Finetuning Multi-modal Models with Heterogeneous Multi-modal MoE Adapter

Anonymous ACL submission

Abstract

Multi-modal models have demonstrated successful applications across various audio-visual and visual-text tasks. To extend the model's applicability to a wider range of downstream tasks, the Parameter-Efficient Fine-Tuning (PEFT) approach has garnered extensive attention as it can efficiently fine-tune pre-trained models by inserting small trainable parameter layers while keeping most of the parameters of the pre-trained model frozen. However, the most existing PEFT method are mainly based on uni-modal model designs and overlook the modal fusion process, which is essential for multi-modal models. In this paper, we introduce HMMoE, a finetuning method tailored for multi-modal interactions that enhances the performance of multi-modal models with minimal trainable parameters. Experimental results across 8 types of downstream tasks, including audio-visual and visual-text tasks, demonstrate the superior performance of our approach over previous methods. The code is provided in the supplementary material.

1 Introduction

The nature of the world is inherently multi-modal, with humans perceiving various types of information through different sensory organs, such as language, images, videos, and sounds. These forms of information often complement and synergize with each other. Consequently, multi-modal models are primarily endowed with the ability to understand and perceive various other modalities (Wang et al., 2023; Bao et al., 2022; Li et al., 2023), such as vision, video, and audio. Multi-modal models can offer more intelligent and comprehensive solutions in practical applications. For instance, in areas like search engines, recommendation systems, and intelligent assistants, these models can better understand user needs and provide more accurate services. However, full parameters finetuning a multi-modal language model demands substantial computational and memory resources, which can be prohibitively expensive for average researchers.

To more efficiently apply pre-trained models to downstream tasks, Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019; He et al., 2022) methods have been proposed. These techniques significantly reduce the cost of fine-tuning by freezing the original parameters and introducing smaller trainable layers, such as Adapters (Houlsby et al., 2019) and Low-Rank Adaptation (LoRA) (He et al., 2022). However, most existing PEFT methods (Luo et al., 2023; Tang et al., 2024; Wang et al., 2024) primarily focus on fine-tuning single-modal inputs and lack specialized structural designs for crucial multi-modal fusion processes. This limitation weakens their effectiveness in multi-modal interaction tasks.

Some previous works (Mustafa et al., 2022; Chen et al., 2023a; Shen et al., 2023) have integrated Mixture of Experts (MoE) (Jacobs et al., 1991; Eigen et al., 2013) methods into PEFT approaches by placing multiple adapters in different transformer layers and using a router to select appropriate expert combinations based on scenarios. This expands encoding capacity while maintaining inference efficiency (Chen et al., 2023a; Yu et al., 2024; Cappellazzo et al., 2024). However, these methods neglect the interactions between different modalities by separating them into distinct MoE adapters, which disregards the mutual guidance of multi-modal features essential for cross-modal tasks. As a result, they struggle to handle complex multi-modal tasks, such as audio-visual applications (Xuan et al., 2020; Tian et al., 2020; Li et al., 2022a; Zhou et al., 2022a).

To enhance modal interactions during the finetuning of multi-modal models, we propose a novel **Heterogeneous Multi-Modal Mixture of Experts Adapter (HMMoE)**. This adapter allows each expert to process inputs from multiple modalities, facilitating multi-modal interactions. Specifically,

inputs are first mapped into a low-rank space via a down-projection, where cross-modal interactions occur, and then fused features are projected back to their original dimensions before proceeding to the next layer. This design enables trainable parameters to effectively adjust the gradients of the frozen pre-trained model by leveraging collaborative interactions among modalities in downstream tasks.

Moreover, we replace the traditional single-expert structure with a heterogeneous expert architecture that includes both conventional adapters and specialized multi-modal interaction experts, such as cross-attention experts for capturing cross-modal information and channel-attention experts for targeted information extraction. Experts are grouped by type, with each group containing multiple identical adapters. A global router assigns weight scores for fusion, while local routers in each group activate the top-k most suitable experts, thereby improving overall model performance.

We integrate the proposed HMMoE modules into existing multi-modal models, and conduct extensive experiments on visual-audio and text-vision tasks. The experimental results show that our model can achieve performance comparable to fine-tuning all parameters while using only 5-7% of the parameters. Furthermore, it demonstrates a substantial improvement over prior Parameter-Efficient Fine-tuning methods, presenting a novel avenue for fine-tuning multi-modal models with limited parameters.

Our main contributions are as follows:

- We propose the **HMMoE** module, which is specially designed to enhance the model's comprehension of cross-modal information through modal interactions within a low-dimensional parameter space.

- We introduce a novel multi-modal heterogeneous experts ensemble and validate that it outperforms traditional single experts in terms of performance.

