EmotionTalk: An Interactive Chinese Multimodal Emotion Dataset With Rich Annotations

Haoqin Sun[1†], Jinghua Zhao[1†], Xuechen Wang[1†], Shiwan Zhao[1*], Jiaming Zhou[1], Hui Wang[1], Xi Yang[2], Yequan Wang[2], Yonghua Lin[2]

[1] College of Computer Science, Nankai University,
[2] Beijing Academy of Artificial Intelligence, Beijing, China,
Correspondence: sunhaoqin@mail.nankai.edu.cn, zhaosw@gmail.com

Abstract

The advancement of Multimodal Emotion Recognition (MER) in Chinese is significantly hindered by the scarcity of high-quality, spontaneous dialogue datasets compared to their English counterparts. In this work, we introduce **EmotionTalk**, the first interactive Chinese multimodal dataset designed to capture the nuance of authentic emotional interplay. Collected from 19 professional actors, the dataset spans 23.6 hours of dyadic conversations across diverse scenarios. A key contribution of EmotionTalk is its multi-grained annotation system, which integrates standard categorical and dimensional labels with fine-grained emotional speaking style captions, enabling research into interpretable emotion analysis. We establish comprehensive benchmarks for emotion recognition and captioning tasks, verifying the dataset's effectiveness and the necessity of multimodal fusion. EmotionTalk serves as a critical resource for bridging the gap in non-English affective computing and is publicly released for the research community.

1 Introduction

Multimodal emotion recognition (MER) has become a key focus in artificial intelligence, integrating speech, vision, and text to capture the complexity of human emotions. It drives advancements in applications like virtual assistants, online education, and mental health monitoring. However, most research relies on English datasets, with Chinese resources remaining scarce. Existing datasets often face issues such as low quality, limited scale, and incomplete modalities, hindering model performance. Therefore, the development of a high-quality Chinese multimodal emotion recognition dataset is of critical importance to advance research in this field.

Traditional emotion recognition tasks include unimodal / multimodal emotion recognition on isolated utterances (Liu et al., 2022a, 2023; Sun et al., 2024a,b) and conversational emotion recognition (Shi et al., 2020, 2023). The former relies on a single modality or integrates multimodal information for emotion recognition. For example, MISA (Hazarika et al., 2020)utilizes modality-invariant and modality-specific representations to fuse multimodal information. DialogueRNN (Majumder et al., 2019) extracts emotional information from conversations by modeling the speaker, context, and emotions within the dialogue. With the in-depth development of emotion recognition research, researchers have gradually introduced emerging tasks such as emotion captioning (Xu et al., 2024; Liang et al., 2024), driven by evolving application scenarios and practical demands. This task is first proposed by SECap (Xu et al., 2024). Subsequently, this task has attracted increasing attention from researchers due to its unique value in interpretability, and has gradually become an important research direction in affective computing.

However, these studies use different datasets, and while they perform well in their respective experiments, directly comparing their performance remains challenging. This is mainly due to significant differences in dataset scale, annotation methods, modality combinations, and dialogue structures, which affect model applicability and generalization. For instance, popular multimodal benchmarks like IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), CMU-MOSEI (Zadeh et al., 2018b), and CH-SIMS (Yu et al., 2020) have been widely used but are primarily in English, with varying emotion category definitions and annotation standards, limiting cross-lingual and cross-cultural applicability. The underlying cause of this situation lies in the dilemma of data acquisition: on one hand, existing Chinese emotional data predominantly originates from film and television resources, which are relatively accessible but of low quality; on the other hand, compared to audiovisual mate-

*Shiwan Zhao is the corresponding author. † These authors contributed equally to the manuscript.

Findings of the Association for Computational Linguistics: ACL 2026, pages 9054–9071

rials, artificially recorded dialogue data can guarantee higher quality standards, thereby enabling the construction of datasets with greater academic research value. However, such high-quality controlled recording data is precisely what constitutes an extremely scarce resource at present. More seriously, emotion captioning research mostly relies on unpublished datasets, leading to a lack of standardized open benchmarks and further hindering research reproducibility and widespread application. Against the backdrop of current data scarcity, we deeply recognize that the importance of high-quality data has become increasingly prominent and cannot be overlooked.

To address these gaps, in this paper, we construct n large-scale interactive Chinese multimodal emotion dataset with fine-grained labels and emotional speaking style captions, **EmotionTalk**, in which the data are contributed by 19 professional actors, ensuring the naturalness and authenticity of the emotion expression. The dataset is in the form of dialogues, containing 23.6 hours of data and 19,250 utterances, along with corresponding labels that support various emotion tasks, including 7 discrete labels, 5 dimensional labels, and 4 caption labels. To the best of our knowledge, EmotionTalk is the first large-scale, comprehensive, recorded interactive Chinese multimodal emotion dataset. We further conduct experiments on unimodal emotion recognition, multimodal emotion recognition, and emotion caption tasks to validate the effectiveness and applicability of the constructed dataset. These experiments not only demonstrate the dataset's performance across different emotion tasks but also highlight its potential to support diverse model development and evaluation.

2 Related Work

Table 1 presents the datasets which are commonly used in the field of multimodal emotion recognition, all of which consist of video, audio and text modalities.

English Datasets: The CMU-MOSEI (Zadeh et al., 2018b) and MELD (Poria et al., 2019) datasets provide large-scale multimodal data sourced from YouTube and TV shows, covering tasks such as discrete emotion classification and continuous sentiment intensity prediction. These datasets are advantageous due to their rich emotional labeling, but they are primarily derived from entertainment content, where emotional expressions tend to be exaggerated. As such, they may not fully capture the natural emotional expressions encountered in real-life situations. In contrast, the CREMA-D (Cao et al., 2014), RAVDESS (Livingstone and Russo, 2018), IEMOCAP (Busso et al., 2008) and MSP-IMPROV (Busso et al., 2016) datasets are based on actor performances and emotion training, with IEMOCAP and MSP-IMPROV consist of conversational data, whereas CREMA-D and RAVDESS record non-dialogue data. These datasets offer higher-quality emotional data. However, these datasets largely rely on pre-written scripts, and their inherent limitations may lead to overly theatrical emotional expressions from actors, lacking the spontaneity found in authentic interactions.

Chinese Datasets: Currently, there have been some preliminary research efforts in the field of multimodal emotion datasets based on Mandarin For example, the CH-SIMS (Yu et al., 2020) and MER-MULTI (Lian et al., 2024) dataset use five continuous emotion labels and six discrete emotion labels respectively, making it suitable for multimodal sentiment analysis on isolated utterances spoken in Mandarin. However, both of them lack dialogue scenarios, overlooking the emotional changes multi-turn interactions. In contrast, datasets like M^{3}ED (Zhao et al., 2022) and MC-EIU$_{ch}$ (Liu et al., 2024) have made progress in terms of dialogue-level data, making it possible for supporting multimodal emotion recognition in conversations. Moreover, M^{3}ED and MC-EIU$_{ch}$ have been significant progress regarding the scale of the data.

Despite numerous advances in emotion recognition, most Chinese datasets still exhibit limitations in terms of scale, data quality, and annotation completeness. Existing Chinese datasets generally focus on relatively simple emotion labels or rely on low-quality data collected from the internet. In contrast, our dataset aims to effectively address these shortcomings by providing 23.6 hours of topic-driven spontaneous emotional dialogues. The dataset not only features high-quality recorded conversations but also includes detailed and comprehensive emotional annotations, making it a valuable asset for MER research and broader emotional dialogue analysis.

