Facial Dynamics in Video: Instruction Tuning for Improved Facial Expression Perception and Contextual Awareness

Jiaxing Zhao[1*], **Boyuan Sun**[2,1*], **Xiang Chen**[1], **Xihan Wei**[1]
[1]Tongyi Lab, Alibaba Group
[2]VCIP, School of Computer Science, Nankai University
zjx244036@alibaba-inc.com, boyuansun@mail.nankai.edu.cn,
xchen.cx@alibaba-inc.com, xihan.wxh@alibaba-inc.com

Abstract

Facial expression captioning has found widespread application across various domains. Recently, the emergence of video Multimodal Large Language Models (MLLMs) has shown promise in general video understanding tasks. However, describing facial expressions within videos poses two major challenges for these models: (1) the lack of adequate datasets and benchmarks, and (2) the limited visual token capacity of video MLLMs. To address these issues, this paper introduces a new instruction-following dataset tailored for dynamic facial expression caption. The dataset comprises 5,033 high-quality video clips annotated manually, containing over 700,000 tokens. Its purpose is to improve the capability of video MLLMs to discern subtle facial nuances. Furthermore, we propose FaceTrack-MM, which leverages a limited number of tokens to encode the main character's face. This model demonstrates superior performance in tracking faces and focusing on the facial expressions of the main characters, even in intricate multi-person scenarios. Additionally, we introduce a novel evaluation metric combining event extraction, relation classification, and the longest common subsequence (LCS) algorithm to assess the content consistency and temporal sequence consistency of generated text. Moreover, we present FEC-Bench, a benchmark designed to assess the performance of existing video MLLMs in this specific task.

Github — https://github.com/Jiaxing-star/FacialDynamic

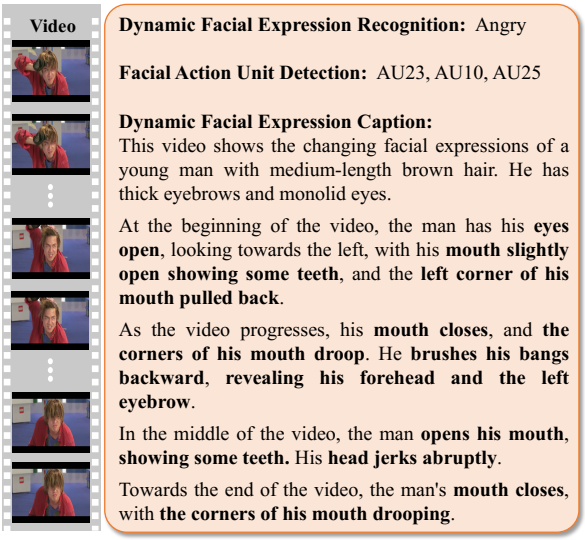


Figure 1: Comparison of annotation styles between different facial expression tasks. Among these, DFER only includes one selected fundamental emotional category, whereas FAUD contains several specific action units. Our DFEC includes detailed facial movements described using natural language.

Introduction

Facial expressions play a crucial role in daily communication and are a vital component of human interaction. Their significance has led to growing interest from researchers in fields such as human-computer interaction (HCI) (Wadley, Kostakos et al. 2022) and mental health (Menefee DS 2022). As researchers have shifted their focus from static images to dynamic video content, Dynamic Facial Expression Recognition (DFER) (Chen et al. 2024a; Li et al. 2024b; Foteinopoulou and Patras 2024) has attracted attention from experts in psychology, computer science, linguistics, neuroscience, and related disciplines due to its wide-ranging applications. The goal of DFER (Lucey, Cohn et al. 2010; Zhao et al. 2011; Jiang et al. 2020) is to classify video sequences into one of several fundamental emotional categories, including neutral, happiness, sadness, surprise, fear, disgust, and anger. However, in practical applications, characters in a video often exhibit a range of emotions at different stages, making discrete emotion classification insufficient for accurately captioning facial expressions. Even multi-class classification fails to adequately capture the subtleties of facial changes.

To address this limitation, we introduce a new task called Dynamic Facial Expression Caption (DFEC). The goal of DFEC is to use natural language to fluently and accurately describe the facial expressions of main characters in videos. In this task, the model is required to generate content that covers all significant facial expressions of the main characters in the video without introducing hallucinations, especially when the video contains subtle, rapid, and meaning-

*These authors contributed equally.

ful changes in expressions, which increases the challenge. Specifically, we leverage the content provided by facial units to describe the facial information of characters in videos. Unlike Facial Action Unit Detection (FAUD) (Zhao, Chu, and Zhang 2016; Shao et al. 2018; Ning, Salah, and Ertugrul 2024; Yuan et al. 2024), which focuses on identifying specific action units, our approach employs temporally coherent natural language, which is more aligned with downstream applications and better suited for large language models. As illustrated in Figure 1, compared to existing DFER and Facial Action Unit Detection (FAUD) tasks, our method provides more detailed and natural captions of facial expressions in videos. While some DFER datasets, such as MAFW (Liu et al. 2023c) and EmoVidCap (Wang et al. 2022a), include descriptions of facial expressions, these descriptions are often overly simplistic and fail to capture the dynamic changes in facial features over time. In this paper, we construct a high-quality DFEC dataset, FDA, containing 5,033 manually annotated videos to help video MLLMs better understand facial expressions in videos.

