Querying as Prompt: Parameter-Efficient Learning for Multimodal Language Model

Tian Liang[1], Jing Huang[1], Ming Kong[1,2], Luyuan Chen[3], Qiang Zhu[1]*

[1] Zhejiang University [2] Hikvision Research Institute
[3] Beijing Information Science and Technology University

[1]{liangtian2022,huangjin9,zjukongming,zhuq}@zju.edu.cn [3]chenly@bistu.edu.cn

Abstract

Recent advancements in language models pre-trained on large-scale corpora have significantly propelled developments in the NLP domain and advanced progress in multimodal tasks. In this paper, we propose a Parameter-Efficient multimodal language model learning strategy, named QaP (Querying as Prompt). Its core innovation is a novel modality-bridging method that allows a set of modality-specific queries to be input as soft prompts into a frozen pre-trained language model. Specifically, we introduce an efficient Text-Conditioned Resampler that is easy to incorporate into the language models, which enables adaptive injection of text-related multimodal information at different levels of the model through query learning. This approach effectively bridges multimodal information to the language models while fully leveraging its token fusion and representation potential. We validated our method across four datasets in three distinct multimodal tasks. The results demonstrate that our QaP multimodal language model achieves state-of-the-art performance in various tasks with training only 4.6% parameters. Code is available at https://github.com/Rainlt/QaP.

1. Introduction

Multimodal Learning (MML) aims to perceive, align, and integrate information from various modalities, such as audio, video, and text, facilitating a more comprehensive understanding of complex scenarios [46]. Recently, large language models (LLMs) trained on extensive textual corpus have shown robust performance in NLP tasks [4, 6, 9, 14, 25, 31, 33, 59], which also influenced the MML domain. Firstly, textual information often contains clear and understandable semantic content, which can be effectively aligned with other modalities. Secondly, many multimodal tasks involve textual outputs, such as Video Captioning and Visual Question Answering (VQA). Furthermore, Transformer-based language models, like BERT [4] and GPT [6], have demonstrated significant reasoning capabilities. As a result, integrating multimodal information into the textual representation space to develop Multimodal Language Models (MLMs) has become a prominent direction in multimodal learning research[51]. Figure 1 illustrates the two mainstream approaches:

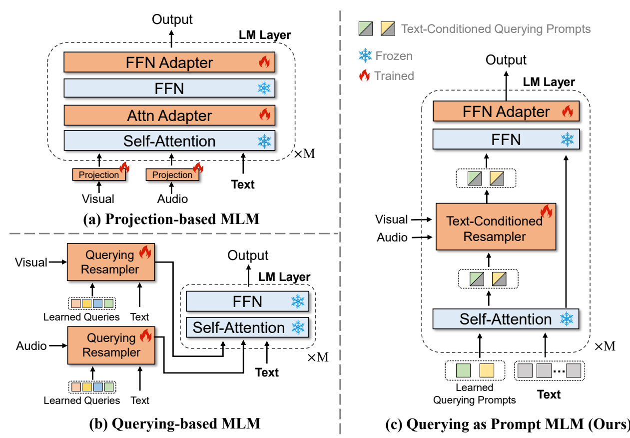

Figure 1. Three kinds of approaches for MLMs: (a) **Projection-based approach**: Aligning modalities through a projection layer, followed by fine-tuning the adaptive layers of Language Model; (b) **Querying-based approach**: Achieving modality alignment through external Querying Resamplers; (c) **Querying as Prompt approach (Ours)**: introducing a set of learnable querying prompts for the MLM model, which can be viewed as both the queries for the embeddings of modalities and the prompts for the text inputs.

- **Projection-based MLM:** As depicted in Figure 1 (a), an intuitive and ideal approach is to project other modalities' embeddings into the textual space to allow LLMs to process them directly. To prevent catastrophic forgetting problems [3], during the fine-tuning of the downstream tasks, the projection-based methods [5, 23, 29, 35, 48]

*Corresponding author.

often follow parameter efficient transfer learning (PETL) [2, 10, 16] , i.e., introduce trainable adapters while keeping the original language model parameters frozen. Although effective multimodal integration, the projection-based methods result in a long input sequence with much text-irrelevant redundant information, creating high computational costs during inferencing.

- **Querying-based MLM:** To keep the information injection to the language models more efficient, as illustrated in Figure 1 (b), Querying-based methods [1, 19, 27, 57] introduce an external resampler (e.g., Q-Former [19]) for each modality to extract text-relevant information, which effectively compresses modality information and facilitates modality bridging. However, the heavy-structured resamplers not only introduce a large number of model parameters but also require additional training.

