Mixture of Experts for Audio-Visual Learning

Ying Cheng[1,2,3] **Yang Li**[1,2] **Junjie He**[1,2] **Rui Feng**[1,2,3*]
[1]School of Computer Science, Fudan University
[2]Shanghai Key Laboratory of Intelligent Information Processing
[3]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
{chengy18, fengrui}@fudan.edu.cn,
{liy23, hejj23}@m.fudan.edu.cn

Abstract

With the rapid development of multimedia technology, audio-visual learning has emerged as a promising research topic within the field of multimodal analysis. In this paper, we explore parameter-efficient transfer learning for audio-visual learning and propose the Audio-Visual Mixture of Experts (*AVMoE*) to inject adapters into pre-trained models flexibly. Specifically, we introduce unimodal and cross-modal adapters as multiple experts to specialize in intra-modal and inter-modal information, respectively, and employ a lightweight router to dynamically allocate the weights of each expert according to the specific demands of each task. Extensive experiments demonstrate that our proposed approach *AVMoE* achieves superior performance across multiple audio-visual tasks, including AVE, AVVP, AVS, and AVQA. Furthermore, visual-only experimental results also indicate that our approach can tackle challenging scenes where modality information is missing. The source code is available at https://github.com/yingchengy/AVMOE.

1 Introduction

Audio-visual learning is a challenging problem that analyzes auditory and visual information simultaneously, which aims at naturally integrating information from multiple modalities to perceive and understand the real-world comprehensively. By leveraging both auditory and visual cues, the models can learn multimodal correlations and solve various tasks such as Audio-Visual Event localization (AVE) [54], Audio-Visual Segmentation (AVS) [65], and Audio-Visual Question Answering (AVQA) [61, 27, 64].

In recent years, audio-visual learning has received steadily increasing attention in the field of multimodal analysis. Some previous works [3, 9, 22, 43, 1] utilize the correspondence between auditory and visual streams to learn multimodal representations in a self-supervised manner. By reducing the reliance on labeled data, self-supervised approaches can scale to larger datasets [14] and improve generalization performance. However, due to the rapid growth of the model size, fine-tuning full parameters of large pre-trained models entails significant computational costs for each downstream task. Recently, some researchers [30, 11] have attempted to inject adapters into frozen audio-visual pre-trained models, so as to reduce trainable parameters when applying to downstream tasks. Although these approaches have achieved satisfactory performance, they only focus on cross-modal attention in adapters to introduce information from other modalities, lacking the prominence of crucial information within the current modality. Besides, not every visual frame and audio segment can promote each other, and only introducing cross-modal adapters may bring irrelevant information during the learning process. As shown in Figure 1, the sound of basketball bounce is completely drowned out by the commentator's voice and the cheers from the audience. In these cases such

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

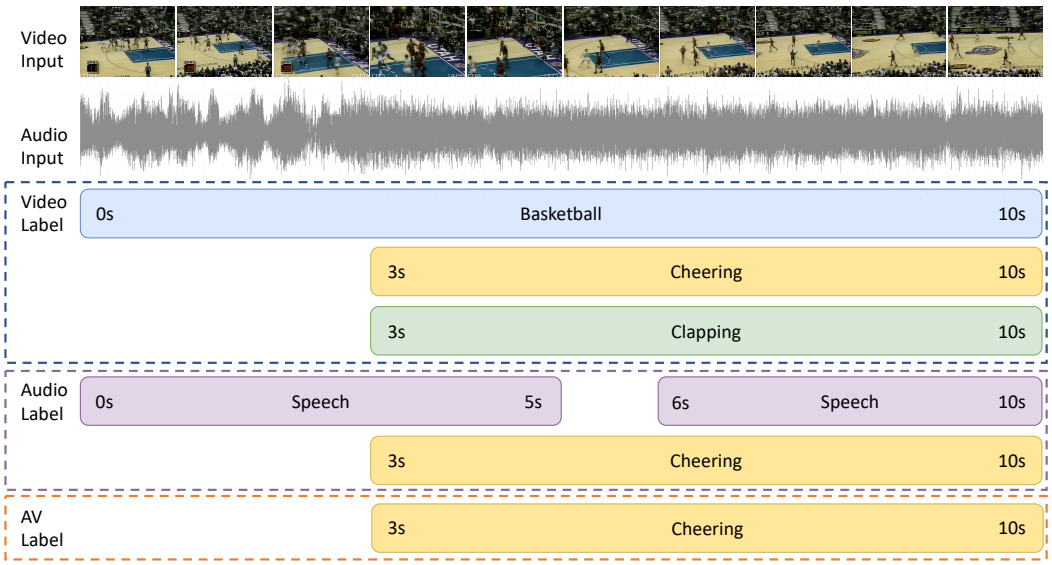

Figure 1: The case that video and audio labels are different. The video label includes *basketball*, *cheering*, and *clapping*, while the audio label includes *speech* and *cheering*.

as off-screen sounds, injecting cross-modal adapters only brings disturbing information, and it is necessary to inject adapters according to the scenario flexibly.

The Mixture of Experts (MoE) is a structural strategy for machine learning models, designed to enhance overall prediction performance by combining the decisions of multiple expert models. Hence, it is appealing to integrate the MoE module into neural networks. Within the MoE strategy, each expert model focuses on different subsets or aspects of the input data, providing predictions for a given input. Additionally, there is a gating mechanism (usually a trainable layer) responsible for determining which expert's predictions should be given more weight based on the input data. This structure allows the model to dynamically adjust its reliance on different experts, thereby optimizing its ability to process various types of data.

In this work, we propose an approach of **M**ixture **o**f **E**xperts for **A**udio **V**isual Learning (**AVMoE**), designed as an efficient and flexible approach to adaptively combine adapter capabilities according to different scenarios. Firstly, we introduce two types of adapters as experts into frozen pre-trained models, i.e., unimodal adapter and cross-modal adapter, to focus on intra-modal and inter-modal information, respectively. Secondly, we introduce a modality-agnostic router to evaluate how relevant each expert is to the characteristics of inputs and assign the weights to each expert. By employing the *AVMoE* scheme, the model can adapt its strategy according to the specific demands of each task, whether it be heavily reliant on visual data, audio data, or requiring a balanced integration of both. We conduct extensive experiments on three audio-visual tasks, i.e., AVE, AVS, and AVQA. Experimental results demonstrate that our approach not only improves the model's performance across a wide range of audio-visual tasks but also makes it robust to variations in the input data. For instance, in scenarios where audio information is obscured or noisy, the model can shift its reliance toward visual cues and vice versa.

The contributions of this paper can be summarized as follows:

- We propose a generic, effective, and flexible approach, *AVMoE*, for applying pre-trained models to audio-visual learning, which can dynamically adjust its strategy according to the specific demands of each task.
- We introduce two types of adapters to focus on both within-modality details and the interactions between modalities. By integrating adapters and MoE, the model can be robust to any variations in the input data, e.g., missing modality or multimodal content mismatch.
- Extensive experiments demonstrate that our method consistently outperforms in diverse audio-visual tasks, showcasing its ability to dynamically prioritize visual or audio cues, enhancing its reliability and effectiveness in real-world applications.

2 Related Work

2.1 Audio-Visual Learning

Audio-visual learning aims to integrate auditory and visual information to enhance the understanding and perception of multimedia scenarios. Some early researchers [3, 22, 43] propose an innovative self-supervised approach that utilizes the correspondence between visual and audio components in video as a supervisory signal to pre-train a multimodal model, and the model can be applied to multiple downstream tasks such as speech enhancement [8, 17, 18], sound localization [60, 32], and action recognition [23, 49]. However, these audio-visual pre-trained models are limited by the ability of temporal localization and complex reasoning.

