Efficient Transfer From Image-Based Large Multimodal Models to Video Tasks

Shidong Cao, Zhonghan Zhao, Shengyu Hao, *Graduate Student Member, IEEE*, Wenhao Chai, Jenq-Neng Hwang, *Life Fellow, IEEE*, Hongwei Wang, *Member, IEEE*, and Gaoang Wang, *Member, IEEE*

***Abstract*—Extending image-based Large Multimodal Models (LMMs) to video-based LMMs always requires temporal modeling in the pre-training. However, training the temporal modules gradually erases the knowledge of visual features learned from various image-text-based scenarios, leading to degradation in some downstream tasks. To address this issue, in this paper, we introduce a novel, efficient transfer approach termed MTransLLAMA, which employs transfer learning from pre-trained image LMMs for fine-grained video tasks with only small-scale training sets. Our method enables fewer trainable parameters and achieves faster adaptation and higher accuracy than pre-training video-based LMM models. Specifically, our method adopts early fusion between textual and visual features to capture fine-grained information, reuses spatial attention weights in temporal attentions for cyclical spatial-temporal reasoning, and introduces dynamic attention routing to capture both global and local information in spatial-temporal attentions. Experiments demonstrate that across multiple datasets and tasks, without relying on video pre-training, our model achieves state-of-the-art performance, enabling lightweight and efficient transfer from image-based LMMs to fine-grained video tasks.**

***Index Terms*—Video understanding, large multimodal model, transfer learning.**

I. Introduction

BENEFITING from vast pre-trained corpora and parameters, large language models (LLMs) [1], [2], [3] have demonstrated remarkable capabilities in general knowledge and linguistic expertise, achieving impressive results in multiple downstream tasks in natural language processing (NLP) field. With their success in NLP, the multi-modal community has seen many efforts utilizing pre-trained LLMs with additional visual encoding modules to accomplish multi-modal understanding. Some of these efforts have achieved fantastic performance in embodied agent [4], image [5], [6] and video [7] question answering tasks. VideoLLaMA [7] exhibits the capability to comprehend videos by integrating visual information. Existing LLM-based video understanding systems [7], [8], [9], [10] transform video information into a natural language question-answering-tokens format by using pre-trained large language models, image encoders, and newly added temporal modeling modules. In terms of temporal modeling, these methods typically adopt the training paradigms of video pre-training and video finetuning, as shown in Fig. 1. In most cases, the goal of video pre-training with Video Caption tasks is to train video-based LMM into a large-scale general knowledge model, providing universal video understanding knowledge.

However, the video pre-training and finetuning scheme does not work well in some situations. Training temporal modules

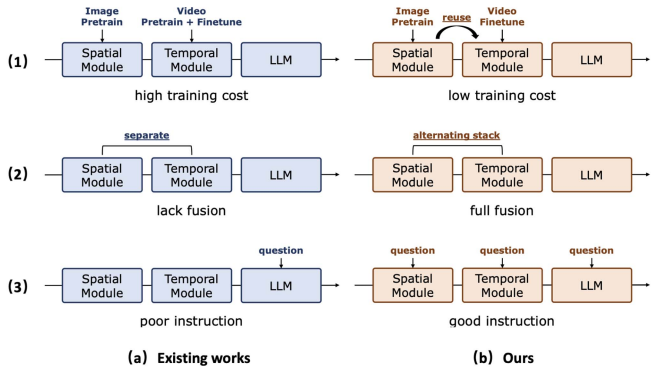

Fig. 1. Different temporal modelings for video LMMs. **(1)** Common models adopt a three-stage approach involving image pre-training, video pre-training, and video finetuning. Training the temporal module during the video pre-training stage requires significant resources and may result in a gap with the finetuning task. Our approach eliminates the need for video pre-training by sharing parameters from spatial modules to temporal modules. **(2)** Common models separate spatial and temporal feature extraction modules in the video feature encoding, leading to high computational complexity and a lack of integration between temporal and spatial features at different levels.
We adopt cyclic spatial-temporal feature interaction yields better performance. **(3)** Due to the absence of video-text pairs in many cases, common models do not perform multi-modal fusion during visual feature extraction. We employ the early multi-modal fusion in temporal modeling to achieve a detailed guide from text instruction.

Received 25 August 2024; revised 14 December 2024; accepted 11 January 2025. Date of publication 3 April 2025; date of current version 28 May 2025. This work was supported by the National Key R&D Program of China under Grant 2022ZD0162000. The guest editor coordinating the review of this article and approving it for publication was Prof. Mengshi Qi. *(Corresponding authors: Hongwei Wang; Gaoang Wang.)*

Shidong Cao, Zhonghan Zhao, and Shengyu Hao are with the Zhejiang University-University of lllinois Urbana Champaign Institute, Zhejiang University, Haining 314400, China (e-mail: 22271126@zju.edu.cn; zhaozhonghan@zju.edu.cn; shengyuhao@zju.edu.cn).

Wenhao Chai and Jenq-Neng Hwang are with the University of Washington, Seattle, WA 98195 USA (e-mail: wenhaochai.19@intl.zju.edu.cn; hwang@uw.edu).

Hongwei Wang is with the Zhejiang University-University of Illinois Urbana Champaign Institute, Zhejiang University, Haining 314400, China, and also with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China (e-mail: hongweiwang@zju.edu.cn).

Gaoang Wang is with the Zhejiang University-University of lllinois Urbana-Champaign Institute, College of Computer Science and Technology, Zhejiang University, Haining 314400, China, also with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: gaoangwang@intl.zju.edu.cn).

Digital Object Identifier 10.1109/TMM.2025.3557692

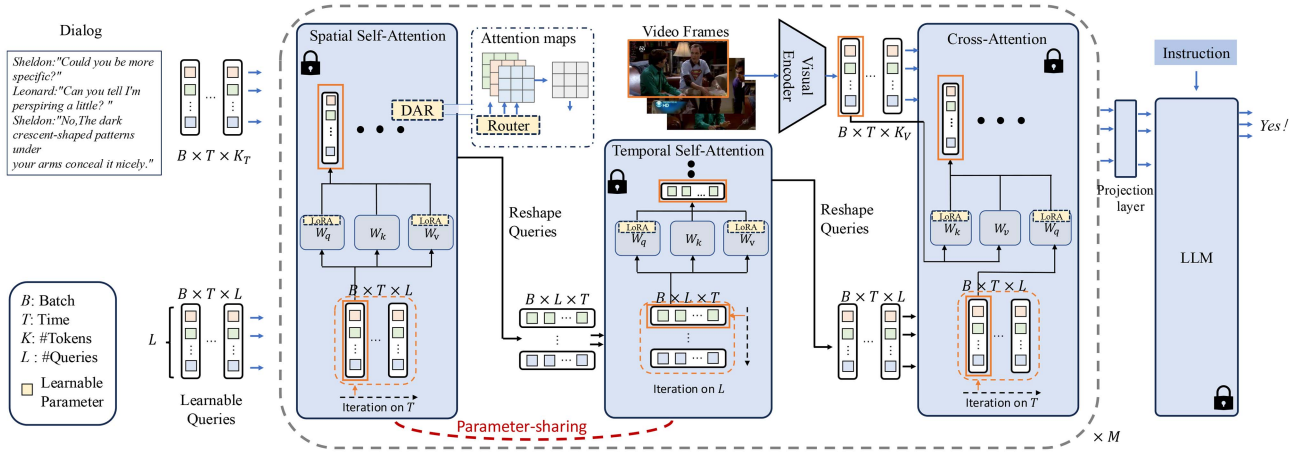

Fig. 2. Overview of MTransLLAMA. First, we perform spatial feature extraction using K_Q pre-defined queries. After applying Spatial Self-Attention, we swap the channels of the queries and proceed to Temporal Self-Attention, where temporal information interaction occurs between tokens with similar semantics across different frames. Following Temporal Self-Attention, we further transform the query channels and extract image features from the input T frames of the video. Finally, we aggregate the tokens from the last layer of the Q-former, concatenate them with instructions, and input them into the frozen LLM to generate the final natural language outputs. We introduce LoRA separately in different attention modules for fine-tuning, while keeping all other parameters frozen. Additionally, we incorporate the DAR module for dynamic routing of attention scope within the attention modules.