In recent years, MLLMs (Chen et al. 2023) have achieved significant success in the perception and understanding of images. Consequently, the perception and understanding of videos have garnered increasing attention from researchers (Cheng et al. 2024; Zhao et al. 2025; Xu et al. 2024; Zhang et al. 2024; Wang et al. 2024). However, when dealing with video inputs, the need to process multiple frames results in a limited number of tokens being allocated to each frame. Even with the highly advanced Qwen2-VL (Wang, Bai et al. 2024) model, when processing video input, the encoding for each frame is limited to only 138 tokens. This token limitation often proves insufficient for encoding detailed information. In videos, the facial region usually occupies only a small portion of the frame, and this lack of detailed encoding capability can significantly impact the performance of our DFEC task.

In this paper, we introduce a novel MLLM named FaceTrack-MM, which is designed to track faces in videos and focus on the main characters' facial expressions in multi-person scenes. Specifically, we integrate a dynamic video face tracking module into the MLLM to accurately model the facial regions of the main characters in videos. The extracted features are then projected and fed into the LLM to generate more precise and detailed captions of the facial region. Besides, We propose a novel evaluation metric Temporal Event Matching (TEM), which combines event extraction, relation classification, and the Longest Common Subsequence (LCS) algorithm to assess the semantic consistency and temporal sequence of generated text. We randomly selected 1,000 samples from the annotated data to form the test set. Using this test set, we constructed a facial expression caption benchmark, FEC-Bench, to compare the performance of 15 open-source and proprietary models across 8 metrics.

We summarize the key contributions of our paper as follows:

1. We construct a high-quality DFEC dataset containing 5,033 manually annotated samples to help video MLLMs better understand and describe facial expressions in videos.
2. We propose a novel FaceTrack-MM, which significantly improves the ability of existing large models to encode facial details.
3. We introduce a new evaluation method, Temporal Event Matching (TEM), to explicitly assess the content consistency and order consistency of the generated text.
4. We construct a FEC-Bench for the DFEC task, laying the foundation for future research in this field.

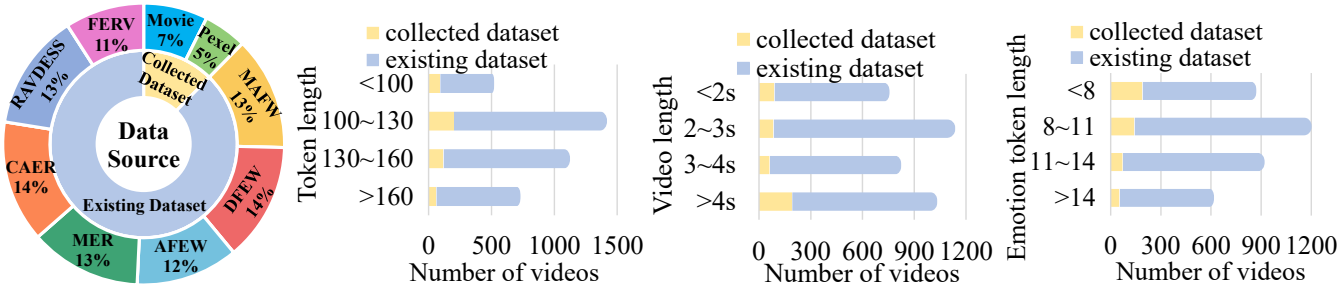
Related Work

Video Caption Dataset

Existing video captioning datasets can be categorized into three groups based on video duration. Short video datasets (Rohrbach et al. 2017; Wang et al. 2019) typically contain video clips ranging from 5 to 30 seconds in length. Longer video datasets (Huang et al. 2020; Krishna et al. 2017; Zhou, Xu, and Corso 2018; Mangalam, Akshulakov, and Malik 2024) include videos that vary in length from 1 to 5 minutes. Very long video datasets (Islam et al. 2024; Fu et al. 2024; Zhou, Shu et al. 2024) consist of videos that can last for hours. However, these datasets often emphasize overall content and action recognition in videos. Specifically, such descriptions typically include specific actions being performed (for example, "A person is drinking coffee."), the presence of certain objects in the scene (for example, "There is a book on the table."), or the development of a sequence of events over time (for example, "The video begins with a man watching TV, then he stands up, walks to a table, and engages in conversation with a woman"). Despite the growing importance of facial expressions in downstream tasks such as generative models (Ding et al. 2023; Li et al. 2024d) and digital human representations (Özacar and Alkhalıfa 2024), there is currently a lack of datasets that specifically focus on detailed descriptions of human facial expressions. Existing datasets often emphasize overall content and action recognition in videos, rather than the details of facial expressions. MAFW (Liu et al. 2023c) and EmVidCap (Wang et al. 2022a) include only very basic facial action annotations. To address this gap, we introduce a dynamic facial expression caption dataset containing 5,033 high-quality video clips, annotated manually with over 50,000 facial expression words and over 700,000 tokens. We have built a comprehensive benchmark based on this dataset, laying the foundation for further exploration in this direction.

Multimodal Large Language Models

Large language models can be broadly divided into proprietary and open categories. Proprietary models—such as GPT-4o (OpenAI 2024), Gemini (Team 2024), Claude-3.5 (Anthropic 2024), and Qwen2-VL (Bai et al. 2023; Wang, Bai et al. 2024)—are developed by commercial entities and restricted by usage policies and intellectual property laws, limiting access, modification, and reuse without explicit permission. Although they perform well across diverse visual tasks, they often underperform in specialized domains like facial description, and their closed nature prevents further optimization or extension.



