MultiERC: A Dataset for Role-Playing Agents to Develop and Evaluate Emotional and Role Consistencies in Multimodal Emotional Expressions

Anonymous ACL submission

Abstract

Multimodal emotional expressions of Role-Playing Agents (RPAs) need to align with both **Emotional Consistency (EC)** and **Role Consistency (RC)** to enable effective multimodal emotional interaction with users, along with the delivery of emotional value and immersive experiences. However, existing datasets do not support such research. On one hand, multimodal emotional dialogue datasets lack role-related factors and semantic descriptions of multimodal data, leading to unstable inference and evaluation. On the other hand, multimodal role-playing datasets mostly adopt synthetic methods, whose generated data lack in-depth integration of real emotional factors. To facilitate research on EC and RC in the multimodal emotional expressions of RPAs, we propose ***MultiERC***, a multimodal role-playing emotional interaction and dialogue dataset, and design a comprehensive and robust evaluation system. Specifically, based on the video materials from the real multimodal emotional dialogue dataset CPED, we further developed role information and detailed semantic descriptions of multimodal emotional expressions. We also introduce an adapted fine-tuning method to train several baselines. Experimental results demonstrate that the proposed dataset and fine-tuning method indeed improve the performance in terms of EC and RC, and highlight that there remains substantial room for improvement in current baselines.[1]

1 Introduction

Recently, Role-Playing Agents (RPAs) have attracted significant attention and expectations in companion scenarios. This is because RPAs possess the potential to serve as safe targets (Winnicott, 2016) for emotional projection, establish connections with users, and further develop into attachment objects, thereby providing emotional value and immersive experiences in companion scenarios. Furthermore, rather than being confined to a single text modality or emotionally barren, an immersive and in-depth interaction tends to be multimodal and rich in emotional expressiveness, including facial expressions, body movements, and speech intonations. Based on this characteristic, for RPAs to engage effectively in multimodal emotional interaction with users, their multimodal emotional expressions should also satisfy both emotional consistency and role consistency, similarly to textual response contents.

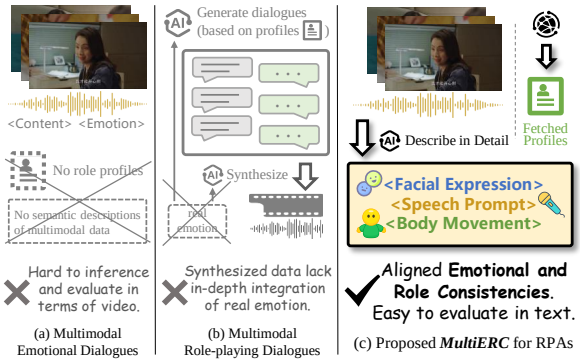

Figure 1: (a) Existing multimodal emotional dialogue datasets. (b) Existing multimodal role-playing dialogues. (c) Proposed dataset has well aligned Emotional and Role Consistencies and is easy to inference and evaluate in text stably and comprehensively.

Emotional Consistency (EC) is typically measured by comparing the predicted contents of RPAs with the groundtruth responses in the context of text conversations (Feng et al., 2025). However, when there are multimodal emotional expressions in model outputs, EC should further measures the degree of emotional consistency across different modalities. Inconsistent emotions across modalities will cause a severe sense of dissonance, which in turn significantly undermines the immersive experience. Therefore, it is entirely necessary to impose requirements on the cross-modal emotional

[1] https://anon-web-design.github.io/MultiERC/

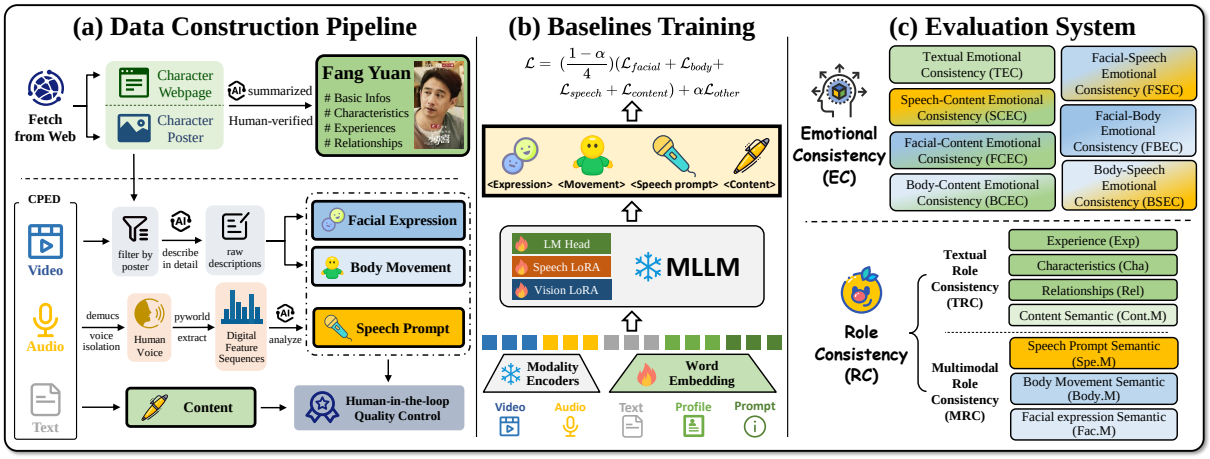

Figure 2: An overview of this paper. (a) Data Construction Pipeline: proposed data construction pipeline includes the acquisition of role profiles, role images, and multimodal detailed semantic descriptions. (b) Baselines Training: we adopt four pairs of special tokens and a crafted loss function. (c) Evaluation System: we designs 14 different kinds of metrics in total for evaluating Emotional Consistency (EC) and Role Consistency (RC).

consistency of RPAs, which is neglected in existing benchmarks of RPAs.

Role Consistency (RC) is a fundamental requirement for RPAs and is typically assessed by multiple dimensions (Tu et al., 2024), including knowledge, personality, behavior, and utterance, among others. However, existing evaluation systems primarily confined to invoke Large Language Models (LLMs) to evaluate the entire model responses in specified dimensions in a generalized manner. This is primarily constrained by earlier RPAs mostly incorporating all multimodal emotional expressions into responses via brackets, rather than adopting a specific format to distinguish them. Moreover, such expressions may also be incomplete. To address this issue, we further require all RPAs to respond in the designated format, and thereby design a set of fine-grained and comprehensive metrics to evaluate RC in multimodal emotional expressions.

Existing datasets also fail to support research on EC and RC. As shown in Figure 1 and Table 1, on the one hand, existing multimodal emotional dialogue datasets lack role-related data (such as role profiles and role images) and semantic descriptions of multimodal data, leading to unstable inference and evaluation, especially in terms of video modality. Specifically, there are currently no open-source video generation models with good performance, nor are there appropriate metrics to accurately evaluate EC and RC from generated role videos. Therefore, it is necessary to convert multimodal data into semantic descriptions, and conduct training, inference, and fair evaluation of EC and RC at the text modality for now. On the other hand, existing multimodal role-playing datasets mostly adopt synthetic methods. Their generated data lack in-depth integration of real emotional factors. Specifically, current work that incorporates emotional factors (Zhang et al., 2025b) only integrates a list of frequent role emotions in profiles to guide synthesis. However, a fixed list of emotions obviously cannot meet the complexity and variability of actual interaction scenarios. In addition, existing work (Feng et al., 2025) has also proven that real data can improve EC, which further confirms the hidden risks that synthetic datasets pose to EC. To address this issue, we construct a new dataset *MultiERC* based on the video materials from the CPED dataset[2] (Chen et al., 2022).

In conclusion, our main contributions can be summarized as follows:

1. We define two crucial consistencies, Emotional Consistency (EC) and Role Consistency (RC), in multimodal emotional expressions of Role-Playing Agents (RPAs) for the first time, and design a comprehensive and robust evaluation system around EC and RC with 14 different metrics in total.
2. We construct *MultiERC*, a multi-turn multimodal role-playing emotional interaction dialogue dataset. To the best of our knowledge,

[2]CPED, a large-scale Chinese Personalized and Emotional Dialogue dataset, contains multi-turn multimodal dialogue data originating from TVs, and several labels including 13 types of emotions, gender, age, Big Five personality traits, 19 types of dialogue acts.

Table 1: Comparison with existing dataset. a, v, and l refer audio, visual, and text respectively. MM refers to MultiModal. EN and CN refer to English and Chinese respectively.

