Y.-B., Sung, Y.-L., Lei, J., Bansal, M., and Bertasius, G. (2023). Vision transformers are parameter-efficient audio-visual learners. In *CVPR*, pages 2299–2309.
- Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). Attention bottlenecks for multi-modal fusion. *NeurIPS*, 34:14200–14213.
- Pan, J., Lin, Z., Zhu, X., Shao, J., and Li, H. (2022). St-adapter: Parameter-efficient image-to-video transfer learning. In *NeurIPS*, pages 26462–26477.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.
- Rao, V., Khalil, M. I., Li, H., Dai, P., and Lu, J. (2022). Dual perspective network for audio-visual event localization. In *ECCV*, pages 689–704.
- Schwartz, I., Schwing, A. G., and Hazan, T. (2019). A simple baseline for audio-visual scene-aware dialog. In *CVPR*, pages 12548–12558.
- Sung, Y.-L., Cho, J., and Bansal, M. (2022). VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, pages 5227–5237.
- Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. (2018). Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263.
- Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. (2023). Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *CVPR*.
- Xia, Y. and Zhao, Z. (2022). Cross-modal background suppression for audio-visual event localization. In *CVPR*, pages 19989–19998.
- Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., and Li, M. (2023). AIM: Adapting image models for efficient video action recognition. In *ICLR*.
- Yun, H., Yu, Y., Yang, W., Lee, K., and Kim, G. (2021). Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*, pages 2031–2041.