- We apply the HHMoE modules to multiple audio-visual and video-text downstream tasks, achieving comparable performance to full fine-tuning with just 5-7% of the parameters, and significantly outperforming existing Parameter-Efficient Fine-tuning methods.

2 Related Work

2.1 Mixture of Experts

Mixture of experts (Jacobs et al., 1991; Eigen et al., 2013; Shazeer et al., 2017) represent a unique neural network architecture by partitioning layer parameters into discrete experts with distinct weights. Throughout both training and inference (Lewis et al., 2021; Li et al., 2022b), inputs are carefully allocated to specific expert weights, engaging only a subset of network parameters. Furthermore, related work (Fedus et al., 2021; Rajbhandari et al., 2022) has enhanced the performance of a mixture of experts layers by routing each example to a single expert rather than to K experts. This approach reduces routing computation while preserving model quality. Unlike traditional methods, where the entire network typically processes each input, this strategy results in significant gains in computational efficiency. In natural language processing (Hazimeh et al., 2021; Mustafa et al., 2022; Zoph et al., 2022; Zhou et al., 2022b; Lepikhin et al., 2021), the mixture of experts approach is widely employed to enhance model performance while maintaining control over the number of computational parameters. This method is also extensively utilized in the field of computer vision (Mustafa et al., 2022; Chen et al., 2024; Lin et al., 2024) and has demonstrated substantial success in visual language tasks. In our model, we not only introduce a mixture of experts structure to increase model capacity like these works, but also promote modal interaction in different scenarios through this method.

2.2 Parameter-Efficient Fine-tuning

Parameter-Efficient Fine-tuning (Houlsby et al., 2019) of large language models is paramount as model sizes continue to grow. Various strategies have been proposed to address this challenge. Low-Rank Adaptation (LoRA) (He et al., 2022; Zhang et al., 2023) reduces parameters by incorporating trainable low-rank matrices into fully connected layers, yielding significant resource savings without extra inference costs. Adapter (Houlsby et al., 2019; Pfeiffer et al., 2020) offers an effective fine-tuning approach by enabling selective modification of pre-trained parameters, thus enhancing resource utilization without compromising model performance. Prompt learning (Lester et al., 2021) presents a promising avenue by leveraging task-specific prompts for fine-tuning with minimal additional parameters. Among these meth-

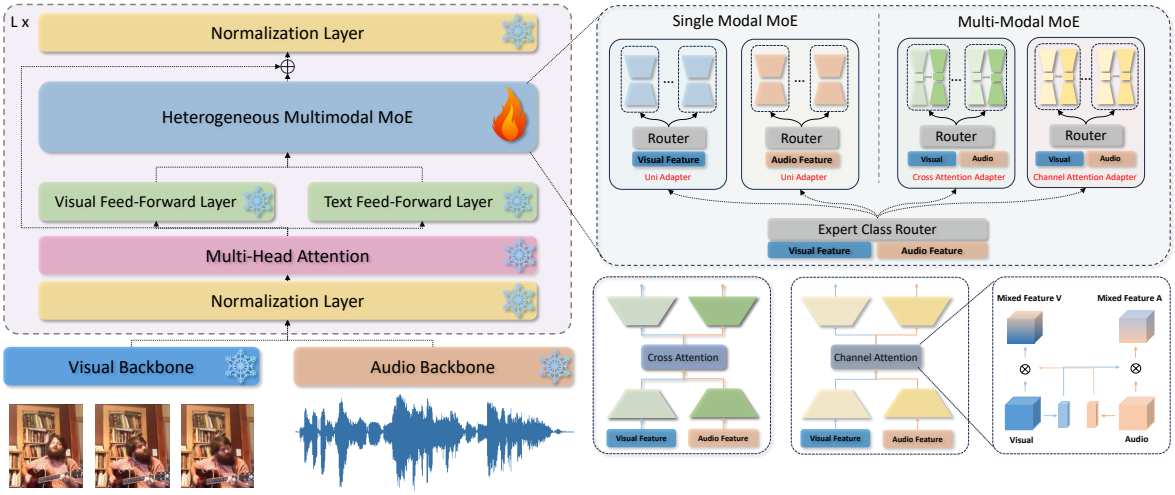


Figure 1: The overall architecture of our proposed method. The left half shows the heterogeneous multi-modal MoE inserted into the pre-trained model as an additional trainable layer. The upper part of the right half shows the overall routing structure. The bottom half on the right shows the internal structure of the heterogeneous experts.

ods, parameter-efficient fine-tuning via low-rank adapters stands out, showing potential for substantial resource savings while potentially improving performance. This inspired us to design the model to perform modal fusion operations in low dimensions to control the number of parameters for overall fine-tuning.