Dataset	Modalities (per turn)	Data Type	Role Profiles	Emotion Labels	Semantic Descriptions of MM data	Language	Dialogue Source
Multimodal Emotional/Empathetic Dialogue Datasets							
MELD (Poria et al., 2019)	(a, v, l)	MM dialogue	✘	✔	✘	EN	"Friends" TV
M3ED (Zhao et al., 2022)	(a, v, l)	MM dialogue	✘	✔	✘	CN	56 TVs
CPED (Chen et al., 2022)	(a, v, l)	MM dialogue	✘	✔	✘	CN	40 TVs
MC-EIU (Liu et al., 2024b)	(a, v, l)	MM dialogue	✘	✔	✘	EN&CN	3 English TVs 4 Chinese TVs
MESC (Chu et al., 2024b)	(a, v, l)	MM dialogue	✘	✔	✘	EN	"In Treatment" TV
AvaMERG (Zhang et al., 2025a)	(a, v, l)	MM dialogue	✘	✔	✘	EN	Text: ED(Li et al., 2021) MM: human acting
Role-Playing Datasets							
Character-LLM (Shao et al., 2023)	(_, _, l)	Pure Text dialogue	✔	✘	✘	EN	Synthesis
CharacterGLM (Zhou et al., 2024)	(_, _, l)	Pure Text dialogue	✔	✘	✘	CN	Real-world, Literary, and Synthesis
CharacterEval-RM (Tu et al., 2024)	(_, _, l)	Pure Text dialogue	✔	✘	✘	CN	Novels
MMRole (Dai et al., 2025)	(_, _, l)	MM based Text dialogue	✔	✘	✘	EN&CN	Synthesis
OmniCharacter-10K (Zhang et al., 2025b)	(a, _, l)	MM dialogue	✔	✘	✘	EN	Synthesis
***MultiERC* (Ours)**	**(a, v, l)**	**MM dialogue**	✔	✔	✔	**CN**	**CPED (Chen et al., 2022)**

this is the first dataset for developing and evaluating EC and RC.

3. We develop and evaluate several baselines with the proposed approach, and organize them into a benchmark. We also conduct two human agreement experiments on proposed data and metrics and prove their effectiveness.

2 Related Work

Existing role-playing datasets can be divided into two aspects: (1) textual role-playing datasets (Wang et al., 2024a; Wu et al., 2024; Zhou et al., 2024; Shao et al., 2023; Li et al., 2023; Chen et al., 2023); (2) multimodal role-playing datasets (Dai et al., 2025; Zhang et al., 2025b). However, these studies are primarily constructed by synthetic methods, and lack in-depth integration of real emotions, especially in multimodal parts. Meanwhile, most of the early multimodal emotional (Poria et al., 2019; Zhao et al., 2022; Chen et al., 2022) or empathetic (Chu et al., 2024b; Zhang et al., 2025a) dialogue datasetslack role-related data and semantic descriptions of multimodal data, which leads to unstable inference and evaluation.

3 Method

3.1 Task Definition

By different input modalities, multimodal emotional interaction can be specifically divided into four subtasks: omni chat, vision-only video chat, audio chat, and text chat. Taking **omni chat** as an example first, given the role profile P_C and role image I_C of character C, RPAs step into shoes of this character and engage in multi-turn multimodal emotional dialogues with user named N_u:

$$e_i = \text{RPAs}(P_c, I_C, N_u, D) \quad (1)$$

$$D = [u_0, e_0, ..., u_i] \quad (2)$$

where D is a dialogue history sequence, with $u_i = [A_i, V_i, T_i]$ as the i-th input from user, including audio A_i, video V_i, and text T_i, and $e_i = [f_i, b_i, s_i, c_i]$ as the i-th turn response from RPAs, including textual response content c_i, facial expression f_i, body movement b_i, and speech prompt (namely speech intonation) s_i. Specifically, u_0 and e_0 are the special input and output cases when $i = 0$, respectively.

Similarly, **vision-only video chat** has the same format but removes audio modality, including user audio input A_i and system speech prompt output s_i. And for **audio chat**, we remove all vision modality from omni chat instead. Finally, when it comes to **text chat**, there are no multimodal input and multimodal emotional expressions in dialogues.

3.2 Data Construction

As shown in Figure 2(a), we construct *MultiERC*, a multi-turn multimodal role-playing emotional interaction dialogue dataset, based on CPED (Chen

et al., 2022). In this section, we briefly introduce the data construction process, leaving details and data examples in Appendix B.

3.2.1 Role Profiles and Images

Role profiles are crucial for all kinds of RPAs. Moreover, role images also assume an equivalently critical function by enabling self-localization of RPAs within videos captured from a third-person perspective. Therefore, we retrieve role information and posters from BaiduBaike[3], and use Deepseek-V3[4] to summarize into profiles containing four primary parts: basic information, characteristics, experiences, and interpersonal relationships. Finally we recruit a team of 20 volunteer students to check and correct the summarized profiles and role images based on webpages.

3.2.2 Multimodal Detailed Semantic Descriptions

For stable and convenient development and evaluation, we propose using detailed semantic descriptions instead of videos, including facial expressions, body movements, and speech prompts, namely intonations. Notably, the generation pipelines for these descriptions exhibit modality-specific adaptations, thereby ensuring high-quality results. Besides, it is not instructed to imitate the speaking styles of different roles, but simply provide the descriptions.

3.2.3 Assemble and Quality Control

In this step, we assemble all components above and retain those dialogues with two speakers only. Leveraging the power of cooperation between human and LLMs, we further conduct human-in-the-loop quality control to filter or modify those low-quality annotations. Finally we divided the dataset into a training set and a test set based on dialogues at a ratio of 9:1. The statistics are as shown in Tables 2 and 3. Besides, detailed statistics of roles such as age, gender and personality trait can be referred to the CPED dataset (Chen et al., 2022).

Table 2: Statistics of train and test splits of *MultiERC*.

Split	# Samples	# Roles	# Dialogues
Train	17072	256	6220
Test	1894	212	691

[3] https://baike.baidu.com/, a Chinese Encyclopedia

[4] Deepseek is a top-tier LLM with up to 671B parameters developed by DeepSeekAI. Link: https://deepseek.ai/

Table 3: Statistics of whole proposed dataset *MultiERC*.

#samples	18966
#dialogues	6911
#roles	256
Avg. of #utterances per role	101.08
Avg. durations of video (seconds)	6.00
Avg. tokens of facial expression part	141.61
Avg. tokens of body movement part	137.40
Avg. tokens of speech prompt part	68.04
Avg. tokens of content part	33.34
Avg. tokens of role profile	520.43

3.3 Baselines Training

We propose using several pairs of special tokens to divide the model output into several parts. We also design a corresponding loss function to balance the learning of various output parts and formatting. Specifically, taking omni chat as an example, the crafted new loss function is:

$$\mathcal{L} = (\frac{1-\alpha}{4})(\mathcal{L}_{facial} + \mathcal{L}_{body} + \mathcal{L}_{speech} + \mathcal{L}_{content}) + \alpha\mathcal{L}_{other} \quad (3)$$

In this way we attach equal importance to each output part. Notably, $\mathcal{L}_{other}$ is primarily exerted on special tokens, thereby balancing the learning of formatting. When it comes to audio chat, since it only has two output parts, the former part of loss function would be changed to $(\frac{1-\alpha}{2})(\mathcal{L}_{speech} + \mathcal{L}_{content})$. Similar change is adapted for vision-only video chat. During main experiment, we set $\alpha = 0.1$, hypothesizing that this configuration can reach the best performance, which is further proved in ablation study. For text chat, as the output has only one part with no required additional format, there is no need for special tokens for division, and thus no modification to the loss function.

3.4 Evaluation System

In order to evaluate Emotional Consistency (EC) and Role Consistency (RC) comprehensively and robustly, we propose an evaluation system. As illustrated in Figure 2(c), it includes 14 different kinds of metrics in total.

3.4.1 Definition of Indicators

During evaluation we essentially use two indicators: LLMScore (Gu et al., 2025) and SEMScore.

LLMScore is implemented following the current mainstream approach (Gu et al., 2025): using a top-tier LLM as judge, providing instructions, scoring criteria, and few-shot examples, and then

Table 4: Seven metrics to evaluate Emotional Consistency (EC) of role-playing agents.

Metrics	Element A	Element B
Textual Emotional Consistency (TEC)	predicted content	target content
Speech-Content Emotional Consistency (SCEC)	predicted speech prompt	predicted content
Facial-Content Emotional Consistency (FCEC)	predicted facial expression	predicted content
Body-Content Emotional Consistency (BCEC)	predicted body movement	predicted content
Facial-Speech Emotional Consistency (FSEC)	predicted facial expression	predicted speech prompt
Body-Speech Emotional Consistency (BSEC)	predicted body movement	predicted speech prompt
Facial-Body Emotional Consistency (FBEC)	predicted facial expression	predicted body movement

Table 5: Seven metrics to evaluate Role Consistency (RC) of role-playing agents. TRC refers to Textual Role Consistency. MRC refers to Multimodal Role Consistency.