To relieve the drawbacks of the aforementioned approaches, we propose an innovative parameter-efficient multimodal learning framework, named Querying as Prompt (QaP), which encompasses two parts: 1) **Querying Prompt**: For each modality, we predefine a learnable querying prompt, serving as both a query for extracting modality information and a prompt for textual information interaction. 2) **Text-Conditioned Resampler**: We incorporate a set of lightweight resamplers into different layers of the language model, which aims to adaptively extract text-informative features from various modalities. Besides, to adapt downstream multimodal learning tasks, we add a lightweight adapter layer for domain adaptation. By introducing a small number of learning parameters, QaP achieves highly efficient modality bridging while fully harnessing the advantages of feature fusion and representation potential of the language model.

We conducted validations on three multi-modal downstream tasks across four datasets, including Music-AVQA [17] for Audio-Video Question Answering (AVQA), TVQA [15] and How2QA [20] for Video Question Answering (VideoQA), and CMU-MOSEI [55] for Multimodal Sentiment Analysis (MSA). The experimental results demonstrate that QaP achieves superior accuracy than both existing fully-finetuned and parameter-efficient methods and is comparable with the methods that include external training data, proving the advantages of modality-bridging effectiveness and task adaptiveness of our approach.

The innovations and contributions of this paper can be summarized as follows:

- We propose a Querying as Prompt strategy for the multimodal language model learning that introduces a set of querying prompts, which serves as both queries for modality information extraction and prompts for textual information interaction.
- We propose a parameter-efficient Text-Conditioned Resampler module to extract text-informative features from different modalities and bridge them to the MLMs.
- We conducted experiments on four datasets involving three downstream tasks, surpassing full-parameter fine-tuning and parameter-efficient baselines with 4.6% trainable parameters.

2. Related Work

2.1. Multimodal Language Model

With the rapid advancement of large-scale pre-trained language models (LMs), multimodal language models (MLMs) have become the mainstream solution for multimodal tasks [41, 42, 44, 49, 50, 58], where the primary consideration lies in bridging the gap between other modalities and the textual modality [1, 5, 19, 22, 23, 27, 29, 35, 43, 48, 57].

Some works intuitively project modalities into the space of text modality. For example, [29] uses a linear projection to map visual modality to text space; [23] employs a learnable prompt as the interface between image features and the language model; and [48] combines the linear mapping with adapter layers. These methods include a long embedding sequence of other modalities into the inputs of MLMs, resulting in considerable computational overhead. To relieve this, [5] makes a pooling operation for the embeddings of other modalities to reduce the increase of the inputs; [35] opts to directly use the [CLS] token as the embedding for modalities. However, these approaches is hard to extract text-informative information from modalities.

To reduce the tokens of other modalities, other methods pre-process the modality information with the guidance of textual information via a query-based structured resampler. For instance, [1] introduces a query-based Perceiver Sampler to sample features of visual modalities into a fixed number of tokens. These sampled features are then integrated into the language model with additional heavy attention layers. [19] proposes Q-Former, a text-conditioned querying transformer pre-trained on image-text pairs. Although query-based methods generate text-related and compressed representations, the external resampler modules require additional pretraining [19].

In this paper, we incorporate the query-based resampler into the layer of MLMs to efficiently bridge multimodal information with the language model. With the full utilization of MLMs' token fusion and representation capabilities, the proposed module is lightweight to the maximum extent.

2.2. Parameter Efficient Transfer Learning

The objective of Parameter-Efficient Transfer Learning (PETL) is to adapt pre-trained models to downstream tasks using a small number of adjustable parameters [2, 8, 10, 12, 28, 56]. [2] first proposed to insert trainable lightweight bottleneck modules between transformer layers to achieve parameter-efficient transfer learning. Inspired by text prompting methods, [21] proposed Prefix Tuning, an efficient structure that prepends a small number of tunable prefix vectors to the keys and values of each layer's multi-head attention. These methods introduce additional computational overhead during inference. [10] proposed LoRA, a method that introduces trainable parameters between transformer layers to learn the low-rank factorization of network weights. The simple linear design enables Lora to merge pre-trained and fine-tuned parameters during inference to eliminate additional inference overhead. Furthermore, some methods based on the Ladder Side [37] do not require backpropagation through the pre-trained model, enhancing training efficiency.

However, the above-mentioned methods often focus on domain adaptation for a single modality rather than modality adaptation. Despite recent efforts exploring the application of existing PETL techniques in visual-text tasks [13, 30, 38, 60], audio-visual tasks [24], the primary emphasis has been on alignment during modality encoding, with limited research on efficiently bridging multiple modalities with the language model. In this paper, our proposed method can efficiently bridge multiple modalities into the language model with limited parameters, enabling improved performance in multimodal tasks.

3. Method

3.1. Overview

In this section, we present our proposed Querying as Prompt (QaP) method to bridge the gap between other modalities with the pre-trained Language Model for MML tasks. As illustrated in Figure 2, with a set of querying prompts and a lightweight Text-Conditioned Resampler module, our approach can make full use of the token fusion and representation capabilities of the pre-trained language model to extract text-informative multimodal information, thereby assisting in accomplishing multimodal tasks efficiently. Below, we present our technical approaches in more detail.