(a) Data source distribution. (b) Token length (c) Video length (d) Emotion token length

Figure 2: Some statistics of proposed DFEC dataset.

In contrast, open large models release their architecture, training data, and weights to promote transparency, reproducibility, and community innovation. Recent years have seen rapid progress in this space. LLaMA-VID (Li, Wang, and Jia 2024) enables hour-long video understanding; LLaVA-OneVision (Li et al. 2024a) advances performance in single-image, multi-image, and video scenarios; Video-LLaMA (Zhang, Li, and Bing 2023) and Video-LLaMA2 (Cheng et al. 2024) jointly model visual and auditory content in videos; mPLUG-Owl2 (Ye et al. 2023b,a) enhances modality collaboration via shared modules; BLIP-2 (Li et al. 2023a) uses Q-Former (Zhang et al. 2023) for vision-language alignment, while BLIP-3 (Xue et al. 2024) adopts more scalable visual token samplers such as perceptual resamplers; and HumanOmni (Zhao, Yang et al. 2025) introduces the first human-centric Omni-MLLM that integrates visual and audio modalities for comprehensive scene understanding.

Nevertheless, most existing open models focus on general scenarios and still encode facial information inadequately, leading to suboptimal performance in fine-grained facial analysis tasks. To address this gap, we propose FaceTrack-MM, a video-based multimodal large language model specifically designed to improve facial representation and reasoning.

Dataset

In this paper, we construct a Dynamic Facial Expression Caption dataset, FDA, to bridge the gap between the broader video understanding research community, which has traditionally focused on narrative progression and main content, and the downstream applications that greatly benefit from detailed captions of facial changes in videos. In this section, we first introduce the sources of the data in our dataset, then describe the annotation process, followed by an analysis of the basic properties of the dataset. Finally, we propose the Temporal Event Matching (TEM), a new evaluation metric for long text captions.

Data Sources

Our data sources are primarily divided into two parts. The first part is extracted from existing datasets, specifically from video datasets related to emotions, to ensure a rich variety of facial expressions. We refer to this as the existing dataset. The second part consists of self-collected data, which we obtained through web scraping. We refer to this as the collected dataset.

Specially, to ensure diversity in the dataset, we sampled 500 videos each from existing datasets such as MAFW (Liu et al. 2023c), DFEW (Jiang et al. 2020), AFEW (Kossaifi et al. 2017), MER (Gómez-Cañón et al. 2021), CAER (Lee et al. 2019), FERV39K (Wang et al. 2022b), and RAVDESS (Livingstone and Russo 2018). Furthermore, during the dataset sampling process, we perform uniform sampling based on emotion categories to maintain an approximately equal distribution of videos across different emotions. Additionally, we collected 500 videos from Pexels Videos using keywords such as “man”, “girl”, “family”, and “woman”, and 500 videos from Cesituwang. Each video is no longer than 20 seconds. We processed these sampled videos by removing the header and footer information and any irrelevant content. Furthermore, we randomly extracted video clips from a large number of famous movies. For each clip, we performed face tracking to ensure that at least one face was present and that the face region occupied more than 5 percent of the frame. We discarded any clips that did not meet these criteria. Ultimately, we extracted 1,000 clips that satisfied the requirements. Subsequently, we conducted a manual screening process to remove videos that contained too many people or static scenes with no movement. We present the statistical data on the sources in Figure 2a, including the proportions of data extracted from each existing dataset and each self-collected dataset.

Annotation Process

Our annotation process is primarily divided into two parts: preliminary annotation generation and the manual correction process. We utilized carefully crafted prompts with GPT-4 (OpenAI 2024) to generate preliminary annotations. The generated preliminary annotations lack detailed captions, especially in facial changes, and contain a significant amount of hallucination. The prompts and some pre-annotated examples can be found in the supplementary material.

During the manual correction process, each video clip is reviewed by three annotators and finally consolidated by a final reviewer. To further ensure annotation reliability, we perform cross-validation among the final reviewers. The an-