Metrics	Evaluate on	Indicator	Belongs to
Experience consistency (Exp)	content	LLMScore	TRC
Characteristic consistency (Cha)	content	LLMScore	TRC
Relationship consistency (Rel)	content	LLMScore	TRC
Content Semantic Similarity(Cont.M)	content	SEMScore	TRC
Speech Prompt Semantic Similarity (Spe.M)	speech prompt	SEMScore	MRC
Body Movement Semantic Similarity (Body.M)	body movement	SEMScore	MRC
Facial expression Semantic Similarity (Fac.M)	facial expression	SEMScore	MRC

requiring it to output scores along with explanations. Since both Doubao Series[5] and GPT Series[6] are used in the experiments of this paper, we employ DeepSeek-V3 as the judge. We also specify the scoring range as 0 to 5.

SEMScore is typically the average of several semantic similarity scores in various researches (Xie et al., 2025), including BLEU-4[7] (Papineni et al., 2002), ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), F1 scores of BERTScore (Zhang et al., 2020). For BERTScore, we specify the state-of-the-art open-source 12-layer Chinese BERT model[8]. Notably, the layer selection experiment of BERTScore (Zhang et al., 2020) is conducted on a machine translation dataset WMT16 (Bojar et al., 2016), and they also claimed that selecting a good layer from the BERT model is important (Zhang et al., 2020). Therefore, to avoid bias from a single layer and improve the robustness of BERTScore, we specify $F1_{12}$ from the final layer of the BERT model, and $F1_8$ from 8th layer (originally selected) of the same model. Since all score are ranging from 0 to 100, the SEMScore can be calculated as:

$$SEMScore = \frac{B_4 + R_1 + R_2 + R_L + F1_8 + F1_{12}}{6} \tag{4}$$

where B_4 refers to BLEU-4, R_1, R_2 and R_L refer to ROUGE-1, ROUGE-2, ROUGE-L respectively. Finally, when calculating average scores, SEMScores are divided by 20 to maintain a consistent value range of 0 to 5 with LLMScore.

3.4.2 Emotional Consistency (EC)

EC is typically measured by comparing the predicted contents of RPAs with those of groundtruth responses in the context of text conversations (Feng et al., 2025), which refer to TEC in this paper. When there are multimodal emotional expressions, EC should further measure the degree of emotional consistency across different modalities. Thus we evaluate EC by seven metrics with LLMScore, as shown in Table 4.

3.4.3 Role Consistency (RC)

As shown in Figure 2 and Table 5, we evaluate RC on two main dimensions: Textual Role Consistency (TRC) and Multimodal Role Consistency (MRC). Among them, TRC is typically required across various benchmarks (Wang et al., 2024b; Tu et al., 2024), while MRC is designed to bridge the gap on multimodal emotional expressions left by previous works.

3.4.4 Average and Overall

When models deviate from the format, required contents fail to be extracted from outputs, preventing the calculation of metrics above. Thereby computing averages overall test set would introduce outliers (NaN value). To ensure that the preceding metrics accurately reflect the corresponding capabilities of model, we calculate partial averages on existing values for proposed metrics. We then incorporate format accuracy when calculating the average scores of EC or RC, thereby transforming the partial averages $Score_{partial}$ into the final

[5] Doubao Series are a set of top-tier LLMs developed by ByteDance. Link: https://www.doubao.com/.

[6] GPT Series are a set of top-tier LLMs developed by OpenAI. Link: https://openai.com/api/.

[7] A n-gram BLEU-n would average from 1-gram to n-gram.

[8] https://huggingface.co/shibing624/text2vec-base-chinese-paraphrase

Table 6: Results of vision-only video chat across all test samples. *trained(subtask)* means one model is trained on specific subtask. The percentages beside some scores are ACC_{format}, which would be multiplied by the scores when calculating averages of EC or RC. ACC_{format} of Exp, Cha, Rel, and Cont.M are the same as TEC and thus omitted. SEMScores are divided by 20 during averaging to maintain a consistent value range with LLMScore.

RPAs	EC					RC							Overall
	TEC	FCEC	BCEC	FBEC	Avg.	Exp	Cha	Rel	Cont.M	Fac.M	Body.M	Avg.	
Doubao-1.5-thinking-vision-pro	2.44(100.00%)	4.30(100.00%)	4.03(100.00%)	4.05(100.00%)	3.71	3.82	4.56	4.33	23.22	24.67(100.00%)	25.38(100.00%)	2.73	**3.22**
GPT-4o	2.25(99.95%)	4.01(99.89%)	3.86(99.89%)	3.93(99.89%)	3.51	3.70	4.47	4.29	22.08	24.82(99.89%)	26.51(99.89%)	2.69	**3.10**
Qwen2.5-VL-Instruct(7B)	2.13(99.84%)	3.50(99.63%)	3.19(99.84%)	3.68(99.79%)	3.12	3.20	3.93	3.76	21.73	23.64(99.79%)	25.42(100.00%)	2.40	**2.76**
Qwen2.5-VL-Instruct(7B) + trained(vision)	2.14(99.68%)	3.41(99.68%)	3.44(99.68%)	4.91(100.00%)	3.47	2.99	3.57	3.52	22.25	36.52(100.00%)	36.07(100.00%)	2.46	**2.97**
Qwen2.5Omni(7B)	1.86(49.73%)	2.88(31.57%)	2.65(27.67%)	2.87(23.86%)	0.81	2.86	3.37	3.39	17.67	19.19(39.97%)	19.15(32.00%)	0.99	**0.90**
Qwen2.5Omni(7B) + trained(vision)	2.14(99.05%)	3.49(99.05%)	3.53(99.05%)	4.92(100.00%)	3.50	2.91	3.49	3.45	22.25	36.48(100.00%)	35.92(100.00%)	2.41	**2.96**
MiniCPM-o-2.6(8B)	2.13(99.95%)	3.52(99.95%)	3.38(99.95%)	3.75(100.00%)	3.19	2.87	3.50	3.37	21.53	24.77(100.00%)	26.95(100.00%)	2.23	**2.71**
MiniCPM-o-2.6(8B) + trained(vision)	2.24(91.29%)	2.93(85.27%)	3.22(87.64%)	4.87(85.27%)	2.88	3.00	3.68	3.58	23.96	37.51(85.69%)	36.49(87.75%)	2.28	**2.58**

overall average $Score_{overtest}$ on the entire testset:

$$Score_{overtest} = Score_{partial} * ACC_{format} = \frac{Score_{sum}}{valid\ values} * \frac{valid\ values}{testset} \quad (5)$$

where $valid\ values$ refer to the number of valid scores, $testset$ refers to the total number of testset samples. Notably, format accuracies of TEC in EC (see Table 4) and other 4 metrics of TRP in RC (see Table 5) are actually the same, since they are all solely calculated on predicted content.

Table 7: Results of text chat across all test samples. This table has the same format as Table 6. ACC_{format} is omitted since no format is required on text chat subtask.

RPAs	EC	RC					Overall
	TEC	Exp	Cha	Rel	Cont.M	Avg.	
Doubao-1.5-thinking-vision-pro	1.91	4.24	4.74	4.52	20.70	3.63	**2.77**
GPT-4o	2.02	4.05	4.59	4.43	20.11	3.52	**2.77**
Qwen2.5-VL-Instruct(7B)	2.03	3.37	4.02	3.91	20.71	3.08	**2.56**
Qwen2.5-VL-Instruct(7B) + trained(text)	2.11	3.07	3.68	3.64	21.38	2.86	**2.49**
Qwen2-Audio-Instruct(7B)	1.86	2.98	3.42	3.39	20.02	2.70	**2.28**
Qwen2-Audio-Instruct(7B) + trained(text)	2.17	3.01	3.73	3.60	20.70	2.85	**2.51**
Qwen2.5-Omni(7B)	1.58	2.89	3.25	3.30	17.66	2.58	**2.08**
Qwen2.5-Omni(7B) + trained(text)	2.14	3.12	3.70	3.69	21.50	2.90	**2.52**
MiniCPM-o-2.6(8B)	1.54	3.08	3.66	3.59	14.38	2.76	**2.15**
MiniCPM-o-2.6(8B) + trained(text)	2.16	2.96	3.57	3.53	21.42	2.78	**2.47**

4 Experiments

4.1 Experiment Settings

We fine-tune Qwen2.5-Omni(7B) (Jin Xu, 2025) and MiniCPM-o-2.6(8B) (Yao et al., 2024) with LoRA (Hu et al., 2021) using 1*A800 80GB GPU on the training set of *MultiERC* for 3 epochs. Besides, several baselines are trained on Qwen2-Audio(7B) (Chu et al., 2024a) and Qwen2.5-VL(7B) (Bai et al., 2025) using 2*4090 24GB GPUs with same configurations. Details of training arguments can be found in Appendix C. During inference, we use the same generation config for all baselines, specifying a temperature of 0.6, top-p of 0.7, top-k of 50, while other settings remain as the default configurations of each base model. Notably, closed-source top-tier LLMs are primarily confined to processing vision-only videos, and fail to directly chat on audio files. Thus we replace all audio input files of the dialogue context by their corresponding speech prompts for these models. Specifically, we select Doubao-1.5-thinking-vision-pro (Guo et al., 2025) and GPT-4o (OpenAI, 2024) for comparison.