3 The Method

3.1 Overview

To illustrate the functionality of the heterogeneous multi-modal mixture of experts adapter (HMMoE), we use the visual-audio task as a representative example. As shown in Figure 1, the HMMoE module is plugged into the transformer structure after the feed-forward layer. Taking ℓ^{th} Layer in Transformer as an example, the HMMoE module accepts visual feature $V^{\ell} \in \mathbb{R}^{B \times S_V \times D}$ and audio feature $A^{\ell} \in \mathbb{R}^{B \times S_A \times D}$ as inputs and outputs features of the same dimension to the next layer. Here, B denotes the batch size, S_A ,and S_V denote the sequence lengths for audio and visual inputs respectively, and D denotes the feature dimension.

Within the HMMoE structure, experts are divided into two groups: single-modal experts and multi-modal experts. The global router determines the weight factor assigned to each group of experts. The coefficients of the combination are assigned by local routers within each group. The structure of the second-level route is used to balance the function of performance and parameter quantity. Visual features and audio features are fused with feature information by multi-modal experts. At the same time, in order to keep the unique feature information of each mode, the feature learning is carried out by their own single-mode experts.

Regarding the overall structure, given the input video feature V^{ℓ} and audio feature A^{ℓ}, the model's output can be expressed as follows:

$$V^{\ell+1} = \mathbf{G}_m^V \sum_{j=1}^{M} \mathbf{W}_j^V \cdot E_j \left(V^{\ell}, A^{\ell}\right) + \mathbf{G}_s^V \sum_{j=1}^{M} \mathbf{W}_j^V \cdot E_j^V \left(V^{\ell}\right) \quad (1)$$

Where $\mathbf{G}$ represents the weight given by the global router, $\mathbf{W}$ represents the weight given by the local router within the group, E represents a single single-modal or multi-modal expert, and M is the number of experts within each group.

3.2 Heterogeneous Expert Group

The purpose of introducing heterogeneous expert groups is to better facilitate the interaction of representations between different modes at the level of different perceptual dimensions. The expert group is divided into two groups: multi-modal experts and single-modal experts. The multi-modal expert receives the feature inputs of the two modes and performs the feature fusion, which is mainly realized by the global attention mechanism expert and the channel attention mechanism expert. We introduced cross-modal attention mechanisms to

enable the model to learn cross-modality transfer information. Single-modal experts, on the other hand, focus on maintaining the extraction of information specific to the mode. The structure design of a heterogeneous expert group ensures the maximum maintenance of the original modal information while extracting the modal interaction information in multiple dimensions.

3.2.1 Multi-modal Expert

The role of the multimodal expert is primarily to facilitate the interaction and fusion of different modes. In order to minimize the number of parameters used within the expert, we use a low-rank decomposition method to map the features to a much smaller dimension and carry out modal feature interactions in this lower dimension. Subsequent experiments show that this method can maintain the performance of the model as much as possible while greatly reducing the number of model parameters. We propose two multi-modal experts to extract features for cross-modal attention mechanisms and channel-dimensional attention mechanisms respectively.

Cross-modal Attention Expert:

In order to establish a comprehensive attention mechanism between the two modals, we introduce a cross-modal attention expert mechanism, enabling the model to effectively capture and integrate intricate relationships and interactions. The input consists of features from two modalities, denoted as $V \in \mathbb{R}^{B\times S_V\times D}$ and $A \in \mathbb{R}^{B\times S_A\times D}$. Initially, these features are mapped to bottleneck dimensions r using projections $\mathcal{W}_{down}$, resulting in two low-dimensional features $\overline{V} \in \mathbb{R}^{B\times S_V\times r}$ and $\overline{A} \in \mathbb{R}^{B\times S_A\times r}$. Subsequently, an interaction attention mechanism is employed for these two low-dimensional features.

Using the Audio-to-Visual attention mechanism as an illustration, the feature representation $\overline{V}$ is transformed via the linear layer $\mathcal{W}_q \in \mathbb{R}^{r\times r}$ to serve as the query. Similarly, the audio features $\overline{A}$ undergo projectios $\mathcal{W}_k \in \mathbb{R}^{r\times r}$ and $\mathcal{W}_v \in \mathbb{R}^{r\times r}$ to calculate the key and value respectively. The dot product between the key and query is computed and then passed through a softmax function to obtain weights. These weights are then dotted with the value to get the output. Combined with residuals for the low-dimensional features, the original feature dimension is then obtained by an up-projection $\mathcal{W}_{up} \in \mathbb{R}^{r\times D}$. The entire procedure for visual feature V can be succinctly articulated using the following formula:

$$\overline{V} = \mathcal{F}_{relu}(V \cdot \mathcal{W}_{down}) \tag{2}$$

$$V_{out} = \left(softmax\left(\frac{\overline{V}\mathcal{W}_q(\overline{A}\mathcal{W}_k)^T}{\sqrt{d_V}} \right) \overline{A}\mathcal{W}_v + \overline{A} \right) \cdot \mathcal{W}_{up} \tag{3}$$

Finally, the cross-modal attention expert introduces the overall attention information of the audio mode into the features of the current visual mode. The process of introducing visual information into audio features is the same.