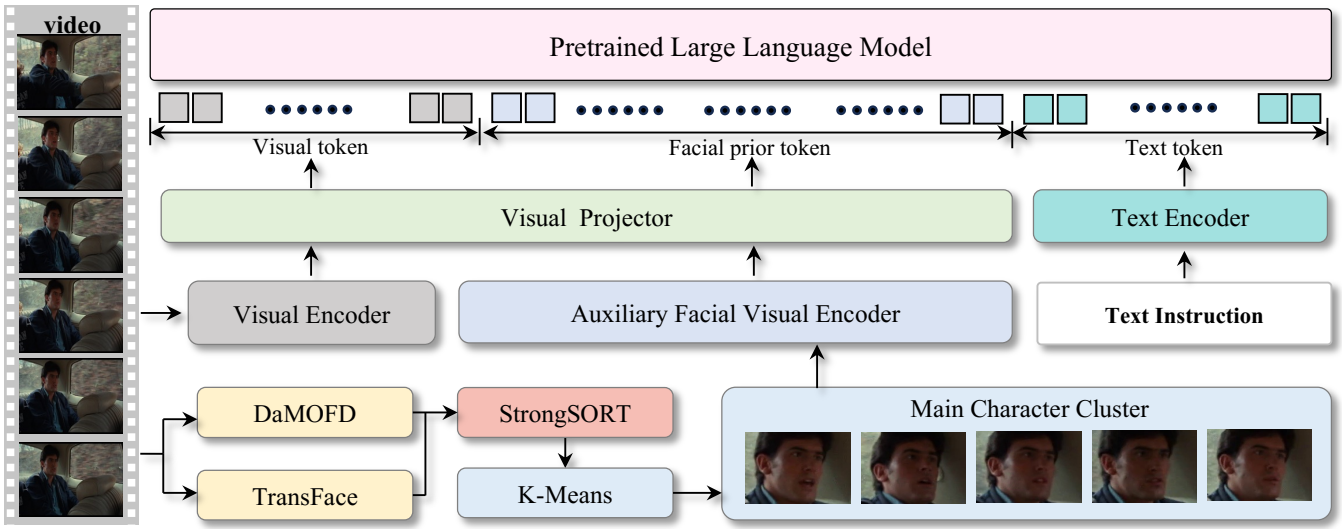


Figure 3: **Architecture of FaceTrack-MM.** Our FaceTrack-MM leverages FaceXFormer (Narayan et al. 2024) as the auxiliary facial visual encoder to extract facial features of the main characters and uses CLIP-ViT-Large (Radford, Kim et al. 2021) as the visual encoder. We utilize the STC module (Cheng et al. 2024) as the visual projector to inject temporal information and use Mistral-7B-Instruct (Jiang et al. 2024) for the pretrained large language model.

notators refer to the pre-generated annotations and modify them to better meet downstream requirements. Specifically, our annotation content is divided into three parts.

External Attributes. We described the external attributes of the individuals in the video, including subtle facial features.

Facial Changes. We focused on detailing the facial changes of the main characters in the video. In this section, we retained certain reasonable inferences generated during the pre-annotation phase, as we believe these inferences enhance the richness of the descriptions. Since these inferences are not objectively present in the video, we marked them with special identifiers.

Summary Sentence. Following the annotation style of the MAFW (Liu et al. 2023c) dataset, we used a single sentence to summarize the content of the video.

During the annotation process, annotators filtered out videos based on the following criteria: static videos, videos containing violent content, and videos with frequent perspective switches. Ultimately, the total number of videos remaining is 5,033. Overall, our dataset explicitly distinguishes between objective facial descriptions and slightly subjective speculative descriptions, while also including a concise one-sentence summary that captures the content. This allows us to select the appropriate annotated content for fine-tuning the model based on our application needs. In the subsequent instruction tuning, dataset attributes, experiments, and benchmark, we use data that includes individual attributes and annotations of the main character's facial changes, with subjective inferences removed.

Dataset Attributes

As mentioned earlier, our data can be categorized into two types based on the source: existing data and self-collected data. Here, we analyze the properties of these two types of data. Comparative analyses with other datasets are provided in the supplementary materials. In Figure 2b, we present the statistics of the number of annotated tokens in the dataset. It can be observed that the average number of tokens in the existing data is 134.7, while the average number of tokens in the self-collected data is 125.3.

In Figure 2c, we present the statistics of the average duration of videos. The average duration of videos from the existing data is 3.13 seconds, while the average duration of videos from the self-collected data is 4.22 seconds. This video length is efficient for capturing subtle facial changes, reducing the computational burden of data processing, facilitating annotation and model training, and making it suitable for real-time applications and emotion recognition.

We use ChatGPT (OpenAI 2023) to extract facial action-related descriptors from the annotations, such as "tilted head", "eyes widened", "smile", etc. Then, we counted the number of keywords in each video to illustrate the level of detail in our annotations regarding facial actions. As shown in Figure 2d, the average number of facial action keywords in the existing data is 10.9, while 9.1 in the self-collected data.

Temporal Event Matching

Traditional evaluation metrics based on n-grams, such as CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), struggle when assessing long video descriptions. This is because there are numerous ways to convey the same meaning in extended text, and many of these equivalent descriptions may have minimal n-gram overlap with the reference text. Moreover, manually assigning quality scores to descriptions is both labor-intensive and subjective. To address these issues, Maaz et al. (Maaz et al. 2024) propose using Chat-

GPT (OpenAI 2023) to rate descriptions on a scale from 1 to 5. However, the specific meanings of these ratings are ambiguous, and the ratings themselves have not been standardized.

Recently, Tarsier (Wang et al. 2024) introduced the AutoDQ metric which extracts text descriptions into a set of events and then calculates precision and recall by evaluating the relationships between the generated text and the reference event set, and vice versa, ultimately deriving the F-measure. While AutoDQ makes significant progress in capturing the semantic consistency and completeness of events, it does not account for the order of events. In many applications, especially those involving temporal and causal relationships, the order of events is crucial. Incorrect event ordering can lead to logically incoherent text, affecting the overall quality. To overcome these limitations, we propose a new evaluation metric named Temporal Event Matching (TEM) that combines event extraction, relation classification, and the longest common subsequence (LCS) algorithm to assess the semantic consistency and event ordering of generated text. TEM calculates the normalized LCS score, and then averages it with the F-measure (Wang et al. 2024) of generated events and reference events to provide a comprehensive evaluation result. By integrating these components, TEM ensures that the generated text not only captures the correct events and their relationships but also maintains the correct order and coherence, providing a more holistic assessment standard.