Dataset	Modality	Dialogue	Sources	Emo-label	Des.	Language	Utts
CMU-MOSI (Zadeh et al., 2016)	a, v, l	No	YouTube	7 Dim.	No	English	2,199
CMU-MOSEI (Zadeh et al., 2018b)	a, v, l	No	YouTube	7 Disc. / 5 Dim.	No	English	22,856
MELD (Poria et al., 2019)	a, v, l	Yes	TVs	7 Disc.	No	English	13,708
CREMA-D (Cao et al., 2014)	a, v, l	No	Act	6 Disc.	No	English	7,442
RAVDESS (Livingstone and Russo, 2018)	a, v, l	No	Act	8 Disc.	No	English	7,356
IEMOCAP (Busso et al., 2008)	a, v, l	Yes	Act	5 Disc.	No	English	7,433
MSP-IMPROV (Busso et al., 2016)	a, v, l	Yes	Act	5 Disc.	No	English	8,438
UniC (Du et al., 2025)	a, v, l	Yes	YouTube	7 Disc. / 3 Dim.	No	English	965
CH-SIMS (Yu et al., 2020)	a, v, l	No	Movies, TVs	5 Dim.	No	Mandarin	2,281
CH-SIMS v2.0 (Liu et al., 2022b)	a, v, l	No	Movies, TVs	7 Dim.	No	Mandarin	4,402
MEC 2017 (Li et al., 2018)	a, v	No	Movies, TVs	8 Disc.	No	Mandarin	7,030
MER-MULTI (Lian et al., 2024)	a, v, l	No	Movies, TVs	6 Disc.	No	Mandarin	3,784
M^3ED (Zhao et al., 2022)	a, v, l	Yes	TVs	7 Disc.	No	Mandarin	24,449
MC-EIU_ch (Liu et al., 2024)	a, v, l	Yes	TVs	7 Disc.	No	Mandarin	11,003
EmotionTalk	a, v, l	**Yes**	**Record**	**7 Disc. / 5 Dim.**	**Yes**	**Mandarin**	**19,250**

Table 1: Summary of multimodal emotion datasets. Disc. = Discrete, Dim. = Dimensional, Des. = Description, and Utts = Utterances.

3 Dataset Description

In this section, we introduce a large-scale, comprehensive, recorded interactive Chinese multimodal emotion dataset, EmotionTalk. We describe data Collection, annotation and statistics in detail.

3.1 Data Collection

Psychological studies indicate that scripted speech often suffers from "read-speech" artifacts (Douglas-Cowie et al., 2007). We move away from traditional scripted methods, instead adopting a theme-driven improvisational performance approach with professional drama actors. This method aims to capture more authentic and natural emotional expressions, effectively simulating spontaneous emotional behavior in real-world scenarios. While this process is more challenging to implement and more time-consuming, its advantages in emotional authenticity are significant.

To ensure a high degree of data diversity, we developed multi-turn dialogue scenarios to simulate genuine human interaction. These scripts are inspired by television plots or generated by Large Language Models (LLMs) and underwent rigorous manual quality assurance. Each scenario involves two characters, designed to capture the dynamic evolution of emotions over time. The scripts encompass a wide range of emotional intensities, from lighthearted conversations to tense, intense conflicts, fully showcasing the richness of emotional expression.

The dataset covers multiple real-life themes, including friendship, family, workplace, and doctor-patient interactions. The friendship theme includes dialogues about joy, conflict, and reconciliation, highlighting support and friction. The workplace theme involves complex emotional scenarios involving collaboration, competition, pressure, and misunderstandings, which elicit emotions such as anxiety, frustration, satisfaction, and anger. Each theme is carefully designed; for instance, the family theme includes scenarios like arguments, holiday gatherings, and farewells, reflecting emotions like warmth, anger, and sadness. The language style varies by theme: friendship dialogues are casual and natural, while workplace exchanges are more formal and serious. This accurately simulates real-world environments, encouraging actors to deliver authentic emotional performances and deepening the emotional layers and immersion of the scripts. It's important to note that actors are encouraged to express genuine emotions based on the theme rather than adhering strictly to a script. To ensure emotional consistency and avoid actors maintaining a single emotion for too long, each dialogue is limited to approximately two minutes.

3.2 Annotation

To ensure the high quality and diversity of the dataset, we design a rigorous data annotation process, incorporating multi-dimensional annotations for emotion categories, emotion intensity, and emotional speaking style caption. The detailed annotation process is outlined below:

Emotion Category: For each sample, we design a multi-step annotation process with cross-validation by *N (N = 5)* annotators. The emotion category annotation is based on the basic emotion theory commonly used in psychological research and covers *K (K = 7)* widely recognized emotion categories: happiness, surprise, sadness, disgust, anger, fear, and neutral. To prevent interference

between different modalities and avoid potential confusion, we follow the modality-independent annotation principle, requiring annotators to view only the current modality information and strictly prohibiting multimodal synchronous annotation. The annotation process follows a predefined sequence, processing text, audio, silent video, and finally video with audio in order. Each emotion annotation consists of a emotion category y_i and a confidence score c_i, of which is set to 0.1, 0.3, 0.5, 0.7, and 0.9, to quantify the annotator's confidence in their judgment. The formula for calculating the weighted confidence score x_k, $k = \{1, \ldots, K\}$ for each category of a sample is as follows:

$$x_k = \frac{1}{N_k} \sum_{i=1}^{N} \mathbb{I}(y_i = k) \cdot c_i, \tag{1}$$

where k represents the emotion category, N_k is the number of annotations for category k, $\mathbb{I}(y_i = k)$ is an indicator function, which equals 1 if the label y_i assigned by annotator i is equal to category k, and 0 otherwise.

Thus, the final emotion category y is calculated as follows:

$$y = \arg\max_k x_k, \tag{2}$$

where $argmax$ represents selecting the category k that corresponds to the maximum weighted confidence x_k as the final category label.

For cases with low confidence or inconsistent annotations, we employ a multi-round negotiation mechanism: first, multiple experienced annotators independently re-evaluate the samples, followed by expert discussions to reach consensus, ensuring the reliability and consistency of annotation quality.

Emotion Intensity: To more accurately quantify the intensity of emotional expressions, we have designed a multimodal-based emotion intensity annotation process aimed at quantitatively labeling the emotional polarity (positive, negative, neutral) and its intensity in utterances. For each audio clip, five annotators will be assigned, and each annotator will evaluate the emotional state as -2 (strongly negative), -1 (weakly negative), 0 (neutral), 1 (weakly positive), or 2 (strongly positive). The annotation results from the five annotators are then averaged to obtain a continuous label that contains emotion intensity information. The final labeling results will be one of the following values: [-2.0, 2.0]. Smaller values indicate higher negativity, while larger values indicate higher positivity. Similarly, for samples with conflicting emotion polarity annotations, we adopt a multi-round negotiation mechanism: 2-3 senior emotion annotation experts engage in thorough discussions to negotiate and reach final decisions on disputed samples, ensuring the authority and consistency of annotation results.

Emotional Speaking Style Caption: A core innovation of this dataset lies in the construction of a four-dimensional, fine-grained speech emotion annotation system that achieves significant breakthroughs in comprehensiveness and precision. Our framework encompasses four refined dimensions—speaker, speaking style, emotion, and overall comprehensive description (Shimizu et al., 2024)—constructing a complete spectrum of features from basic vocal qualities (e.g., warmth, clarity) and high-level speaking styles (e.g., rhythm, intonation) to deep emotional semantics. To strictly mitigate LLM hallucinations, we implement a two-stage verification protocol. First, human experts annotated discrete attributes (e.g., 'Style: Fast-paced', 'Pitch: High'). DeepSeek-R1 (Guo et al., 2025) is restricted to performing controlled paraphrasing only, converting these structured attributes into natural sentences without inferring new information. Second, the 10% manual inspection 7 focused on fact-checking: ensuring no hallucinated adjectives (e.g., describing 'trembling voice' when the attribute was absent) are introduced. The 98.5% agreement confirms the captions remain grounded in human perception.