Recently, some researchers have made efforts for complicated audio-visual downstream tasks. Audio-Visual Event Localization (AVE [54]) aims at classifying and localizing events in the given video, and most prior works [57, 59, 62, 37] introduce interactions to utilize multimodal cues from modality-specific pre-trained models. Audio-Visual Video Parsing (AVVP [53]) expands the task of localizing one event to multiple events and removes the restriction that audio and visual events are definitely consistent. Previous works [63, 56] usually propose multi-scale hybrid networks and aggregate features in a weakly-supervised manner. Audio-Visual Segmentation (AVS [65]) aims to localize and segment the masks of sounding objects at the pixel level. To tackle this task, most prior methods learn correspondences between audio and visual patches and perform mask decoding by variational auto-encoder [39], contrastive conditional latent diffusion [38], vision transformer [31], bidirectional generation [15]. Audio-Visual Question Answering (AVQA [61]) requires the model to correctly answer the questions by understanding both audio and visual modalities comprehensively. Previous works tackle this task by spatiotemporal grounding model [26], missing modality recalling [44], and multi-scale feature fusion [7].

Although these researches have achieved satisfactory performance, designing models for distinct tasks separately is often expensive and time-consuming. In order to improve efficiency and generalization, some recent works have explored parameter-efficient transfer learning for audio-visual downstream tasks. LAVish [30] proposes to introduce cross-modal adapters into frozen pre-trained transformers for audio-visual data. DG-SCT [11] proposes dual-guided spatial-channel-temporal attention mechanisms as cross-modal prompts injected into pre-trained models. Unlike these prior methods focus on cross-modal adapters, we introduce uni-modal adapters and the MoE scheme to combine multimodal information dynamically.

2.2 Mixture of Experts

Mixture of Experts (MoE) [19] is a framework that has been extensively studied and applied in various domains of machine learning. Some works explore the role of MoE in the research fields of Natural Language Processing (NLP) [51, 24, 13, 10, 68] and Computer Vision (CV) [47, 35]. These works typically regard the Feed-Forward Network (FFN) layers as experts and choose the most relevant experts for the given task. Some efforts have been made to improve the routing strategy by differentiable selection [16], linear assignment [25], simple hashing algorithm [48], dense-to-sparse strategy [42], and expert choice [67].

Recently, some works have extended the original MoE scheme to enhance its applicability in various advanced learning paradigms. In the field of multi-task learning, Ma et al. [36] and Chen et al. [6] have successfully integrated MoE schemes to deal with the complexities of learning multiple tasks simultaneously. These approaches leverage expert-specific knowledge to effectively address the unique requirements of each task, thus improving the performance and efficiency of models.

Similarly, the MoE scheme has also been adapted for multimodal learning. Mustafa et al. [41] explore the usage of MoE in combining visual and textual information, demonstrating significant improvements in tasks that require understanding from both image and text inputs. Shen et al. [52] further employ MoE to scale both the text-based and vision-based feed-forward networks, showing the robustness and flexibility in handling diverse data sources. Cao et al. [4] and Akbari et al. [2] also contribute to this field by developing MoE-based models that alternate between different modalities, effectively capturing the complex interactions between them. However, these approaches only investigate the effectiveness of MoE in vision-language models, to the best of our knowledge, the MoE scheme has not been explored for audio-visual learning.

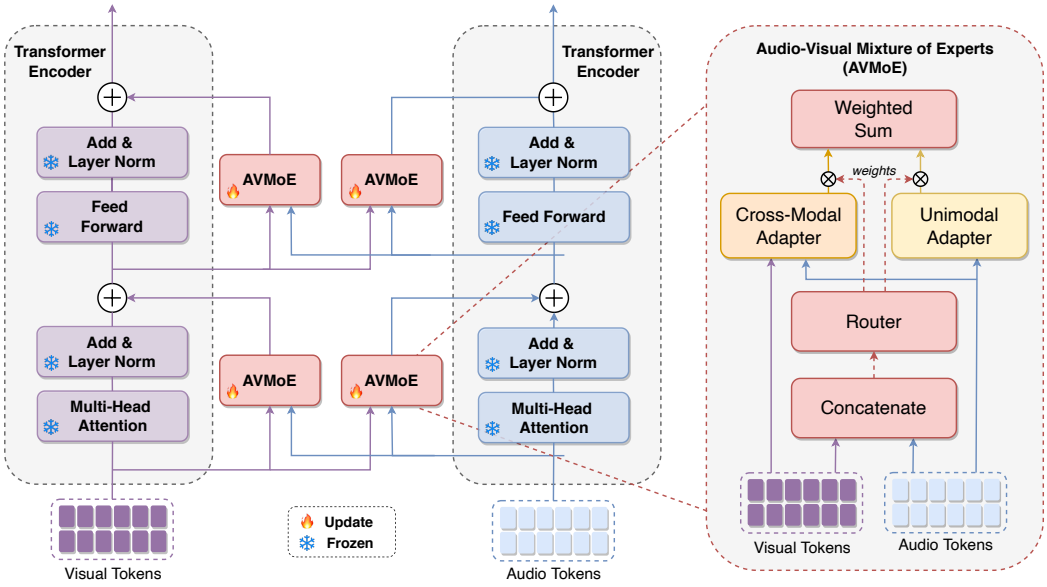

Figure 2: Method Overview. We propose to inject trainable adapters with the MoE scheme into the frozen pre-trained backbones for dynamically adjusting the strategy. The AVMoE module mainly consists of a router layer and two types of adapters, the router ingests concatenated multimodal tokens and allocates the weights for adapters, and the output of AVMoE module is the weighted sum of all the adapters' predictions.

3 Method

In this section, we present our approach *AVMoE* as shown in Figure 2, which adaptively integrates multiple types of adapters into the frozen pre-trained model using the audio-visual mixture of experts (MoE) framework. The structure of our framework consists of the frozen pre-trained model (Sec. 3.1) and trainable MoE modules (Sec. 3.2), where the MoE module contains two types of adapters and a router. In the following, we will describe each component in more detail.

3.1 Frozen Pre-trained Model

In this work, we aim to adapt pre-trained models to audio-visual tasks with minimal changes to the original parameters. Hence, we add a few trainable parameters while keeping the pre-trained model frozen. Following previous work [11], we use the standard Swin-Transformer (Swin-T) [33] and audio transformer (HTS-AT) [5] as our pre-trained backbones. Each layer of these models consists of Multi-Head Attention (MHA), Feed Forward Network (FFN), and Layer Normalization with Residual Connections (Add & Layer Norm).

3.2 Audio-Visual Mixture of Experts

In this subsection, we describe how our proposed *AVMoE* can flexibly adjust frozen pre-trained models to various audio-visual downstream tasks. As illustrated in the right of Figure 2, the AVMoE module contains three primary components: two types of adapters as experts, each specializing in different aspects of the multi-modal data, and a router for ingesting concatenated tokens and allocating the corresponding weights of experts.

3.2.1 Router

The router layer is responsible for activating the appropriate experts based on the input features. To be specific, after obtaining visual embedding v_t and audio embedding a_t, we concatenate them as the input i_t of the AVMoE router. Following previous works [13], we employ Multi-Layer Perceptrons (MLPs) as the architecture of the router R, which produces weights by using the softmax function. The weights for Cross-Modal Adapter (CMA) and Unimodal Adapter (UA) are computed as follows:

$$w_{\text{CMA}} = \frac{\exp\left(r_{\text{CMA}}(i_t)\right)}{\exp\left(r_{\text{CMA}}(i_t)\right) + \exp\left(r_{\text{UA}}(i_t)\right)}, w_{\text{UA}} = \frac{\exp\left(r_{\text{UA}}(i_t)\right)}{\exp\left(r_{\text{CMA}}(i_t)\right) + \exp\left(r_{\text{UA}}(i_t)\right)} \tag{1}$$

where r_{CMA} and r_{UA} are the outputs of the router for CMA and UA. The produced weights determine the importance of each adapter's prediction to the final output.