As shown in Algorithm 1, our evaluation metric calculation is structured into five key steps. In the first step, drawing on the methodology of AutoDQ, we utilize the ChatGPT model to extract a set of events, denoted as E_G from the generated text and E_R from the reference text, respectively. This initial extraction process aims to identify and isolate the critical events within both texts, setting the stage for subsequent comparison and assessment.

In the second step, we leverage ChatGPT once more to evaluate the relationship between each event in E_G and each event in E_R. These relationships are categorized into three primary types: Same Meaning, Opposite Meaning, and No Relation. By classifying these relationships, we can achieve a more precise understanding of the semantic alignment between the generated and reference texts.

Thirdly, to quantify the similarity between the two sets of events, we employ the Longest Common Subsequence (LCS) algorithm, a dynamic programming technique that efficiently determines the proportion of the longest subsequence of E_R found within E_G. This ratio, referred to as the LCS score, serves as an indicator of the consistency between the generated and reference texts at the event level.

In the fourth step, we calculate the F-measure (Wang et al. 2024), a composite metric that balances precision and recall, offering a comprehensive view of performance. Finally, we integrate the LCS score with the F-measure to derive a combined score, which stands as our ultimate evaluation criterion for the quality of the generated text. The specific prompts are detailed in the supplementary materials.

Algorithm 1: Temporal Event Matching

Require: Generated Text G, Reference Text R
Ensure: Evaluation Score S

Event Extraction: Extract events from G and R to get E_G and E_R
Relation Classification: Classify relations in E_G and E_R
function LCS-SCORE(E_G, E_R)
    Initialization: Initialize $m \leftarrow |E_G|$, $n \leftarrow |E_R|$
    Array Creation: Create a 2D array L of size $(m+1) \times (n+1)$
    for $i \leftarrow 0$ to m **do**
        Column Initialization: $L[i][0] \leftarrow 0$
    end for
    for $j \leftarrow 0$ to n **do**
        Row Initialization: $L[0][j] \leftarrow 0$
    end for
    for $i \leftarrow 1$ to m **do**
        for $j \leftarrow 1$ to n **do**
            if $E_G[i-1] == E_R[j-1]$ **then**
                Match Found: $L[i][j] \leftarrow L[i-1][j-1]+1$
            else
                No Match: $L[i][j] \leftarrow \max(L[i-1][j], L[i][j-1])$
            end if
        end for
    end for
    Return Normalized LCS Score: return $\frac{L[m][n]}{m}$
end function
Calculate LCS Score: $lcs \leftarrow$ LCS-SCORE(E_G, E_R)
F-Measure Calculation: Calculate precision, recall, and F-measure using E_G and E_R
Comprehensive Evaluation Score: $S \leftarrow \frac{lcs+F\text{-measure}}{2}$
Return Final Score: return S

Method

In this section, we first introduce the motivation behind our work. We then describe the architecture of our model and finally detail the instruction tuning process.

Motivation

Since existing video MLLMs primarily focus on understanding the general content and narrative progression of videos, their ability to encode detailed facial information is limited. Different from directly using existing facial prior expert to provide facial feature encoding for large language models (Li et al. 2024c), we propose a method that accurately models facial regions in videos with specific face tracking module, resulting in precise facial description results.

Architecture

Overall As shown in Figure 3, our model primarily consists of a pre-trained visual encoder, a visual projector, a dynamic face extraction module, an auxiliary facial visual encoder, and a large language decoder. Specifically, during

the training process, for the input video, we first use a pre-trained image encoder to extract features from the selected frames. Then, we use the STC (Cheng et al. 2024) module to map these features to the text domain. Additionally, for the input video, we use the dynamic face extraction module to extract facial information of the main characters. These facial features are subsequently processed by an auxiliary facial visual encoder to extract visual features. These visual features are also mapped to the text domain. Finally, the visual features, prior features, and text features are combined and fed into the large-scale language model for text generation.

The pre-trained visual encoder, visual projector, and large-scale language model all follow the design of VideoLLaMA2 (Cheng et al. 2024). The visual encoder uses CLIP-ViT-Large (Radford, Kim et al. 2021), the visual projector is the STC module, and the large-scale language model employs Mistral-7B-Instruct (Jiang et al. 2024). We use the FaceXFormer (Narayan et al. 2024) architecture, designed for face analysis, as the auxiliary facial visual encoder.

Dynamic Video Face Tracking Module Compared to directly using a face feature extractor, which may result in multiple faces being detected and faces mismatching among frames, we incorporate the Face Detection and Multi-Object Tracking techniques to consistently and accurately track facial features of the main characters across the video, which can be divided into the following steps:

Face Feature Extraction via DaMOFD (Liu et al. 2023d) & TransFace (Dan et al. 2023). We first downsample the video to a uniform frame rate of 16 fps and get the raw video frames $\{f_t\}_{t=1}^{T}$. Subsequently, we perform face keypoint detection approach DaMOFD (Liu et al. 2023d) and face feature extraction approach TransFace (Dan et al. 2023) on each raw video frame in (1).

$$\begin{cases} \mathcal{P}_t = \text{DaMOFDx}(f_t') & \text{(landmark detection)} \\ \phi_t = \text{TransFace}(f_t') & \text{(feature extraction)} \end{cases}, \quad (1)$$

where $\mathcal{P}_t \in \mathbb{R}^{68\times 2}$ denotes the extracted facial landmarks, and $\phi_t \in \mathbb{R}^d$ represents facial features.