3.2. Language Model with QaP

Text token inputs with Querying Prompts. Given the sequence of text tokens $X_t \in \mathbb{R}^{(T_t \times D_t)}$, where T_t is the number of tokens and D_t is the dimension of tokens. We incorporate a set of learnable *Querying Prompts* along with the text embeddings as the initial input of the language model. Specifically, assuming there are k modalities are introduced in addition to the text modality, Querying Prompts can be represented as $\mathbf{q} = [q_1, ..., q_k]$, where each vector corresponds to a specific modality and with the same dimension of the text feature. Thus, for the LM model with L layers, the inputs for each layer $X^{(l)}$ can be expressed as:

$$X^{(l)} = [\mathbf{q}^{(l)}; X_t^{(l)}] \tag{1}$$

where $\mathbf{q}^{(l)}$ represents the Querying Prompt of the l-th layer, and $X_t^{(l)}$ represents the text embedding for the l-th layer.

Each q_i of $\mathbf{q}$ corresponds with the i-th modality and can be treated as a prompt to inject modality-specific information for textual information interaction. In comparison to projection-based methods, our model introduces only a small number (the number of modalities k) of input lengths, alleviating the computational burden associated with directly inputting unsampled multimodal feature sequences.

MLM layer with TCR. Note that in the initial state, querying prompts do not contain the specific content of modal information. We incorporate a query-based *Text-Conditioned resampler* (TCR) module into LM layers to achieve adaptive text-relevant information extraction from modal features.

We first briefly review the operation flow of the standard transformer-based LM layer. For the l-th layer, given the textual input, $X_t^{(l)}$, the LM layer first employs a Multi-Head Self-Attention (MSA) layer for token integration, followed by subsequent processing through a Feedforward Neural Network (FFN) layer:

$$\begin{aligned} X_{ao}^{(l)} &= X_t^{(l)} + MSA(X_t^{(l)}) \\ X_t^{(l+1)} &= X_t^{(l)} + FFN(X_{ao}^{(l)}) \end{aligned} \tag{2}$$

Here the $X_{ao}^{(l)}$ represents the attention output. Note that, for the sake of conciseness, we skip the descriptions of the Layer Norm and Multi-Head mechanism. Furthermore, for completeness, we define the MSA operation below:

$$MSA(X_t) = Softmax((X_t W_q)(X_t W_k)^T)(X_t W_v) \tag{3}$$

where W_q, W_k, and W_v denote learnable mapping matrices.

We add the TCR module behind the MSA layer, for each querying prompt to incorporate multimodal information into their representations, which is then fed into the subsequent FFN layer along with the text feature. Formally, given the concatenated text input and querying prompts $X^{(l)}$ and all the multimodal features $X_i^{(l)}, i = 1, \ldots, k$, the operation before the FFN layer of the MLM layer with TCRs are as follows:

$$\begin{aligned} [\mathbf{q}_{int}^{(l)}, X_{int}^{(l)}] &= MSA(X^{(l)}), \\ \hat{\mathbf{q}}_{int}^{(l)} &= TCR(q_{int,i}^{(l)}, X_i^{(l)}, ...), \quad i = 1, \ldots, k \\ X_{ao}^{(l)} &= X^{(l)} + Concat([\hat{\mathbf{q}}_{int}^{(l)}, X_{int}^{(l)}]) \end{aligned} \tag{4}$$

where TCR_i represents the Text-Conditioned Resampler module for the i-th modality, int indicates intermediate feature. And we will introduce the detailed structure of the module in Section 3.3.

Through the aforementioned operations, our proposed method utilizes the MSA layer for the information propagation between modal-specific querying prompts and the text