during video pre-training tends to gradually erase the knowledge of visual features acquired from diverse image-text-based scenarios. This leads to reduced transfer performance when there is a large domain difference between downstream tasks and the video pre-training dataset. When the downstream tasks, like the video sarcasm detection, significantly differ from the pre-trained video captioning tasks, the extracted general visual features during the video pre-training phase lack fine-grained information to guide the reasoning for downstream tasks, leading to poor performance. In addition, the transfer performance decreases when the video scenes of the downstream tasks greatly differ from those in the pre-training phase. Due to the amount of video-text pre-training data volume being much smaller than the image-text pre-training data volume, the ability of video-based models to generalize across different video scenes degrades. Furthermore, downstream task datasets are usually small-scale, making training video-based LMMs unaffordable.

To address the issues above, in this paper, we introduce MTransLLAMA, a new efficient temporal modeling framework that transfers pre-trained image-based LMMs [6] to video understanding tasks with only small-scale datasets, as shown in Fig. 2. Our method enables fewer trainable parameters and achieves faster adaptation and higher accuracy than pre-training video-based LMMs. We opt to use a pre-trained image-based LMM to achieve better scene generalization and adopt early multi-modal fusion to capture fine-grained visual information in videos based on guided text, thereby facilitating task transfer. Specifically, we employ a pre-trained image-text Q-former for spatial modeling and introduce a low-rank adaptation (LoRA) [11] in the attention layer. For temporal modeling, we directly transfer most of the Q-former's attention layer parameters and perform attention operations across different dimensions, followed by fine-tuning with Lora. To better capture the global and local information of the data, we also employ a dynamic routing-based transfer technique [12] to achieve dynamic shifting of sample attention.

Our main contributions can be summarized as follows:

- We introduce a new temporal modeling framework of video-based LMMs, primarily focused on efficiently transferring to small-scale, fine-grained, domain-specific downstream datasets. By efficiently transferring a pre-trained image LMM into a video-based LMM model, our approach eliminates the need for a video pre-training phase and achieves better generalization.
- We propose a novel video multi-modal fusion structure that integrates spatial modeling, temporal modeling, and text-visual fusion within a single module, lifting the efficiency of spatial-temporal information fusion and multi-modal integration.
- We introduce a new transfer method that dynamically routes the attention weights in the Transformer across samples. Compared to traditional transfer techniques, this method is more effective in handling data with varying attention lengths.
- We demonstrate that our approach excels in training and inference costs, tunable parameters, computational complexity, and data requirements. It achieves state-of-the-art performance on various tasks such as Ego-centric QA and intent detection in multiple small-scale specific video scene datasets.

II. Related Work

A. Conversation and Speech Video Understanding

The task of understanding dialogue and speeches in videos has always been a widely explored issue [13], [14], [15], [16], [17]. Most of the datasets [18], [19] and methods in this field primarily focus on emotion analysis for specific tasks, identifying the emotions, humor, or sarcasm of the speaker from dialogue, speech videos, and subtitle texts. It has been indicated by Chauhan et al. [20] that there is a close connection between

sentiment analysis [21], [22], sarcasm detection [23], and humor detection [24].

Many methods [20], [25], [26] utilized pre-trained Convolutional Neural Network (CNN) [27] and Bidirectional Encoder Representations from Transformers (Bert) [28] to extract textual and video features, followed by the use of a Multi-Modal Transformer [29] for feature fusion. However, these methods required training attention models from scratch, and it has been shown that attention mechanisms perform poorly with limited data. Additionally, these models are hindered by a lack of domain-specific and general knowledge, failing to fully understand advanced linguistic expertise and visual information. With the development of LLMs achieving significant success in the NLP field, InstructERC [30] proposed a method using LLMs to address emotion recognition tasks. By leveraging the knowledge of large models, it achieved commendable results, but it only considered the textual modality and did not utilize the video modality. Furthermore, it required finetuning the LLM, which incurs high training costs.

B. Video Large Language Models

Large language models such as ChatGPT [2] and LLAMA [1] have achieved tremendous success in the NLP field. Many works (like LLAVA [5] and GPT-4 [2]) attempt to integrate other modalities such as vision and audio to enhance understanding [31], [32]. LLaVA uses adapters [33] to map image information tokens directly into the natural language representation space. BLIP-2 [6], leveraging a frozen LLM and Image Encoder, introduces a feature extraction module Q-former for multi-modal fusion, based on text-image contrastive learning. This significantly enhances the capability of multi-modal feature representation. Building on the success of image and LLM integration, many works [7], [8], [10] have started focusing on applying LLMs to video understanding, especially in video question answering (VQA) tasks (like VideoLLAMA [7]). VideoLLAMA introduced a temporal fusion Q-former for modeling temporal information. VideoChat [10] uses a Video Foundation Model [34] to obtain video representations and align them with LLMs. However, all these models underwent pre-training on extensive video datasets, making the training costs unaffordable for most researchers and practitioners. Our proposed MTransLLAMA obviates the need for video pre-training and costly temporal modules, achieving a transition from an image-based LMM to a video-based LMM. Moreover, these models' video pre-training datasets are primarily VQA datasets, which are not efficient for other downstream video understanding tasks like emotion analysis with a large domain gap.

C. Temporal Modeling

With the advancement of AI, more attention is shifting from image to video understanding, including tasks like video question answering [35], [36], [37] and video action recognition [38], [39], [40]. Unlike image models, due to the temporal nature of videos, models require temporal modules for video understanding. With the advent of the Transformer, more work has utilized attention modules as the backbone for video understanding. With the rise of Image pre-trained models, ViT, CLIP [41], and their variants (like Deter [42] and DINO [43]) have been introduced, achieving state-of-the-art performance in image classification [44], [45], [46], object detection [47], [48], [49], and segmentation [50]. The visual patterns they learn from large-scale image-text pre-training data are of great value. However, due to the quadratic complexity growth of Transformers, large temporal convolution modules have introduced significant training costs for video understanding.

To overcome these limitations, a strategy known as parameter-efficient transfer [11], [33] has become increasingly popular in NLP, aiming to fine-tune only a few parameters while keeping large pre-trained models frozen to achieve strong performance. As large ViTs develop, these techniques are being introduced into Computer Vision. However, these works [5], [51] either focus on tuning pre-trained image models for image task finetuning or adjusting pre-trained video models for downstream video tasks. Consequently, many works have transferred image-based pre-trained models to the video domain. Some methods [52] focus on prompt or sampling modeling, while other methods [53] design temporal modules as intermediate structures, as illustrated in the middle. Yang et al. (AIM) [54] repurpose CLIP's attention modules through adapters for efficient video transfer. However, these methods do not explore multiple modalities, and the temporal attention of image patches lacks interpretability. We propose Multi-Modal Query Fine-tuning, achieving modeling with a multi-modal Q-former without temporal modules.