Multi-Object Tracking via StrongSORT (Du et al. 2023). Equipped with $\mathcal{P} = \{\mathcal{P}_t\}_{t=1}^{T}$ and $\phi = \{\phi_t\}_{t=1}^{T}$, we then applied the StrongSORT (Du et al. 2023) multi-object tracking algorithm to extract target trajectories.

$$\mathcal{T} = \{\tau_k\}_{k=1}^{K} = \text{StrongSORT}(\mathcal{P}, \phi), \quad (2)$$

where K is the number of target trajectories. The StrongSORT algorithm provides robust trajectory information; each trajectory includes the position, area, and motion path of each target.

K-means-based Main Trajectory Selection. We assume that the main characters in a video would occupy a relatively larger area and remain consistent throughout a trajectory. Therefore, we first calculate the average area and the average cosine similarity as attributes for all frames in each trajectory $\tau_k \in \mathcal{T}$:

$$\begin{cases} A_k = \frac{1}{n}\sum_{i=1}^{n} \text{Area}_i \\ S_k = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \cos(\phi_i, \phi_j) \end{cases} \quad (3)$$

where n is the number of frames in the trajectory τ_k. The set of attributes of $\mathcal{T}$ can be denoted as $\mathcal{A} = \{(A_k, S_k)\}_{k=1}^{K}$. We then perform K-means clustering with 2 cluster centers on $\mathcal{T}$ according to $\mathcal{A}$, getting the main character cluster $\mathcal{T}^m$ and the background character cluster $\mathcal{T}^b$, respectively.

Through this process, we can obtain one or multiple main facial trajectories. To ensure that the extracted video frames contain the faces of the main characters while maintaining as even a distribution as possible, during the frame extraction process, if a frame does not contain the face of a main character, we replace it with the nearest frame that does contain the face of a main character. Finally, with the main facial trajectories $\mathcal{T}^m$, we further leveage the FaceXFormer (Narayan et al. 2024) to extract facial prior tokens.

Instruction Tuning

Since our data is entirely manually annotated, it is of high quality but relatively limited in quantity. We fine-tuned the model using LoRA (Hu et al. 2022) on the VideoLLaMA2-7B (Cheng et al. 2024) base model. We generated 100 similar instructions using ChatGPT and manually selected 20 of them, some of which are shown in the supplementary material. For each annotated video, we randomly selected one of the 20 instructions to form the instruction data.

Experiments

Implementation Details

Our code is based on VideoLLaMA2 (Cheng et al. 2024). The LoRA (Hu et al. 2022) parameters we used are set as follows: lora_r is set to 64, lora_alpha is set to 128. The batch size is set to 32, and we train for 3 epochs. Each video is sampled to 16 frames. During training, we use bfloat16 precision to improve computational efficiency and model performance. The entire training process is conducted on 8 NVIDIA A100 GPUs.

FEC-Bench

We randomly selected 1,000 samples from our constructed dataset to form the test set. In addition to the metrics we proposed, we employed a variety of evaluation metrics, including four ChatGPT-related metrics (Maaz et al. 2024), the n-gram based metric CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and AutoDQ (Wang et al. 2024), to comprehensively assess the performance of a large number of video MLLMs, including proprietary models. Due to the lack of instruction tuning specifically on our dataset, the generated text does not always match the format of the reference text. This discrepancy leads to biased results when using ChatGPT for evaluation. Therefore, before evaluating the generated text using ChatGPT, we first use ChatGPT to convert the generated text into a similar format as the reference text without altering its content. The specific prompts used and the effects before and after conversion are detailed in the supplementary materials. This extensive benchmarking not only provides a robust comparison framework but also aids in identifying the strengths and weaknesses of each model. The detailed results of this comprehensive evaluation

Method	LLM Size	VideoChatGPT Scores				N-gram Based		Event Based	
		Correctness	Detail	Context	Temporal	CIDEr	Rouge-L	AutoDQ	TEM
GPT4-O (OpenAI 2024)	-	4.18	3.98	4.50	3.92	0.225	0.179	0.410	0.291
GPT4-O* (OpenAI 2024)	-	4.22	3.97	4.48	3.90	0.264	0.213	0.432	0.303
Claude3.5-Sonnet (Anthropic 2024)	-	4.11	3.97	4.41	3.85	0.212	0.197	0.420	0.298
Claude3.5-Sonnet* (Anthropic 2024)	-	4.13	4.01	4.49	4.05	0.243	0.228	0.442	0.307
VideoLLaMA (Zhang, Li, and Bing 2023)	7B	3.60	3.67	3.84	3.50	0.189	0.196	0.303	0.199
VideoChat (Li et al. 2023b)	7B	3.47	3.52	3.92	3.38	0.251	0.192	0.344	0.229
VideoChat2 (Li et al. 2023b)	7B	3.70	3.56	4.16	3.52	0.202	0.229	0.311	0.231
Chat-UniVI (Jin et al. 2023)	7B	3.64	3.63	4.21	3.61	0.189	0.231	0.396	0.261
LLaVA-Next-Video (Zhang et al. 2024)	7B	4.19	4.07	4.39	4.04	0.250	0.249	0.395	0.276
ShareGPT4Video (Chen et al. 2024b)	7B	4.24	4.13	4.35	4.09	0.192	0.205	0.394	0.278
LLaMA-VID (Li, Wang, and Jia 2024)	7B	3.95	4.01	4.22	3.71	0.195	0.231	0.339	0.241
VideoLLaMA2 (Cheng et al. 2024)	7B	4.17	4.02	4.47	3.93	0.253	0.266	0.344	0.258
PLLaVA (Xu et al. 2024)	7B	4.21	4.15	4.37	4.08	0.268	0.250	0.393	0.257
ST-LLM (Liu et al. 2023a,b)	7B	4.00	3.98	4.31	3.94	0.213	0.238	0.321	0.240
Qwen2-VL (Wang, Bai et al. 2024)	7B	4.23	4.16	4.52	4.02	0.204	0.233	0.422	0.309
Tarsier (Wang et al. 2024)	7B	3.59	3.50	4.07	3.41	0.143	0.185	0.415	0.292
LLaVA-OneVision (Li et al. 2024a)	7B	3.68	3.47	4.10	3.42	0.115	0.165	0.379	0.275
Ours	**7B**	**4.42**	**4.30**	**4.60**	**4.26**	**0.418**	**0.473**	**0.483**	**0.364**

