MicroEmo: Time-Sensitive Multimodal Emotion Recognition with Subtle Clue Dynamics in Video Dialogues

Liyun Zhang
Institute for Datability Science, Osaka University
Osaka, Japan
liyun.zhang@lab.ime.cmc.osaka-u.ac.jp

Zhaojie Luo
School of Biological Science and Medical Engineering,
Southeast University
Nanjing, China
luozhaojie@seu.edu.cn

Shuqiong Wu
SANKEN, Osaka University
Osaka, Japan
wu@am.sanken.osaka-u.ac.jp

Yuta Nakashima[†]
Institute for Datability Science, Osaka University
Osaka, Japan
n-yuta@ids.osaka-u.ac.jp

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable multimodal emotion recognition capabilities, integrating multimodal cues from visual, acoustic, and linguistic contexts in the video to recognize human emotional states. However, existing methods have limitations. They neglect to capture the dynamics of local subtle clues in facial features and also do not leverage the contextual dependencies of the utterance-aware temporal segments in videos. These shortcomings somewhat restrict their expected effectiveness. In this work, we propose MicroEmo, a time-sensitive MLLM aimed at directing attention to the local facial clue dynamics and the contextual dependencies of utterance-aware video clips. Our model incorporates two key architectural contributions: (1) a global-local attention visual encoder that integrates global frame-level timestamp-bound image features with local facial features of temporal dynamics of subtle clues; (2) an utterance-aware video Q-Former that captures multi-scale and contextual dependencies by generating visual token sequences for each utterance segment and for the entire video then combining them. The experiments demonstrate that in a new Explainable Multimodal Emotion Recognition (EMER) task that exploits multi-modal and multi-faceted clues to predict emotions in an open-vocabulary (OV) manner, MicroEmo demonstrates its effectiveness compared with the latest methods.

CCS Concepts

• **Computing methodologies → Machine learning**; • **Information systems → Multimedia information systems**.

Keywords

Multimodal open-vocabulary emotion recognition challenge (MER-OV 2024), Multimodal large language models (MLLMs), Emotional clue feature fusion, Utterance-aware dependency.

1 Introduction

Multimodal Large Language Models (MLLMs) have shown exceptional capabilities in integrating multimodal information (e.g., visual, acoustic, and linguistic modalities) and understanding the context to generate text descriptions [30]. Many MLLMs with outstanding performance can be adapted to different tasks by tuning on specific datasets. However, there still remains a gap in the expected effectiveness. This is because video understanding requires a full perception and deep reasoning of visual/acoustic details and temporal dynamics for models [7, 34], particularly in the multimodal emotion recognition task.

For example, in a new Explainable Multimodal Emotion Recognition (EMER) task [17] that exploits multi-modal and multi-faceted clues to predict emotions in an open-vocabulary (OV) manner. This not only requires the integration of multimodal information but also demands the in-depth capture of emotion-related subtle information. Emotion is a highly sensitive state form that undergoes dynamic changes. The emotions in each utterance of the video evolve dynamically over time, which is crucial information that cannot be overlooked. It necessitates the design of specialized modules within the model to capture and leverage these elements, enabling the provision of detailed recognition of subtle emotional changes without compromising contextual accuracy. However, current existing methods ignore capturing local facial features of temporal dynamics of subtle clues and do not leverage the contextual dependencies of the utterance-aware temporal segments in the video, thereby limiting their expected effectiveness to a certain extent when tuning MLLMs to Multimodal Emotion Recognition tasks.

To this end, we propose MicroEmo, a novel time-sensitive MLLM aimed at directing attention to the local facial clue dynamics and the contextual dependencies of utterance-aware video clips. MicroEmo incorporates two key modules, which consists of i) a global-local attention visual encoder that first extracts and integrates global frame-level timestamp-bound image features with local facial features of temporal dynamics of subtle clues, and ii) an utterance-aware video Q-Former that that models each utterance time segment feature via an utterance-aware sliding window and combines it with whole features to produce multi-scale fused tokens for the image.