To promote the balanced usage of experts, inspired by previous works [13, 20], we introduce Gaussian noise to the router's output at the training stage, thereby promoting a more balanced use of experts. This method prevents the model from over-relying on certain experts, forcing the model to explore a wider range of experts:

$$g' = g + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

where ϵ is Gaussian noise with a standard deviation σ.

3.2.2 Audio-Visual Adapters

To enhance the efficiency and performance of audio-visual models, we follow previous works [30, 11], adding a few trainable parameters (i.e., adapters) into each layer of frozen pre-trained backbones. However, different from them, we design two types of adapters, i.e., CMA and UA. As shown in Figure3, the main differences between them are multimodal feature fusion and self-attention. These adapters are injected into the frozen pre-trained model via the MoE strategy.

Cross-Modal Adapter. Firstly, to introduce interactions and integrate information from multiple modalities, we follow *LAVisH* adapter [30] as our cross-modal adapter. As shown in the left of Figure3, the cross-modal adapter consists of audio-visual latent token compression, audio-visual feature fusion, and bottleneck module.

We propose to utilize latent tokens, which are small, randomly initialized vectors that serve as placeholders for condensed information from the audio and visual modalities. For layer l, we have m latent tokens for audio, denoted as L_a^l, and m latent tokens for visual, denoted as L_v^l. These tokens are significantly fewer in number compared to the total number of tokens from the audio or visual inputs, which helps in reducing the computational complexity.

In this step, the adapter uses cross-modal attention to compress the information from all the audio or visual tokens into the corresponding latent tokens. This is achieved by applying the cross-modal attention operation between the input tokens and the latent tokens of the same modality. The compressed latent tokens capture the most relevant information from the original tokens, which is essential for efficient information transfer between modalities.

For audio and visual inputs, the compression is given by:

$$S_a^l = f_c(L_a^l, X_a^l, X_a^l), \quad S_v^l = f_c(L_v^l, X_v^l, X_v^l) \tag{3}$$

where f_c is the cross-attention function, X_a^l and X_v^l are the audio and visual tokens at layer l, and S represents the resulting latent summary tokens.

After compressing the modality-specific information into latent tokens, the adapter fuses the compressed information from one modality with the full set of tokens from the other modality. This is done through another round of cross-modal attention, which allows the model to integrate audio and visual cues effectively. For cross-modal fusion:

$$X_{av}^l = f_c(X_a^l, S_v^l, S_v^l), \quad X_{va}^l = f_c(X_v^l, S_a^l, S_a^l) \tag{4}$$

where X_{av}^l and X_{va}^l represent the newly computed audio and visual representations that incorporate information from both modalities.

The final step in the adapter involves a lightweight module that refines the fused audio-visual representations into more discriminative features suitable for downstream tasks. This module consists of a bottleneck structure with a down-projection layer (θ^{down}), a non-linear activation function (σ), and an up-projection layer (θ^{up}). For the audio-visual and visual-audio pathways:

$$Z_{av}^l = \theta^{up}(\sigma(\theta^{down}(X_{av}^l))), \quad Z_{va}^l = \theta^{up}(\sigma(\theta^{down}(X_{va}^l))) \tag{5}$$

where Z_{av}^l and Z_{va}^l are output features that are passed to subsequent layers or used for task-specific predictions.

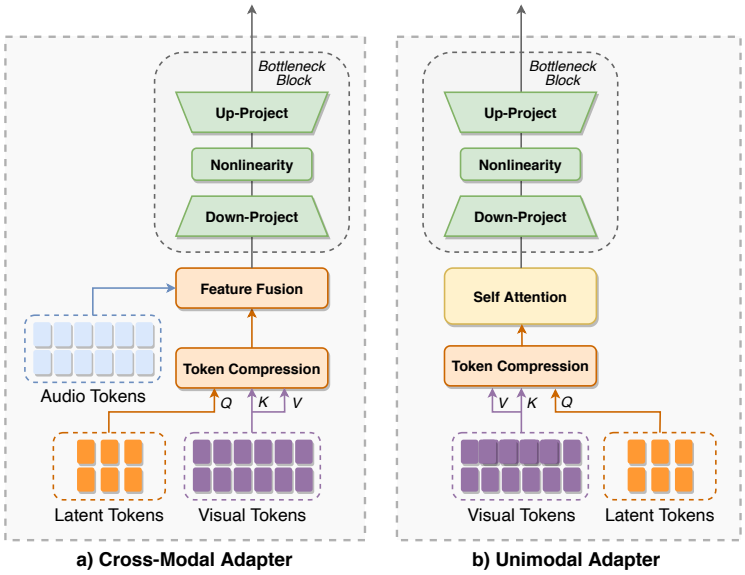

Figure 3: The structure of adapters. a) Cross-Modal Adapter (CMA), consisting of token compression, multimodal feature fusion and bottleneck block; b) Unimodal Adapter (UA), consisting of token compression, self-attention and bottleneck block.

Unimodal Adapter. While cross-modal adapters effectively introduce interactions between audio and visual tokens, they may not always be suitable for scenarios where auditory and visual content do not perfectly align, potentially introducing irrelevant information. Additionally, some multimodal data might primarily rely on one modality, or another modality may be completely missing, making inter-modal interactions less effective. Hence, it is also necessary to introduce intra-modal interactions and design unimodal adapters, which can address these issues by providing targeted improvements that compensate for the limitations of cross-modal adapters. As shown in the right of Figure3, after obtaining compressed tokens S_a^l and S_v^l, the self-attention layer ingests them and outputs newly computed representations.

$$X_a^l = f_s(X_a^l, S_v^l, S_v^l), \quad X_v^l = f_s(X_v^l, S_a^l, S_a^l) \tag{6}$$

where f_c is the cross-attention function, X_a^l and X_v^l represent the output audio and visual representations, respectively.

In this work, we propose to utilize MoE strategy to introduce intra- and inter-modal interactions synergistically. By integrating these two types of adapters, our model can dynamically adjust the attention paid to unimodal and cross-modal information according to the scenario.

4 Experimental Analysis

To demonstrate the capabilities of our proposed *AVMoE* in solving diverse audio-visual tasks, we conduct extensive experiments on four downstream tasks, i.e., Audio-Visual Event Localization (AVE), Audio-Visual Video Parsing (AVVP), Audio-Visual Segmentation (AVS), and Audio-Visual Question Answering (AVQA). More details about audio-visual tasks and datasets will be illustrated in the Appendix.

4.1 Audio-Visual Event Localization

For the AVE task, we adopt the overall segment-wise accuracy of predicted event categories as the evaluation metric, and the event label of each video segment is required to be predicted in a fully-supervised manner. As shown in Table 1, we compare our proposed *AVMoE* with previous methods on the test set of AVE dataset. We focus on *LAVisH* [30] and *DG-SCT* [11], as they also explore parameter-efficient fine-tuning methods like ours and have achieved impressive results on this benchmark.

Table 1: **Audio-Visual Event Localization.** Comparison with previous methods in a fully-supervised manner. The performance is evaluated on the test set of AVE dataset with classification accuracy. The visual and audio encoders are pre-trained on the ImageNet and Audioset, respectively. † means that no official code was provided to report some of the baseline-specific metrics.