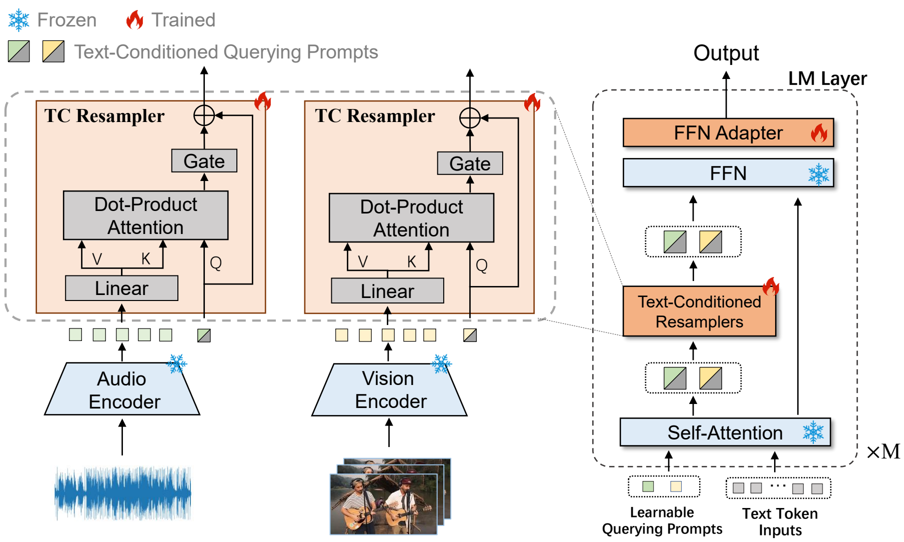

Figure 2. Illustration of our Querying as Prompt method in the pre-trained Language Model. The Learnable Querying Prompts, along with text embedding, are jointly input into the language model. Text information is integrated into the Querying Prompt through the frozen self-attention layer, resulting in a Text-conditioned Querying Prompt. Subsequently, the Text-conditioned Querying Prompt interacts with multimodal embeddings through the Text-conditioned Resampler module, facilitating Modality Adaptation. After obtaining text-relevant multimodal information, it undergoes Domain Adaptation by combining with text embedding through a FFN layer and an Adapter layer.

modality while incorporating text-relevant multimodal information into the querying prompts. The layer-wise information interaction further mitigates information loss caused by the compressed representation.

FFN Adapter. The aforementioned *Querying Prompts* and *Text-Conditioned Resamplers* are primarily for modality adaptation. To achieve domain adaptation when transferring the model to downstream tasks, we introduce a Parallel Adapter [8] for the FFN layer. Specifically, the FFN Adapter includes a learnable down-projection linear layer, a non-linear activation function, a dropout layer, and an up-projection linear layer. Additionally, a learnable Gate is applied to control the size of the adapter. The operations of the FFN layer and the FFN Adapter can be expressed as follows:

$$\begin{aligned} X^{(l+1)} &= X_{ao}^{(l)} + FFN(X_{ao}^{(l)}) + Adapter(X_{ao}^{(l)}) \\ Adapter(x) &= g_a * f((XW_{down})W_{up}) \end{aligned} \tag{5}$$

3.3. Text-Conditioned Resampler

In this part, we provide a detailed technical description of the Text-Conditioned Resampler (TCR) modules. In essence, a TCR is an information extraction module based on gated cross-attention, which extracts text-conditioned information from the corresponding modality through the model-specific querying prompt vector. Next, we first briefly introduce the extraction of modal-specific inputs and then provide a detailed description of the TCR.

Modal-specific Inputs. Initially, for each modality, we generate the embeddings from raw data through pre-trained model-specific feature encoders. We hereby illustrate with Video and Audio, which are the primary modalities introduced in our experiments on multimodal learning.

For the video modality, given the raw video input $V \in \mathbb{R}^{(T_V \times W \times H \times 3)}$, where T_V is the frame number, H and W are the width and height of frames with 3 channels. The video encoder typically encodes each video frame, yielding a visual feature sequence $X_V \in \mathbb{R}^{(T_V \times D_V)}$, denoted as:

$$X_V = Encoder_V(V, \theta_V) \tag{6}$$

where θ_V represents the visual encoder parameters.

For the audio modality, given the raw audio spectrogram input $A \in \mathbb{R}^{(T_A \times C_A)}$, where T_A is the audio span with C_A dimensions. The audio encoder first divides the audio into N segments of length T_C, where $T_A = N \times T_C$, and then encodes each segment to obtain an audio feature sequence $X_A \in \mathbb{R}^{(T_A \times D_A)}$, denoted as:

$$X_A = Encoder_A(A, \theta_A) \tag{7}$$

where θ_A represents the audio encoder parameters.

Considering the findings of [29], it has been demonstrated that the modality embeddings generated by the encoders with enriched textual supervision pre-training are more transferable to the textual space. Thus, in this paper, we chose pre-trained CLIP [32] and CLAP [45] as the Visual and Audio encoders.

Text-Conditioned Resamplers Text-Conditioned Resamplers (TCR) are a set of modules for extracting text-related information from a specific modality. Given the modal-specific embedding as input, a TCR module sequentially achieves three purposes: (i) *Representation Transformation*: aligning the modal embedding with the representation space of the text modality; (ii) *Modality Information Injection*: extracting text-related modal information via query; and (iii) *Modal Intensity Control*, regulating the strength of modality information through a gate.