III. Methodology

In this section, we provide a detailed introduction to our proposed MTransLLAMA framework, which efficiently transfers the image-based LMM to handle video understanding tasks. It preserves the extensive pre-training knowledge of the image-text model, resulting in improved transfer generalization. MTransLLAMA achieves early multi-modal fusion and temporal-spatial cyclic fusion by reusing the image-text model. To endow the image-text model with the ability to understand temporal information and reduce video pre-training costs, MTransLLAMA proposes a channel swapping method that reuses multi-modal attention module parameters in the image-text model (detailed in Section III-B and Section III-C). To better capture local and global semantics of data, MTransLLAMA introduces a spatial-temporal domain dynamic routing transfer for multiple modalities (discussed in Section III-D).

A. Overview of MTransLLAMA

Given the context $\mathbf{C}$ (including dialogues, instructions, questions, etc.) and the corresponding video $\mathbf{v} \in \mathbb{R}^{T\times 3\times H\times W}$ comprising T frames sampled from the video sequence, each of size $H \times W$, our proposed MTransLLAMA aims to transfer an image-based LMM to generate answers $\mathbf{C}_o$ in natural language form with the combination of video and text. The proposed MTransLLAMA has three key components, including a multi-modal spatial-temporal early fusion module, a multi-modal query temporal reusing module, and a dynamic attention

routing module. The goal of each module is briefly described as follows.

1) Multi-Modal Spatial-Temporal Early Fusion: Traditional video-based LMMs usually adopt the following mechanism for question answering (QA), i.e.,

$$\mathbf{C}_o = \text{LLM}(\mathbf{E}_t, f_{\boldsymbol{\theta}}(\mathbf{E}_v, \mathbf{q})), \tag{1}$$

where $\mathbf{C}_o$ is the answer generated by LLM; $f_{\boldsymbol{\theta}}$ is the spatial-temporal module with learnable parameters $\boldsymbol{\theta}$; $\mathbf{E}_t$, $\mathbf{E}_v$ and $\mathbf{q}$ are text embeddings, visual embeddings and learnable queries, respectively. Unlike the conventional methods, our proposed fusion mechanism extracts the spatial-temporal information of the dialogue and instruction from early layers and interacts with visual features, which can be formulated as follows,

$$\mathbf{C}_o = \text{LLM}(\mathbf{E}_t, f_{\boldsymbol{\theta}}(\mathbf{E}_t, \mathbf{E}_v, \mathbf{q})), \tag{2}$$

Besides, as shown in Fig. 1, traditional video-based LMMs place the temporal model modules entirely after the spatial model modules. However, the output of the spatial models consists of high-level visual features, which lack the interaction of low-level spatial-temporal information. Our method iteratively cycles through spatial and temporal modules, accomplishing the spatial-temporal fusion of multi-modal information at various granularities.

2) Multi-Modal Query Temporal Reusing: For LLM-based video understanding methods [7], [10] that deal with temporal information, a separately trained temporal modeling module is typically required. However, pre-training temporal modules consumes vast data and computational resources, making it unaffordable to learn f_θ in downstream video understanding tasks. Additionally, the video pre-training phase might compromise the rich image-text information, leading to a reduction in transferability. To address the above challenge, our innovative strategy involves reusing the pre-trained self-attention layers in multi-modal models for temporal modeling. Interestingly, we find the pre-trained spatial self-attention weights can work well for temporal modeling with efficient low-rank adaptation (LoRA). In addition, we adopt a channel swapping strategy to further reduce the cost of temporal modeling.

3) Dynamic Attention Routing: Existing research [12], [55] shows that in certain vision and language tasks, multi-modal reasoning often requires visual attention from different receptive fields. To enable the model to understand both the high-level semantics and local relations, we propose a dynamic attention routing (DAR) strategy in our designed multi-modal spatial-temporal model f_θ. The router can automatically select cooperative attention masks. We not only perform dynamic routing on spatial attentions of the multi-modal data but also on temporal attentions.

4) Training Loss: During training, we aim for the LLM's output answer $\mathbf{C}_o$ to match our ground truth answer $\mathbf{C}_a$. We use the cross-entropy loss between the embedding token G of the ground truth answer $\mathbf{C}_a$ and the probability distribution $\mathbf{p}$ of the output answer $\mathbf{C}_o$, following the same training process as all LLM workflows. The loss is shown as follows.

$$\mathcal{L} = -\sum_i \mathbf{G}_i \log(\mathbf{p}_i), \tag{3}$$

where

$$\mathbf{G} = \text{Tokenizer}(\mathbf{C}_a), \tag{4}$$

and Tokenizer maps natural language to the output tokens of the LLM.

B. Multi-Modal Spatial-Temporal Early Fusion

Given the context $\mathbf{C}$ and the corresponding video $\mathbf{v}$, we first extract the text and visual embeddings like BLIP-2. Text embeddings $\mathbf{E}_t \in \mathbb{R}^{K_T \times D}$ are derived from $\mathbf{C}$ using a Bert Tokenizer, where D represents the embedding size, and K_T denotes the sequence length after padding. The visual embedding $\mathbf{E}_v \in \mathbb{R}^{T \times K_V \times D}$ are obtained frame by frame using the Vision Transformer (ViT) as described by:

$$\mathbf{E}_v = \{\mathbf{x}_i = \text{ViT}(\mathbf{v}_i) \mid \forall i = 1, \ldots, T\} \tag{5}$$

where K_V is the number of tokens after encoding for each frame image.

Then, these video representations $\mathbf{E}_v$ are fed into the Q-former along with the queries $\mathbf{q} \in \mathbb{R}^{L \times D}$, and text embedding $\mathbf{E}_t$ for attention operations, allowing the multi-modal information to be represented by queries $\mathbf{q} \in \mathbb{R}^{T \times L \times D}$, where L is the number of queries for each image.

At the first MTransLLAMA layer, we update the original multi-modal query $\mathbf{q}_0$ with spatial and temporal position embeddings as:

$$\mathbf{q}^0_{i,l} = \mathbf{q}_{i,l} + \mathbf{e}^t_i + \mathbf{e}^s_l, \tag{6}$$

where $\mathbf{q}_{i,l}$ is the original query with sequence index $l \in (0, L)$ and the frame index i; and $\mathbf{e}^t$ and $\mathbf{e}^s$ are the learnable temporal and spatial positional embedding, respectively.

During the process of extracting video features, each layer of our feature extractor is composed of spatial self-attention, temporal self-attention, and cross-attention. By interacting text and image attention, we achieve the early fusion of text and visual embedding.

Due to the repetitive structure of the Q-former, we implement spatial-temporal recurrent feature extraction, which facilitates the interaction of spatial-temporal features at various granularities. This approach has been shown to be beneficial in previous research [36].

C. Multi-Modal Query Temporal Reusing

The reuse of pre-trained self-attention weights in the temporal modeling includes the channel swapping strategy and the efficient low-rank adaptation (LoRA), as shown in Fig. 3, which are demonstrated in this subsection in detail.

1) Channel Swapping: We employ an image-text Q-former for video representation and a frozen projection layer to align with LLM Tokens. For a batch B of data, we initially flatten B and T channels in queries, i.e., from $B \times T \times L \times D$ to $BT \times L \times D$ within each attention module of Q-former.