Method	Visual Encoder	Audio Encoder	Trainable Params (%) ↓	Total Params (M) ↓	Acc ↑
AVT [29]	VGG-19	VGGish	6.8	231.5	76.8
MPN [62]	VGG-19	VGGish	N/A	N/A	77.6
PSP [66]	VGG-19	VGGish	0.8	217.4	77.8
DPNet† [46]	VGG-19	VGGish	N/A	N/A	79.7
AVEL [54]	ResNet-152	VGGish	2.7	136.0	74.0
AVSDN [28]	ResNet-152	VGGish	5.7	140.3	75.4
CMRAN [59]	ResNet-152	VGGish	10.7	148.2	78.3
MM-Pyramid [63]	ResNet-152	VGGish	25.0	176.3	77.8
CMBS [58]	ResNet-152	VGGish	6.6	216.7	79.7
LAVisH [30]	ViT-B-16 (shared)		4.4	107.2	75.3
LAVisH [30]	ViT-L-16 (shared)		4.3	340.1	78.1
AVMoE (Ours)	ViT-B-16 (shared)		31.8	150.4	76.4
AVMoE (Ours)	ViT-L-16 (shared)		32.6	483.1	79.2
LAVisH [30]	Swin-V2-B (shared)		4.4	114.2	78.8
LAVisH [30]	Swin-V2-L (shared)		2.7	238.8	81.1
AVMoE (Ours)	Swin-V2-B (shared)		41.1	206.6	79.4
AVMoE (Ours)	Swin-V2-L (shared)		39.4	374.4	81.5
LAVisH [30]	Swin-V2-L	HTS-AT	30.6	374.9	78.6
DG-SCT [11]	Swin-V2-L	HTS-AT	43.6	461.3	82.2
AVMoE (Ours)	Swin-V2-L	HTS-AT	34.9	404.0	**82.6**

For a fair comparison, we train our *AVMoE* with shared pre-trained visual backbones (ViT-B, ViT-L, Swin-V2-B, Swin-V2-L) and separate audio-visual backbones (HTS-AT and Swin-V2-L). Firstly, it can be seen that our method achieves consistently better performance than *LAVisH* across different shared backbones, which indicates that our method can be adapted to different backbones. We note that the structure of Swin-V2-L is better than ViT, since Swin-V2-L allows hierarchical feature representation and more efficient computation by constraining self-attention within local windows and progressively expanding them. Secondly, when utilizing distinct pre-trained backbones for each modality, our model outperforms *LAVisH* and *DG-SCT* by 4.0% and 0.4%, respectively. It should be noted that the number and proportion of trainable parameters is less than that of *DG-SCT* with complicated attention mechanisms, showcasing the efficiency and superiority of our model. Lastly, we also observe that the audio encoder pre-trained on Audioset can further improve the performance, which indicates that combining different modalities is beneficial for audio-visual learning.

4.2 Audio-Visual Video Parsing

Compared to the AVE task, the AVVP task is a more complex and challenging multimodal task involving analyzing and understanding audio and visual information within a video to identify and categorize events. This task aims to segment a video into temporal events and predict their modality-specific categories, which may include actions that are audible, visible, or both. A key challenge in AVVP is that audio and visual modalities may not be fully correlated. For instance, due to limitations in camera field-of-view or occlusions, some events may occur in only one modality. Therefore, effectively integrating audio and visual information while mitigating the interference of irrelevant modality information is crucial.

As shown in Table2, our *AVMoE* has significant advantages over *DG-SCT* and *MGN* on the AVVP task. For example, the audio-visual event parsing metric at segment-level exceeds DG-SCT by 2%, which indicates that the scheme of MoE is suitable for balancing weight assignment for multi-modal information.

Table 2: **Audio-Visual Video Parsing.** Comparison with previous methods on the test set of LLP dataset. We conduct comparative experiments on two kinds of annotation: segment-level and event-level. The **Type** computes the average audio, visual, and audio-visual event evaluation results, and the **Event** considers all audio and visual events for each sample to compute the F-score.

Method	Segment-level					Event-level				
	A	V	AV	**Type**	**Event**	A	V	AV	**Type**	**Event**
AVE[54]	49.9	37.3	37.0	41.4	43.6	43.6	32.4	32.6	36.2	37.4
AVSDN[28]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN[53]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MGN[40]	60.7	55.5	50.6	55.6	57.2	51.0	52.4	44.4	49.3	49.2
DG-SCT[11]	59.0	59.4	52.8	57.1	57.0	49.2	**56.1**	46.1	50.5	49.1
AVMoE(Ours)	**62.1**	**60.0**	**54.4**	**58.8**	**59.0**	**51.8**	55.7	**47.6**	**51.7**	**50.2**

Table 3: **Audio-Visual Segmentation.** Comparison with previous methods under the S4 and MS3 settings of AVSBench dataset. $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{F}}$ denote the mean $\mathcal{J}$ and $\mathcal{F}$ metric values over the whole dataset.

Method	Visual Encoder	Audio Encoder	Trainable Params (%) ↓	Total Params (M) ↓	Setting			
					S4		MS3	
					$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
AVS [65]	PVT-v2	VGGish	58.7	174.5	78.7	87.9	54.0	64.5
LAVisH [30]	Swin-V2-L (shared)		14.0	266.4	80.1	88.0	49.8	60.3
LAVisH [30]	Swin-V2-L	HTS-AT	47.6	389.7	78.0	87.0	49.1	59.9
DG-SCT [11]	Swin-V2-L	HTS-AT	61.5	594.8	80.9	89.2	53.5	64.2
AVMoE (ours)	Swin-V2-L	HTS-AT	54.4	501.2	**81.1**	**89.7**	**54.5**	**68.7**

4.3 Audio-Visual Segmentation

To explore whether the model can associate the visual regions with the sound components, we compare our approach with state-of-the-art models on the AVS task. The Jaccard index $\mathcal{J}$ [12] and F-score $\mathcal{F}$ are used as the evaluation metrics, where $\mathcal{J}$ and $\mathcal{F}$ denote the region similarity and contour accuracy, respectively. As illustrated in Table 3, our approach *AVMoE* significantly outperforms *DG-SCT* and *LAVisH* under both S4 and MS3 settings, which verifies that the scheme of MoE is also beneficial for audio-visual segmentation. Our model shows more significant improvements under the MS3 setting of multi-sound sources, indicating that the model with *AVMoE* strategy can deal with more complex scenarios.

Figure 4 shows the qualitative results of our *AVMoE* model and *DG-SCT* on the AVS task. As can be seen, under the S4 setting, our model locates and segments the sounding objects more accurately than *DG-SCT*. For the case of ambulance siren (column #1), *DG-SCT* mistakenly locates the ambulance and the car nearby at the same time, while our model can exclude the car that hasn't produced sound. For the case of dog barking (column #3), our *AVMoE* model contours the object shapes (e.g., tail) more perfectly. Under the MS3 setting, compared to *DG-SCT* which cannot locate all sound sources in some cases, we can almost locate and segment each sound source. It can be seen that even in the case of three sound sources (column #4), we can still segment each instrument well, which verifies the effectiveness of our *AVMoE* model.

4.4 Audio-Visual Question Answering

We also want to explore whether our model can generalize to complex audio-visual tasks. To this end, we conduct experiments on the AVQA task and compare our model with previous leading methods. The AVQA task contains three categories of questions (audio, visual, and audio-visual), which involve questions related to audio (AQ), visual (VQ), or both audio and visual information simultaneously (AVQ). The multiple forms of questions challenge the adaptability to various modalities and the spatio-temporal reasoning ability of each model.