For a TCR module embedded in the l-th layer, given the i-th feature sequence input X_i, it needs to be transformed into the representation space of the language model. We simply employ a linear transformation to achieve this, denoted as:

$$f_i^l = W_i^l X_i + b_i^l \tag{8}$$

Next, we employ a dot-producted cross attention to extract textual-related information from the transformed modal features through the corresponding querying prompt, where q corresponds to the modality-associated querying prompt vector q_i^l, k and v corresponds to the transformed modal feature f_i^l. The computation process is expressed as:

$$h_i^l = \sum Softmax(q_i^l \cdot {f_i^l}^T) \cdot {f_i^l}^T \tag{9}$$

Finally, a learnable gate is employed to control the input intensity of modality information:

$$\hat{q}_i^l = q_i^l + h_i^l \cdot g_i^l \tag{10}$$

With the querying prompt vector as a carrier, we adaptively inject multimodal information into different layers of the language model. An entire module introduces only a linear layer, a parameter-freed dot-producted cross attention, and a gating unit. The parameter-efficient modules sufficiently leverage the information fusion and reasoning capabilities of the language model and integrate multimodal information into the language model efficiently. In comparison to existing Query-based methods, such as Q-Former [19] and Flamingo [1], which require the introduction of a large number of parameters for text-related information sampling and injection, our approach introduces significantly fewer training parameters, making it easier for training on downstream tasks.

4. Experiments

4.1. Downstream Tasks and Datasets

Our experiments revolve around three multimodal downstream tasks: Audio-Visual Question Answering (AVQA), Video Question Answering (VideoQA), and Multimodal Sentiment Analysis (MSA). Among them, the AVQA task involves the Music-AVQA dataset [17], which encompasses three modalities, making it the primary focus of our research. For the other two tasks, we conducted experiments on the How2QA [20] and TVQA datasets [15] for VideoQA, as well as the CMU-MOSEI dataset [55] for MSA. The following provides a detailed introduction to the mentioned tasks:

Audio-Visual Question Answering: Music-AVQA [17] is a large-scale dataset requiring comprehensive multimodal understanding and spatiotemporal reasoning over audio-visual scenes. The dataset comprises 9,288 videos with an average length of 60 seconds each. The videos encompass 22 musical instruments, resulting in a total duration exceeding 150 hours and 45,867 QA pairs. Following the approach outlined in [17], we split the dataset into training, validation, and testing sets with 32,087, 4,595, and 9,185 QA pairs, respectively. And we evaluate our model based on answer prediction accuracy.

Video Question Answering: In comparison to AVQA, VideoQA often does not require the involvement of the Audio modality in the question-answering process. Consequently, we selected two VideoQA datasets to supplement the validation of our model. We employed the How2QA [20] and TVQA [15] datasets. Specifically, How2QA comprises 28k video clips and 38k questions, while TVQA consists of 22k video clips and 153k questions. Following [48], we partitioned How2QA into 35k/3k for training/validation, and TVQA into 122k/15k/15k for training/validation/testing, respectively. It is noteworthy that due to our inability to access the testing set labels of TVQA, we compare the results on the validation set.

Multimodal Sentiment Analysis: We employed the CMU-MOSEI dataset [55], widely utilized in multimodal sentiment analysis tasks. The CMU-MOSEI dataset consists of 22,856 movie review video clips sourced from YouTube, featuring 1,000 narrators expressing opinions on 250 topics. Each video includes corresponding audio and transcript text. For each clip, there is a sentiment polarity annotation in the range of (-3, +3), indicating the degree of positive or negative emotion. We followed the dataset partitioning strategy of the [55] to delineate the datasets into training, validation, and test sets. Our evaluation metrics, consistent with [55], include mean absolute error (MAE), Pearson correlation (Corr), accuracy (Acc-2), and F1 score.

4.2. Implementation Details

We employed the DeBERTa-V2-XLarge [9] model as our language model, featuring 24 transformer layers and a hidden dimension of D=1536. For visual data, we utilized the CLIP ViT-L/14 [32] model to encode video frames and the CLAP model [45] to encode audio clips into embeddings. Following [48] and [17], We sample 10 frames for each video and 10 clips for audio. Otherwise, we set the model's hyperparameters based on ablation studies. Specifically, the number of querying prompts for AVQA and MSA tasks is set to 1, and for the VideoQA task, it is set to 2. The Text-conditioned Resampler (TCR) was inserted into the first 12

Method	Finetune Encoder	Trainable Params↓	Accuracy ↑
AVSD [34]	✓	N/A	68%
Pano-AVQA [53]	✓	N/A	70%
AVQA [17]	×	10.6 M	71.52%
PSTP-Net [18]	×	**4.3M**	73.52%
Lavish [24]	✓	21.09 M	77.17%
Ours	×	40 M	**78.41%**

Table 1. Comparison with other works after finetuning on Music-AVQA. Our work outperforms other approaches without finetuning the audio encoder and visual encoder.

layers of the language model, with the gates in the first 6 layers initialized to 1 and the gates in the subsequent 6 layers initialized to 0. Additionally, following [8], we set a scale factor of 3 for the gates.

Regarding the training strategy, we conducted 20 epochs of training for all the downstream datasets, utilizing a learning rate of 3e-5. We implemented a linear warm-up for the initial 10% of iterations, succeeded by a linear decay of decreasing to 0 over the subsequent 90%, in accordance with the approach outlined in [48]. More information about implementation details and evaluation metrics will be provided in the supplementary.