4.2 Main Results

We report the average results of Emotional Consistency (EC) and Role Consistency (RC) on four different subtasks of multimodal emotional interaction across all test samples for each evaluated role-playing agent. As shown in Tables 6, 7, 8 and 9, fine-tuning on our dataset with the proposed approach brings consistent improvements on overall score and various metrics across subtasks for most models, validating the effectiveness and robustness of proposed dataset and training method.

Detailed results further indicate that different base models may have different adaptability on specific subtasks. Specifically, fine-tuned *Qwen2.5-Omni(7B)* exhibits consistent improvements across all metrics and subtasks, highlighting the strong adaptability of the base model on all modalities. The other omni model, *MiniCPM-o-2.6(8B)*, performs comparably in audio and omni chat fine-tuning but yields less satisfactory under vision and text chat. This is probably because *MiniCPM-o-2.6(8B)* relies more on audio modality, since its

Table 8: Results of audio chat across all test samples. This table has the same format as Table 6.

RPAs	EC			RC						Overall
	TEC	SCEC	Avg.	Exp	Cha	Rel	Cont.M	Spe.M	Avg.	
Doubao-1.5-thinking-vision-pro	2.30(100.00%)	4.45(100.00%)	3.38	3.94	4.61	4.38	21.49	38.44(100.00%)	3.18	**3.28**
GPT-4o	2.27(99.95%)	4.29(99.95%)	3.28	3.70	4.47	4.32	21.64	39.53(99.95%)	3.11	**3.19**
Qwen2-Audio-Instruct(7B)	2.12(14.20%)	3.38(0.84%)	0.16	3.11	3.55	3.54	20.34	29.33(1.53%)	0.32	**0.24**
Qwen2-Audio-Instruct(7B) + trained(audio)	2.06(99.00%)	3.27(97.73%)	2.62	2.90	3.50	3.36	20.85	42.96(98.42%)	2.56	**2.59**
Qwen2.5-Omni(7B)	1.80(63.57%)	3.01(54.33%)	1.39	2.87	3.35	3.40	17.61	20.00(71.70%)	1.48	**1.43**
Qwen2.5Omni(7B) + trained(audio)	2.24(95.88%)	3.55(95.88%)	2.77	3.10	3.79	3.69	22.13	55.43(100.00%)	2.79	**2.78**
MiniCPM-o-2.6(8B)	2.13(99.21%)	2.94(99.16%)	2.51	2.81	3.36	3.32	21.56	28.52(99.47%)	2.38	**2.45**
MiniCPM-o-2.6(8B) + trained(audio)	2.30(99.95%)	3.04(99.52%)	2.66	3.02	3.69	3.57	22.47	49.43(99.58%)	2.77	**2.72**

Table 9: Results of omni chat across all test samples. This table has the same format as Table 6.

RPAs	EC							
	TEC	SCEC	FCEC	BCEC	FSEC	BSEC	FBEC	Avg.
Doubao-1.5-thinking-vision-pro	2.50(100.00%)	4.41(100.00%)	4.33(100.00%)	4.11(100.00%)	4.58(100.00%)	4.21(100.00%)	4.05(100.00%)	4.03
GPT-4o	2.26(100.00%)	4.26(100.00%)	3.99(100.00%)	3.88(100.00%)	4.25(100.00%)	4.12(100.00%)	3.93(100.00%)	3.81
Qwen2.5-Omni(7B)	1.74(29.62%)	2.83(14.20%)	2.87(17.74%)	2.75(15.96%)	3.15(15.05%)	3.41(14.68%)	3.00(10.61%)	0.45
Qwen2.5-Omni(7B) + trained(omni)	2.17(98.94%)	3.47(98.94%)	3.36(98.94%)	3.47(98.94%)	4.80(100.00%)	4.74(100.00%)	4.93(100.00%)	3.83
MiniCPM-o-2.6(8B)	2.14(99.52%)	3.15(99.52%)	3.01(99.47%)	2.85(99.52%)	3.21(99.47%)	3.02(99.47%)	3.27(99.47%)	2.93
MiniCPM-o-2.6(8B) + trained(omni)	2.22(100.00%)	3.43(100.00%)	2.93(100.00%)	3.27(100.00%)	4.56(100.00%)	4.49(100.00%)	4.89(100.00%)	3.69

RPAs	RC								Overall
	Exp	Cha	Rel	Cont.M	Spe.M	Fac.M	Body.M	Avg.	
Doubao-1.5-thinking-vision-pro	3.76	4.52	4.27	23.60	31.89(100.00%)	25.46(100.00%)	26.29(100.00%)	2.56	**3.29**
GPT-4o	3.65	4.42	4.27	22.09	35.57(100.00%)	26.31(100.00%)	27.47(100.00%)	2.56	**3.19**
Qwen2.5-Omni(7B)	2.79	3.27	3.39	16.25	18.31(18.85%)	19.96(28.25%)	19.03(23.23%)	0.53	**0.49**
Qwen2.5-Omni(7B) + trained(omni)	2.93	3.51	3.51	22.42	52.45(100.00%)	36.45(100.00%)	36.20(100.00%)	2.46	**3.14**
MiniCPM-o-2.6(8B)	2.63	3.22	3.21	21.43	30.55(99.68%)	24.72(99.63%)	26.17(99.68%)	2.02	**2.48**
MiniCPM-o-2.6(8B) + trained(omni)	2.95	3.59	3.57	23.37	52.10(100.00%)	37.12(100.00%)	36.31(100.00%)	2.51	**3.10**

speech decoder directly connects to the main backbone (Yao et al., 2024), unlike *Qwen2.5-Omni(7B)* which uses a Talker module as a bridge (Jin Xu, 2025). Furthermore, fine-tuned *Qwen2-Audio(7B)* shows comprehensive improvement in text chat but only gains in format accuracy for audio chat despite higher overall scores. This indicates its good adaptability to pure-text role-playing but poor audio generalization, which reflected in extremely low ACC_{format}. Fine-tuning such poorly adapted model risks catastrophic forgetting (Zhai et al., 2023) and overfitting problems, thus potentially reducing performance. When it comes to *Qwen2.5-VL-Instruct(7B)*, most *EC* metrics have improved, but the *Exp*, *Cha*, and *Rel* metrics in *RC*, which are related to role knowledge, have decreased to a comparable level with other trained models. This likely stems from knowledge confusion in a single LoRA module after extensive role-specific fine-tuning, restricting generalization upper bound. This also highlights the necessity of a more effective and robust multi-role training approach in the future. Finally, although there are still some gaps between our trained baselines and closed-source top-tier multimodal LLMs, both of them show substantial room for improvement.

In summary, experimental results demonstrate the effectiveness of proposed dataset *MultiERC* and fine-tuning approach. But it is also highlighted that there remains substantial room for improvement in EC and RC metrics for all models.

4.3 Ablation Study

As shown in Table 10, during module-wise ablation, as expected, the *Overall* score exhibit declines. Specifically, without new loss, models learn less from special tokens and content but more on multimodal expressions, lowering ACC_{format}, *TEC* and *TRC* (including *Exp*, *Cha*, *Rel*, and *Cont.M*). Excessive focus on expressions also causes overfitting, reducing performance. Ablating special tokens disables proposed new loss, shifting attention more to format than expressions and content. This mitigates expression overfitting but leads to content underfitting. Notably, ablating profile consistently causes unstable training, confirming its necessity

Table 10: Ablation study results across all test samples, where *Ours* refers to activating role profiles, special tokens, and new loss with $\alpha = 0.1$. The minus sign '-' means removing certain components. All ablations are conducted on Qwen2.5-Omni(7B) under omni chat subtask. The percentage has the same meaning as Table 6.