Channel-Attention Experts:

In a variety of cross-modal downstream tasks, relying solely on overall attention mechanisms often fails to capture the fine-grained information specific to each modality. To address this, a channel-attention expert is introduced to place particular emphasis on the channel dimension, ensuring that this detailed information is adequately represented.

Channel-attention expert receives inputs from two modal features denoted as $V \in \mathbb{R}^{B\times S_V\times D}$ and $A \in \mathbb{R}^{B\times S_A\times D}$. The whole process is again illustrated through the audio-to-video channel attention process. In order to obtain fine-grained interaction information of dimensions, feature A is averaged pooled along the S_A dimension, and then multiplied with feature V element-by-element. Following that, feature V is mapped to a low-dimensional space with $\mathcal{W}_{down} \in \mathbb{R}^{D\times r}$ in a similar way as previously mentioned, so that the number of parameters in the model is kept as low as possible. The remaining processing methods are similar to cross-modal attention experts. The final mapped feature undergoes element-wise multiplication with the channel attention weight and is residual-connected with the original values, yielding the output feature V_{out}. The entire process can be formally described as follows:

$$\text{Attn=Sigmoid}(AvgPool_s\left(AvgPool_s(A) \cdot V\right)) \tag{4}$$

$$V_{out} = \mathcal{F}_{relu}(V \cdot \mathcal{W}_{down}) \cdot \mathcal{W}_{up} \cdot (1 + Attn) \tag{5}$$

Single-modal Experts Over-reliance on cross-modal information can impair the model's ability to extract unique features from each modality, ultimately degrading overall performance. To address this, we introduce single-modal experts with simplified structures to preserve the feature extraction capability of individual modes.

The single-modal expert is similar to the traditional adapter. It consists of two fully connected layers. For the input single modal feature $V \in \mathbb{R}^{B \times S_V \times D}$, the feature dimension D is first mapped to a lower bottleneck dimension r through a down-projection $\mathcal{W}_{down} \in \mathbb{R}^{D \times r}$, and an up-projection $\mathcal{W}_{up} \in \mathbb{R}^{r \times D}$ maps the feature back to the original dimension. Combined with the residual connection of the original features, this process can be represented using the following formula:

$$V_{out} = V + \mathcal{F}_{relu}(V \cdot \mathcal{W}_{down}) \cdot \mathcal{W}_{up} \quad (6)$$

The processing of audio features follows the same approach. To maintain the uniqueness of each modality, the single-modal experts for visual and audio features are kept distinct.

3.3 Routing Method

The aim of the routing method is to select the most suitable experts for processing input features. In our proposed model design, there are two types of routers: the global router is employed to assign appropriate weights to different types of experts, while the local router is utilized to select the top-k experts within each group of experts.

Global Router: To fully leverage the strengths of each group of experts, the global router is tasked with assigning appropriate weight coefficients to the different expert groups. This involves learning how to effectively determine the weight coefficients for various combinations of experts.

Instead of a fixed weight coefficient, we set up a learnable weight allocation method, which proved to be reasonable in subsequent analysis. The model has different preferences for different levels of experts, and the adaptive parameter setting can dynamically adapt to this preference. We set the global route weight G to control the weights assigned to different groups of experts. In simple terms, for the input feature x, the weight coefficient can be expressed as: $G_{s,m}(x) = SoftMax(W_{gr}(x))$, where W_{gr} represents a series of learnable linear mapping operations.

Local Router: Local routers are used to select the most appropriate combination of experts within each group for feature assignment. Multi-modal expert groups can be expressed as $\overline{E}_m = [E_1, E_2, \ldots, E_N]$, where N is the total number of experts in each group. With a learnable weight W^{lr}, the weight coefficient for each expert is obtained:

$$P(x)_i = \frac{e^{W_i^{lr} x}}{\sum_{j=1}^{N} e^{W_j^{lr} x}} \quad (7)$$

Therefore, each feature in the expert treatment of each group is processed by the top-k expert with the highest probability, and the weighted sum is calculated according to the softmax result of the probability:

$$Group(x) = TopK(P(x)_i \cdot E(x)_i) \quad (8)$$

Under the joint action of the global router and the local router, the experts of each group are reasonably assigned appropriate weights to process the input features.