4.3. Main Results

In this section, we present the comparisons between our approach and other methods on three downstream tasks. Since some language-model-based modality bridging methods [5, 23, 29, 35] are primarily applied to image-text tasks, we reproduced and compared these methods in the AVQA task, which includes three modalities simultaneously.

4.3.1 Audio-Video Question Answering

Comparison with the State-of-the-Art:
We fine-tuned our method on the Music-AVQA dataset [17] and compared it with existing works. As shown in Table 1, it is evident that our approach surpasses previous methods while introducing only a small number of trainable parameters. This demonstrates the effectiveness of our approach in bridging multi-modal information into the language model. Notably, Lavish, also a Parameter Efficient Transfer Learning method, serves as a primary point of comparison. Unlike our work, Lavish incorporates adapters into the encoder for fine-tuning, significantly increasing the training time due to the online extraction of video and audio features. Our method not only achieves a 1.2% improvement in accuracy compared to Lavish but also exhibits a notable increase in training speed by reducing the time spent on redundant encoding of multi-modal features.

Method	Trainable Params↓	Accuracy ↑
Full Parameters	890M	77.73%
Limber †[29]	**2M**	72.79%
PromptFuse †[29]	7M	75.72%
MAGMA†[5]	30M	77.44%
eP-ALM†[35]	45M	75.09%
Ours	40M	78.41%
Ours$_{hf}$	45M	**78.69%**

Table 2. Comparison with other efficient bridging methods for multimodal language model. Our method surpasses other approaches, including the full-parameters fine-tuned method. Ours$_{hf}$ means we used the same hierarchical feature as eP-ALM. †: the reimplemented version.

Comparison with Other Modality Adaption Methods:
For some other parameter efficient modality bridging methods [5, 29, 29, 35], we reimplemented them on the AVQA dataset [17]. For a fair comparison, we employed the same Visual encoder (CLIP [32]) and Audio encoder (CLAP [45]), as well as the language model (DeBERTa-V2 [9]) for all method. "†" represents the reimplemented version:

eP-ALM†: eP-ALM[35] utilizes the [CLS] token of the video encoder as the video embedding, while we use frame-wise encoding with CLIP [32] and perform average pooling on all frame embeddings to obtain the video embedding. Additionally, we separately extract the last 6 layers' features of CLIP as hierarchical inputs. Moreover, eP-ALM benefits from the use of adapters [35]. Therefore, we also apply the Adapter method to eP-ALM†in both the Attention and FFN layers.

Limber†: Limber[29] bridges multi-modal features with the language model using only one trainable linear layer. In our replication, we set up a linear layer for both Audio and Visual to achieve dimension alignment and modality bridging.

PromptFuse†: which is equivalent to PromptFuse [23] and use Prompt Tuing (N=10). Following [35], we applied a linear layer before inputting audio/visual into the model.

MAGMA†: which is equivalent to MAGMA [5] and using an adapter. Following [35], we freeze the encoder for better performance.

As shown in Table 2, Compared with other modality bridging methods, we observe that our approach achieves better results with a comparable number of trainable parameters. We achieve a 1% improvement over the best-performing MAGMA and surpass the performance of fully parameterized training. It is noteworthy that when employing the same hierarchical visual features as eP-ALM, without pooling but using the text-conditioned resampler to bridge into the language model, we achieve an accuracy of 78.69%, demonstrating the excellent performance of our

Method	Extra Data	Trainable Params↓	How2QA↑	TVQA↑
SiaSamRea [52]	✓	-	84.1%	-
Just Ask [47]	✓	157M	85.3%	-
Frozenbilm$_{full}$	✓	890M	87.5%	79.1%
Frozenbilm [48]	✓	**30M**	86.7%	**82.4%**
Ours	×	40M	**94.5%**	80.19%

Table 3. Comparison with other works on VideoQA task. The primary comparison is with Frozenbilm since it utilizes the same video features and language model as our approach. Frozenbilm$_{full}$ represents full-parameters finetuned Frozenbilm.

text-conditioned resampling method.

4.3.2 Video Question Answering

We fine-tuned the VideoQA task and compared our results with other methods. In this case, we utilized the same language model and visual encoder as [Frozenbilm]. As shown in Table 3, on the How2QA dataset, we achieved state-of-the-art performance, even surpassing the pre-trained Frozenbilm with additional multi-modal data. On the larger TVQA dataset, we still achieved competitive performance compared to the pre-trained Frozenbilm.