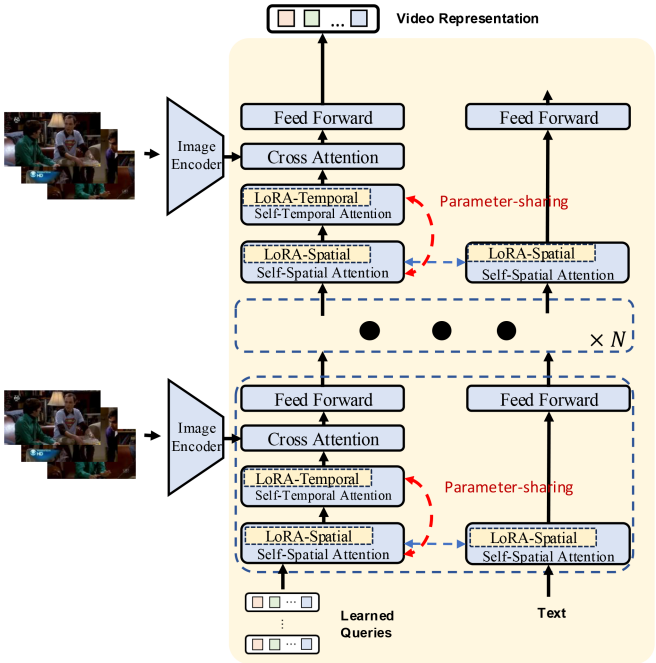

Fig. 3. Overview of Parameter Reuse: The temporal module is implemented by reusing the attention block parameters from the original Q-former, with small-parameter LoRA fine-tuning applied.

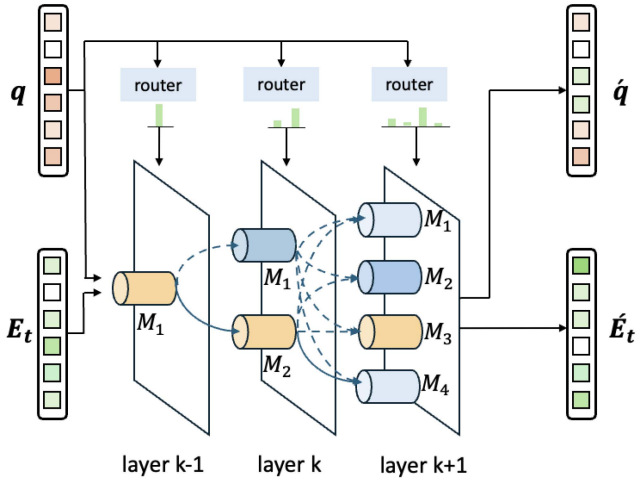

Fig. 4. Overview of Dynamic Attention Routing: when using self-attention to fuse features between the text and the query, we apply masks $\mathbf{M}$ with different receptive fields to control the attention scope. As the attention progresses through layers, the selectable routing options increase, and the query is used to learn the routing probabilities α to choose the routing path.

This process allows the video representation to undergo spatial feature attention operations, enabling information interchange among different sequence number queries j within a frame i after they pass through cross-attention and self-attention in Q-former.

In the spatial attention module, we adopt self-attention, denoted as S-ATN, for query and text embeddings in the m-th layer, i.e.,

$$\mathbf{E}_t{}', \mathbf{q}' = \text{S-ATN}^L(\mathbf{E}_t \cup \mathbf{q}), \tag{7}$$

where the superscript L for S-ATN^L denotes that the attention is across the L dimension. As shown in Fig. 2, when interacting with visual embeddings, we calculate cross-attention, denoted as C-ATN, between visual embeddings and query embeddings, which can be formulated as follows,

$$\mathbf{q}' = \text{C-ATN}^L(\mathbf{E}_v, \mathbf{q}). \tag{8}$$

In the temporal module, we reuse the parameters from the spatial attention module to reduce the cost. In this module, we reshape the batch data from $BT \times L \times D$ to $BL \times T \times D$ and calculate attentions across the T dimension as follows,

$$\mathbf{q}' = \text{S-ATN}^T(\mathbf{q}), \tag{9}$$

where S-ATN^T represents the temporal self-attention operator.

The channel swapping can utilize temporal modeling and spatial modeling to handle the similarity of information, thereby efficiently processing video temporal information with minimal parameters.

2) LoRA Fine-Tuning: To ensure that the new video representation is understandable by the frozen LLM, inspired by efficient fine-tuning techniques in NLP, we adopt a low-rank matrix adaptation (LoRA) approach for fine-tuning the Q and V matrices of the Query attention in Q-former while keeping all other parameters frozen:

$$\begin{aligned} \mathbf{W}_{qs} &= \mathbf{W}_q + LoRA_s(\mathbf{W}_q) \\ \mathbf{W}_{qt} &= \mathbf{W}_q + LoRA_t(\mathbf{W}_q) \end{aligned} \tag{10}$$

where $\mathbf{W}$ is the weight of Q-former parameters. s and t represent the temporal and spatial attention layers, respectively.

D. Dynamic Attention Routing (DAR)

As shown in Fig. 4, in each self-attention layer, we introduce a dynamic attention routing (DAR), which is a new routing selection mechanism that enables the model to understand both the high-level semantics and local relations. We initially design masks $\{\mathbf{M}_0, \ldots, \mathbf{M}_l\}$ with different receptive fields to accommodate varying degrees of attention interaction. As the number of attention layers increases, we utilize more routing choices, simulating the modal inconsistencies of different semantic levels.

In the k-th attention layer, the router can route the cooperative attention masks group $\{\mathbf{M}_0, \ldots, \mathbf{M}_{p_k-1}\}$, where p_k is the number of mask matrices in the k-th layer of DAR. We average the attention weights with different masks using the routing layer to obtain the final representation after attention:

$$DAR(\mathbf{q}) = \frac{[\mathbf{W}_q(\mathbf{E}_t \cup \mathbf{q})][\mathbf{W}_k(\mathbf{E}_t \cup \mathbf{q})]^T}{\sqrt{D_h}} \otimes \sum_{j=0}^{p_k-1} \alpha_j \mathbf{M}_j, \tag{11}$$

where α is the routing probability. And q_m is queries in the attention module. $\otimes$ represents the Hadamard product.

The routing probability α of the k-th attention layer can be obtained by conditional computation on the input. α can adjust the weighting of subsequent attention routing to achieve control

over different receptive fields.

$$\alpha = \text{MLP}(\text{APool}(\mathbf{q})) \tag{12}$$

where $\text{APool}(\cdot)$ refers to the 1D adaptive average pooling performed over all the embeddings of patches in the image; and MLP is a two-layer multi-layer perceptron.

IV. Experiments

In this section, we describe our experimental setup and present the results of our experiments. To validate the effectiveness of our rapid transfer method, we conducted experiments on multiple datasets. These include video intent analysis datasets and traditional VQA datasets.

The video intent analysis task, aimed at detecting the emotional intent of speakers, differs from the existing tasks of video-based LMMs. It requires attention to the speaker's emotions, a requirement that significantly deviates from video-based LMM's video pre-training tasks. We demonstrate quantitative results using the Video Sarcasm Detection Dataset MUStARD [23] and the Video Humor Detection Dataset UR-Funnyv2 [18]. These datasets encompass tasks across various conversational video scenarios, each sample consisting of a video segment and its corresponding dialogue. Our experiments on these datasets effectively showcase the capability of our method in analyzing dialogue intent in datasets with significant domain differences.

The VQA dataset is commonly used in video-based LMMs for instruction fine-tuning. To measure our method's transferability to unknown downstream VQA tasks, we fine-tuned it solely on the CLEVRER-MC [56] dataset and selected four diverse tasks from MVBench [57] for testing: Object Existence, Moving Direction, Moving Count, Moving Attribute. These tasks are complex temporal analysis challenges that cannot be solved by examining just a few frames of an image. Similarly, we conducted experiments on the qaEgo4d [58] and youcook2 [59] datasets. Although these QA tasks are relatively common, the scenarios differ significantly from those in traditional video pre-training datasets.