RPAs	EC							
	TEC	SCEC	FCEC	BCEC	FSEC	BSEC	FBEC	Avg.
Ours	2.17(98.94%)	3.47(98.94%)	3.36(98.94%)	3.47(98.94%)	4.80(100.00%)	4.74(100.00%)	4.93(100.00%)	3.83
Ours - new_loss	2.12(96.57%)	3.46(96.57%)	3.40(96.57%)	3.49(96.57%)	4.78(96.57%)	4.73(96.57%)	4.93(96.57%)	3.71
Ours - spe_token	2.12(99.05%)	3.54(99.05%)	3.41(99.05%)	3.51(99.05%)	4.87(99.05%)	4.80(99.05%)	4.91(99.05%)	3.84
Ours - profile	1.89(97.62%)	3.52(97.62%)	3.82(97.62%)	3.75(97.62%)	4.47(97.62%)	4.47(97.62%)	4.70(97.62%)	3.71
Ours - profile - new_loss	1.84(97.31%)	3.48(97.20%)	3.84(97.15%)	3.84(97.15%)	4.49(97.15%)	4.54(97.15%)	4.70(97.15%)	3.71
Ours - profile - spe_token	1.66(68.69%)	3.39(67.32%)	3.04(62.25%)	3.30(65.84%)	3.69(62.25%)	3.82(62.25%)	4.03(62.25%)	2.10
Ours ($\alpha = 0.01$)	2.19(72.60%)	3.40(72.60%)	3.22(72.60%)	3.40(72.60%)	4.78(72.60%)	4.73(72.60%)	4.93(72.60%)	2.76
Ours ($\alpha = 0.2$)	2.14(99.52%)	3.45(99.52%)	3.40(99.52%)	3.46(99.52%)	4.81(100.00%)	4.76(100.00%)	4.94(99.52%)	3.84

RPAs	RC								Overall
	Exp	Cha	Rel	Cont.M	Spe.M	Fac.M	Body.M	Avg.	
Ours	2.93	3.51	3.51	22.42	52.45(100.00%)	36.45(100.00%)	36.20(100.00%)	2.46	**3.14**
Ours - new_loss	2.88	3.42	3.46	21.96	52.19(100.00%)	36.61(100.00%)	36.03(100.00%)	2.39	**3.05**
Ours - spe_token	2.90	3.33	3.41	21.20	52.16(100.00%)	36.06(100.00%)	35.44(100.00%)	2.40	**3.12**
Ours - profile	2.44	3.03	3.01	20.19	38.46(98.68%)	31.37(100.00%)	29.71(99.79%)	2.03	**2.87**
Ours - profile - new_loss	2.42	2.98	2.97	19.94	37.62(98.05%)	31.22(99.74%)	29.53(98.57%)	2.00	**2.85**
Ours - profile - spe_token	2.22	2.78	2.75	19.20	36.42(68.06%)	29.25(85.22%)	27.44(68.53%)	1.34	**1.72**
Ours ($\alpha = 0.01$)	2.93	3.56	3.53	22.30	52.42(100.00%)	36.47(100.00%)	35.82(100.00%)	2.05	**2.40**
Ours ($\alpha = 0.2$)	2.91	3.48	3.46	22.41	52.55(100.00%)	36.72(100.00%)	35.97(100.00%)	2.45	**3.14**

Table 11: Results of human agreement experiment on description quality. κ refers to Quadratic Weighted Fleiss' Kappa coefficient. $\bar{s}$ refers to the average score.

	facial expression	body movement	speech prompt
$\bar{s}$	4.148	4.133	4.572
κ	0.5051	0.4093	0.4392

Table 12: Quadratic Weighted Fleiss' Kappa between human scores and LLMScores. *human* refers to inter-human agreement, while *human&llm* refers to human-LLM agreement.

	TEC	**SCEC**	**FCEC**	**BCEC**	**FSEC**
human	0.4721	0.6142	0.7133	0.5141	0.5907
huamn&llm	0.4748	0.6152	0.6393	0.4741	0.5775
	BSEC	**FBEC**	**Exp**	**Cha**	**Rel**
human	0.5773	0.8208	0.4324	0.5109	0.5983
huamn&llm	0.5647	0.7822	0.4030	0.4814	0.5604

for RPAs. Furthermore, a lower $\alpha = 0.01$ allows for more training on content, which is advantageous for *TEC* and *TRC*. However, reduced focus on special tokens leads to a decline in ACC_{format}, and overfitting may also manifest in multimodal expressions. In contrast, a higher $\alpha = 0.2$ leads to opposite behavior.

4.4 Human Agreement Experiment

We recruit four graduate students to conduct human agreement experiments on data quality and evaluation metrics that use LLMScore as an indicator. Human raters are asked to score: (1) the semantic consistency between three synthesized descriptions and real-world videos over 50 randomly selected samples in the whole dataset; (2) on each metric with the same instruction as LLMScore, over 20 randomly selected samples in main results. All scores are ranging from 0 to 5. Since the scoring results are ordinal data evaluated by multiple raters, we use the Quadratic Weighted Fleiss' Kappa coefficient (Fleiss, 1971) for consistency evidence. Average scores in Table 11 demonstrate substantial consistency between synthesized descriptions and real videos, with moderate confidence (greater than 0.4). Results in Table 12 also indicates that LLMScore has moderate agreement with human on all proposed metrics. Details of ethics are in Appendix A.2.

5 Conclusion

Emotional Consistency (EC) and **Role Consistency (RC)** are both crucial for Role-Playing Agents (RPAs) to engage effectively in multimodal emotional interaction with users. However, existing datasets and evaluation systems do not support such research. Therefore, we propose ***MultiERC***, a multimodal role-playing emotional interaction and dialogue dataset, and design a comprehensive and robust evaluation system. We also introduce a fine-tuning method for baselines. Experimental results finally demonstrate the effectiveness.

Limitations

We acknowledge the language and genre bias in proposed dataset *MultiERC* since all video materials are from Chinese TV Series. Besides, although proposed evaluation system pragmatically proxies multimodal generation via semantic descriptions to evaluate stably, a more comprehensive evaluation system on video-generation model is necessary in the future. To address these limitations, we would soon develop a multilingual expansion of our dataset based on various dialogue video materials. We are also planning to complete the evaluation system by developing robust metrics for emotional and role consistencies in videos generated by several top-tier video-generation LLMs. This initiative could then facilitate the advancement of open-source video-generation role-playing models in the near future.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024a. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2024b. Towards multimodal emotional support conversation systems. *Preprint*, arXiv:2408.03650.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *Preprint*, arXiv:2408.04203.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6218–6240, Albuquerque, New Mexico. Association for Computational Linguistics.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 178 others. 2025. Seed1.5-vl technical report. *Preprint*, arXiv:2505.07062.

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

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2021. Knowledge bridging for empathetic dialogue generation. *Preprint*, arXiv:2009.09708.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024a. Generative expressive conversational speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4187–4196, New York, NY, USA. Association for Computing Machinery.

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

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 23*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2024. Promptttts++: 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.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Donald W. Winnicott. 2016. 159transitional objects and transitional phenomena: A study of the first not-me possession1. In *The Collected Works of D. W. Winnicott: Volume 4, 1952-1955*. Oxford University Press.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3271–3290, Bangkok, Thailand. Association for Computational Linguistics.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *Preprint*, arXiv:2309.10313.

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025a. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. *Preprint*, arXiv:2502.04976.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang, and Yongbin Li. 2025b. Omnicharacter: Towards immersive role-playing agents with seamless

speech-language personality interaction. *Preprint*, arXiv:2505.20277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

A Ethics

A.1 Ethics in Potential Risks

This work adheres to the ACL Code of Ethics, ensuring ethical compliance throughout all stages of the research. The *MultiERC* dataset was constructed based on CPED (Chen et al., 2022), a partially open-source dataset (video materials are closed-source), under the approval of all related authors of CPED. Other data, such as role profiles and images, are all publicly available, with rigorous quality control to prevent privacy risks. We acknowledge the potential biases present in data and have taken proactive measures to ensure diversity and reduce these biases. Furthermore, we recognize that RPAs may generate responses that could misleadingly appear as actual statements made by real individuals. To prevent misunderstanding, it is essential to explicitly indicate that such content is simulated and does not represent genuine speech. Moreover, the training baselines of this research is developed based on LlamaFactory (Zheng et al., 2024) and follows the Apache 2.0 license. Finally, we confirm that Ai Assistants, such as Doubao[9], are only used for translation, and the generated results would be further verified by authors, preventing directly used in any part of the writing.

A.2 Ethics in Human Agreement

We recruited five graduate students who were unfamiliar with one another to evaluate the consistency between three detailed semantic descriptions and the corresponding expressions in 50 randomly selected video samples from the test set. They were asked to give a score ranging from 0 to 5 for the consistency, without being informed that these descriptions were actually the ground truth of our proposed dataset. Before the experiment, we specified a clear scoring criterion for these evaluators:

- 0 points: The detailed semantic description is *completely inconsistent* with the video;
- 3 points: The detailed semantic description is *basically consistent* with the video;
- 5 points: The detailed semantic description is *perfectly consistent* with the video.