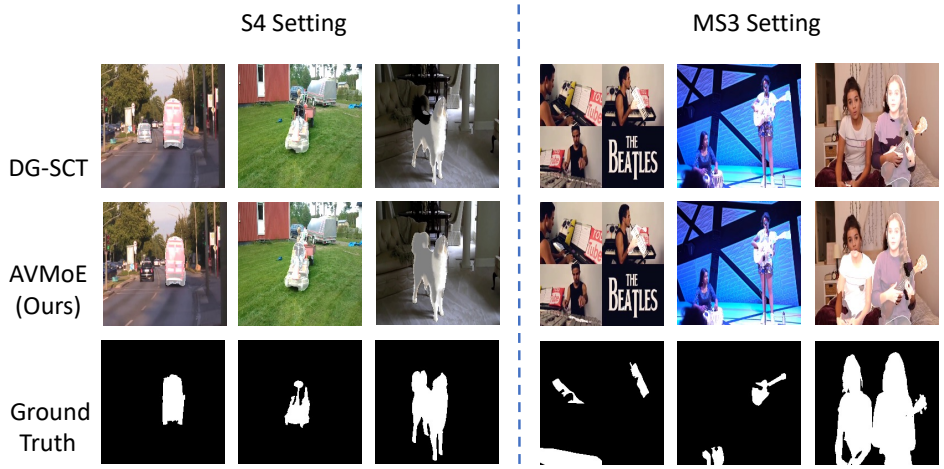

Figure 4: Qualitative examples of *AVMoE* and *DG-SCT* under the S4 setting and MS3 setting of the AVS task.

Table 4: **Audio-Visual Question Answering.** Comparison with previous methods on the test set of MUSIC-AVQA dataset. We report accuracy on three types of questions, i.e., Audio Question (AQ), Visual Question (VQ), and Audio-Visual Question (AVQ). LAVisH* denotes our implementation version of LAVisH.

Method	Visual Encoder	Audio Encoder	Trainable Params (%)↓	Total Params (M)↓	AQ	VQ	AVQ	Avg
AVSD [50]	VGG-19	VGG-like	N/A	N/A	68.5	70.8	65.5	67.4
Pano-AVQA [64]	Faster RCNN	VGG-like	N/A	N/A	70.7	72.6	66.6	68.9
ST-AVQA [27]	ResNet-18	VGG-like	11.2	94.4	74.1	74.0	69.5	71.5
LAVisH [30]	Swin-V2-L(shared)		8.4	249.8	75.7	80.4	70.4	74.0
LAVisH* [30]	Swin-V2-L	HTS-AT	28.7	367.4	75.4	79.6	70.1	73.6
DG-SCT [11]	Swin-V2-L	HTS-AT	50.0	520.2	77.4	81.9	70.7	74.8
AVMoE (ours)	Swin-V2-L	HTS-AT	42.7	456.6	**77.6**	**82.7**	**71.9**	**75.7**

Table 4 reports the experimental results on the Music-AVQA dataset[27] across three distinct scenarios. Notably, our method exhibits superior generalization abilities and consistently achieves State-Of-The-Art (SOTA) performance in all scenarios. Especially in the challenging scenarios of AVQ, our *AVMoE* outperforms *DG-SCT* by 1.2% (71.9% v.s. 70.7%)with fewer trainable and total parameters, which verifies the effectiveness and efficiency of our approach. The dynamic adaptability of MoE strategy not only establishes the robustness of audio-visual models but also illustrates the potential of leveraging the flexible strategy to enhance performance in complex audio-visual tasks. The comprehensive experimental results across four audio-visual downstream tasks illustrate that our approach achieves significant improvements while efficiently transferring from frozen backbones to various downstream tasks.

5 Ablation Studies

5.1 Number of Experts

As shown in Table 5, we first investigate the impact of the number of experts (i.e., cross-modal adapters and uni-modal adapters) on the tasks of AVE, AVS, and AVQA. It can be seen that increasing the number of experts consistently enhances the model's performance, thereby validating the effectiveness of the MoE strategy. Besides, we observe that the performance of our model with only one CMA and one UA still exceeds that of *LAVisH*. Especially under the MS3 setting of Audio-Visual Segmentation (AVS) task, it has improved by 7.2% compared with *LAVisH* (67.5% v.s. 60.3%). We conjecture that introducing more experts may further improve the abilities of the model. However, due to the limitation of GPU memory, we did not discuss the results in this work.

Table 5: **AVMoE design.** CMA and UA denote cross-modal adapter and uni-modal adapter, respectively.

Num. Experts		AVS (S4)		AVS(MS3)		AVQA (Acc.)				AVE (Acc.)
CMA	UA	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	AQ	VQ	AVQ	Avg	
1	1	80.2	88.6	51.8	67.5	75.7	81.4	70.9	74.7	81.3
1	2	80.6	89.1	53.1	68.1	76.3	82.0	71.4	75.1	81.8
2	1	80.8	89.2	53.6	68.4	76.2	82.1	71.6	75.1	82.3
2	2	**81.1**	**89.7**	**54.5**	**68.7**	**77.6**	**82.7**	**71.9**	**75.7**	**82.6**

Table 6: **Modality Ablation.** "A" and "V" denote audio and visual, respectively. All models are well-trained on audio-visual modalities and tested on different modalities.

Methods	Modalities	AVE(Acc.)	AVS($\mathcal{M}_{\mathcal{F}}$)	AVQA(Acc.)
DG-SCT	V	72.1	79.1	67.3
	A+V	82.2	89.2	74.8
AVMoE (Ours)	V	78.4	85.4	73.2
	A+V	82.6	89.7	75.7

5.2 Modality Ablation

In order to explore how our well-trained *AVMoE* model handles different modality inputs, we conduct experiments of our model and *DG-SCT* on visual-only data and audio-visual data. As shown in Table 6, we compare our *AVMoE* model with *DG-SCT* on AVE, AVS, and AVQA tasks. It can be seen that our approach achieves significant performance on the test scenarios of visual-only data and audio-visual data. We observe that, our *AVMoE* model, when compared to *DG-SCT*, there is no significant decrease in accuracy on visual-only data. These results indicate that our model with the MoE scheme can still achieve good performance even in the absence of audio information. Although *DG-SCT* designs a more complex architecture of dual-guided spatial-channel-temporal attention mechanism in adapters, only introducing cross-modal adapters may bring irrelevant information in these challenging cases.

Moreover, to explore whether the model can dynamically activate experts, we visualize the activation weight of each expert varies across parallel expert layers. The activation ratio of experts is determined by calculating the ratio of each expert's selection frequency in each MoE layer to the total number of tokens. The visualization and more details will be discussed in Appendix.

6 Conclusion

In this paper, we develop a flexible framework that introduces Mixture of Experts (MoE) for audio-visual learning. Specifically, we inject unimodal adapters and cross-modal adapters as experts into the frozen pre-trained model and our *AVMoE* can dynamically combine the abilities of adapters according to different scenarios through the MoE strategy. Experimental results illustrate that our *AVMoE* has achieved greatly improved performance on complex tasks (such as AVQA and the MS3 setting of AVS). The ablation studies and qualitative results demonstrate that our model benefits significantly from the *AVMoE* strategy. However, due to computational resource limitations and efficiency considerations, we only investigate a small number of experts in the MoE scheme. In subsequent work, we will explore how to efficiently introduce more experts into our model.

7 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62172101), in part by the Science and Technology Commission of Shanghai Municipality (No.23511100602), and supported by the Postdoctoral Fellowship Program of CPSF (No. GZC20230483).

References

[1] Triantafyllos Afouras et al. "Self-supervised learning of audio-visual objects from video". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020, pp. 208–224.

[2] Hassan Akbari et al. "Alternating gradient descent and mixture-of-experts for integrated multimodal perception". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 79142–79154.

[3] Relja Arandjelovic and Andrew Zisserman. "Look, listen and learn". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 609–617.

[4] Bing Cao et al. "Multi-modal gated mixture of local-to-global experts for dynamic image fusion". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 23555–23564.

[5] Ke Chen et al. "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 646–650.

[6] Tianlong Chen et al. "Adamv-moe: Adaptive multi-task vision mixture-of-experts". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 17346–17357.