Benefiting from the attention to local facial features of temporal dynamics of subtle clues and fully capturing multi-scale and contextual dependencies of utterances in videos, MicroEmo shows its superiority in capturing more subtle and deeper emotional information to enable the provision of detailed recognition of subtle emotional changes without compromising contextual accuracy. Finally, audio is also processed by a pre-trained encoder and whole audio Q-Former to obtain audio tokens, these multimodal tokens are concatenated with transcribed speech and instructions to feed into the LLM for result generation. Our contributions to this paper are listed as follows:

- We propose a novel global-local attention visual encoder, which integrates global frame-level timestamp-bound image features with local facial features of temporal dynamics of subtle clues. This facilitates the model to capture more subtle and deeper emotional information to output richer and more reliable results.
- We propose a novel utterance-aware video Q-Former, which captures multi-scale and contextual dependencies by generating visual token sequences for each utterance segment and for the entire video then combining them. This ensures the accuracy and consistency of contextual dependencies when the model provides detailed recognition of subtle emotional changes.
- We conduct preliminary qualitative experiments to demonstrate the effectiveness of MicroEmo compared with the latest methods in the Explainable Multimodal Emotion Recognition (EMER) task.

2   Related works

Multimodal Large Language Model (MLLM) has attracted much attention due to its adaptability to adapt to different tasks by tuning on specific datasets, e.g., tuning on EMER [17] for multimodal emotion recognition. Numerous studies have endeavored to integrate LLMs with a video encoder for video understanding [19, 22, 25, 26]. They typically employ open-source LLMs [1, 29] to encode the video into vision tokens compatible with the LLMs. Such as VideoChat [12] uses a video transformer to encode video features and then implements a Query Transformer (Q-Former) [11] for video tokens. Video-llama [34] uses a vision transformer (ViT) [4] with an image Q-Former to encode frames and employs a video Q-Former for temporal modeling. TimeChat [25] incorporates frame timestamps to make the model time-sensitive and uses a video Q-Former to establish inter-frame temporal dependencies.

However, these methods ignore capturing local features of temporal dynamics and do not leverage the contextual dependencies of

important segments in videos, thereby limiting their expected effectiveness to a certain extent. In contrast, our MicroEmo proposes a global-local attention visual encoder and an utterance-aware video Q-Former to capture subtle local features of temporal dynamics and contextual dependencies in videos, such as utterance segment context of emotion recognition, enabling the provision of subtle reasoning results without compromising contextual accuracy.

3   Methodology

We present MicroEmo, a time-sensitive MLLM featuring two novel modules: a global-local attention visual encoder and an utterance-aware video Q-Former. These modules enable MicroEmo to capture more subtle and deeper emotional information in videos. We verify the effectiveness of our model on the new task EMER and the new style dataset EMER-Fine [17] it pioneered.

3.1   Overview

MicroEmo is mainly composed of a global-local attention visual encoder, an utterance-aware video Q-Former for image, and a large language model, as depicted in Figure 1. Given an input video, the global-local attention visual encoder first extracts and integrates global frame-level timestamp-bound image features with local facial features of temporal dynamics of subtle clues. Next, the utterance-aware video Q-Former models each utterance time segment feature via an utterance-aware sliding window and combines it with whole features to produce multi-scale fused tokens for the image. Audio is also passed through a pre-trained encoder and whole audio Q-Former to obtain audio tokens. Finally, these multimodal tokens are concatenated with transcribed speech and instructions, which are then fed into the LLM to generate responses.

3.2   Global-Local Attention Visual Encoder

First, for the global frame-level features, we use the time-aware frame encoder inspired by InstructBLIP [3] and Timechat [25] as shown in Figure 1. Given an input video, we use a pre-trained ViT [27] to encode each frame and obtain frame features. Subsequently, we use the keypoint indices of the MediaPipe [21] face mesh via extracting landmarks of each frame to define different facial regions as subtle emotional clue regions, which are predefined on facial anatomy. Then, the patches are combined to create masks associated with subtle emotional clue regions. we implemented a local clue attention mechanism based on the generated mask. This allows each patch to selectively attend to pertinent local emotional clue regions, facilitating the extraction of fine-grained facial local features. By emphasizing spatially relevant areas, we enhance the model's capacity to discern subtle facial emotional clues, which are crucial for accurate emotion recognition.