Additionally, we further provided some well-crafted and out-of-domain scoring examples for evaluators to reference. And we've informed all human participants of potential risks, and taken reasonable measures to reduce risks, including anonymization, confidentiality agreement, the announcement of data usage, etc. We finally conducted this experiment under approval of ethics review board from institution.

B Details of Data Construction Pipeline

As shown in Figure 2(a), we construct *MultiERC*, a multi-turn multimodal role-playing emotional interaction dialogue dataset, based on CPED (Chen et al., 2022).

B.1 Role Profiles and Role Images

Given that the CPED dataset (Chen et al., 2022) is all from Chinese TV Series, we retrieve all role-related data from BaiduBaike[10], a Chinese encyclopedia platform. We filter out invalid roles and corresponding dialogues, and provide the webpages of characters to Deepseek-V3 to summarize the

[9] https://www.doubao.com/
[10] https://baike.baidu.com/

profiles with four parts, including basic information, characteristics, experiences, and interpersonal relationships. Finally we recruit a team of 20 volunteer students from school to check and correct the summarized profiles and role images based on webpages.

B.2 Multimodal Detailed Semantic Descriptions

For stable and convenient development and evaluation, we propose using detailed semantic descriptions instead of videos, including facial expressions, body movements, and speech prompts, namely intonations. Notably, the generation pipelines for these descriptions exhibit modality-specific adaptations, thereby ensuring high-quality results.

Specifically, we first separate all videos into audios and vision-only videos. For vision-only videos, we first extract several frames from each, with assigned $fps = 4$, $max_videolen = 50$, and $min_videolen = 2$. Then we utilize GPT4o-mini[11] (OpenAI, 2024), a high-cost-performance closed-source MLLM, to capture and summarize facial expressions and body movements from these frames for each video. For audios, due to the fact that there is currently no closed-source high-performance MLLM that can extract and describe speech tone of assigned audio files in Chinese, well and cost-effectively, we propose another novel approach for speech prompt generation, inspired by PromptTTS++ (Shimizu et al., 2024) and NCSSD (Liu et al., 2024a). First, we use Demucs (Rouard et al., 2023) to separate the vocals from the original audios which might comprise background music, sound effects, and other noises. And we initially filter out audios with an $SNR < 4$. Second, we use PyWORLD[12] to extract the numerical feature sequences of *pitches* and *volumes* from the vocal files. Subsequently, we calculate the *speeds* by dividing the durations of the vocal segments by the numbers of words uttered within those segments. To ensure data quality, we filter out error-prone vocal files with a *speed* value less than 0.5. Third, we further normalize and standardize the numerical feature sequences on the entire dataset to obtain new ones with a value range of $[-1, 1]$. In this way, the describer LLM would have a more macroscopic understanding of the numerical values to obtain more accurate descriptions. Finally, we use Deepseek-V3[13] to analyze the numerical feature sequences to obtain the final description.

Besides, during the pipeline above, it is not instructed to imitate the speaking styles of different roles, but simply provide the detailed semantic descriptions of multimodal data.

B.3 Assemble and Quality Control

In this step, we assemble all components above and only retain the sentences that contain all three detailed semantic descriptions simultaneously, then filter to retain those dialogues with two speakers only. We further conduct human-in-the-loop quality control, leveraging the power of cooperation between human and LLMs. Specifically, we first instruct Doubao-seed-1-6-vision-250815[14] to filter out those error samples, and give suggestions on correcting, then we instruct volunteers to correct these samples with suggestions. After two rounds of filtering and correction, we acquire a high-quality dataset in the end. Finally, we divided the dataset into a training set and a test set based on a ratio of 9:1.

C Training Arguments

We primarily fine-tuned the Qwen2.5-Omni(7B) (Jin Xu, 2025) and MiniCPM-o-2.6 (Yao et al., 2024) with LoRA (Hu et al., 2021) using 1*A800 80GB GPU on the training set of *MultiERC* for 3 epochs, to develop our training baselines for omni chat. Besides we also trained several baselines on Qwen2-Audio(7B) (Chu et al., 2024a) and Qwen2.5-VL(7B) (Bai et al., 2025) using 2*4090 24GB GPUs for comparison. Notably, we used llamafactory for training. For data preparation, we set $video_fps$ as 2.0, $image_max_pixels$ and $video_max_pixels$ both as 50176, $video_max_length_total$ as 38. We also adopt a balanced frame sampling strategy for videos in multi-turn dialogues. Specifically, we calculated the proportion of each video's duration to the total video duration, and determined the number of frames to sample for each video by multiplying this proportion by the $video_max_length_total$. During training, we set $lora_rank$ as 8, $lora_alpha$ as 16, $lora_dropout$ as 0.05 for all training baselines. We

[11] https://openai.com/api/

[12] A fast and high-quality vocoder which parameterizes speech into three components: pitch contour, harmonic spectral envelope, aperiodic spectral envelope. Here we primarily use the first two features.

[13] https://deepseek.ai/

[14] https://www.doubao.com/

also adopt a learning rate equals to 1e-5 with cosine scheduler type and warm-up ratio as 0.1. Notably we activated a validation set randomly sampled from train set with a ratio of 0.1, which means only 15363 samples are actually trained in each epoch.

D Prompt Templates

D.1 Prompts for Multimodal Detailed Semantic Descriptions

We provide a unified system prompt *"You are a useful AI assistant."*, and different user prompts for instructing GPT to give multimodal detailed semantic descriptions on videos and audios, as shown in Fig 4 and Fig 5.

D.2 Prompts for Agents

For agent implementation of EMRPA, we provide role profile and role image in system prompt, leaving all multi-turn multimodal dialogue history in subsequent messages. For example, in multimodal dialogue history of omni chat, user would input *"{content}\n{vision-only video}\n{audio}\n"* each turn, and assistant would response in the required format. It has similar form in video chat and audio chat, but remove corresponding missed modalities. And for text chat, assistant directly outputs *content* without any format in each turn. The system prompts for different modality input are as shown in Fig 6.

D.3 Prompts for LLMScores

We provide a unified system prompt *"You are a grader who is good at scoring. Your scoring range is 0-5, and when scoring, you only need to output an integer between 0 and 5 without giving an explanation."*, and different user prompts for instructing GPT to score during evaluation, as shown in Fig 7, Fig 8, Fig 9.

人物基本信息
方圆是电视剧《小欢喜》中的主要角色之一，他是一位中年男性，出生于1973年8月17日，血型为AB型。

人物性格
方圆性格开朗，善于与人交往，对待朋友和家人都非常真诚和友好。他与妻子童文洁关系融洽，两人在生活中相互扶持，共同面对各种挑战。他对儿子方一凡实行宽容政策，只希望儿子健康成长，在成绩方面没有过高要求，把儿子当朋友，理解和支持儿子的演艺梦。他还与好友季胜利、乔卫东等人保持着良好的关系，经常与他们交流生活中的点滴。方圆情绪稳定，也很有责任心，而且能很快适应新环境。方圆虽然经历了中年危机，但他依然保持积极向上的态度，即使改行成为网约车司机甚至兼职影视剧配音，也会努力在新工作中寻找自己的价值。他的生活哲学是乐观面对一切，即使遇到困难也能够从容应对。

人物背景
方圆原本是一家医疗器械公司的中层管理人员，后因公司政策调整失业，成为网约车司机，同时还兼职影视剧配音补贴家用。他生活态度乐观，喜欢花草鸟兽，对待家庭和孩子非常宽容，尤其是在教育儿子方一凡方面，他更注重孩子的健康成长而非成绩。

人物关系
方圆是童文洁的丈夫，方一凡的父亲，同时还是林磊儿的姨父。方圆与乔卫东是好友，与季胜利是发小。

Basic Information of the Character
Fang Yuan is one of the main characters in the TV series “A Little Joy”. He is a middle-aged man born on August 17, 1973, with an AB blood type.

Personality
Fang Yuan has a cheerful disposition and excels in interpersonal communication, treating friends and family with utmost sincerity and kindness. He shares a harmonious relationship with his wife, Tong Wenjie, with whom he supports each other in life and confronts various challenges together. He adopts a tolerant approach toward his son, Fang Yifan, prioritizing the boy’s healthy growth over academic performance. Regarding his son as a friend, he understands and supports Fang Yifan’s dream of pursuing a career in acting. He also maintains good relationships with close friends such as Ji Shengli and Qiao Weidong, often exchanging life stories with them.
Emotionally stable and highly responsible, Fang Yuan adapts quickly to new environments. Despite experiencing a midlife crisis, he remains positive. Even after changing careers to become a ride-hailing driver and taking on part-time work as a film and television dubber, he strives to find meaning in his new roles. His life philosophy is to face everything optimistically and handle difficulties with composure.