4 Experiments

4.1 Experiment Setup

In this section, we mainly introduce our model training methods. We insert the HMMoE module into the encoder layer of the pre-trained multi-modal model, initializing both single-modal and multi-modal experts. During training, we freeze the original model parameters and train only the parameters of the HMMoE layers and classification head parameters. For comparison, we also evaluate the performance of traditional PEFT methods including series-adapter (Houlsby et al., 2019), parallel adapter (He et al., 2021), LoRA (Hu et al., 2021) and LoRA-FA (Zhang et al., 2023).

Downstream tasks are mainly divided into visual-audio tasks and text-visual tasks. The visual-audio task is based on the pre-trained Swin-T (Liu et al., 2021) and HT-SAT (Chen et al., 2022) models, two representative traditional visual and audio encoders, performed on the AVE (Tian et al., 2018; Xia and Zhao, 2022), AVVP (Tian et al., 2020), AVQA (Li et al., 2022a) and AVS (Zhou et al.) tasks. The text-visual tasks including MSVD and MSRVTT (Xu et al., 2016) are implemented based on the pre-trained VALOR (Chen et al., 2023b) model, where VALOR is a typical dual-tower encoder facilitates the mixed processing of encoded information of multiple modes. In addition, we also explore the experimental effect of the proposed method on the model whose structure is pre-trained based on the MOE method. We test VQA and NLVR tasks based on representative VLMO (Bao et al., 2022) models. Our method was trained on an A800 GPU, and the training parameter settings are the same as the base modeland. More experimental details are in the supplementary materials.

Method	Param.(M)	AVE	AVVP		AVQA				AVS-S4	
		Acc	segment	event	AQ	VQ	AVQ	Avg	mIoU	F
Full-finetune	*313(100%)*	82.2	52.8	46.1	77.4	81.9	70.7	74.8	80.9	89.2
Lora	20(6.3%)	79.8	52.6	45.9	75.4	81.3	70.5	74.3	79.8	88.1
Lora-FA	20(6.3%)	79.5	52.5	46.0	75.1	80.9	70.7	74.7	79.2	87.9
Series-Adapter	20(6.3%)	79.9	52.0	45.9	76.3	81.9	70.2	74.2	80.2	88.6
Parallel-Adapter	20(6.3%)	80.2	52.3	45.3	**76.9**	81.7	71.1	74.9	80.1	88.8
Ours	20(6.3%)	**81.1**	**53.4**	**46.8**	76.7	**82.4**	**71.3**	**75.1**	**80.9**	**89.3**

Table 1: Performance comparison of visual-audio tasks based on Swin-T and HT-SAT encoders: a comparison with traditional methods with equal parameters.

4.2 Implementation Detail

4.2.1 Visual-Audio Tasks

In this series of experiments, we seamlessly integrate our HMMoE module into the well-established Swin-T and HT-SAT model layers during training. For the Audio-Visual Event (AVE) task, our HMMoE module is synergistically combined with CMBS, and overall accuracy in predicting event categories is used as the evaluation metric. In the Weakly-Supervised Audio-Visual Video Parsing (AVVP) task, our module is paired with MGN, and performance is assessed using segment-level and event-level metrics across all event types (audio, visual, and audio-visual). For the Audio-Visual Segmentation (AVS) task, the HMMoE module is integrated with the original AVS model, with performance evaluated using mean Intersection over Union (mIoU) and F-score. Finally, in the Audio-Visual Question Answering (AVQA) task, our HMMoE module is incorporated into the ST-AVQA framework, with answer prediction accuracy as the primary evaluation metric.

4.2.2 Text-Visual Tasks

In this set of experiments, we employ the VALOR model as a pre-trained foundation to tackle open-ended video QA tasks on the MSRVTT-QA and MSVD-QA datasets, using QA accuracy as the evaluation metric. We also assess vision-language classification tasks using the VLMo model, which is pre-trained using the MoE method. These tasks include the NLVR task on the NLVR2 dataset and the VQA task on the VQA2 dataset, both evaluated based on QA accuracy.

4.2.3 Comparison setting

In these tasks, our HMMoE module was configured with single-modal, cross-modal attention, and channel-attention expert groups, each containing two experts, with the rank (r) of the experts set to 32 and top-2 router method. In order to make a fair comparison with traditional fine-tuning methods including Lora, Lora-FA, serial-adapter and parallel-adapter, we adjust the value of the low-rank mapping dimension r to ensure that the number of parameters used by various methods is consistent.