Furthermore, the image Q-Former compresses the frame tokens. The Q-Former takes learnable queries as input, which interacts with the frame features and facial local features via cross-attention and updates the initial queries respectively. The global and local results are fused and further processed to obtain final visual tokens [11]. During visual token extraction, we add the frame-level timestamp, e.g., `"This frame is sampled at 0.0s."`, as a condition to the Q-Former to fuse the visual and timestamp information for binding temporal dynamics to the visual tokens.
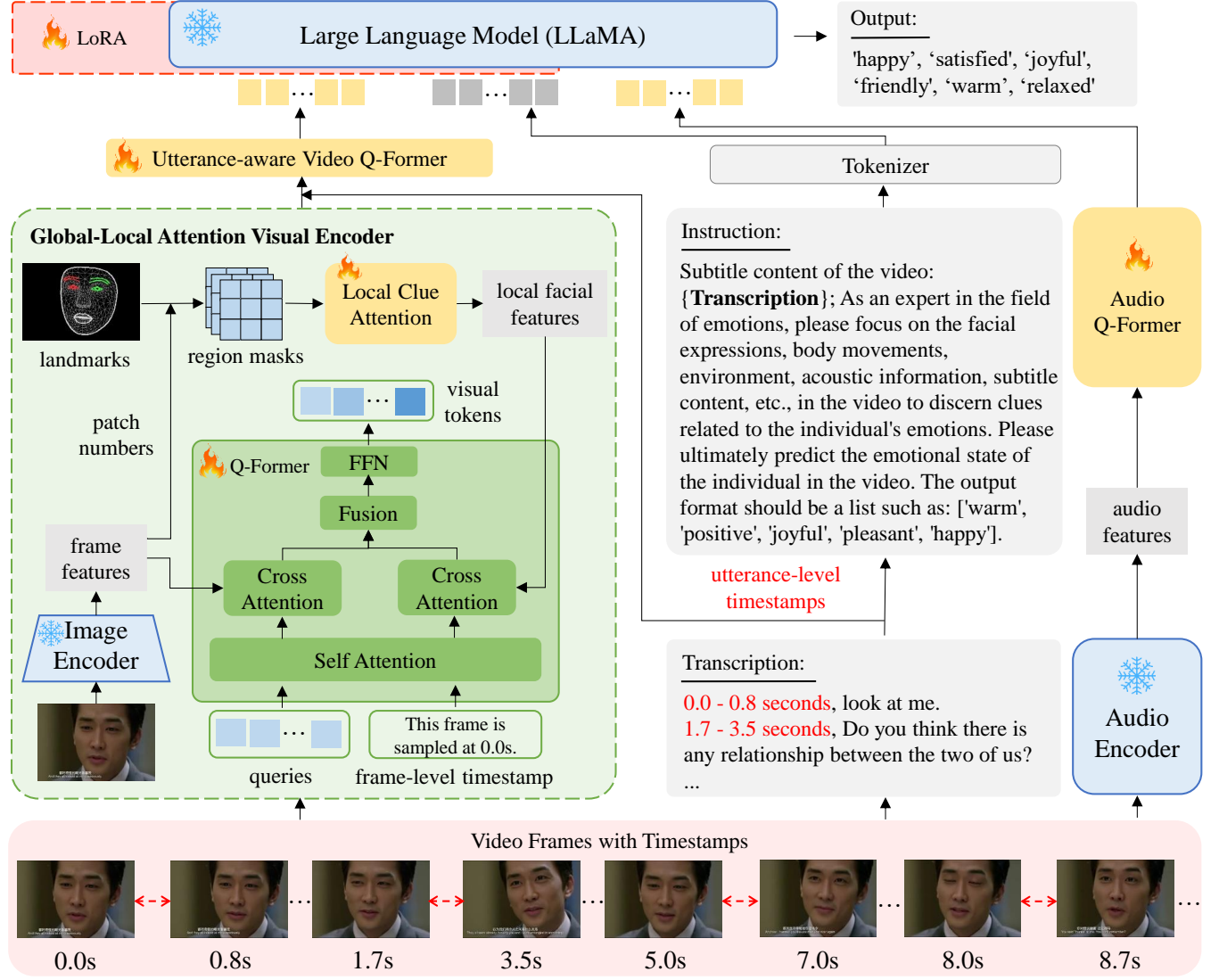
Figure 1: The overall architecture of MicroEmo. Input a sequence of video frames along with their timestamps, (a) Global-local attention visual encoder first extracts and integrates global frame-level timestamp-bound image features with local facial features of temporal dynamics of subtle clues. Then (b) utterance-aware video Q-Former combines utterance-level timestamps to generate visual tokens for each utterance segment via an utterance-aware sliding window and combines them with whole tokens extracted by a global video Q-Former to produce multi-scale fused visual tokens. Audio is also passed through a pre-trained encoder and whole audio Q-Former to obtain audio tokens. Finally, these multimodal tokens are concatenated with transcribed speech and instructions, which are then fed into a (c) Large Language Model to generate responses.