Background
Originally a middle-level manager at a medical device company, Fang Yuan lost his job due to corporate policy adjustments. He then became a ride-hailing driver while taking on part-time dubbing work to support his family. An optimistic person with a love for nature (plants, birds, and animals), he is particularly tolerant toward his family and children. In educating his son, Fang Yifan, he focuses more on the child’s well-being than academic achievements.

Relationships
Fang Yuan is the husband of Tong Wenjie, the father of Fang Yifan, and the uncle-in-law of Lin Leier. He is friends with Qiao Weidong and has been childhood friends with Ji Shengli.

Figure 3: An example of role profile and corresponding role image.

你是一位数据标注师，给定一个人物海报图片和一个视频图像序列，其中视频图像序列是从视频片段中按时间先后顺序截取出来的，截取时间间隔为{time_interval}秒，总共{total_frames}张，总时长为{duration_time}秒，而人物海报图片中指定的人物是你需要关注并进行标注的对象。你的任务是在给定视频片段（以视频图像序列展示）中找到海报图片上指定的人物（一般是视频片段中的说话者），并对该人物在视频片段（以视频图像序列展示）中的**面部表情**和**肢体动作**进行细致的标注。

标注的注意事项：
1. 标注**面部表情**时，要求准确并清楚地描述标注对象的“五官细节”及其变化过程，其中“五官细节”包括但不限于眉毛、眼睛、嘴角、嘴型、面部肌肉等各方面的细节。描述“五官细节”的变化过程时需要以动态的视角进行观察，并用最可能准确的词语将变化过程的全部细节描述出来。
2. 标注**肢体动作**时，要求准确并清楚地描述标注对象的“四肢动作”及其变化过程，其中“四肢动作”包括但不限于左右手臂、左右手势、双腿、双脚、体位姿态等各方面的细节。描述“四肢动作”的变化过程时需要以动态的视角进行观察，用最可能准确的词语将变化过程的全部细节描述出来。
3. 描述“五官细节”或“四肢动作”时不要分点陈列，要求用一句话自然地将所有要点描述完成。
4. 标注时务必以视频图像序列为准，不能擅自添加不存在的细节，也不能错过任何存在的细节。
5. 请注意海报图片上指定的人物在给定视频片段（以视频图像序列展示）中的**面部表情**和**肢体动作**随着时间推移可能存在若干次变化，因此在标注**面部表情**和**肢体动作**时需要清楚地描述出全部的变化过程及其细节，并要求使用表示先后顺序的连接词让描述更准确清晰，如“最初”、“然后”、“接着”、“随后”、“最后”等。但禁止对视频图像序列的每张图都进行一次标注，要求对该视频图像序列只进行一次标注。
6. 请注意在给定视频片段（以视频图像序列展示）中可能会出现多个不同的人物，且不一定每一张图像都包含海报图片上指定的人物，因此在标注时请先找出海报上指定的人物再进行标注，要求直接忽略视频图像序列中不存在标注对象的无关图片，千万不要找错标注对象（标注对象一般是视频片段中的说话者），也不要描述其他无关人物，但可以适当提及相关人或物品以丰富标注对象与周围事物的交互细节。
7. 标注时请忽略可能出现的字幕、人物着装、背景环境等无关细节，专注于对海报图片上指定人物在给定视频片段（以视频图像序列展示）中的**面部表情**和**肢体动作**进行细致描述，而且不需要给出任何注释或推理过程，直接按照返回格式给出最终的标注即可。标注的描述语句要求使用中文，且用语通顺无误，每项标注字数在100字以内即可。
8. 输出的标注内容中不要出现“在视频片段中”、“指定人物的面部表情”、“从画面中可以看到”之类的无关用语，也不要用“指定人物”、“海报人物”等词语来指代标注对象，请直接给出**面部表情**和**肢体动作**的标注内容。

给定的人物海报图片如下：
{poster_img}

给定的视频片段（以视频图像序列展示）如下：
{frames_of_video}

返回格式：

面部表情：...
肢体动作：...

You are a data annotator. Given a character poster image and a video image sequence (extracted from a video clip in chronological order at {time_interval}-second intervals, totaling {total_frames} images with a duration of {duration_time} seconds), the character specified in the poster is the object you need to focus on and annotate. Your task is to locate the specified character from the poster (typically the speaker in the video clip) within the given video clip (represented by the video image sequence) and provide detailed annotations of the character's facial expressions and body movements in the clip.

Annotation Guidelines:

When annotating facial expressions, accurately describe the "facial feature details" of the annotated object and their dynamic changes, including but not limited to eyebrows, eyes, mouth corners, mouth shape, and facial muscles. Observe from a dynamic perspective and use the most accurate terms to describe all details of the change process.
When annotating body movements, accurately describe the "limb actions" of the annotated object and their dynamic changes, including but not limited to movements of the left/right arms, gestures, legs, feet, and postures. Observe from a dynamic perspective and use the most accurate terms to describe all details of the change process.
When describing "facial feature details" or "limb actions," avoid bullet points; instead, naturally integrate all key points into a single sentence.
Base annotations strictly on the video image sequence; do not invent non-existent details or miss any existing details.
Note that the specified character’s facial expressions and body movements may undergo multiple changes over time in the video clip (video image sequence). Clearly describe all change processes and details using sequence connectives (e.g., "initially," "then," "next," "subsequently," "finally") to ensure accuracy and clarity. Do not annotate each frame individually; provide a single annotation for the entire sequence.
The video image sequence may contain multiple characters, and not every frame may include the specified character from the poster. First identify the specified character before annotating; ignore frames without the target and avoid describing irrelevant characters, though you may appropriately mention interacting persons or objects to enrich contextual details.
Ignore irrelevant details such as subtitles, clothing, and background environments. Focus solely on detailed descriptions of the specified character’s facial expressions and body movements in the video clip (video image sequence). Provide no annotations or reasoning—simply follow the output format. Use Chinese, ensure smooth phrasing, and keep each annotation under 100 words.
Do not include irrelevant phrases like "in the video clip," "the specified character’s facial expressions," or "as seen in the frame." Avoid using terms like "specified character" or "poster character" to refer to the target; directly describe the facial expressions and body movements.

The given character poster image is as follows:
{poster_img}

The given video clip (displayed as a video image sequence) is as follows:
{frames_of_video}

Return format:
Facial expressions: ...
Body movements: ...

Figure 4: The template of user prompt for describing facial expressions and body movement from given frames of video.

给定一段人声音频的基本信息，包括总时长、文本内容、情绪、语速、音调、音量等，你需要写出符合要求的、正确的提示语以供TTS模型根据提示语和文本内容重新生成这段人声音频。要求用自然简洁的语言写出提示语，并尽量在一句话内涵盖所有给定信息，包括情绪、语速、音调、音量等特征及其起伏变化。注意提示语是对TTS模型的一种提示，不是文本内容，提示语中也不能明显地提及文本内容。要求提示语使用中文，字数在30到50字以内即可。要求提示语中省略人称代词“你我他”等。

人声音频各基本信息的含义：
1. 该段人声音频包含{len_utts}个分句，总时长的单位为秒，文本内容、情绪和语速是按分句顺序提取的列表，列表长度等于分句数量。语速列表中的数值无量纲，已经过归一化和标准化处理，服从均值为0方差为1的标准正态分布。
2. 音调序列和音量序列都是从人声音频中同步分帧并计算出的数值特征序列，每帧计算出一个音调数值和一个音量数值，相邻两帧间隔{frame_interval_sec}秒，因此两个序列长度相等且一一对应。音调和音量中的数值都无量纲，且已经过归一化和标准化处理，服从均值为0方差为1的标准正态分布。平均音调和平均音量则分别是音调序列和音量序列取平均而得。在描述整体音调音量高低时可以参考平均值，在描述音调音量起伏变化特征时则需要参考原序列。
3. 对于服从标准正态分布的数值（语速、音调、音量），认为在-1到1之间都是正常的、标准的，其中0到1之间是略高，-1到0之间是略低；大于1的则认为偏高，大于3的是非常高；小于-1的则认为偏低，小于-3的是非常低。

示例一：
人声音频的基本信息：
```json
…
```
提示语：以中性的语气、平稳且标准的音量、音调偏高并略带起伏地娓娓道来，最后音调突然夹得很高。

示例二：…

示例三：
人声音频的基本信息：
```json
…
```
提示语：以中性语气，语速先慢后快再平稳，音量正常且略有起伏，音调整体偏低沉，起伏间略显平稳，传达出稳妥可靠的态度。

以上示例仅供参考。

当前音频：
人声音频的基本信息：
```json
{{
  "总时长": {duration_ttl},
  "文本内容": {content_texts},
  "情绪": {emotions},
  "语速": {speeds},
  "音调序列": {pitches_10ms},
  "平均音调": {pitches_avg},
  "音量序列": {volumes_10ms},
  "平均音量": {volumes_avg},
}}
```
提示语：

Given basic information about a human voice audio, including total duration, text content, emotion, speaking speed, pitch, and volume, you need to write appropriate and correct prompts for a TTS model to regenerate the audio based on the prompts and text content. The prompts should be natural and concise, preferably covering all given information—such as emotional tone, speaking speed, pitch, volume, and their fluctuations—in a single sentence. Note that the prompts are instructions for the TTS model, not the text content itself, and should not explicitly mention the text content. The prompts must be in Chinese, between 30 and 50 characters, and omit personal pronouns like “he, she".