Table 1: Comparison of different methods on average ChatGPT scores (Correctness, Detail, Context, Temporal, Consistency) and n-gram based metrics (CIDEr, ROUGE-L), event-based metrics (AutoDQ, our proposed TEM) for 1000 videos, categorized by model type (proprietary vs. open-source). * denotes the use of in-context learning during evaluation.

are presented in Table 1. We observe that even with carefully tuned prompts and context learning, existing proprietary large-scale models and open-source community large-scale models still struggle to achieve excellent performance in the task of video facial expression description. In contrast, our method, after additional modeling of the main character's facial information and instruction tuning on DFEC, significantly outperforms all other methods on the FEC-Bench.

Ablation Study

In Table 2, we conduct experiments to evaluate different designs of facial feature extraction modules, with VideoLLaMA2 as our baseline. First, we performed instruction tuning on our training set using the original VideoLLaMA2 model, which yielded a noticeable improvement—approximately a 7 percent gain—on two event-based evaluation metrics, demonstrating the value of task-specific adaptation. Next, we used RetinaFace to extract facial regions from each frame and concatenated the detected faces as part of the input; when more than two faces appeared in a single frame, we randomly selected two for concatenation to keep the input dimension manageable. These face crops were then encoded by a CLIP model to obtain visual features, but this variant showed almost no improvement over the tuned baseline. Replacing RetinaFace with our dynamic face tracking module led to clear and consistent gains across all metrics, indicating that temporally coherent face selection better supports expression understanding. Finally, substituting the general-purpose CLIP encoder with FaceXFormer—a specialized facial feature extractor designed for expressive and contextual modeling—further boosted performance and achieved the best results among all tested variants.

Method	VC Avg.	AutoDQ	TEM
VideoLLaMA2 (Cheng et al. 2024)	4.15	0.344	0.258
Instruction Tuning	4.28	0.422	0.321
with RetinaFace Face Det	4.31	0.426	0.324
with Dynamic Face Tracking	4.37	0.471	0.357
Ours	**4.40**	**0.483**	**0.364**

Table 2: Ablation Study on Dynamic Face Tracking Module. VC Avg. represents the average VideoChatGPT score.

Conclusion

In this paper, we tackle the challenges of dynamic facial expression captioning using video Multimodal Large Language Models (MLLMs). We introduce a new instruction-following dataset containing 5,033 high-quality video clips with over 700,000 tokens, designed to enhance MLLMs' ability to capture subtle facial differences. We also propose FaceTrack-MM, a model that efficiently tracks and focuses on the main character's facial expressions in complex scenes using a limited number of tokens. Additionally, we develop a novel evaluation metric that combines event extraction, relation classification, and the Longest Common Subsequence (LCS) algorithm to assess the quality of generated captions. Finally, we present FEC-Bench, a benchmark for evaluating MLLMs performance in facial expression captioning task. Our results demonstrate significant improvements in capturing and describing dynamic facial expressions, contributing to the advancement of this field and providing valuable resources for future research.

References

Anthropic. 2024. Claude-3.5.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Chen, H.; Huang, H.; Dong, J.; Zheng, M.; and Shao, D. 2024a. FineCLIPER: Multi-modal Fine-grained CLIP for Dynamic Facial Expression Recognition with AdaptERs. *arXiv preprint arXiv:2407.02157*.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; et al. 2024b. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. *arXiv preprint arXiv:2406.04325*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Muyan, Z.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Dan, J.; Liu, Y.; Xie, H.; Deng, J.; Xie, H.; Xie, X.; and Sun, B. 2023. TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective. arXiv:2308.10133.

Ding, Z.; Zhang, C.; Xia, Z.; Jebe, L.; Tu, Z.; and Zhang, X. 2023. DiffusionRig: Learning Personalized Priors for Facial Appearance Editing. In *CVPR*.

Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; and Meng, H. 2023. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*.

Foteinopoulou, N. M.; and Patras, I. 2024. EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition. In *IEEE FG*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*.

Gómez-Cañón, J. S.; et al. 2021. Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Processing Magazine*, 38: 106–114.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, G.; Pang, B.; Zhu, Z.; Rivera, C.; and Soricut, R. 2020. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*.

Islam, M. M.; Ho, N.; Yang, X.; Nagarajan, T.; Torresani, L.; and Bertasius, G. 2024. Video ReCap: Recursive Captioning of Hour-Long Videos. *arXiv preprint arXiv:2402.13250*.