4.3 Main Results

4.3.1 Performance Comparison

Based on the experimental results of table 1, our HMMoE method shows significant superiority for visual-audio tasks. In each task, our model achieved the best performance in both accuracy and overall performance. Specifically, in the AVE task, our model performs well and achieves the highest accuracy compared to other methods. In the AVVP and AVS-S4 tasks, our model also achieved the best scores, showing its superiority in handling audio and video tasks. This superiority is mainly due to the novel approach and effective strategy adopted in our model. By combining the Swin-T and HT-SAT encoders, our model is able to better capture the relationship between audio and video information, enabling more accurate task processing. At the same time, our model also performs

Method	Param.(M)	MSRVTT	MSVD
Full-finetune	*315(100%)*	44.5	54.9
Lora	16(5.1%)	43.7	53.5
Lora-FA	16(5.1%)	43.0	53.1
Series-Adapter	16(5.1%)	43.6	54.1
Parallel-Adapter	16(5.1%)	44.1	54.2
Ours	16(5.1%)	**45.2**	**55.6**

Table 2: Performance of text-visual tasks based on VALOR: comparison with traditional methods with equal parameters.

well in terms of parametric efficiency and performance balance, further highlighting its advantages in visual-audio tasks.

Based on the experimental results in the table 2 and 3, our method achieves the best performance on two different text-visual tasks. On the VALOR, our HMMoE module improves the MSRVTT and MSVD tasks by 0.7 percentage points, respectively, compared to other methods. On the VLMO benchmark, our model outperformed its closest competitor by 0.6 percentage points on the VQA task and 0.8 percentage points on the NLVR task. This improvement is due to the modal interaction features in our model, which effectively leverage the interaction between text and visual information. This enables our model to capture complex multi-modal relationships more accurately, enhancing both task performance and generalization. Compared to traditional PEFT methods, our model excels in cross-modal feature fusion, offering a clear advantage in text-visual tasks.

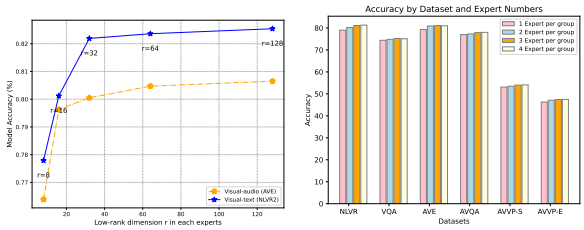
4.4 Ablation Study

4.4.1 Expert module ablation

In order to demonstrate that the design of single-modal experts and multi-modal experts can improve performance, we conducted an ablation experiment on the type of expert. For the downstream tasks of NLVR2, VQA, MSVD-QA, and MSRVTT-QA, we ablated single-modal experts, cross-attention experts, and channel-attention experts one by one. As the results in Table 5, introducing cross-attention experts generally improves model performance compared to using only single-modal experts. Additionally, fine-tuning performance can be further enhanced by incorporating channel-attention experts, which provide targeted dimension information extraction capabilities. The results show that the structure of expert combination is more robust in downstream tasks.

Model	Param.(M)	VQA	NLVR
Full-finetune	*360(100%)*	76.2	82.7
Lora	19(5.3%)	73.8	80.9
Lora-FA	19(5.3%)	73.6	80.5
Series-Adapter	19(5.3%)	74.4	81.1
Parallel-Adapter	19(5.3%)	74.6	81.4
Ours	19(5.3%)	**75.2**	**82.2**

Table 3: Performance on text-visual tasks based on VLMO: comparison with traditional methods with equal parameters.



(a) Rank ablation (b) Expert number ablation

Figure 2: The graph (a) shows the relationship between the low-rank dimension r and model performance, and the graph (b) shows the relationship between the number of experts in each group and model performance.

4.4.2 Low-dimention rank ablation

In our HMMoE method, the dimension r of the low-rank mapping plays a crucial role in reducing the original feature dimension to a lower-dimensional space. Decreasing r can help reduce the model's training parameters, but setting r too low may lead to a significant loss of original feature information, resulting in degraded performance. As shown in Figure 2a, the overall performance of the model increases rapidly as r increases. However, once r reaches around 32, further increases in r yield diminishing returns in performance improvement. Therefore, it is important to select r within a reasonable range: setting r too low can result in the loss of critical feature details while setting it too high can lead to unnecessary resource consumption without significant performance gains.

4.4.3 Experts number ablation

In general, the performance of the model increases as the number of experts in each group increases. As shown in Figure 2b, the rate of performance improvement diminishes with the addition of more experts. Since our goal is to fine-tune the entire multimodal model with minimal overhead, it is preferable to limit the number of experts to a low level, such as 2 or 3, to maintain performance without significantly increasing the model's complexity. This approach allows us to effectively balance performance gains with the number of parameters, ensuring an efficient trade-off.