3.3 Statistics

Our dataset comprises 744 dialogues with a total of 19,250 unimodal samples covering text, audio, and video modalities. The audio data spans 23.6 hours with an average segment length of 4.4 seconds, while the text data contains 469,387 characters with approximately 24 characters per sentence on average. To support comprehensive emotion analysis, we provide independent and detailed emotion labels for each modality (text, audio, video) while constructing multimodal fusion labels, forming a hierarchically rich and comprehensively covered annotation framework.

3.3.1 Emotional Distribution

Our dataset manifests distinct emotional distributions across modalities, underscoring the necessity of multimodal perception. Within the audio modal-

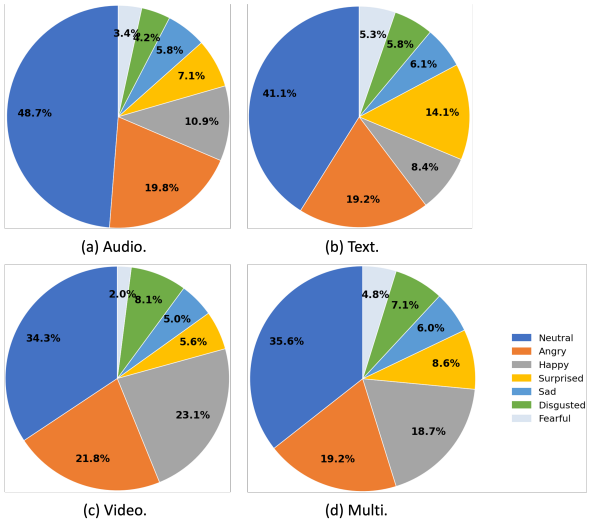


Figure 1: Data distribution across different modalities.

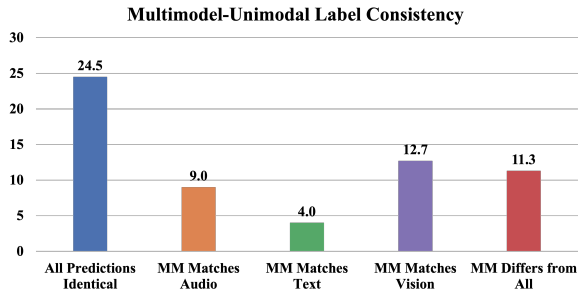


Figure 2: Multimodal–Unimodal Emotion Label Consistency. MM denotes the annotated emotion label for the multimodal input. Others denote annotated emotion labels for the corresponding unimodal inputs.

ity, neutral (48.7%) and anger (19.8%) represent the prevailing sentiments, though categories such as happiness, surprise, sadness, disgust, and fear maintain significant representation. The textual modality exhibits a similar concentration, with neutral (41.1%), anger (19.2%), and surprise (14.1%) acting as the dominant categories. Notably, the visual modality offers a more diversified emotional landscape. In unimodal video scenarios, neutral (34.3%), happiness (23.1%), and anger (21.8%) are the most frequent. However, when transitioning to an integrated multimodal consensus, the distribution shifts slightly: neutral (35.6%), anger (19.2%), and happiness (18.7%) become the primary labels. This distributional variance illuminates the unique contribution of each modality to the final emotional percept and mirrors the complexity of emotional expression in naturalistic conversations, thereby providing a diverse sample space for robust model training.

Dataset	Fleiss' Kappa (κ)
MELD	0.43
IEMOCAP	0.48
MSP-IMPROV	0.49
M^3ED	0.59
MC-EIU_ch	0.54
EmotionTalk	**0.78**

Table 2: Inter-Annotator Agreement Comparison.

3.3.2 Cross-modal Label Divergence

A defining characteristic of our dataset is the occurrence of label divergence—where the same sample conveys conflicting emotional signals across different modalities. For instance, a speaker's verbal content may remain neutral, while their vocal prosody conveys anger and their facial expressions manifest sadness. Such cross-modal inconsistencies reflect the multi-layered and often paradoxical nature of human communication. This phenomenon emphasizes the limitations of unimodal sentiment analysis and highlights the indispensable value of multimodal fusion. By capturing these inter-modal conflicts, our dataset enables researchers to investigate emotional consistency and dissonance, providing a rigorous foundation for developing models that can navigate complex, real-world affective signals.

As illustrated in Figure 2, we analyze the consistency between aggregated multimodal labels and their unimodal counterparts. Only 24.5% of samples reach a unanimous agreement across all modalities, indicating that cross-modal semantic alignment remains a formidable challenge for the majority of the data. Among partially aligned cases, multimodal labels converge more frequently with vision (12.7%) and audio (9.0%) than with text (4.0%). Furthermore, 11.3% of samples exhibit no alignment with any single unimodal label, suggesting that multimodal integration often results in a "holistic" emotional judgment that transcends individual modality predictions. These findings suggest that in our specific conversational context, acoustic and visual cues exert a more dominant influence on the final emotional decision than textual information.

3.4 Inter-rater Agreement Analysis

Fleiss' Kappa is a statistical measure assessing inter-rater agreement when multiple evaluators cat-

Speech Model	Speech(Four)	Multimodal(Four)	Speech(All)	Multimodal(All)	Mean
WavLM-Base	72.05 / 71.18	61.41 / 61.24	60.01 / 58.33	47.96 / 46.80	60.36 / 59.39
Whisper-Base	71.09 / 70.84	63.09 / 63.09	62.79 / 60.36	48.69 / 47.54	61.42 / 60.46
Wav2vec 2.0-Large	73.42 / 74.55	62.51 / 62.71	63.52 / 61.88	51.10 / 48.88	62.64 / 62.01
Wav2vec 2.0-Base	76.54 / 76.05	63.47 / 63.96	65.62 / 63.53	52.15 / 51.66	64.45 / 63.80
WavLM-Large	76.58 / 75.73	64.31 / 64.50	63.73 / 61.91	52.36 / 51.00	64.25 / 63.29
Whisper-Large	76.03 / 75.62	66.62 / 66.52	64.68 / 63.16	52.62 / 51.00	64.99 / 64.07
Hubert-Base	78.63 / 79.07	**71.25 / 71.29**	64.99 / 64.86	58.44 / 57.75	68.33 / 68.24
Hubert-Large	**82.42 / 82.28**	70.29 / 70.81	**70.65 / 69.32**	**60.95 / 60.28**	**71.08 / 70.67**
Text Model	**Text(Four)**	**Multimodal(Four)**	**Text(All)**	**Multimodal(All)**	**Mean**
Vicuna-7B	59.05 / 54.90	51.25 / 48.88	40.90 / 38.98	37.79 / 36.00	47.25 / 44.69
RoBERTa-Base	57.19 / 56.97	50.83 / 51.33	47.17 / 44.17	39.04 / 38.96	48.56 / 47.86
LERT-Base	59.81 / 59.32	51.09 / 51.60	47.54 / 44.78	40.64 / 39.88	49.77 / 48.90
Sentence-BERT	59.88 / 59.27	52.95 / 53.26	47.02 / 44.68	40.49 / 38.99	50.09 / 49.05
DeBERTa-Large	61.59 / 59.85	49.11 / 48.82	49.90 / 45.94	43.50 / 42.02	51.03 / 49.16
RoBERTa-Large	61.49 / 60.28	52.68 / 53.05	47.23 / 44.21	42.82 / 41.83	51.06 / 49.84
BERT-Base	**63.85 / 61.69**	49.17 / 49.87	**52.10** / 48.75	46.07 / 44.33	52.80 / 51.16
BLOOM-7B	62.00 / 60.87	52.62 / 52.98	50.47 / 48.84	**46.96** / 46.01	53.01 / 52.18
ChatGLM2-6B	62.21 / 61.08	53.41 / 54.01	50.52 / 48.95	45.96 / **46.12**	53.03 / **52.54**
Baichuan-7B	61.52 / 59.77	**58.59 / 58.56**	51.42 / **49.11**	43.44 / 41.93	**53.74** / 52.34
Visual Model	**Visual(Four)**	**Multimodal(Four)**	**Visual(All)**	**Multimodal(All)**	**Mean**
Data2vec-Base	32.84 / 25.76	37.81 / 31.84	43.55 / 39.62	38.78 / 35.26	38.25 / 33.12
VideoMAE-Base	58.59 / 56.64	53.83 / 52.11	48.06 / 43.09	42.09 / 38.46	50.64 / 47.58
VideoMAE-Large	64.47 / 62.31	61.86 / 60.05	51.05 / 49.08	49.42 / 44.37	56.70 / 53.95
Dinov2-Large	64.86 / 64.42	61.86 / 60.01	55.40 / 53.48	53.67 / 47.68	58.95 / 56.40
CLIP-Base	69.07 / 69.17	66.11 / 66.19	53.14 / 52.62	47.85 / 48.89	59.04 / 59.22
EVA-02-Base	72.72 / 72.53	65.79 / 65.16	54.72 / 54.47	49.79 / 49.78	60.76 / 60.49
Dinov2-Giant	73.42 / 73.03	68.17 / 66.16	57.55 / 57.26	48.74 / 47.77	61.97 / 61.06
CLIP-Large	**77.38 / 77.33**	**74.60 / 74.58**	**64.31 / 61.57**	**57.55 / 56.69**	**68.46 / 67.54**