The proposed global-local attention visual encoder can fully integrate global timestamp-bound frame features with local facial features of temporal dynamics of subtle clues. Benefiting from the attention to temporal dynamics of subtle clues of the videos, MicroEmo shows its superiority in capturing more subtle and deeper emotional information to facilitate richer and more reliable emotion recognition results.

3.3 Utterance-Aware Video Q-Former

Since frames are encoded independently, we can use a sliding video Q-Former to model temporal relationships across frames and enhance the feature fusion in the temporal dimension [25]. Here, we propose a novel utterance-aware video Q-Former (yellow block in Figure 1), which combines utterance-level timestamps to design a dynamic sliding window with a length of each utterance segment and within each window utilizing the video Q-Former extract video

Liyun Zhang, Zhaojie Luo, Shuqiong Wu, and Yuta Nakashima[†]

tokens from frame features. Meanwhile, we also extract video tokens from whole frame features via a global video Q-Former and then combine them to produce multi-scale fused tokens.

It is crucial to capture the context of each conversation segment, which is the key to the model's understanding of emotional fluctuations and changes. We use this way to fully capture multi-scale and contextual dependencies, maintaining the accuracy and consistency of contextual dependencies when the model provides detailed recognition of subtle emotional changes. Notably, the method maintains its effectiveness even in videos with minimal or no dialogue content, thus ensuring robustness across diverse types of video material. Finally, we use a linear layer to transform the dimension of video tokens to match the dimension of the LLM embedding space.

3.4 Large Language Model

We concatenate inputs from different modalities, i.e., the video tokens T_V, audio tokens T_A, and text query tokens T_{L_q} including transcribed speech and instruction. They are fed into a large language model to generate reasonable and expected responses T_{L_a}. They have the same token embedding dimension. The training of the MLLM typically utilizes a two-stage training framework. The first stage pre-trains the model using large-scale image-audio-text pairs for vision-audio-language mapping [14, 20, 24, 31]. The second stage finetunes the model with instruction data for instruction following. Considering computing efficiency, we reuse the checkpoints of the existing open-source models after the first stage of training, e.g., the vision-language and audio-language branches from [34], conducting only instruction tuning. During the training procedure, we utilize the language modeling loss for generating target answers T_{L_a} and Θ is the trainable parameter, which serves as the objective function:

$$L = -logP_\Theta(T_{L_a}|T_V, T_A, T_{L_q}) \quad (1)$$

To better adapt the LLM to multimodal video tasks, we apply the parameter-efficient fine-tuning method, i.e., LoRA [8] to train.

4 Experiments

4.1 Implementation Details

We use ViT-G/14 from EVA-CLIP [27] as the image encoder, Imagebind [5] as the audio encoder, and LLaMA-2 (7B) [30] as the language foundation model. The parameters of the image Q-Former are initialized from InstructBLIP's checkpoint, while the video Q-Former for vision and audio Q-Former for audio is initialized from Video-LLaMA's checkpoint. We finetune our MicroEmo on train-set of EMER-Fine [17] dataset for 3 epochs, using a batch size of 1, with a single NVIDIA A40 (49G) machine. The parameters of ViT, Imagebind and LLM are frozen, while those of image Q-Former, local clue attention module, video Q-Former, audio Q-Former and linear layer (not shown in the diagram due to space constraints) are tuned.

4.2 Experiment Setup

Tasks, Datasets and Evaluation Metrics We evaluate our model on a new Explainable Multimodal Emotion Recognition (EMER) task

[17], test-set of EMER-Fine as the test dataset, the new $Accuracy_S$ and $Recall_S$ in EMER as metrics based on 'gpt-3.5-turbo' API, and MLLM such as VideoChat [12], Video-LLaMA [34] and Chat-UniVi [9], etc. as baselines for evaluation.

Table 1: Main results on the test-set of EMER-Fine [17].