A. Datasets

a) MUStARD [23]: The Multimodal Sarcasm Detection Dataset (MUStARD) stands as the sole available resource for sarcasm detection in conversational videos. Sourced from well-known TV shows such as Friends and The Big Bang Theory, MUStARD comprises audiovisual utterances annotated with sarcasm labels. Each target utterance is linked to historical dialogues, providing essential context for comprehending sarcastic remarks. We conducted experiments on two commonly used partitioning schemes, including the original 5-fold partition and the train-dev-test partition. We labeled the dataset using the train-dev-test partition as MUStARD*.

b) UR-Funnyv2 [18]: For multimodal humor detection, we employ the UR-Funnyv2 dataset, gathered from TED talk videos and featuring three modalities. Like MUStARD, this dataset includes context preceding the target punchline. Additionally, we generate a reduced version of UR-Funnyv2 by truncating its training set. Detailed split statistics for these video intent analysis datasets are provided in Table I.

c) CLEVRER-MC [56]: The CLEVRER dataset is designed for evaluating computer vision and language understanding systems on dynamic scene understanding and causal reasoning. Unlike the original CLEVR dataset, CLEVRER focuses on video understanding, featuring synthetic video scenes with various physical interactions, such as object collisions and movements.

d) MVBench [57]: MVBench is a standardized benchmark for testing large video models, comprising 20 tasks that are highly relevant to temporal reasoning, such as Reasoning, Direction, Recognition, and more. It provides a standardized measure to test the temporal understanding capabilities of VideoLMM. In this context, we selected four of these tasks for testing.

e) qaEgo4d [58]: qaEgo4d is a video question-answering dataset focused on egocentric (first-person) video footage. It is part of the larger Ego4D dataset, which contains extensive video recordings captured from the wearer's perspective, offering a unique viewpoint that emphasizes the individual's interactions with their environment. More details are shown in Table II.

f) youcook2 [59]: youcook2 is the largest task-oriented, instructional video dataset in the vision community. It contains 2000 long untrimmed videos from 89 cooking recipes; on average, each distinct recipe has 22 videos. The procedure steps for each video are annotated with temporal boundaries and described by imperative English sentences. The videos were downloaded from YouTube and are all in the third-person viewpoint.

B. Experimental Setup

We conduct all experiments using a Frozen Vicuna-7B model [1] and a Q-former model [6] pre-trained on image-text datasets. In all experiments, we integrate LoRA into both the query and value projection layer in each attention module of the Q-former. Only the parameters of LoRA are trained while Queries and the LLM are frozen.

We train our model using the Pytorch framework on Nvidia-A40, with 48 GB dedicated memory in each GPU. Similar to BLIP-2, during our data processing stage, we used ViT and BERT to extract image and text features, respectively. Throughout the training process, we kept the parameters of the feature extraction models frozen. We use pre-trained Q-former feature extraction models and fine-tune them with LoRA during training. All other weights are frozen.

We train our model using the AdamW [70] optimizer with learning rate $= 0.0001$, betas $= (0.9, 0.95)$, weight decay $= 0.05$. For Dataset UR-Funnyv2, qaEgo4d, CLEVRER-MC and youcook2. We train our model with hyper-parameter batchsize $= 2$, window size $= 4$, epoch $= 25$. The hyper-parameter on MUStARD and MUStARD* dataset is batchsize $= 4$, windows size $= 8$, epoch $= 90$.

C. Results, Discussion and Analysis

In this section, we compare our MTransLLAMA model with various unimodal and multimodal baselines.

TABLE I
STATISTICS ABOUT THE THREE VIDEO INTENT ANALYSIS DATASETS USED IN OUR EXPERIMENTS

Dataset	#Samples	Train			Dev			Test		
		# Pos	# Neg	Total	# Pos	# Neg	Total	# Pos	# Neg	Total
MUStARD	690	276	276	552	0	0	0	69	69	138
MUStARD∗	690	207	207	414	69	69	138	69	69	138
UR-Funnyv2	7614	3810	3804	7614	494	486	980	490	504	994

MUStARD is the 5-fold split and MUStARD* is the train-dev-test split for the same dataset. These datasets have significant domain differences compared to the traditional training data of video-based LMMs.

TABLE II
STATISTICS ABOUT THE THREE DATASETS FOR OUT-OF-PRETRAINING SCENES

Dataset	Train		Dev		Test	
	# Videos	# Questions	# Videos	# Questions	# Videos	# Questions
qaEgo4d	997	10746	162	1913	166	1854
youcook2	1333	10337	0	0	457	3492
CLEVRER-MC	10000	149660	0	0	800	800

These datasets exhibit significant scene differences compared to the traditional training data of video-based LMMs. CLEVRER-MC and youcook2 only has a train/test split, and the Moving Direction task is a zero-shot task annotated by MVBench.

1) Generalization to Out-of-Pretraining Tasks: In order to test the performance when the downstream task significantly differs from the pretraining task, we conducted tests on intent analysis tasks. Although the scenarios in these datasets frequently appear in the pretraining data, they are often pre-trained using video captioning tasks, which differ significantly from the intent analysis. Intent analysis requires more attention to changes in facial expressions and the integration of textual information.

The tests include three datasets, UR-Funnyv2, MUStARD, and MUStARD*. We compare our model with state-of-the-art (SOTA) methods, including models designed specifically for sarcasm and humor detection **MuLOT** [63] and **MIL** [62], models focusing on multimodal information fusion **AGM** [64], **SimMMDG** [65], **I2MCL** [66], and existing state-of-the-art video understanding models based on large language models **VideoLLAMA** [7],**VideoChat** [10], **VideoLLAMA2** [68],**VideoChat2** [57], **LLaVA-Video** [67] and **InternVL2** [69],

Since the test sets of MUStARD and UR-Funnyv2 are balanced, we use binary accuracy as the evaluation metric. The results are shown in Table III. Based on the results, we have the following observations. MTransLLAMA is clearly superior to single-modal methods. Benefiting from the complementary information of multimodal data, MTransLLAMA improves 19.6% and 7.7% on accuracy compared with visual and textual methods, respectively, on MUStARD. For Humor Detection, MTransLLAMA achieved a 9.8% lead over unimodal methods. On the one hand, compared with only using the data of video modality, it is relatively more effective in detecting the intent expressed in the highly semantic text. On the other hand, as an important unit of expressing intent, videos can significantly improve the performance of MTransLLAMA.

In contrast to other multi-modal methods that utilize all three modalities of text, video, and audio, our MTransLLAMA approach specifically focuses on the text and video modalities and keeps the audio branch frozen. MTransLLAMA achieves 2.1%, 2.1%, and 1.6% improvements in the accuracy of sarcasm detection and humor detection on UR-Funny2, MUStARD, and MUStARD*. This result demonstrates that MTransLLAMA is particularly adept at recognizing intent data.

In comparison to other large language model-based methods, our MTransLLAMA approach demonstrates significant improvements in the same LoRA fine-tuning scenarios. Additionally, in terms of training costs, our model does not specifically model temporal modules, saving a considerable amount of video pre-training resources. Specifically, we achieve enhancements of 1.2%, and 2.1% without employing video pre-training on MUStARD. These results clearly indicate the effectiveness of our method compared to the state-of-the-art methods.