Table 3: We report unimodal results. Four means that four emotion labels are used: happy, angry, sad, and neutral. All means that all emotion labels are used. The values reported in each cell denote ACC / F1.

egorize items. The formula for Fleiss' Kappa is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{3}$$

where P_o represents the observed proportion of agreement, and P_e represents the expected proportion of agreement under random conditions. In our dataset, the Fleiss' Kappa values are 0.79 for audio, 0.66 for text, 0.73 for silence video, and 0.78 for multimodal, indicating substantial to almost perfect agreement across all modalities.

To quantitatively assess the annotation quality, we compare the inter-rater agreement of EmotionTalk with existing datasets. As shown in Table 2, EmotionTalk achieves a Fleiss' Kappa of 0.78 for overall, surpassing M^{3}ED (0.59). This demonstrates that our multi-round negotiation mechanism effectively mitigates the ambiguity inherent in spontaneous emotional dialogues.

4 Experiments

In this section, we evaluate our dataset across a variety of tasks, including unimodal / multimodal emotion recognition, multimodal emotion analysis, and emotional speaking style captioning. We build our experimental pipeline upon the MerBench (Lian et al., 2024), which provides a standardized setup for benchmarking multimodal models.

Specifically, in the continuous setting, we focus on a binary classification task that distinguishes between positive and negative emotions, where samples with scores below 0 are labeled as negative, and those above 0 as positive. For the first three tasks, accuracy and f1-score are used as the evaluation metric, while for the emotional speaking style captioning, we adopt BLEU$_4$, ROUGE$_L$, METEOR, SPIDEr, FENSE, BERTScore and CLAPScore for evaluation. To facilitate reproducibility, we document all experimental settings in Appendix table IV, including hyperparameter tuning strategies, optimizer selection, and the values of all key training parameters.

4.1 Unimodal Emotion Recognition

This section reports the emotion recognition performance of different feature extractors on the corresponding modalities, as shown in Table 3.

Features	Algorithms	Fusion	Multimodal(Four)	Multimodal(All)	Mean
	MCTN	Frame-level	69.26 / 68.75	48.99 / 48.53	59.13 / 58.64
	MFM	Frame-level	80.71 / 80.50	60.13 / 52.79	70.42 / 66.65
Hubert-Large	GMFN	Frame-level	83.02 / 83.05	66.46 / 63.68	74.74 / 73.37
	MMIM	Uttrance-level	82.25 / 82.33	65.47 / 63.93	73.86 / 73.13
Baichuan-7B	MISA	Uttrance-level	82.57 / 82.54	69.86 / 69.53	76.22 / 76.04
	TFN	Uttrance-level	83.02 / 82.99	70.13 / 70.03	76.58 / 76.51
CLIP-Large	MulT	Frame-level	**83.22 / 83.20**	69.41 / 69.55	76.32 / 76.38
	MFN	Frame-level	82.77 / 82.73	67.86 / 67.24	75.32 / 74.99
	Attention	Uttrance-level	81.80 / 81.77	70.46 / 70.81	76.13 / 76.29
	LMF	Uttrance-level	82.38 / 82.34	**71.23 / 70.84**	**76.81 / 76.59**

Table 4: We report multimodal results for the EmotionTalk dataset.

	Decoder	$BLEU_4$	$ROUGE_L$	METEOR	SPIDEr	FENSE	BERTScore	CLAPScore
Speaker	Transformer-based	0.011	0.397	0.204	0.229	0.842	0.974	0.860
	GPT-2	**0.020**	**0.430**	**0.212**	0.256	0.765	0.976	**0.899**
	Qwen-2	0.009	0.414	0.205	**0.258**	**0.846**	**0.977**	0.878
Style	Transformer-based	0.065	0.517	0.313	0.339	0.512	0.985	0.895
	GPT-2	0.075	0.510	0.298	0.350	**0.611**	0.987	0.850
	Qwen-2	**0.127**	**0.564**	**0.339**	**0.482**	0.523	**0.988**	**0.912**
Emotion	Transformer-based	0.032	0.366	0.191	0.276	0.932	0.973	0.843
	GPT-2	0.014	**0.399**	0.147	0.235	0.903	0.972	0.818
	Qwen-2	**0.058**	0.361	**0.199**	**0.353**	**0.942**	**0.975**	**0.853**
Overall	Transformer-based	0.018	0.469	0.233	0.230	0.921	0.980	0.878
	GPT-2	0.015	0.462	0.214	0.227	0.890	0.980	0.849
	Qwen-2	0.033	**0.535**	**0.268**	0.121	0.562	0.984	0.885
	Qwen2.5-Omni	0.000	0.345	0.155	0.183	0.889	0.983	0.896
	Qwen3-Omni-30B-A3B	**0.107**	0.462	0.224	**0.574**	**0.936**	**0.987**	**0.905**

Table 5: Automatic captioning results. Transformer-based, GPT-2, and Qwen-2 use Hubert as speech encoder.

Feature Extractor: To assess the performance of our dataset, we employ a comprehensive suite of pre-trained baseline models across different modalities. Specifically, for the speech modality, we utilize Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and Whisper (Radford et al., 2023). For the text modality, our selection includes Vicuna-7B (Chiang et al., 2023), LERT (Cui et al., 2022), DeBERTa (He et al., 2020), BERT (Devlin et al., 2019), Sentence-BERT (Reimers and Gurevych, 2019), BLOOM-7B (Workshop et al., 2022), RoBERTa (Liu et al., 2019), ChatGLM2 (Du et al., 2021) and Baichuan-7B (Yang et al., 2023). For the visual modality, we adopt Data2Vec (Baevski et al., 2022), VideoMAE (Tong et al., 2022), EVA-02 (Fang et al., 2024), CLIP (Radford et al., 2021), and DINOv2 (Oquab et al., 2023).