Jiang, A. Q.; et al. 2024. Mixtral of Experts. arXiv:2401.04088.

Jiang, X.; Zong, Y.; Zheng, W.; Tang, C.; Xia, W.; Lu, C.; and Liu, J. 2020. DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild. arXiv:2008.05924.

Jin, P.; Takanobu, R.; Zhang, C.; Cao, X.; and Yuan, L. 2023. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv preprint arXiv:2311.08046*.

Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; and Pantic, M. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65: 23–36.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*, 706–715.

Lee, J.; Kim, S.; Kim, S.; Park, J.; and Sohn, K. 2019. Context-aware emotion recognition networks. In *ICCV*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, M.; Zhang, X.; Fan, C.; Liao, T.; and Xiao, G. 2024b. Dual-STI: Dual-path Spatial-Temporal Interaction Learning for Dynamic Facial Expression Recognition. *Information Sciences*, 120953.

Li, Y.; Dao, A.; Bao, W.; Tan, Z.; Chen, T.; Liu, H.; and Kong, Y. 2024c. Facial Affective Behavior Analysis with Instruction Tuning. arXiv:2404.05052.

Li, Y.; Wang, C.; and Jia, J. 2024. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models.

Li, Z.; Cao, M.; Wang, X.; et al. 2024d. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. In *CVOR*.

Liu, R.; Li, C.; Ge, Y.; Shan, Y.; Li, T. H.; and Li, G. 2023a. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*.

Liu, R.; Li, C.; Tang, H.; Ge, Y.; Shan, Y.; and Li, G. 2023b. ST-LLM: Large Language Models Are Effective Temporal Learners. *https://arxiv.org/abs/2404.00308*.

Liu, Y.; Dai, W.; Feng, C.; Wang, W.; Yin, G.; Zeng, J.; and Shan, S. 2023c. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. arXiv:2208.00847.

Liu, Y.; Deng, J.; Wang, F.; Shang, L.; Xie, X.; and Sun, B. 2023d. DamoFD: Digging into Backbone Design on Face Detection. In *The Eleventh International Conference on Learning Representations*.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*.

Lucey, P.; Cohn, J. F.; et al. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 94–101.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2024. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NIPS*, 36.

Menefee DS, J. C., Ledoux T. 2022. The Importance of Emotional Regulation in Mental Health. *American Journal of Lifestyle Medicine*.

Narayan, K.; VS, V.; Chellappa, R.; and Patel, V. M. 2024. FaceXFormer: A Unified Transformer for Facial Analysis. *arXiv preprint arXiv:2403.12960*.

Ning, M.; Salah, A. A.; and Ertugrul, I. O. 2024. Representation Learning and Identity Adversarial Training for Facial Behavior Understanding. arXiv:2407.11243.

OpenAI. 2023. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI. 2024. GPT-4O system card.

Özacar, K.; and Alkhalıfa, M. 2024. DigiHuman: A Conversational Digital Human with Facial Expressions. *Turkish Journal of Science and Technology*, 19(1): 25–37.

Radford, A.; Kim, J. W.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.; Larochelle, H.; Courville, A.; and Schiele, B. 2017. Movie description. *IJCV*, 123: 94–120.

Shao, Z.; Liu, Z.; Cai, J.; and Ma, L. 2018. Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment. In *European Conference on Computer Vision*, 725–740. Springer.

Team, G. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Wadley, G.; Kostakos, V.; et al. 2022. The Future of Emotion in Human-Computer Interaction. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*.

Wang, H.; Tang, P.; Li, Q.; and Cheng, M. 2022a. Emotion Expression With Fact Transfer for Video Description. *IEEE Transactions on Multimedia*, 24: 715–727.

Wang, J.; Yuan, L.; Zhang, Y.; and Sun, H. 2024. Tarsier: Recipes for Training and Evaluating Large Video Description Models. arXiv:2407.00634.

Wang, P.; Bai, S.; et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 4581–4591.

Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; and Zhang, W. 2022b. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos. arXiv:2203.09463.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. arXiv:2404.16994.

Xue, L.; et al. 2024. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models. arXiv:2408.08872.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2023b. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. arXiv:2311.04257.

Yuan, K.; Yu, Z.; Liu, X.; Xie, W.; Yue, H.; and Yang, J. 2024. AUFormer: Vision Transformers are Parameter-Efficient Facial Action Unit Detectors. In *ECCV*. Springer.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, Q.; Zhang, J.; Xu, Y.; and Tao, D. 2023. Vision Transformer with Quadrangle Attention. *arXiv preprint arXiv:2303.15105*.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; and PietikäInen, M. 2011. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9): 607–619.

Zhao, J.; Sun, B.; Chen, X.; et al. 2025. LLaVA-Octopus: Unlocking Instruction-Driven Adaptive Projector Fusion for Video Understanding. *arXiv preprint arXiv:2501.05067*.

Zhao, J.; Yang, Q.; et al. 2025. HumanOmni: A Large Vision-Speech Language Model for Human-Centric Video Understanding. *arXiv preprint arXiv:2501.15111*.

Zhao, K.; Chu, W.-S.; and Zhang, H. 2016. Deep Region and Multi-Label Learning for Facial Action Unit Detection. In *CVPR*.

Zhou, J.; Shu, Y.; et al. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264*.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32.