[7] Zailong Chen et al. "Question-Aware Global-Local Video Understanding Network for Audio-Visual Question Answering". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[8] Ying Cheng et al. "Improving multimodal speech enhancement by incorporating self-supervised and curriculum learning". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 4285–4289.

[9] Ying Cheng et al. "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 3884–3892.

[10] Nan Du et al. "Glam: Efficient scaling of language models with mixture-of-experts". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5547–5569.

[11] Haoyi Duan et al. "Cross-modal Prompts: Adapting Large Pre-trained Models for Audio-Visual Downstream Tasks". In: *Advances in Neural Information Processing Systems* 36 (2024).

[12] Mark Everingham et al. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88 (2010), pp. 303–338.

[13] William Fedus, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39.

[14] Jort F Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[15] Dawei Hao et al. "Improving audio-visual segmentation with bidirectional generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 3. 2024, pp. 2067–2075.

[16] Hussein Hazimeh et al. "Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29335–29347.

[17] Zili Huang et al. "Investigating self-supervised learning for speech enhancement and separation". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6837–6841.

[18] Kuo-Hsuan Hung et al. "Boosting self-supervised embeddings for speech enhancement". In: *arXiv preprint arXiv:2204.03339* (2022).

[19] Robert A Jacobs et al. "Adaptive mixtures of local experts". In: *Neural computation* 3.1 (1991), pp. 79–87.

[20] Albert Q Jiang et al. "Mixtral of experts". In: *arXiv preprint arXiv:2401.04088* (2024).

[21] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. "Cooperative learning of audio and video models from self-supervised synchronization". In: *Advances in Neural Information Processing Systems*. 2018, pp. 7763–7774.

[23] Haoyuan Lan, Yang Liu, and Liang Lin. "Audio-visual contrastive learning for self-supervised action recognition". In: *arXiv preprint arXiv:2204.13386* 1 (2022).

[24] Dmitry Lepikhin et al. "Gshard: Scaling giant models with conditional computation and automatic sharding". In: *arXiv preprint arXiv:2006.16668* (2020).

[25] Mike Lewis et al. "Base layers: Simplifying training of large, sparse models". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 6265–6274.

[26] Guangyao Li, Wenxuan Hou, and Di Hu. "Progressive Spatio-temporal Perception for Audio-Visual Question Answering". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 7808–7816.

[27] Guangyao Li et al. "Learning to answer questions in dynamic audio-visual scenarios". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19108–19118.

[28] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. "Dual-modality seq2seq network for audio-visual event localization". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 2002–2006.

[29] Yan-Bo Lin and Yu-Chiang Frank Wang. "Audiovisual transformer with instance attention for audio-visual event localization". In: *Proceedings of the Asian Conference on Computer Vision*. 2020.

[30] Yan-Bo Lin et al. "Vision transformers are parameter-efficient audio-visual learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2299–2309.

[31] Yuhang Ling et al. "TransAVS: End-to-End Audio-Visual Segmentation with Transformer". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 7845–7849.

[32] Tianyu Liu et al. "Induction Network: Audio-Visual Modality Gap-Bridging for Self-Supervised Sound Source Localization". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 4042–4052.

[33] Ze Liu et al. "Swin transformer v2: Scaling up capacity and resolution". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 12009–12019.

[34] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[35] Yuxuan Lou et al. "Cross-token modeling with conditional computation". In: *arXiv preprint arXiv:2109.02008* (2021).

[36] Jiaqi Ma et al. "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1930–1939.

[37] Tanvir Mahmud and Diana Marculescu. "Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 5158–5167.

[38] Yuxin Mao et al. "Contrastive conditional latent diffusion for audio-visual segmentation". In: *arXiv preprint arXiv:2307.16579* (2023).

[39] Yuxin Mao et al. "Multimodal variational auto-encoder based audio-visual segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 954–965.

[40] Shentong Mo and Yapeng Tian. "Multi-modal grouping network for weakly-supervised audio-visual video parsing". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 34722–34733.

[41] Basil Mustafa et al. "Multimodal contrastive learning with limoe: the language-image mixture of experts". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 9564–9576.

[42] Xiaonan Nie et al. "Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate". In: *arXiv preprint arXiv:2112.14397* (2021).

[43] Andrew Owens and Alexei A Efros. "Audio-visual scene analysis with self-supervised multisensory features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 631–648.

[44] Kyu Ri Park, Youngmin Oh, and Jung Uk Kim. "Enhancing Audio-Visual Question Answering with Missing Modality via Trans-Modal Associative Learning". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 5755–5759.

[45] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[46] Varshanth Rao et al. "Dual perspective network for audio-visual event localization". In: *European Conference on Computer Vision*. Springer. 2022, pp. 689–704.

[47] Carlos Riquelme et al. "Scaling vision with sparse mixture of experts". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8583–8595.

[48] Stephen Roller, Sainbayar Sukhbaatar, Jason Weston, et al. "Hash layers for large sparse models". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17555–17566.

[49] Pritam Sarkar and Ali Etemad. "Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 8. 2023, pp. 9723–9732.

[50] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. "A simple baseline for audio-visual scene-aware dialog". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12548–12558.

[51] Noam Shazeer et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer". In: *arXiv preprint arXiv:1701.06538* (2017).

[52] Sheng Shen et al. "Scaling vision-language models with sparse mixture of experts". In: *arXiv preprint arXiv:2303.07226* (2023).

[53] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. "Unified multisensory perception: Weakly-supervised audio-visual video parsing". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer. 2020, pp. 436–454.

[54] Yapeng Tian et al. "Audio-visual event localization in unconstrained videos". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 247–263.

[55] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[56] Yu Wu and Yi Yang. "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1326–1335.

[57] Yu Wu et al. "Dual attention matching for audio-visual event localization". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6292–6300.

[58] Yan Xia and Zhou Zhao. "Cross-modal background suppression for audio-visual event localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19989–19998.

[59] Haoming Xu et al. "Cross-modal relation-aware networks for audio-visual event localization". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 3893–3901.

[60] Hanyu Xuan et al. "A proposal-based paradigm for self-supervised sound source localization in videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1029–1038.

[61] Pinci Yang et al. "Avqa: A dataset for audio-visual question answering on videos". In: *Proceedings of the 30th ACM international conference on multimedia*. 2022, pp. 3480–3491.

[62] Jiashuo Yu, Ying Cheng, and Rui Feng. "Mpn: Multimodal parallel network for audio-visual event localization". In: *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2021, pp. 1–6.

[63] Jiashuo Yu et al. "Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing". In: *Proceedings of the 30th ACM international conference on multimedia*. 2022, pp. 6241–6249.

[64] Heeseung Yun et al. “Pano-avqa: Grounded audio-visual question answering on 360deg videos”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2031–2041.

[65] Jinxing Zhou et al. “Audio-visual segmentation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 386–403.

[66] Jinxing Zhou et al. “Positive sample propagation along the audio-visual event line”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8436–8444.

[67] Yanqi Zhou et al. “Mixture-of-experts with expert choice routing”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 7103–7114.

[68] Barret Zoph et al. “St-moe: Designing stable and transferable sparse expert models”. In: *arXiv preprint arXiv:2202.08906* (2022).

A Audio-Visual Tasks and Datasets

To demonstrate the capabilities of our proposed *AVMoE* in solving diverse audio-visual tasks, we conduct extensive experiments on four downstream tasks, i.e., Audio-Visual Event Localization (AVE), Audio-Visual Video Parsing (AVVP), Audio-Visual Segmentation (AVS), and Audio-Visual Question Answering (AVQA). The transferring framework is shown in Figure 5.