The experimental results in Table 3 validate the EmotionTalk dataset as a high-quality, discriminative benchmark for multimodal emotion recognition. A critical observation is the consistent performance scaling across different model capacities; for instance, as the speech encoder scales from Wav2vec 2.0-Base to Large, the accuracy in the $Speech(All)$ task improves from 59.98% to 67.04%. This positive correlation between model parameters and performance serves as a proxy for the dataset's high signal-to-noise ratio, proving that the labels are grounded in objective features that larger models can progressively extract. Furthermore, the significant performance drop observed when shifting from the four-class set to the full-label set—where even high-capacity models like Dinov2-Giant experience a decline of 15.87% in accuracy—highlights the dataset's granularity and challenge. It demonstrates that EmotionTalk is not a simplistic sentiment corpus but a complex dataset that effectively captures fine-grained emotional variances. We intentionally benchmark these results using a diverse array of established, representative architectures rather than the most recent large-scale models to ensure a stable and reproducible reference. By doing so, we shift the focus from validating model-specific performance to assessing the dataset's inherent difficulty and modal distribution. These baselines confirm that while the dataset provides sufficient discriminative

power, the "Multimodal" performance remains a non-trivial bottleneck, positioning EmotionTalk as a valuable resource for future research in cross-modal interaction and fusion strategies.

4.2 Multimodal Analysis

Table 4 presents the performance of various multimodal fusion algorithms on the EmotionTalk dataset using the optimal encoder from each modality—HuBERT-Large (speech), Baichuan-7B (text), and CLIP-Large (visual). The fusion methods are categorized into frame-level (e.g., MFN (Zadeh et al., 2018a), GMFN (Zadeh et al., 2018b), MCTN (Pham et al., 2019), MFM (Tsai et al., 2018), and MulT (Tsai et al., 2019)) and utterance-level (e.g., TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), and the Attention mechanism (Vaswani et al., 2017)) strategies, enabling a comparative analysis of their effectiveness in multimodal emotion recognition.

Several key observations can be made. First, in the "Multimodal(Four)" task, most models achieve competitive performance, with MulT reaching a peak F1-score of 83.22%. This highlights the advantage of cross-modal attention mechanisms in capturing tightly aligned multimodal features. Second, a significant performance degradation (averaging approximately 12%) is observed across all models when transitioning to the "Multimodal(All)" task. This discrepancy underscores the complexity and nuance of Chinese emotional expressions within our dataset, posing a substantial hurdle for current SOTA models. Notably, LMF achieves the superior performance in the all-class setting and the overall mean metrics (71.23% / 76.59%), suggesting that low-rank tensor decomposition is more robust in modeling diverse modal interactions and handling complex class distributions. Furthermore, utterance-level fusion strategies generally exhibit more stable performance than frame-level methods in complex scenarios. In summary, while our dataset serves as a high-quality resource for multimodal training, its emotional intricacy presents a formidable challenge for future research in deep cross-modal alignment.

4.3 Emotional Speaker Style Captioning

As shown in Table 5, the experimental results for the automatic captioning task on the EmotionTalk dataset further demonstrate the corpus's multi-faceted value as a rich resource for multimodal understanding. By evaluating models across four descriptive dimensions—Speaker, Style, Emotion, and Overall—we observe that the dataset successfully encodes complex acoustic and stylistic signatures that are learnable by generative architectures. For instance, the Qwen-2 decoder, utilizing Hubert as a speech encoder, achieves a competitive SPIDEr score of 0.482 in style description and 0.353 in emotion-specific captioning, indicating that the dataset's fine-grained annotations can support high-level semantic generation. It is important to reiterate that our selection of baselines, ranging from standard Transformer-based models to more recent variants like Qwen3-Omni-30B-A3B, is intended to establish a robust and interpretable reference point for the research community rather than to pursue maximum performance through exhaustive tuning. The notable discrepancy between high BERTScore values (e.g., 0.988 for Qwen-2) and more conservative overlap-based metrics like BLEU further highlights that the EmotionTalk dataset encourages models to capture the semantic essence of spoken affect rather than simple verbatim replication. Ultimately, these results serve as a diagnostic baseline, proving that the dataset is a sophisticated benchmark for resolving the intricate relationship between acoustic signals and multi-dimensional descriptive metadata.

Our fine-grained annotation system transcends traditional classification by providing a structured semantic blueprint for emotional understanding. Unlike simple categorical labels, our four decoupled dimensions—speaker, style, emotion, and overall—act as dense semantic anchors that enable modular flexibility and controllable generation, such as isolating "style" for prosody analysis. Crucially, these multi-layered annotations foster explainable emotion captioning by articulating the specific acoustic evidence behind a perceived state (e.g., "rapid breathing indicating anxiety"), thereby transforming raw signals into human-interpretable justifications. This framework ensures the EmotionTalk dataset is not merely a collection of metadata, but a valuable toolkit for advancing interpretable and generative affective science.

5 Conclusion

This paper presents EmotionTalk, a large-scale Chinese multimodal emotion dataset designed to bridge the gap between data quantity and annotation granularity. Distinguished from ex-

isting scripted corpora, EmotionTalk comprises 23.6 hours of topic-guided spontaneous dialogues recorded by 19 professional actors. This methodological choice effectively captures the nuances of authentic emotional interplay often lost in structured performances. Beyond its scale, the dataset introduces a novel contribution to the field: fine-grained emotional speaking style captions, which enable more interpretable and generative affective computing tasks. We validate the dataset's quality and utility through rigorous quality assurance processes and comprehensive benchmarks on multiple tasks. As a robust and versatile resource, EmotionTalk not only serves as a testing ground for multimodal algorithms but also opens new avenues for research in human-like conversational AI.

6 Limitations

The EmotionTalk dataset serves as an important resource in the field of conversational emotion recognition, providing a valuable experimental foundation for related research. However, when conducting in-depth analysis, it is necessary to examine several of its characteristics in order to more comprehensively evaluate its applicability across different research scenarios. First, it is worth noting that this dataset has a relatively limited scale, containing only 19 participants and 23.6 hours of multimodal data. Although the dataset demonstrates excellence in ensuring data quality and annotation precision, the limited sample size may to some extent affect the generalization ability of models trained on this dataset when applied to broader populations and diverse conversational scenarios.

These characteristics are not fundamental flaws of the dataset, but rather products of its specific research design, pointing researchers toward future development directions. For example, in future research, one could consider using the EmotionTalk dataset as an initial validation set and combining it with other larger-scale or more robust datasets to construct conversational emotion recognition systems with greater robustness and generalizability.

7 Ethics Statement

This study is conducted in accordance with rigorous ethical guidelines to ensure the protection of participants' rights and well-being. All recordings take place in a quiet indoor environment, where professional actors engage in natural, emotionally diverse, and logically coherent dialogues based on predefined emotional themes and content outlines. The annotation cost for each data sample is 0.2 RMB.

To preserve participant privacy, all data are anonymized by removing personal identifiers and replacing them with coded labels. The dataset is released under the CC BY-NC 4.0 license, which prohibits commercial use and supports ethical research practices. Data are securely stored, and access is restricted to authorized researchers for academic purposes only.

In conclusion, this study demonstrates a strong commitment to ethical standards, encompassing informed consent, the protection of personal privacy, appropriate compensation, and the responsible dissemination of data, thereby safeguarding participant rights and supporting ethical scientific advancement.