Model	Avg	Accuracy$_S$	Recall$_S$
Audio + Subtitle			
Qwen-Audio [2]	38.66	46.97	30.35
OneLLM [6]	40.56	42.55	38.56
SECap [32]	52.78	61.36	44.19
SALMONN [28]	51.28	54.17	48.38
Video + Subtitle			
Otter [10]	37.72	42.22	33.22
VideoChat [12]	46.34	41.49	51.19
Video-LLaMA [34]	40.97	39.44	42.50
Video-LLaVA [18]	42.75	45.38	40.13
VideoChat2 [13]	43.83	49.24	38.42
OneLLM [6]	53.40	58.65	48.14
LLaMA-VID [15]	50.90	51.69	50.11
mPLUG-Owl [33]	48.84	48.33	49.34
Video-ChatGPT [23]	46.12	50.00	42.25
Chat-UniVi [9]	53.20	54.29	52.11
Audio + Video + Subtitle			
SECap + mPLUG-Owl	64.42	57.95	70.90
SALMONN + Video-ChatGPT	59.71	53.48	65.93
SECap + Video-ChatGPT	58.43	51.60	65.26
SECap + Chat-UniVi	60.38	51.39	69.37
SALMONN + mPLUG-Owl	65.15	55.28	**75.03**
SALMONN + Chat-UniVi	62.64	55.84	69.44
AffectGPT [16]	61.75	62.03	61.46
Ours	**66.21**	**63.82**	68.59
EMER (Multi) [17]	79.31	80.91	77.70

4.3 Main Results

We evaluate our MicroEmo and some existing MLLMs for EMER task as shown in Table 1, where, the baseline methods except for AffectGPT [16] are zero-shot performance results. AffectGPT is only tuned on the EMER-Fine [17] dataset in the second stage without pre-training of the first stage, i.e., pre-training the model using large-scale image-audio-text pairs for vision-audio-language mapping, which is the same as our experimental setting. EMER (Multi) [17] can be viewed as a post-processed upper bound approximating the ground truth, serving as a reference point here. The experimental results indicate that our proposed model outperforms the baseline methods overall, demonstrating the superiority and effectiveness of our approach. Incorporating large-scale multimodal data alignment pre-training in the first stage is expected to result in further performance improvements.

Table 2: Ablation study for key modules.

Model	Avg	Accuracy$_S$	Recall$_S$
MicroEmo	66.21	63.82	68.59
w/o utterance-aware V-QF	56.01	52.53	59.49
w/o global-local AVE	54.57	50.43	58.71
w/o all proposed modules	46.21	49.75	47.98

4.4 Ablation Study

We conducted an ablation study to evaluate the key novel modules, i.e., the global-local attention visual encoder (AVE) and utterance-aware video Q-Former (V-QF). As shown in Table 2, removing either of these modules decreases the overall performance, highlighting their effectiveness and reasonability in the architecture.

5 Conclusion

We propose MicroEmo that directs attention to the local facial clue dynamics and fully captures multi-scale and contextual dependencies of utterances in videos, and tuning on EMER. MicroEmo demonstrates capabilities in capturing subtle emotion to output richer and more reliable results without compromising contextual accuracy. In the future, we will apply the utterance-aware video Q-Former to acoustic modality and integrate emotion in audio into multimodal emotion recognition task based on MLLM.

Acknowledgments

References

[1] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[2] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919* (2023).

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=vvoWPYqZJA

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition* (Oct 2020).

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, KalyanVasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. (May 2023).

[6] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26584–26595.

[7] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[8] HuEdward J., Yulong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv: Computation and Language,arXiv: Computation and Language* (Jun 2021).

[9] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13700–13710.

[10] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. arXiv:2305.03726 [cs.CV] https://arxiv.org/abs/2305.03726

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

[12] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

[13] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

[14] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. 2023. VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models. (Nov 2023).

[15] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043* (2023).

[16] Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024. AffectGPT: Dataset and Framework for Explainable Multimodal Emotion Recognition. *arXiv preprint arXiv:2407.07653* (2024).

[17] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401* (2023).

[18] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[19] Haogeng Liu, Qihang Fan, Tingkai Liu, Linjie Yang, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. 2023. Video-Teller: Enhancing Cross-Modal Generation with Fusion and Decoupling. (Oct 2023).

[20] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems* 36 (2024).

[21] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, MichaelD. Hays, Fan Zhang, C.K. Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Hua Wang, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *Cornell University - arXiv,Cornell University - arXiv* (Jun 2019).

[22] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video Assistant with Large Language model Enhanced abilitY. (Jun 2023).

[23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).

[24] Shuhuai Ren, Lei Li, Xuanchang Ren, Guangxiang Zhao, and Xu Sun. 2022. Delving into the Openness of CLIP. (Jun 2022).

[25] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14313–14323.

[26] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2023. 6MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. (Jul 2023).

[27] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* (2023).

[28] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289* (2023).

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[31] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.544

[32] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion

Liyun Zhang, Zhaojie Luo, Shuqiong Wu, and Yuta Nakashima[†]

captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19323–19331.

[33] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).

[34] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 543–553.