4.4.4 Router method ablation

To evaluate the impact of the top-k policy on the local router, we conducted tests comparing various routing methods, including the top-1 policy, which can be considered a hard routing mode. As shown

Module			AVE	AVVP		AVQA				AVS-S4	
Single	Cross	Channel	Acc	Segment	Event	AQ	VQ	AVQ	Avg	mIoU	F
✓	-	-	80.1	53.1	46.1	74.8	81.8	70.6	74.6	79.8	88.5
✓	✓	-	80.9	53.6	46.9	75.0	82.3	71.1	74.9	80.7	89.0
✓	✓	✓	**81.1**	**54.1**	**47.4**	**76.7**	**82.4**	**71.3**	**75.1**	**80.9**	**89.3**

Table 4: Ablation results of the expert module in visual-audio tasks (Single for single-mode experts, Cross for cross-attention experts, and Channel for channel attention experts).

in Table 7, with all other conditions held constant, the routing method using the top-2 policy outperforms that of the top-1 policy. This indicates that combining multiple experts of the same type can enhance the model's comprehension capabilities.

4.5 Heterogeneous Effect Analysis

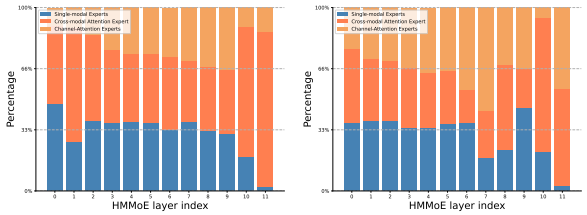
To verify the effectiveness of our newly introduced heterogeneous expert module, we conducted a comparative analysis of its performance. The results demonstrate that the performance of models with mixed heterogeneous experts surpasses that of models combining multiple single-expert models. This improvement highlights that the enhanced model performance is attributed not only to the increased number of experts but also to the innovative heterogeneous structural design. The integration of single-mode information into other modes via cross-attention and channel attention mechanisms contributes significantly to the mode fusion process while maintaining a low parameter count.

4.6 Experts Distribution Analysis

Our HMMoE model structure is incorporated into the transformer layers, and the choice weight of each group of experts varies at different levels. Experimental analysis shows that the model tends to maintain a balanced selection of experts at the lower-level layers. However, at higher-level layers, where layers have a more significant impact on model performance, the model tends to prioritize multi-modal experts. As illustrated in 3, the model exhibits a preference for experts focused on channel-attention and cross-attention, indicating that integrating more cross-modal auxiliary information helps the model make better classification decisions.

Module			NLVR	VQA	MSVD	MSRVTT
Single	Cross	Channel	Acc	Acc	Acc	Acc
✓	-	-	80.4	74.4	54.1	43.6
✓	✓	-	81.0	75.1	54.2	44.6
✓	✓	✓	**82.2**	**75.2**	**55.6**	**45.2**

Table 5: Ablation results of the expert module in text-visual tasks (Single for single-mode experts, Cross for cross-attention experts, and Channel for channel attention experts).



(a) Visual-audio task (b) Text-visual task

Figure 3: Distribution of expert utilization across different layers for each expert type.

5 Conclusion

In conclusion, we have introduced a novel Heterogeneous Multi-modal Mixture of Experts Adapter (HMMoE) to mitigate the absence of modality interaction during the fine-tuning phase of multi-modal models. Our approach extends the input of each expert from a single modality to multiple modalities, enabling effective cross-modal interactions within each expert. By mapping inputs to a low-rank space for interaction and subsequently back to their original dimensions, our method facilitates efficient gradient adjustments of the frozen pre-trained model parameters. Additionally, we have transitioned from the traditional single-expert structure to a heterogeneous expert framework that integrates various interaction types, including cross-attention experts and channel-attention experts. This more diverse architecture allows our model to better capture and process the intricate relationships within multi-modal data. The experimental results highlight the effectiveness and advantages of our proposed module, showing significant improvements in managing complex multi-modal scenarios.

References

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Umberto Cappellazzo, Daniele Falavigna, and Alessio Brutti. 2024. Efficient fine-tuning of audio spectrogram transformers via soft mixture of adapters. *arXiv preprint arXiv:2402.00828*.

Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. 2024. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1110–1119.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. 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)*, pages 646–650. IEEE.

Peikun Chen, Fan Yu, Yuhao Liang, Hongfei Xue, Xucheng Wan, Naijun Zheng, Huan Zhou, and Lei Xie. 2023a. Ba-moe: Boundary-aware mixture-of-experts adapter for code-switching speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023b. Valor: Vision-audio-language omni-perception pretraining model and dataset. *Preprint*, arXiv:2304.08345.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *CoRR*, abs/1312.4314.