8 Acknowledgement

This work has been supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yiming Cui, Wanxiang Che, Shijin Wang, and Ting Liu. 2022. Lert: A linguistically-motivated pre-trained language model. *arXiv preprint arXiv:2211.05344*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, and 1 others. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025. Unic: a dataset for emotion analysis of videos with multimodal and unimodal labels: Q. du et al. *Language resources and evaluation*, 59(3):2857–2892.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

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

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ya Li, Jianhua Tao, Björn Schuller, Shiguang Shan, Dongmei Jiang, and Jia Jia. 2018. Mec 2017: Multimodal emotion recognition challenge. In *2018 first Asian conference on affective computing and intelligent interaction (ACII Asia)*, pages 1–5. IEEE.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2024. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*.

Ziqi Liang, Haoxiang Shi, and Hanhui Chen. 2024. Aligncap: Aligning speech emotion captioning to human preferences. *arXiv preprint arXiv:2410.19134*.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. 2024. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*.

Yang Liu, Haoqin Sun, Geng Chen, Qingyue Wang, Zhen Zhao, Xugang Lu, and Longbiao Wang. 2023. Multi-level knowledge distillation for speech emotion recognition in noisy conditions. In *Proc. Interspeech 2023*, pages 1893–1897.

Yang Liu, Haoqin Sun, Wenbo Guan, Yuqi Xia, and Zhen Zhao. 2022a. Discriminative feature representation based on cascaded attention network with adversarial joint loss for speech emotion recognition. In *Interspeech 2022*, pages 4750–4754.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022b. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258.

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

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

Steven R Livingstone and Frank A Russo. 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*, 13(5):e0196391.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sadat Shahriar and Yelin Kim. 2019. Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE.

Xiaohan Shi, Sixia Li, and Jianwu Dang. 2020. Dimensional emotion prediction based on interactive context in conversation. In *INTERSPEECH*, pages 4193–4197.

Xiaohan Shi, Xingfeng Li, and Tomoki Toda. 2023. Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation. In *Proc. Interspeech*, pages 765–769.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2024. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12672–12676. IEEE.

Haoqin Sun, Shiwan Zhao, Xiangyu Kong, Xuechen Wang, Hui Wang, Jiaming Zhou, and Yong Qin. 2024a. Iterative prototype refinement for ambiguous speech emotion recognition. In *Proc. Interspeech 2024*, pages 3200–3204.

Haoqin Sun, Shiwan Zhao, Xuechen Wang, Wenjia Zeng, Yong Chen, and Yong Qin. 2024b. Fine-grained disentangled representation learning for multimodal emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11051–11055. IEEE.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are

data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. Qwen2 technical report, 2024. *URL https://arxiv. org/abs/2407.10671*.

Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2019. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 6685–6689. IEEE.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

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

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*.

A Datasheets for datasets

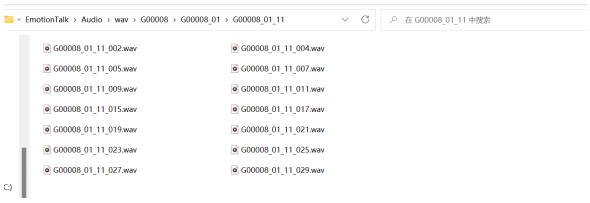
A.1 Dataset Snapshots

The dataset comprises 744 dialogues, encompassing a total of 19,250 utterances for each unimodal modality—text, audio, and video. The audio data span approximately 23.6 hours, with an average duration of 4.4 seconds per utterance.

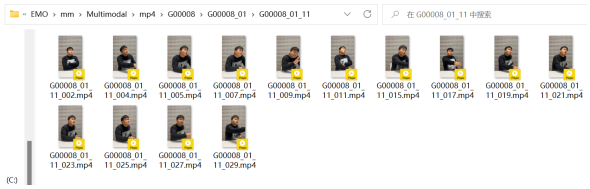
Each utterance is stored as an individual JSON file following a unique naming convention in the format: "<group_No>_<session_No>_<Speaker_id>_<Utt_No>.json". Corresponding audio and video files are named identically, with the extensions ".wav" and ".mp4" respectively: "<group_No>_<session_No>_<Speaker_id>_<Utt_No>.wav" and "<group_No>_<session_No>_<Speaker_id>_<Utt_No>.mp4". The samples of audio and video files in EmotionTalk are shown in Fig 3.

Attribute	Category	Count / Stats
Gender	Male Female	10 (52.6%) 9 (47.4%)
Age	Range (Mean ± SD)	24 – 40 (29.6 ± 4.5)
Education	Master's Bachelor's Associate	2 15 2
Origin	Northern China (e.g., Beijing, Hebei) Southern China (e.g., Zhejiang, Guangdong)	15 4

Table 6: Demographic Summary of the 19 Professional Actors. The balanced gender ratio and diverse origins ensure a comprehensive coverage of Mandarin speech patterns.



(a) Examples of audio file samples.



(b) Examples of video file samples.

Figure 3: Snapshots of audio and video samples in the EmotionTalk dataset. All files are named following a consistent and structured format.

A.2 Participants Demographics

To ensure diversity and minimize demographic bias, we recruited 19 professional actors with a balanced gender ratio, comprising 10 males and 9 females. The participants' ages range from 24 to 40 years (Mean=29.6, SD=4.5), covering the primary demographic groups relevant to our workplace, family, and social scenarios. They originate from 8 different provinces across China, including Beijing, Hebei, Zhejiang, and Guangdong. In terms of educational background, all actors hold higher education degrees, with 17 possessing Bachelor's or Master's degrees and 2 holding Associate degrees. Linguistically, while all dialogues are conducted in Mandarin, the speech patterns encompass a natural spectrum ranging from standard Putonghua to Putonghua with regional accents, preserving the authentic linguistic diversity of their geographical origins. Detailed demographic statistics are provided in Table 6.

A.3 Data Format

Each utterance in the EmotionTalk dataset is associated with a corresponding ".jsonl" file, which contains detailed sample-level annotations. These annotations include not only the basic information such as the emotion label, speaker identity, and transcript, but also rich metadata that describes the expressive characteristics of the utterance. Specifically, the 'style_cap' and 'emo_cap' provide fine-grained natural language descriptions of the speaker's prosody (e.g., 'rapid breathing', 'trembling voice', 'high pitch') and nuanced emotional states. Unlike discrete labels which only provide classification targets, these textual descriptions serve as dense semantic anchors, designed to support Large Language Model (LLM) based emotion understanding, generation, and explainable AI tasks. The detailed annotation fields are listed in Table 7.

A.4 Data Distribution

In this study, to make full use of the data and ensure both effective model training and fair evaluation, the dataset is divided into training, validation, and test sets in an approximate ratio of 8:1:1. Specifically, 80% of the data is used for training the model to learn effective feature representations, 10% is allocated for validation to assist in model selection and prevent overfitting during training, and the remaining 10% is reserved as the test set to evaluate the model's generalization performance. Crucially, to rigorously evaluate the model's generalization capabilities on unseen subjects, we adopted a strict speaker-independent split strategy. Unlike random shuffling which may lead to identity leakage, our dataset is partitioned by actor groups. Specifically, dialogues from sessions G01 and G12 (involving 4 distinct actors) are selected for the validation set, while sessions G03 and G15 (involving another 4

distinct actors) are reserved for the test set. None of the actors in the validation or test sets overlap with those in the training set. This setup ensures that the evaluation metrics reflect the model's ability to learn emotion-agnostic features rather than overfitting to specific speaker identities.