2) Generalization to Out-of-Pretraining Scenes: In order to test the performance when the scene of the downstream task significantly differs from the pretraining scene, we conducted tests on the qaEgo4d, CLEVRER-MC and youcook2 dataset. These datasets feature first-person perspectives or rare scenarios, which differ significantly from the scenes in the video pre-training datasets.

To facilitate fair comparison and understanding, we employ MVBench [57] for testing on CLEVRER-MC. This approach involves converting the generated answers into multiple-choice format, thereby enabling objective evaluation. This method allows for quantifiable comparisons and mitigates the subjectivity often associated with open-ended answers, ensuring a more standardized assessment of the VQA model's performance. We evaluated performance on the qaEgo4d and youcook2 datasets using accuracy, ROUGE, and METEOR metrics. We compare our model with state-of-the-art (SOTA) methods, including models designed specifically for VQA and video understanding models based on large language models.

We tested our model on four tasks in CLEVRER-MC. As shown in Table IV, The results show that InternVL achieves

TABLE III
PERFORMANCES ON THE VIDEO INTENT ANALYSIS DATASETS

Dataset	Modality	UR-Funnyv2	MUStARD	MUStARD*	Extra Data	Tunable.p
Unimodal Methods						
Timesfomer [60]	V	56.2	54.3	56.5	✗	125M
SVM+Bert [28]	T	58.9	66.2	65.3	✗	400M
EfficientNet [61]	V	55.4	55.0	55.7	✗	10M
Sarcasm and Humor Detection Methods						
MSEA [20]	V+T	—	71.7	—	✓	2M
MUStARD++ [25]	V+A+T	—	71.7	—	✓	2.5M
MIL* [62]	V+A+T	—	71.2	71.7	✗	4M
MuLOT* [63]	V+A+T	—	70.5	71.1	✗	5M
Multi-Modal Fusion Methods						
LMF [60]	V+A+T	65.2	62.4*	63.9	✗	1.5M
MULT [29]	V+A+T	60.9	64.6*	63.2	✗	4.5M
AGM [64]	V+A+T	65.0	—	—	✗	10M
SimMMDG [65]	V+A+T	65.6	68.9*	72.5	✗	25M
I2MCL [66]	V+A+T	65.1	64.3*	65.2	✗	40M
LLM + Zero-shot						
VideoLLAMA [7]*	V+A+T	40.1	44.2	35.6	✓	–
VideoChat [10]*	V+A+T	42.1	37.4	37.7	✓	–
LLaVA-Video [67]	V+T	57.5	54.3	50.7	✓	–
VideoLLAMA2 [68]	V+A+T	58.0	52.1	53.0	✓	–
InternVL2 [69]	V+T	52.1	55.7	52.1	✓	–
LLM + Fine-Tuning						
VideoLLAMA [7]	V+A+T	66.7	70.5	69.6	✓	0.5B
VideoChat [10]	V+A+T	66.3	71.1	71.1	✓	0.5B
VideoLLAMA2 [68]	V+A+T	67.1	71.5	71.1	✓	0.7B
InternVL2 [69]	V+T	67.7	72.7	72.5	✓	0.8B
MTransLLAMA (Ours)	V+T	68.2	73.9	**74.6**	✗	2.5M
MTransLLAMA (Ours)	V+A+T	**68.3**	**74.6**	73.3	✗	2.5M

Since the test sets are balanced, only binary classification accuracy is reported. Tunable.p represents tunable parameters. Models and digits with * denotes results from reproductions conducted by the authors. Extra data refers to whether additional annotations were made to the dataset for training. In the modality column, T, V, and A represent Text, Video, and Audio, respectively, "Tunable.p" represents the number of parameters trained during instruction tuning. Black and blue highlighting indicate the best and second-best performance, respectively.

TABLE IV
ACCURACY ON THE CLEVRER-MC DATASET

Tasks	MD	MC	MA	OE
VideoLLAMA	22.5	22.5	32.5	48.0
VideoChat	25.5	20.5	42.5	53.0
VideoChat2	23.0	42.0	58.5	58.0
VideoLLAMA2	35.0	46.0	55.0	53.5
LLaVA-Video	41.5	44.5	79.0	61.0
InternVL2	**56.0**	**86.5**	**89.5**	**95.5**
MTransLLAMA (Ours)	26.5	46.0	40.0	53.0

MD, MC, MA, OE represents "Moving Direction", "Moving Count", "Moving Attribute" and "Object Existence", respectively. Best performance is highlighted in bold.

strong performance on the CLEVRER-MC dataset, and newer models like LLaVA-Video and VideoLLaMA also demonstrate impressive results. However, without the use of video pre-training, and by utilizing only 1% of the training video dataset and 0.5% of trainable parameters, our model achieves performance on par with VideoLLaMA2 on the "MA" (Moving Attribute) and "OE" (Object Existence) tasks, while demonstrating performance similar to or exceeding that of VideoChat2 and VideoLLaMA on the "MD" (Moving Direction) and "MC" (Moving Count) tasks. Despite other LLM-based methods, except for VideoLLaMA, not keeping the LLM frozen during training and instead fine-tuning the LLM using LoRA—which resulted in improved performance—our model still maintains significantly lower costs.

As shown in Table V, on both the qaEgo4d and youcook2 datasets, our model achieved state-of-the-art performance. Compared to traditional VQA methods, our model's accuracy on the qaEgo4d dataset is not as high as HCRN. However, this is because HCRN is a non-generative approach that requires a predefined answer dictionary and utilizes a greater number of video frames. Moreover, our method ranks just below the state-of-the-art InternVL2 among LLM-based methods.

TABLE V
PERFORMANCES ON THE VIDEO QUESTION ANSWERING DATASETS ARE EVALUATED USING ACCURACY (ACC), AS WELL AS METEOR (M) AND ROUGE, TWO COMMON TEXT TRANSLATION METRICS, TO REPRESENT THE RESULTS

Dataset	qaEgo4d			youcook2	
Model	Acc	M	ROUGE	M	ROUGE
Video Understanding Methods					
All models take 16 frames as input.					
SimpleVQA	9.7	18.3	27.7	—	—
HCRN∗ [71]	10.3	17.2	25.7	—	—
JustAsk∗ [72]	9.6	17.8	26.7	—	—
Longformer [73]	6.7	16.9	24.4	—	—
VideoBert [74]	—	—	—	12.0	28.8
AT+Video [75]	—	—	—	17.8	36.5
All models take 4 frames as input.					
VideoChat2	1.6	15.5	13.5	18.0	34.2
VideoLLAMA2	1.9	15.1	12.9	17.9	35.5
InternVL2	**10.7**	18.7	**29.2**	17.7	35.1
LLaVA-Video	2.3	15.7	13.1	17.2	36.1
MTransLLAMA (Ours)	10.1	**18.9**	28.1	**18.2**	**38.6**

* represents non-generative question answering. Best performance is highlighted in bold.