A.1 Audio-Visual Event Localization

AVE task is defined as an event that is both visible and audible in a video segment[54]. The AVE dataset focuses on combining audio and visual modalities to temporally mark the boundaries of all audio-visual events. It contains 4,143 videos covering 28 event categories, with each video lasting up to 10 seconds and including at least one 2s long audio-visual event. For this task, we first divide the video into 10 consecutive time segments. Then, we use the frozen transformers (e.g., ViT, Swin or HTS-AT) optimized by the AVMoE adapters to extract the audio and visual features that incorporate cross-modal interaction information from these segments. Last, we predict the audio-visual events by feeding the concatenated audio-visual features through two MLP layers, as shown in Figure 5. Referring to previous methods[54, 46, 58, 66], we calculate the proportion of time segments that are correctly predicted and use this as a standard to evaluate the performance of our method.

A.2 Audio-Visual Video Parsing

AVVP task aims to parse videos into temporal event segments and label them as audible, visible, or both. For this task, we use the Look, Listen, and Parse (LLP) dataset[53], which contains 11,849 videos from different domains, covering 25 categories. LLP is a semi-supervised annotation dataset, and each video has video-level event annotations. Only 1,849 randomly selected videos have second-by-second annotations for audio and visual events. We made improvements based on MGN[40] by using *AVMoE* to enhance the Transformer layers in it, as shown in Figure5. Referring to previous methods[53, 40], F-scores are used to evaluate segment-level and event-level predictions of audio, visual, and audiovisual events which are predicted by our method.

A.3 Audio-Visual Segmentation

AVS task aims to achieve pixel-level segmentation of visual objects based on audio. We conduct experiments on the AVSBench[65], a dataset containing both single-source segmentation and multiple-source segmentation. The single-source dataset contains 4,932 semi-supervised videos. During the training process, only the first sample frame of the video is fully labeled, but during evaluation, all video frames need to be predicted. The multi-source dataset contains 424 fully supervised videos with every frame labeled. For this task, we optimize the frozen transformer with trainable adapters to replace the pre-trained visual encoder and audio feature extractor of the U-Net used in AVS[65]. Then, the audio and visual features are fused and ingested into the decoder for prediction, as shown in Figure5. Referring to previous methods[65], we calculate the mean Intersection-over-Union (mIoU) of the predicted segmentation and the ground truth masks to evaluate our method.

A.4 Audio-Visual Question Answering

AVQA task aims to answer questions about different visual objects, sound sources and their associations in the video. We conduct experiments on the MUSIC-AVQA dataset[27], which contains more than 45K question-answer pairs covering 33 different question templates. For this task, we utilize the trainable adapter to optimize the frozen Transformer and then replace the pre-trained visual and audio encoders of the baseline proposed in the AVQA dataset, as shown in Figure5. Referring to previous methods[27], the answer prediction accuracy is used to evaluate the performance of methods.

B Audio-Visual Embeddings

To enhance the fusion of audio-visual features, we follow the previous work [11], which utilizes pre-trained vision and audio transformers to encode visual and audio inputs, respectively. The encode process is as follows.

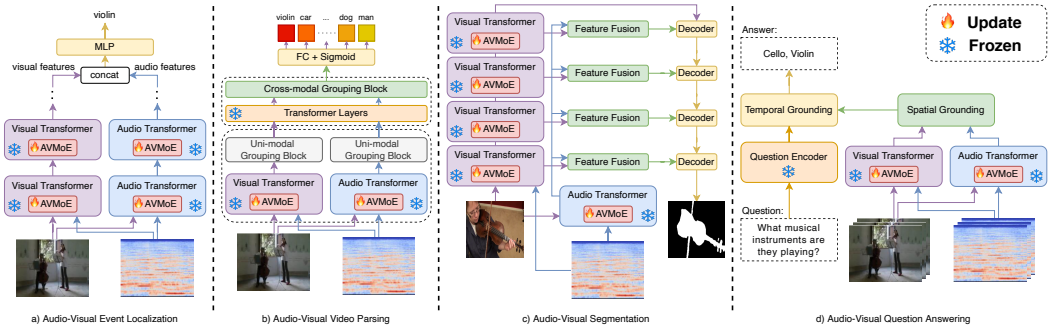

Figure 5: Applying AVMoE to audio-visual downstream tasks: audio-visual event localization, audio-visual video parsing, audio-visual segmentation, and audio-visual question answering. The purple and blue modules are frozen pre-trained visual and audio models, and the red modules are our proposed trainable adapter modules, and the remaining modules are set up with reference to the baseline model of these tasks, with some parameters trainable.

Visual embedding. For the given video sequence, we sample a set of RGB visual frames $\{V_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$, where H and W denote the height and width of frames, respectively. These video frames are sampled at the rate of 1 fps and then decomposed into n non-overlapping patches. The patches are treated as individual tokens for sequence processing by the transformer model, and we can obtain the visual embeddings $v_t \in \mathbb{R}^{n\times d}$, where d denotes the embedding dimension.

Audio embedding. For the given audio stream, we first employ Short-Time Fourier Transform (STFT) to convert waveforms into spectrograms $\{A_t\}_{t=1}^{T} \in \mathbb{R}^{T\times L\times F}$, where L and F denote the time and frequency bins, respectively. Similarly, the spectrograms are also divided into k non-overlapping patches and then projected into audio embeddings $a_t \in \mathbb{R}^{k\times d}$.

C Implementation Details

For the above dataset, we extract visual frames at one frame per second and audio frames from video files at 8,000fps. We use the pre-trained Swin-V2-Large[34] as the visual transformer module with a spatial resolution of 192×192, and the pre-trained HTS-AT[5] as the audio transformer module. All of their parameters are frozen. We use the gradient accumulation method to accommodate the large parameters of the model, with weight updates after every 16 batches of training. For AVE and AVVP tasks, 32 latent tokens and a downsampling factor of 8 are used in the AVMoE adapter. For the AVS task on AVSBench and the AVQA task on MUSIC-AVQA, we use two latent tokens and set the downsampling factor and the number of group convolutions to 8 and 4, respectively.

For all of our experiments, we utilize the Adam[21] optimizer to train our models and set the scheduler to make the learning rate decay to 0.35 times its original value after every 3 epochs. We set the learning rate of the AVMoE adapter to 5e-4, while the learning rates of the final prediction layer are the same as previous work [11], 5e-6 for AVE, 3e-4 for AVVP, 3e-4 for the S4 setting of AVS, 1.5e-4 for the MS3 setting of AVS, and 1e-4 for AVQA. For audio pre-processing, we compute audio spectrograms via the PyTorch[45] kaldi fbank with 192 triangular mel-frequency bins and set frameshift over 5.2ms. For the task with audio-visual shared Transformer, we spliced the input channels of the audio spectrogram by tensor replication into 3 channels to match the input dimensions of the SwinV2 model.

For these audio-visual downstream tasks, all experiments are conducted on 8x NVIDIA 3090 (24G) GPUs, and the batch size on a single GPU varies depending on the parameters of the models. The one GPU batch size is set to be 1 for AVE and AVVP tasks, and 2 for AVS and AVQA tasks. Besides, we employ a gradient accumulation strategy to mitigate the computational resource limitations. To avoid overfitting, we set the early-stop hyper-parameterization to end the training process after five epochs with no improvement in model performance. Most of the experiments converge to the best results in about ten epochs.