B Feature Extraction

B.1 Models

To comprehensively evaluate the proposed dataset, we conduct extensive experiments on three tasks: unimodal emotion recognition, multimodal emotion recognition / sentiment analysis and emotional speaker style captioning. For the unimodal and multimodal emotion recognition tasks, we employ a range of state-of-the-art models as feature extractors to obtain representations from each modality. Then, we select several high-quality features as the foundation for multimodal fusion. For the emotional speaker style captioning task, we utilize three types of decoders, including transformer, GPT-2 and Qwen-2, to assess the quality and utility of the dataset. The details of the models are provided in Table 15.

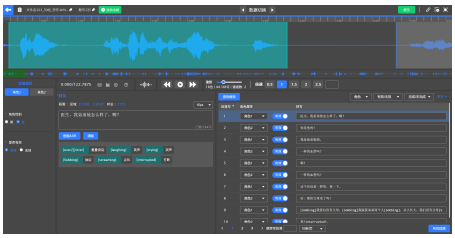
B.2 Hyperparameters and computing resources

Key training hyperparameters for different models are summarized in Table 9, Table 10, and Table 11. All models are trained using the AdamW optimizer.

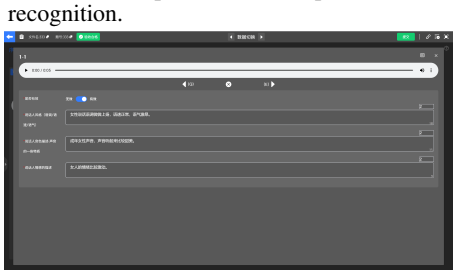
The experiments based on the Qwen-2 decoder are conducted on an NVIDIA A800 GPU, while all other experiments are performed using an NVIDIA GeForce RTX 3090 GPU.

C Annotation Website

To improve the efficiency and consistency of the annotation process, we developed a customized web-based annotation platform (Fig. 4). The platform is designed with built-in quality control features, including context-aware playback (allowing annotators to hear previous turns for better context understanding) and real-time logical checks (preventing conflicting labels). Fig. 4(a) shows the interface for discrete emotion categorization, while Fig. 4(b) illustrates the captioning interface where annotators describe speaking styles. A multi-stage review mechanism was implemented on this platform, where senior experts could review and resolve ambiguous samples, ensuring high inter-annotator agreement.



(a) Annotation platform of the speech emotion recognition.



(b) Annotation platform of the emotional speaking style captioning.

Figure 4: Overview of the annotation platform interface.

D Extra Experiment Results and Analysis

We select a candidate pool of four superior models for each individual modality, subsequently conducting experiments on their random combinations to assess multimodal synergy. Table 12 reports the multimodal emotion recognition results. All combinations adopt the LMF algorithm for fusion. Among the configurations, the combination of RoBERTa-Base (text), HuBERT-Large (speech), and Dinov2-Giant (visual) achieves the best overall performance, with the highest score in the Discrete (Four) setting (83.23%) and the highest average (81.87%). Notably, different model combinations yield comparable performance on the continuous labels, while their results on discrete tasks vary considerably, underscoring the impact of feature selection. These findings confirm that even under the same fusion ßstrategy, the choice of multimodal features can significantly affect the overall performance of multimodal fusion.

D.1 Results of Emotion Recognition in Conversation (ERC)

In the ERC task, a pronounced modality discrepancy is observed, with speech-based architectures consistently outperforming their text-based counterparts by a significant margin, as detailed in Table 13 (Hazarika et al., 2018; Yeh et al., 2019; Majumder et al., 2019; Ghosal et al., 2019). Our experimental results yield several critical insights into the proposed Chinese multimodal dataset.

Name	Description
emotion	Emotion label.
Confidence_degree	Annotator's self-rated confidence in the emotion label.
Continuous_label	5-dimensional sentiment labels.
speaker_id	Unique speaker identifier.
emotion_result	Final aggregated emotion label.
Continuous label_result	Final averaged sentiment labels aggregated from five annotators.
content	Transcript of the utterance.
startTime	Utterance start time in the session.
endTime	Utterance end time in the session.
duration	Total duration of the utterance.
emo_cap	Caption describing the type and intensity of the expressed emotion.
spe_cap	Caption describing the speaker's voice quality.
style_cap	Caption describing speaking style.
caption_1 – caption_5	Emotional speaking style caption.
file_path	Relative path to the audio file.

Table 7: Description of Sample-Level Annotations

	Angry	Disgusted	Fearful	Happy	Neutral	Sad	Surprised	Total
Train	2950	1142	672	2986	5377	919	1367	15413
Validation	409	95	125	360	675	111	133	1908
Test	339	134	125	246	801	123	161	1929
Total	3698	1371	922	3592	6853	1153	1661	19250

Table 8: Statistics of the data distribution across the training, validation, and test sets.

Hyperparameter	Four (Unimodal/Multimodal)	All (Unimodal/Multimodal)
Learning Rate	1e-3	1e-5
L2 Regularization Weight	1e-5	1e-5
Batch Size	32	32
Epochs	100	100

Table 9: Training hyperparameters used for the unimodal and multimodal models in Table 2 on the EmotionTalk dataset. "Four" refers to using four emotion labels (happy, angry, sad, neutral), while "All" refers to using full labels.

Model	Hidden Dim	Dropout	Learning Rate	Grad Clip
MCTN	64 – 256	0.0 – 0.3	1e-3	0.6 – 1.0
MFM	128 / 256	0.0 – 0.7	1e-3	-1.0
GMFN	128 / 256	0.0 – 0.7	1e-3	-1.0
MMIN	64 – 256	0.0 – 0.3	1e-3	0.6 – 1.0
MISA	64 – 256	0.2 – 0.5	1e-4	-1.0 – 1.0
TFN	64 / 128	0.2 – 0.5	1e-3	-1.0
MulT	64 – 256	0.0 – 0.3	1e-3	0.6 – 1.0
MFN	128 / 256	0.0 – 0.7	1e-3	-1.0
Attention	64 – 256	0.2 – 0.5	1e-5	-1.0
LMF	32 – 256	0.2 – 0.5	1e-5	-1.0

Table 10: Key training hyperparameters used for each multimodal model in Table 3 on the EmotionTalk dataset.

Decoder	Batch Size	Epochs	Learning Rate	Weight Decay	Warmup
Transformer-based	8	15	1.7e-05	3.0e-04	0
GPT-2	8	15	1.7e-05	3.0e-04	0
Qwen-2	4	6	1e-4	0.0	1,000

Table 11: Training hyperparameters for each decoder in Table 5.

Multimodal						
Text	Speech	Visual	**Discrete(Four)**	**Discrete(All)**	**Continuous**	**Mean**
Baichuan-7B	Hubert-Base	CLIP-Large	81.31	69.10	**93.35**	81.25
RoBERTa-Base	Hubert-Large	Dinov2-Giant	**83.23**	**69.21**	93.16	**81.87**
RoBERTa-Large	WavLM-Large	Dinov2-Large	78.13	65.01	93.10	78.75
ChatGLM2-6B	W2v 2.0-Large	CLIP-Base	73.82	63.50	92.26	76.53

Table 12: We utilize the LMF for multimodal fusion.

Modality	Model	UA	WA	F1
Speech	CMN	58.61	76.60	61.64
	ICON	56.98	77.20	62.36
	DialogueRNN	65.32	78.76	68.37
	DialoguGCN	64.83	78.69	69.24
Text	CMN	32.93	60.70	30.51
	ICON	33.26	60.55	30.75
	DialogueRNN	34.03	61.09	31.13
	DialoguGCN	34.79	61.70	32.27

Table 13: Performance comparison of different models across modalities (Speech and Text) on the Emotion Recognition in Conversation (ERC) task.