TABLE VI
COMPARISON IN PARAMETERS AND TIME EFFICIENCY BETWEEN OUR METHOD AND OTHER VIDEO-BASED LMMS

Method	V.P	Complexity	#T.P	F.LLM	F.Proj	#V.T
VideoLLAMA	✓	s^2f^2	1B	✓	✗	sf
VideoLLAMA2	✓	s^2f^2	7B	✗	✗	sf
VideoChat	✓	s^2f^2	196M	✓	✗	sf
VideoChat2	✓	s^2f^2	7B	✗	✗	sf
InternVL2	✓	s^2f^2	8B	✗	✗	sf
LLaVA-Video	✗	s^2f^2	7B	✗	✗	sf
Ours	✗	s^2+f^2	2.5M	✓	✓	32

The abbreviations "V.P" (video pretraining), "Complexity" (algorithm complexity), "#T.P" (number of tunable parameters), "F.LLM" (frozen LLM), "F.Proj" (frozen projection layer), and "#V.T" (number of visual tokens) represent the corresponding characteristics, respectively. The parameters f and s represent the number of frames and the number of tokens per frame, respectively. In the "#V.T" section, for other methods, the number of visual tokens input to the LLM increases linearly with the number of frames f, and the number of tokens per single frame s is generally greater than or equal to 32.

As shown in Table VI, we also compare our method with different video-based LMMs. In terms of efficiency, the main advantage of our approach is the savings during the Video Pretraining phase, which requires a large amount of video data, such as WebVid-2M, and consumes resources nearly a hundred times more than the subsequent fine-tuning phase. Moreover, as the frame number increases, our model has lower memory usage and algorithmic complexity compared to other models. Additionally, our model has the fewest tunable parameters, and by freezing the projection layer and the large language model, it becomes easier to train and fine-tune.

D. Ablation Study

To probe the effectiveness of each component in MTransLLAMA, we conduct ablation experiments. All the experiments are implemented by Q-former and LLAMA-7B.

Our goal is to add a few tunable parameters to the frozen space-only model and close the gap to the video pre-trained model. To validate the effectiveness of Multimodal Query Temporal Reusing (MQT), we conducted experiments by removing the MQT module and only using the spatial information extractor to concatenate frames of images. As shown in Table VII, the performance of the frozen space-only model rapidly declined, despite using more tokens and spending more time in this setting. This indicates that our temporal adaptation introduces strong temporal modeling to the space-only model. These results successfully validate the effectiveness of our proposed MQT module.

TABLE VII
ABLATION STUDY ON THE MUSTARD AND QAEGO4D DATASETS

Dataset	UR-Funnyv2			qaEgo4d		
Model	Acc	Recall	Prec	Acc	METEOR	ROUGE
w/o. Fusion	63.4	62.6	64.2	10.0	18.7	27.8
w/o. DAR	67.4	64.6	69.1	9.8	18.5	27.5
w/o. MQT	58.3	53.9	59.9	9.0	17.5	25.8
MTransLLAMA (Ours)	**68.2**	**64.8**	**70.2**	**10.1**	**18.9**	**28.1**

To better observe the impact of the proposed method on the results, we conducted experiments on text and visual modalities. Experiments show our method is beneficial to the results. Best performance is highlighted in bold.

We also evaluate UniFusion and DAR in Table VII. To measure the effectiveness of UniFusion, we controlled the input modalities to observe the impact on performance. Our experiments included dual-modal training with and without early modality fusion using UniFusion. Compared to other fusion methods, our UniFusion system performed the best, highlighting the importance of early modality interaction for information exchange. Early fusion refers to our approach of inputting textual information into the Q-former for modality fusion. Compared to previous methods that align vision to text, this approach extracts features more effectively and significantly reduces the token usage of the LLM, making it highly efficient in conserving computational resources. Our early fusion method shows a noticeable improvement in UrFunnyv2, a humor detection dataset, where enhanced text analysis is crucial. It also performs well in qaEgo4d, where it better integrates instructions with key visual regions.

E. Case Study

We conducted case studies on the performance of MtransLLAMA across various datasets, exploring its transfer capabilities in downstream video datasets, including task transfer and scene transfer. As shown in Fig. 5, we performed a comparative evaluation of MtransLLAMA with previous methods. In the sarcasm detection task, MTransLLAMA only inputs the question into the LLM and the dialogue into the multimodal fusion module, while other methods require inputting both the question and the dialogue. For sarcasm detection, MTransLLAMA provides excellent responses based on video information and textual dialogue, producing fewer hallucinations. In the youcook2 dataset, MTransLLAMA also demonstrated strong generalization capabilities in datasets with rare scenes.

Leonard: I never would have identified the fingerprints of string theory in the aftermath of the Big Bang.
Sheldon: My apologies. What's your plan?
Sheldon: It's just a privilege to watch your mind at work.

Pair with the past dialogue and video, is it sarcastic when < Sheldon: It's just a privilege to watch your mind at work >?

VideoLLAMA:
The video shows a young man standing in front of a large glass door, This requires more information to determine whether it is sarcastic.
VideoChat2:
No.
Ours:
Yes! it is sarcastic.

Pair with the video, What happened in the last 5 seconds?

VideoLLAMA:
The hand of a person was holding a pot of food.
VideoChat2:
A person is cooking food in a stainless steel pot.
Ours:
Add the tomatoes to the pot.

Fig. 5. Case Study: The first clip is extracted from the MUStARD dataset, and the second clip is from the youcook2 dataset. We demonstrate the outputs of different LMMs.

V. Conclusion

Our model performs well when encountering scenes and tasks that were not seen during video pretraining. However, if the downstream task closely resembles the original pretraining dataset, such as CLEVRER-MC, our model's performance falls short compared to the original video analysis models.

In this paper, we propose a novel LMM-based video understanding model that addresses the challenge of poor transferability and generalization in traditional video understanding models. This model achieves efficient transfer from an image-text LMM to a video-text LMM by incorporating temporal capabilities through parameter sharing and channel swapping in a pre-trained image-text model. Final experiments demonstrate that our model achieves state-of-the-art performance in multiple small-scale specific video scene datasets.

References

[1] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[2] Y. Liu et al., "Summary of ChatGPT-related research and perspective towards the future of large language models," *Meta- Radiol.*, 2023, Art. no. 100017.
[3] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
[4] Z. Zhao et al., "See and think: Embodied agent in virtual environment," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 187–204.
[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, vol. 36, pp. 34892–34916.
[6] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
[7] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proc. 2023 Conf. Empir. Methods Natural Lang. Process.: Syst. Demonstr.*, Y. Feng and E. Lefever Eds., Singapore, Dec. 2023, pp. 543–553. [Online]. Available: https://aclanthology.org/2023.emnlp-demo.49/
[8] E. Song et al., "MovieChat: From dense token to sparse memory for long video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18221–18232.
[9] E. Song et al., "Moviechat: Question-aware sparse memory for long video question answering," 2024, *arXiv:2404.17176*.
[10] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards detailed video understanding via large vision and language models," in *Proc. 62nd Annu. Meet. Assoc. Comput. Linguistics*, L.-W. Ku, A. Martins, and V. Srikumar Eds., Bangkok, Thailand, Aug. 2024, pp. 12585–12602. [Online]. Available: https://aclanthology.org/2024.acl-long.679/
[11] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[12] Y. Zhou et al., "TRAR: Routing the attention spans in transformer for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2074–2084.

[13] W. Han et al., "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interact.*, 2021, pp. 6–15.

[14] P. P. Liang et al., "Multibench: Multiscale benchmarks for multimodal representation learning," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021.

[15] X. Zhang, Y. Chen, and G. Li, "Multi-modal sarcasm detection based on contrastive attention mechanism," in *Proc. Natural Lang. Process. Chin. Comput.: 10th CCF Int. Conf.*, 2021, pp. 822–833.

[16] A. Potamianos, E. Fosler-Lussier, E. Ammicht, and M. Perakakis, "Information seeking spoken dialogue systems— Part II: Multimodal dialogue," *IEEE Trans. Multimedia*, vol. 9, pp. 550–566, 2007.