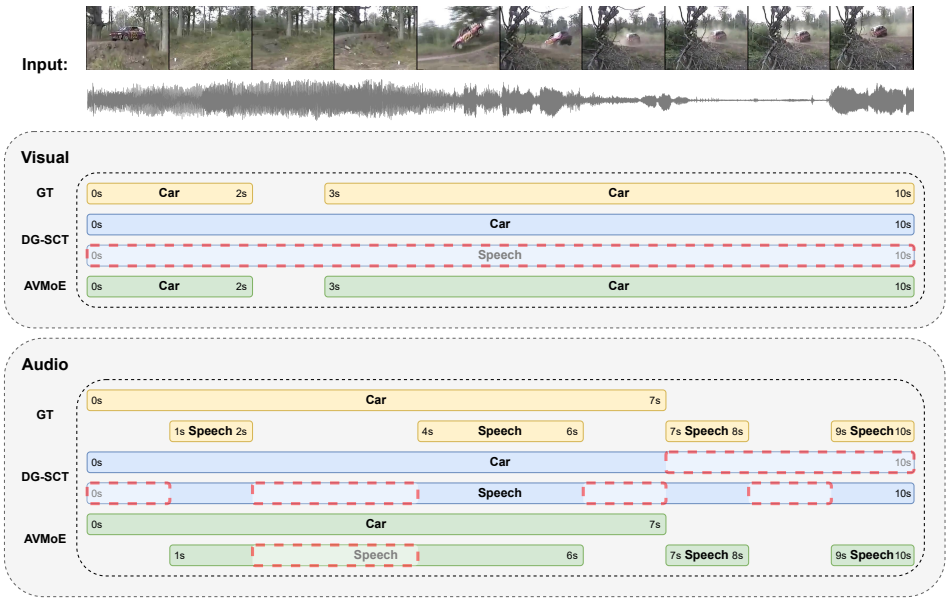

Figure 6: The qualitative experimental results on the AVVP task.

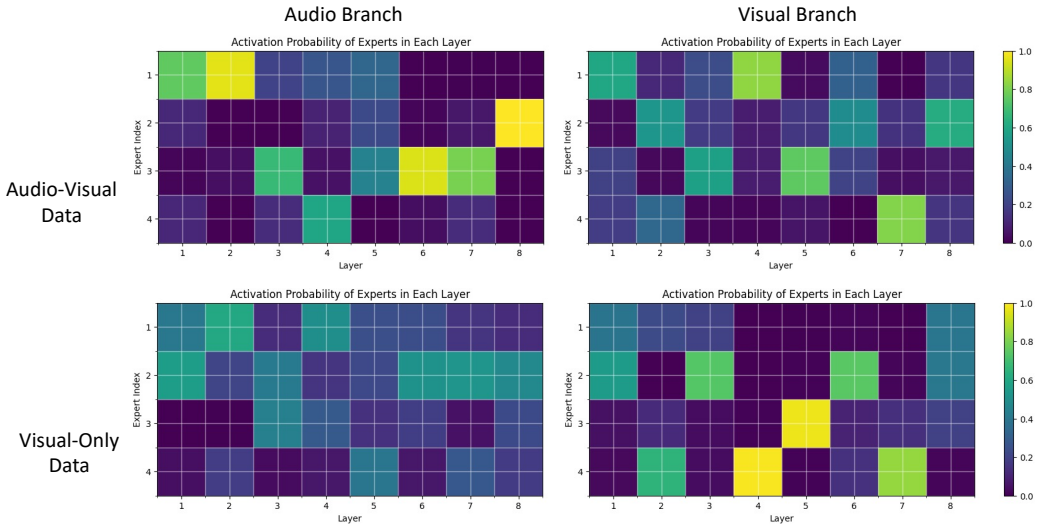

Figure 7: The activation probability of experts. Experts #1 and #2 denote cross-modal adapters, and experts #3 and #4 denote unimodal adapters.

D Qualitative Experiments of Audio-Visual Video Parsing

Figure 6 shows the qualitative experimental results on the AVVP task. In this example, the video label only includes car, and the audio label includes car and speech, which are not exactly the same. We observe that the visual predictions of *DG-SCT* are affected by the audio features, resulting in a visual prediction of speech, which does not exist in the video frame. On the other hand, our *AVMoE* retains unimodal information and has MoE modules to adjust the weight distribution for audio-visual features. This avoids the interference of inconsistent information and obtains the correct result. As for the audio predictions, our model can better localize the temporal boundaries of events compared to *DG-SCT*, which reveals that our model is capable of dealing with complex scenarios.

E The activation probability of experts

The activation visualization of experts is shown in Figure 7. It can be observed that: 1)In the visual branch, The distribution of expert activation is relatively even, while in the audio branch, the activation probability of the crossmodal adapter is higher. We hypothesize that this may be due to the visual information plays a vital role in the processing of audio-visual data. 2) When the model deals with visual-only data, the activation probability of the unimodal adapter in the visual branch becomes higher, which demonstrates the ability of our *AVMoE* to dynamically combine adapters according to different scenarios.

In addition, we observe that when the model is processing visual-only data, the activation probability of the unimodal adapter in the visual branch becomes higher, which proves the ability of our *AVMoE* to dynamically activate adapters according to different scenarios.

F Feature Visualization

Figure 8 visualizes the learned audio and visual features by employing t-SNE [55]. Each point represents the feature of an individual audio or visual event, with each color indicating a specific category. The visualizations reveal that the features extracted by the proposed *AVMoE* model are more compact within classes and more distinct between classes, which demonstrates that our *AVMoE* can learn discriminative features for each modality across various audio-visual downstream tasks.

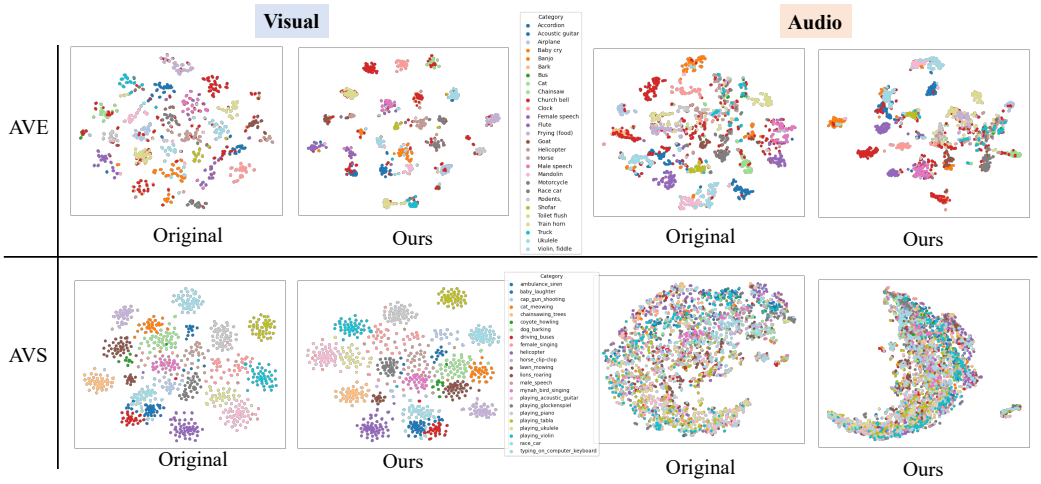

Figure 8: Qualitative visualizations of visual and audio features of original and Ours on AVE and AVS tasks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our scope and contributions in the abstract and introduction Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We analyze the issues with our method based on experimental results in Section 4 and explore our limitations in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not introduce any new theories.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We first describe the pipeline of our method in Section 3, and then detail the implementation details and training parameters in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the URL of our code in abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental settings, benchmark and training details can be found in Section 4, and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Benchmarks and previous methods do not provide any suggestions or practices regarding the experiment statistical significance. The cost of a single experiment is very expensive, which means that the number of experiments that meet statistical significance is unacceptable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments compute resources are reported in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: 3D object detection in point clouds is mainly used for academic research and technological development, without involving scenes that have an impact on society, specific groups, and is not directly associated with any potential negative social applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not involve generative models, pretrained models, and do not have a risk of abuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We adhere to the usage restrictions of all assets and correctly cite their sources.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[NA]

    Justification: We do not release new assets now, and the code will be open-sourced after the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[NA]

    Justification: We do not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.