Method	MAE↓	Corr↑	ACC-2↑	F1↑
LMF [26]	0.623	0.700	-/82.0	-/82.1
TFN [54]	0.593	0.677	-/82.5	-/82.1
MFM [39]	0.568	0.703	-/84.4	
ICCN [36]	0.565	0.704	-/84.2	-/84.2
MulT [40]	0.580	0.713	-/82.5	-/82.3
Self-MM [52]	0.530	0.765	82.81/85.17	82.67/83.97
MMIM [7]	0.526	0.772	82.24/85.97	82.66/85.94
UniMSE [11]	**0.523**	0.773	85.86/87.50	85.79/87.46
Ours	0.529	**0.825**	**86.95/88.03**	**90.87/90.90**

Table 4. Results on CMU-MOSEI. For Acc-2 and F1, we have two sets of results: non-negative/negative (left) and positive/negative (right). The best results are marked in bold.

4.3.3 Multimodal Sentiment Analysis

We validated our method on the MSA task with other approaches, most of which are trained with traditional paradigms. UniMSE [11] proposed to directly concat multimodal features with text embedding in language model with linear fusion. UniMSE was trained jointly on four sentiment analysis datasets. As shown in Table 4, in our case, using only one dataset, our MAE metric closely aligns with UniMSE, while Corr exceeds it by 1.4%, ACC-2 non-negative surpasses it by 1.1%, ACC-2 negative exceeds it by 0.5%, F1 non-negative surpasses it by 5.1%, and F1 negative exceeds it by 2.5%. These results convincingly demonstrate the superior performance of our method in multi-modal sentiment analysis tasks.

Used Linear	Params ↓	Accuracy ↑
Normal Attention	245M	**78.61%**
Linear dim+Proj. Query	75M	78.01%
Linear dim+Proj. Key	75M	76.14%
Linear dim+Proj. Value	75M	77.53%
Proj. KV	**40M**	78.41%

Table 5. Ablation on the linear projection used for Text-Conditioned Resampler. Linear dim: the linear layer for dimension alignment. Proj. KV: Key and value shared linear layer.

Insert Layer	1-12	12-24	1-24	1-24*	1-12*
Accuracy	**78.41%**	76.64%	77.56%	77.98%	76.82%

Table 6. Ablation on insert layer of Text-Conditioned Resampler. '*' indicates that the insertion is performed only in the layers with even indices.

4.4. Ablation Study

Text-Conditioned Resampler's Linear setting: To align the parameter count of the Text-conditioned Resampler with normal parameter efficient method [2, 8], we experimented with the linear layer in the Attention layer of the Text-conditioned Resampler (TCR). Since it is necessary to align the dimensions of audio and visual features with the language model, at least one linear layer is required for dimension alignment. Additionally, another linear layer is used to map Query, Key, and Value separately. Through experimentation in Table 5, we found that the mapping for Query showed the most significant improvement, indicating that the TCR requires a linear layer for modality space mapping when interacting with querying prompts and multi-modal information. Therefore, we use a shared linear layer for Key and Value, achieving both dimension alignment and modality space mapping. In the end, with a trainable parameter count of only 40M, we achieved comparable performance with the normal attention method.

Gate	Accuracy↑
w/o Gate	78.07%
1-0 Gate	**78.41%**
All_1 Gate	78.32%
All_0 Gate	76.93%
Tanh Gate	76.66%

Table 7. Ablation on gate initialization. 1-0 gate initialization signifies that half of the layers' gates are initialized to 1, while the other half is initialized to 0. Tanh gate indicates the application of the tanh function to the gate.

Insert Layer: We conducted experiments on the number of layers where the Text-conditioned Resampler is inserted. As shown in Table 6, the best performance was achieved when inserted in layers 1-12, indicating that the model needs to receive and process multi-modal information early in the process. Moreover, even when inserting features in layers 12-24 (similar to the reimplemented eP-ALM[35]), our accuracy still improved by 1.5% compared to eP-ALM, further validating the effectiveness of our approach.

Gate initialization: We conducted an ablation experiment on the initialization settings of the gates. As shown in Table 7, in this experiment, the "1-0 Gate" setting initializes the gates to 1 in the first 6 layers and 0 in the subsequent 6 layers, showing a slight improvement compared to the "All_1" setting where all layers are initialized to 1. However, initializing all gates to 0 or using the $Tanh$ function on the gates during "All_0" initialization significantly reduces model performance. This highlights the model's need for joint multi-modal information in the early layers.

Number of Querying Token: We conducted experiments on the number of querying tokens associated with a specific modality. As shown in Figure 3, good performance was achieved when the querying token quantity was set to 1, 2, or 5, with the best performance observed when the quantity was equal to 1, and sharing querying prompts between two modalities results in effectiveness reduction. Besides, directly concatenating the multimodal features and text embeddings for the input without Querying as Prompt will also degrade the performance by 1%. This indicates that our approach can significantly compress multi-modal information, effectively reducing the computational load on the language model. A detailed comparison of computational effectiveness will be provided in the supplementary material.