[17] A.-A. Liu et al., "Counterfactual visual dialog: Robust commonsense knowledge learning from unbiased training," *IEEE Trans. Multimedia*, vol. 26, pp. 1639–1651, 2024.

[18] M. K. Hasan et al., "UR-Funny: A multimodal language dataset for understanding humor," in *Proc. 2019 Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, K. Inui, J. Jiang, V. Ng, and X. Wan Eds., Hong Kong, China, Nov. 2019, pp. 2046–2056. [Online]. Available: https://aclanthology.org/D19-1211/

[19] S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguistics*, A. Korhonen, D. Traum, and L. Márquez Eds., Florence, Italy, Jul. 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050/

[20] D. S. Chauhan, S. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4351–4360.

[21] J. Li, M. Zhang, D. Ji, and Y. Liu, "Multi-task learning with auxiliary speaker identification for conversational emotion recognition," 2020, *arXiv:2003.01478*.

[22] X. Song, L. Huang, H. Xue, and S. Hu, "Supervised prototypical contrastive learning for emotion recognition in conversation," in *Proc. 2022 Conf. Empir. Methods Natural Lang. Process.*, Y. Goldberg, Z. Kozareva, and Y. Zhang Eds., Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 5197–5206. [Online]. Available: https://aclanthology.org/2022.emnlp-main.347/

[23] S. Castro et al., "Towards multimodal sarcasm detection (an _Obviously_ perfect paper)," in *Proc. 57th Annu. Meeting Assoc. Computat. Linguistics*, 2019, pp. 4619–4629. [Online]. Available: https://aclanthology.org/P19-1455

[24] M. M. Islam and T. Iqbal, "Multi-GAT: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1729–1736, Apr. 2021.

[25] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, "A multimodal corpus for emotion recognition in sarcasm," in *Proc. 13th Lang. Resour. Eval. Conf.*, N. Calzolari et al. Eds., Marseille, France, Jun. 2022, pp. 6992–7003. [Online]. Available: https://aclanthology.org/2022.lrec-1.756/

[26] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[27] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, J. Burstein, C. Doran, and T. Solorio Eds., Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[29] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguistics*, A. Korhonen, D. Traum, and L. Márquez Eds., Florence, Italy, Jul. 2019, pp. 6558–6569. [Online]. Available: https://aclanthology.org/P19-1656/

[30] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang, "InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task LLMs framework," 2023, *arXiv:2309.11911*.

[31] W. Chai and G. Wang, "Deep vision multimodal learning: Methodology, benchmark, and trend," *Appl. Sci.*, vol. 12, no. 13, 2022, Art. no. 6588.

[32] X. Wang, G. Wang, W. Chai, J. Zhou, and G. Wang, "User-aware prefix-tuning is a good learner for personalized image captioning," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, Springer, 2023, pp. 384–395.

[33] R. He et al., "On the effectiveness of adapter-based tuning for pretrained language model adaptation," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, C. Zong, F. Xia, W. Li, and R. Navigli Eds., Aug. 2021, pp. 2208–2222. [Online]. Available: https://aclanthology.org/2021.acl-long.172/

[34] A. Arnab et al., "VIVIT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.

[35] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Invariant grounding for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2928–2937.

[36] J. Xiao et al., "Video as conditional graph hierarchy for multi-granular question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 2804–2812.

[37] J. Lin et al., "VideoXUM: Cross-modal visual and textural summarization of videos," *IEEE Trans. Multimedia*, vol. 26, pp. 5548–5560, 2024.

[38] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4768–4777.

[39] M. Wang, J. Xing, and Y. Liu, "ActionClip: A new paradigm for video action recognition," 2021, *arXiv:2109.08472* .

[40] Z. Qi, R. Zhu, Z. Fu, W. Chai, and V. Kindratenko, "Weakly supervised two-stage training scheme for deep video fight detection model," in *Proc. IEEE 34th Int. Conf. Tools Artif. Intell.*, 2022, pp. 677–685.

[41] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[42] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Euro. Conf. Comput. Vis.*, Springer, 2020, pp. 213–229.

[43] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[44] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[45] C. Zhang, H. Bai, and Y. Zhao, "Fine-grained image classification by class and image-specific decomposition with multiple views," *IEEE Trans. Multimedia*, vol. 25, pp. 6756–6766, 2023.

[46] Z. Wu et al., "Conditional consistency regularization for semi-supervised multi-label image classification," *IEEE Trans. Multimedia*, vol. 26, pp. 4206–4216, 2024.

[47] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[48] J. Xu, Y.-L. Li, and S. Wang, "AdaZoom: Towards scale-aware large scene object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 4598–4609, 2023.

[49] Z. Yao and L. Wang, "Boundary information progressive guidance network for salient object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 4236–4249, 2022.

[50] S. Minaee et al., "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, 3523–3542, Jul. 2022.

[51] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[52] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2Video: Mastering video-text retrieval via image clip," 2021, *arXiv:2106.11097*.

[53] H. Luo et al., "Clip4Clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[54] T. Yang et al., "AIM: Adapting image models for efficient video action recognition," 2023, *arXiv:2302.03024*.

[55] Y. Tian, "Dynamic routing transformer network for multimodal sarcasm detection," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguist.*, Toronto, ON, Canada, 2023, pp. 2468–2480. [Online]. Available: https://aclanthology.org/2023.acl-long.139

[56] K. Yi et al., "CLEVRER: Collision events for video representation and reasoning," 2019, *arXiv:1910.01442*.

[57] K. Li et al., "MVBench: A comprehensive multi-modal video understanding benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22195–22206.

[58] L. Bärmann and A. Waibel, "Where did i leave my keys?-Episodic-memory-based question answering on egocentric videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1560–1568.

[59] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32 .

[60] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguistics*, I. Gurevych and Y. Miyao Eds., Melbourne, Australia, Jun. 2018, pp. 2247–2256. [Online]. Available: https://aclanthology.org/P18-1209/

[61] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[62] Y. Zhang et al., "Learning multi-task commonness and uniqueness for multi-modal sarcasm detection and sentiment analysis in conversation," *IEEE Trans. Artif. Intell.*, vol. 5, no. 3, pp. 1349–1361, Mar. 2024.

[63] S. Pramanick, A. Roy, and V. M. Patel, "Multimodal learning using optimal transport for sarcasm and humor detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 3930–3940.

[64] H. Li et al., "Boosting multi-modal model performance with adaptive gradient modulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 22214–22224.

[65] H. Dong, I. Nejjar, H. Sun, E. Chatzi, and O. Fink, "SimMMDG: A simple and effective framework for multi-modal domain generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023.

[66] Y. Zhou, X. Wang, H. Chen, X. Duan, and W. Zhu, "Intra-and inter-modal curriculum for multimodal learning," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 3724–3735.

[67] Y. Zhang et al., "Video instruction tuning with synthetic data," 2024, *arXiv:2410.02713*.

[68] Z. Cheng et al., "VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs," 2024, *arXiv:2406.07476*.

[69] Z. Chen et al., "InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24185–24198.

[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[71] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for multimodal video question answering," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3027–3050, 2021.

[72] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1686–1697.

[73] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[74] C. Sun et al., "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7464–7473.

[75] J. Hessel, B. Pang, Z. Zhu, and R. Soricut, "A case study on combining ASR and visual features for generating instructional video captions," in *Proc. 23rd Conf. Comput. Natural Lang. Learn.*, M. Bansal and A. Villavicencio Eds., Hong Kong, China, Nov. 2019, pp. 419–429. [Online]. Available: https://aclanthology.org/K19-1039/