William Fedus, Barret Zoph, and Noam M. Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. 2021. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. *Preprint*, arXiv:2110.04366.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022a. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022b. Branch-train-merge: Embarrassingly parallel training of expert language models. *CoRR*, abs/2208.03306.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. 2023. Towards efficient visual adaption via structural reparameterization. *ArXiv*, abs/2302.08106.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. 2022. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pages 18332–18346. PMLR.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023. Scaling vision-language models with sparse mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11329–11344. Association for Computational Linguistics.

Yiwen Tang, Ray Zhang, Zoey Guo, Xianzheng Ma, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. 2024. Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5171–5179.

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. 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*, pages 436–454. Springer.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. *Preprint*, arXiv:1803.08842.

Jifeng Wang, Kaouther Messaoud, Yuejiang Liu, Juergen Gall, and Alexandre Alahi. 2024. Forecast-peft: Parameter-efficient fine-tuning for pre-trained motion forecasting models. *Preprint*, arXiv:2407.19564.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19175–19186. IEEE.

Yan Xia and Zhou Zhao. 2022. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. 2020. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 279–286.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Dong Wang, Huchuan Lu, and You He. 2024. Boosting continual learning of vision-language models via mixture-of-experts adapters. *arXiv preprint arXiv:2403.11549*.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. 2023. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *Preprint*, arXiv:2308.03303.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Avsbench: A pixel-level audio- visual segmentation benchmark.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022a. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. 2022b. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Limitation Analysis

Although the HMMoE model proposed in this paper allows for fine-tuning multimodal models with low parameters and maintaining good performance, it still has certain limitations. The primary challenge lies in the incorporation of the MoE (Mixture of Experts) architecture into the model. The number of experts needs to be adjusted in accordance with the model's scale and the parameter volume of the specific task to achieve the optimal balance between performance and efficiency. Additionally, exploring method to further fine-tune multi-modal models in the audio-visual domain using fewer parameters is also a direction worth pursuing.

B Heterogeneous Expert Comparative Experiment

The performance comparison between the heterogeneous multi-modal expert structure and the traditional homogeneous expert structure is analyzed experimentally. Under the same parameter scale, the introduction of heterogeneous expert model can effectively improve the model's perception ability on multi-modal information.

NLVR			
Single Expert	Cross Expert	Channel Expert	Acc
6	-	-	81.1
2	2	2	**82.2**
VQA			
Single Expert	Cross Expert	Channel Expert	Acc
4	-	-	74.9
2	1	1	**75.2**
AVE			
Single Expert	Cross Expert	Channel Expert	Acc
3	-	-	79.8
1	1	1	**81.0**

Table 6: Evaluation of heterogeneous experts' efficiency across NLVR, VQA and AVE tasks.

C Routing Top-K Policy Comparison

Selecting the top-2 routing policy can obviously improve the relative performance of the model.

Router method	MSVD	NLVR	AVE	VQA
Top-1	54.7	80.5	79.9	74.8
Top-2	**55.6**	**82.2**	**81.1**	**75.2**

Table 7: Comparison of performance metrics across different local router methods.

D Implementation details

For the **VQA** task, the model is trained with a batch size of 32, a learning rate of 3×10^{-5}, a warmup step ratio of 0.1, and for 10 epochs.The model architecture used is *vlmo_base_patch16*, with an image size of 480 and a drop path rate of 0.15.

For the **NLVR2** task, the model is trained with a batch size of 48, a learning rate of 5×10^{-5}, a warmup step ratio of 0.1, and for 10 epochs. The model architecture used is *vlmo_base_patch16*, and sharded training is not used.

For the **MSRVTT-QA** task, the model is trained with a learning rate of 2×10^{-5}, a training video sample number of 8, a test video sample number of 8, and lasts for 40 epochs.

For the **MSVD-QA** task, the model is trained with a learning rate of 1×10^{-5}, a training video sample number of 8, a test video sample number of 8, and for 5 epochs.

For the **AVE** task, the model is trained with a batch size of 8, a learning rate of 5×10^{-4}, and for 10 epochs.

For the **AVQA** task, the model is trained with a batch size of 8, a learning rate of 1×10^{-4}, and for 10 epochs.

For the **AVVP** task, the model is trained with a batch size of 8, a learning rate of 3×10^{-4}, and for 10 epochs.

For the **S4** setting of the **AVS** task, the model is trained with a batch size of 8, a learning rate of 3×10^{-4}, and for 10 epochs. For the **MS3** setting of the **AVS** task, the model is trained with a batch size of 2, a learning rate of 1.5×10^{-4}, and for 10 epochs.

All of our experiments are trained on one NVIDIA A800 GPU.