Modality	Model	UA	WA	F1
Speech	BiGRU	58.51	71.99	59.20
	AVEF	56.52	72.71	60.78
	DEP	58.41	72.07	62.14
	EAMT	59.11	73.34	62.61
Text	BiGRU	31.59	57.96	28.68
	AVEF	31.88	57.64	28.80
	DEP	32.53	57.83	29.00
	EAMT	32.47	58.04	29.33

Table 14: Performance comparison of different models across modalities on the Emotion Prediction in Conversation (EPC) task.

First, a stark performance disparity exists between modalities: the SPEECH modality achieves a peak F1 score of 69.24% via DialogueGCN, whereas the TEXT modality reaches a maximum of only 32.27%. This suggests that in our corpus, affective cues are primarily encoded within prosodic features—such as pitch contours, intensity, and rhythm—rather than semantic tokens. This phenomenon likely reflects the inherent lexical ambiguity of spontaneous Chinese conversation, where the same semantic string can convey polar-opposite emotions depending on tonal inflection.

Second, context-aware architectures, specifically DialogueRNN and DialogueGCN, demonstrate clear superiority over memory-based models like CMN and ICON. This highlights the necessity of modeling complex inter-speaker dependencies and the non-linear flow of information in multi-party dialogues. Specifically, the graph-based approach of DialogueGCN proves most effective, suggesting that emotional states in our dataset are better captured through relational modeling rather than simple temporal recurrence. Furthermore, the persistent gap between Weighted Accuracy (WA) and Unweighted Accuracy (UA)—most notably in the text modality where UA fluctuates between 31-34%—underscores a challenging class imbalance that mirrors the natural, long-tailed distribution of emotions in real-world human interactions.

D.2 Results of Emotion Prediction in Conversation (EPC)

For the EPC task, we observe similar performance trends across modalities, though with an overall decrease in raw metrics compared to ERC, as shown in Table 14. The transition from recognition to prediction entails a significant increase in task complexity, characterized by a universal performance degradation across all evaluated models (Shahriar and Kim, 2019; Shi et al., 2020, 2023).

In the EPC setting, the SPEECH modality maintains its overwhelming dominance. The EAMT model achieves a peak F1 score of 62.61 in speech, more than double the 29.33 achieved in text. This substantial gap indicates that prosodic features in our dataset serve as critical "emotional precursors"; they provide early acoustic signals of shifting affective states that precede explicit semantic expression.

Architecturally, the EAMT model consistently outperforms BiGRU, AVEF, and DEP across all metrics (UA, WA, and F1). This suggests that the attention-based cross-modal mechanisms in EAMT are better equipped to handle the predictive lag inherent in EPC by capturing long-range temporal

dependencies and subtle contextual cues. The fact that EAMT maintains relatively high WA (73.34) in speech despite the task's difficulty further validates the dataset's utility as a robust benchmark for forecasting emotional trajectories in complex conversational environments.

All models are trained using the Adam optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-5. The batch size is set to 16, and models are trained for a maximum of 30 epochs (except for DialogueRNN, which is trained for 100 epochs). A StepLR scheduler is employed to decay the learning rate by a factor of 0.1 every 10 epochs in ERC.

E Declaration of AI-Assisted Writing

The authors used LLMs solely for the purposes of language polishing, grammatical correction, and stylistic refinement of the manuscript. All scientific conceptualization, data collection, experimental analysis, and the formulation of original ideas were conducted entirely by the human authors. The authors maintain full responsibility for the content, accuracy, and integrity of the final work.

Speech Model	Link	License
Whisper-Base (Radford et al., 2023)	huggingface.co/openai/whisper-base	Apache License 2.0
Whisper-Large (Radford et al., 2023)	huggingface.co/openai/whisper-large-v2	Apache License 2.0
WavLM-Base (Chen et al., 2022)	huggingface.co/microsoft/wavlm-base	CC BY-SA 3.0
Wav2vec 2.0-Base (Baevski et al., 2020)	huggingface.co/TencentGameMate/chinese-wav2vec2-base	MIT License
Wav2vec 2.0-Large (Baevski et al., 2020)	huggingface.co/TencentGameMate/chinese-wav2vec2-large	MIT License
WavLM-Large (Chen et al., 2022)	huggingface.co/microsoft/wavlm-large	CC BY-SA 3.0
Hubert-Large (Hsu et al., 2021)	huggingface.co/TencentGameMate/chinese-hubert-large	MIT License
Hubert-Base (Hsu et al., 2021)	huggingface.co/TencentGameMate/chinese-hubert-base	MIT License
Text Model	**Link**	**License**
Vicuna-7B (Chiang et al., 2023)	huggingface.co/CarperAI/stable-vicuna-13b-delta	CC BY-NC-SA 4.0
LERT-Base (Cui et al., 2022)	huggingface.co/hfl/chinese-lert-base	Apache License 2.0
DeBERTa-Large (He et al., 2020)	huggingface.co/microsoft/deberta-v3-large	MIT License
BERT-Base (Devlin et al., 2019)	huggingface.co/google-bert/bert-base-chinese	Apache License 2.0
Sentence-BERT (Reimers and Gurevych, 2019)	huggingface.co/sentence-transformers/paraphrase-multilingual -mpnet-base-v2	Apache License 2.0
BLOOM-7B (Workshop et al., 2022)	huggingface.co/bigscience/bloom-7b1	BigScience Responsible AI License 1.0
ChatGLM2-6B (Du et al., 2021)	huggingface.co/THUDM/chatglm2-6b	Apache License 2.0
RoBERTa-Large (Liu et al., 2019)	huggingface.co/hfl/chinese-roberta-wwm-ext-large	Apache License 2.0
RoBERTa-Base (Liu et al., 2019)	huggingface.co/hfl/chinese-roberta-wwm-ext	Apache License 2.0
Baichuan-7B (Yang et al., 2023)	huggingface.co/baichuan-inc/Baichuan-7B	
Visual Model	**Link**	**License**
Data2vec-Base (Baevski et al., 2022)	huggingface.co/facebook/data2vec-vision-base	Apache License 2.0
VideoMAE-Base (Tong et al., 2022)	huggingface.co/MCG-NJU/videomae-base	CC BY-NC 4.0
EVA-02-Base (Fang et al., 2024)	https://huggingface.co/timm/eva02_base_patch14_224.mim_in22k	MIT License
VideoMAE-Large (Tong et al., 2022)	huggingface.co/MCG-NJU/videomae-large	CC BY-NC 4.0
CLIP-Base (Radford et al., 2021)	huggingface.co/openai/clip-vit-base-patch32	Apache License 2.0
Dinov2-Large (Oquab et al., 2023)	huggingface.co/facebook/dinov2-large	Apache License 2.0
Dinov2-Giant (Oquab et al., 2023)	huggingface.co/facebook/dinov2-giant	Apache License 2.0
CLIP-Large (Radford et al., 2021)	huggingface.co/openai/clip-vit-large-patch14	Apache License 2.0
Captioning Model	**Link**	**License**
Transformer-based (Lewis et al., 2019)	huggingface.co/fnlp/bart-base-chinese	Apache License 2.0
GPT-2 (Lagler et al., 2013)	huggingface.co/uer/gpt2-chinese-cluecorpussmall	Apache License 2.0
Qwen-2 (Yang et al.)	huggingface.co/Qwen/Qwen2-7B	Apache License 2.0
Qwen2.5-Omni (Xu et al., 2025a)	https://huggingface.co/Qwen/Qwen2.5-Omni-7B	Apache License 2.0
Qwen3-Omni-30B-A3B (Xu et al., 2025b)	huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct	Apache License 2.0

Table 15: An overview of the models employed across different tasks.