What is Querying Prompt focusing on? As shown in Figure 4, We conducted a visual analysis of the attention weights of the Text-conditioned Resampler (TCR) on the AVQA dataset [17]. For the first two examples, it is evident that the Querying as Prompt method effectively suppresses text-irrelevant background frames, demonstrating a higher quality of sampling. In the third example, since there is no "banjo" object in the video frames corresponding to the question, the background frame of the last frame attains the highest attention weight. This further substantiates that our Querying Prompt approach can sample multimodal information corresponding to the understanding of the text.

4.5. Limitations

Our approach uses a full token-fusion self-attention mechanism to transfer text information to the Querying Prompt. However, due to limited resources and time, we only validate our method on a bidirectional language model but without the exploration of the decoder-only autoregressive language model yet. Moreover, our method is also expected to extend to the LLMs with larger scales and other modalities such as point clouds and depth maps, although only low additional overhead is incurred when increasing the number of modalities. We believe our proposed approach is also applicable to the aforementioned fields and will keep working on them in the future.

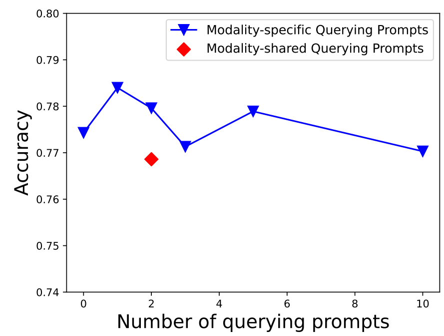

Figure 3. Comparison of different numbers of querying prompts. When the value is 0, it means directly concatenating the multi-modal features and inputting them in the beginning.

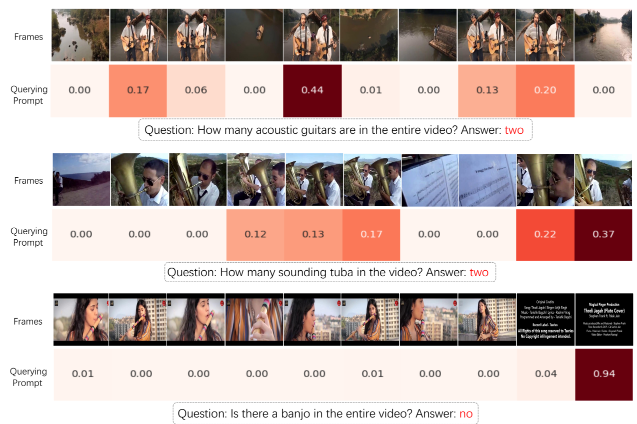

Figure 4. Visualization of the attention weights between the Querying Prompt and frames.

5. Conclusion

We present "Querying as Prompt", a parameter-efficient multimodal learning framework that bridges the modality gaps in multimodal language models. Our framework uses two novel components: *Querying Prompts* and *Text-conditioned Resamplers* to enable the pretrained language model to absorb multimodal information with limited training parameters. We evaluate our approach on four multi-modal datasets and it outperforms existing methods with similar or fewer parameters without extra data for training.

Acknowledgements. This work is supported by the National Science and Technology Major Project (2022ZD0115904).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2, 5

[2] Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*, 2019. 2, 7

[3] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[5] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021. 1, 2, 6

[6] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694, 2020. 1

[7] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021. 7

[8] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 2, 4, 6, 7

[9] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 1, 5, 6

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

[11] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*, 2022. 7

[12] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 2

[13] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023. 3

[14] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 1

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2, 5

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2

[17] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 2, 5, 6, 8

[18] 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*, pages 7808–7816, 2023. 6

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 5

[20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2, 5

[21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[22] Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. Lmeye: An interactive perception network for large language models. *arXiv preprint arXiv:2305.03701*, 2023. 2

[23] Sheng Liang, Mengjie Zhao, and Hinrich Schütze. Modular and parameter-efficient multimodal fusion with prompting. *arXiv preprint arXiv:2203.08055*, 2022. 1, 2, 6

[24] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 3, 6

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[26] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018. 7

[27] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023. 2

[28] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task

fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021. 2

[29] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022. 1, 2, 4, 6

[30] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 3

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5, 6

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1

[34] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019. 6

[35] Mustafa Shukor, Corentin Dancette, and Matthieu Cord. ep-alm: Efficient perceptual augmentation of language models. *arXiv preprint arXiv:2303.11403*, 2023. 1, 2, 6, 8

[36] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8992–8999, 2020. 7

[37] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022. 3

[38] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 3

[39] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018. 7

[40] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, page 6558. NIH Public Access, 2019. 7

[41] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[42] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023. 2

[43] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[44] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[45] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4, 5, 6

[46] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[47] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021. 7

[48] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. 1, 2, 5, 6, 7

[49] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3081–3089, 2022. 2

[50] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

[51] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1

[52] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. *Advances in Neural Information Processing Systems*, 34:26462–26474, 2021. 7

[53] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021. 6

[54] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion net-

work for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 7

[55] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 2, 5

[56] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 2

[57] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[58] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[59] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3