Meanings of basic information for human voice audio:
The audio contains {len_utts} clauses. The total duration is in seconds. The text content, emotions, and speaking speeds are lists extracted in clause order, with the list length equal to the number of clauses. The values in the speaking speed list are dimensionless, normalized, and standardized to follow a standard normal distribution with a mean of 0 and a variance of 1. Both the pitch and volume sequences are numerical feature sequences calculated from synchronized frames of the audio, with one pitch and one volume value per frame and an interval of {frame_interval_sec} seconds between adjacent frames. Thus, the two sequences are of equal length and correspond one-to-one. The values in the pitch and volume sequences are dimensionless, normalized, and standardized to follow a standard normal distribution. The average pitch and average volume are the means of the pitch and volume sequences, respectively. When describing overall pitch and volume levels, refer to the averages; when describing fluctuations, refer to the original sequences.
For values following a standard normal distribution (speaking speed, pitch, volume), those between -1 and 1 are considered normal and standard, with 0 to 1 being slightly high and -1 to 0 slightly low; values above 1 are considered high, and those above 3 are very high; values below -1 are low, and those below -3 are very low.

Example 1:
Basic information of voice audio: …
Prompt: In a neutral tone, with steady and standard volume, slightly high and slightly fluctuating pitch, speaking slowly and smoothly, with a sudden high pitch at the end.

Example 2: …

Example 3:
Basic information of voice audio: …
Prompt: In a neutral tone, speaking speed starts slow, then speeds up, then becomes steady, with normal and slightly fluctuating volume, overall low and steady pitch, conveying a secure and reliable attitude.

The above examples are for reference only.

Current audio:
Basic information of voice audio:
```json
{{
  “total duration": {duration_ttl},
  “text content": {content_texts},
  “emotion": {emotions},
  “speed": {speeds},
  “pitch sequence": {pitches_10ms},
  "pitch average": {pitches_avg},
  "volume sequence": {volumes_10ms},
  "volume average": {volumes_avg},
}}
```
Prompt:

Figure 5: The template of user prompt for describing speech prompt from given numerical feature sequences.

你现在要扮演{sys_role_name}，角色信息如下：
{role_profile}
你的人物形象如下：
{placeholder_img}
接下来我将扮演{usr_role_name}与你对话，我的人物形象如下：
{placeholder_img}
对话中每轮我都会输入一句话和对应的一段音频及视频（以图像序列展示），你需要给出符合角色特点的回复。回复中要包含面部表情、肢体动作、语音提示和回复内容，并分别包含在<facial_expression> </facial_expression>，<body_movement> </body_movement>，<speech_prompt> </speech_prompt>和<content> </content>之间。例如：
```
<facial_expression>面部表情在这</facial_expression>
<body_movement>肢体动作在这</body_movement>
<speech_prompt>语音提示在这</speech_prompt>
<content>回复内容在这</content>
```
现在对话开始。

You are now to play the role of {sys_role_name} with the following character information:
{role_profile}
Your character image is as follows:
{placeholder_img}
Next, I will play {usr_role_name} and converse with you. My character image is as follows:
{placeholder_img}
In each round of the conversation, I will input a sentence along with a corresponding audio clip and video (displayed as an image sequence). You need to provide a response that matches the character's traits, including facial expressions, body movements, speech prompts, and reply content, enclosed within <facial_expression> </facial_expression>, <body_movement> </body_movement>, <speech_prompt> </speech_prompt>, and <content> </content> respectively. For example:
```
<facial_expression>Facial expression here</facial_expression>
<body_movement>Body movement here</body_movement>
<speech_prompt>Speech prompt here</speech_prompt>
<content>Reply content here</content>
```
Now let the conversation begin.

(a)

你现在要扮演{sys_role_name}，角色信息如下：
{role_profile}
你的人物形象如下：
{placeholder_img}
接下来我将扮演{usr_role_name}与你对话，我的人物形象如下：
{placeholder_img}
对话中每轮我都会输入一句话和对应的一段视频（以图像序列展示），你需要给出符合角色特点的回复。回复中要包含面部表情、肢体动作和回复内容，并分别包含在<facial_expression> </facial_expression>，<body_movement> </body_movement>和<content> </content>之间。例如：
```
<facial_expression>面部表情在这</facial_expression>
<body_movement>肢体动作在这</body_movement>
<content>回复内容在这</content>
```
现在对话开始。

You are now to play the role of {sys_role_name} with the following character information:
{role_profile}
Your character image is as follows:
{placeholder_img}
Next, I will play {usr_role_name} and converse with you. My character image is as follows:
{placeholder_img}
In each round of the conversation, I will input a sentence and a corresponding video (displayed as an image sequence). You need to provide a response that aligns with the character's traits. The response should include facial expressions, body movements, and reply content, enclosed within <facial_expression> </facial_expression>, <body_movement> </body_movement>, and <content> </content> respectively. For example:
```
<facial_expression>Facial expression here</facial_expression>
<body_movement>Body movement here</body_movement>
<content>Reply content here</content>
```
Now let the conversation begin.

(b)

你现在要扮演{sys_role_name}，角色信息如下：
{role_profile}
接下来我将扮演{usr_role_name}与你对话。对话中每轮我都会输入一句话和对应的一段音频，你需要给出符合角色特点的回复。回复中要包含语音提示和回复内容，并分别包含在<speech_prompt> </speech_prompt>和<content> </content>之间。例如：
```
<speech_prompt>语音提示在这</speech_prompt>
<content>回复内容在这</content>
```
其中你所生成的语音提示和回复内容都将会指导后续TTS模型合成语音，因此注意不要生成与语音提示无关的信息。

现在对话开始。

You are now to play the role of {sys_role_name} with the following character information:
{role_profile}
Next, I will play {usr_role_name} to have a conversation with you. In each round of the conversation, I will input a sentence and a corresponding audio clip. You need to provide a response that matches the character's traits, including a speech prompt and reply content, enclosed within <speech_prompt> </speech_prompt> and <content> </content> respectively. For example:
```
<speech_prompt>Speech prompt here</speech_prompt>
<content>Reply content here</content>
```
Note that both the speech prompt and reply content you generate will guide the subsequent TTS model to synthesize speech, so avoid generating information irrelevant to the speech prompt.

Now let the conversation begin.

(c)

你现在要扮演{sys_role_name}，角色信息如下：
{role_profile}
接下来我将扮演{usr_role_name}与你对话。对话中每轮你都需要给出符合角色特点的回复。

现在对话开始。

You are now to play the role of {sys_role_name} with the following character information:
{role_profile}
Next, I will play {usr_role_name} to have a conversation with you. You need to provide a response that matches the character's traits in each round.

Now let the conversation begin.

(d)

Figure 6: The template of system prompt for EMRPA in (a)omni chat, (b)vision-only video chat, (c)audio chat, (d)text chat.

给定一段文本和一个情感标签，你需要识别给定文本中蕴含的情感，并评价其与给定情感标签的一致性。要求在输出时只输出0~5之间的整数，不需要给出解释，其中0表示完全不一致，5表示完全一致。

示例：
```
# 示例一
文本：安迪姐姐你出国太久了，你不知道现在网络有多可怕，你没看见我刚才给你发的小视频，你看他们都怎么说，你 而且现在楼主只贴了你的照片，楼里那些人都是在骂你的，我觉得这帖子针对你的知道吗，我现在下班了，我先回家接着干，你也赶紧回来，咱俩一起商量
情感标签：沮丧
得分：5

# 示例二
文本：这是给你的生日惊喜啊 罗宾今天是你30岁生日吧，你要不要拆一下看看是什么东西呢？
情感标签：中性
得分：3

# 示例三
文本：这不写着呢吗 卖给王叔叔啦 妈咪 我想吃点心 怎么样 给你不就得了嘛 还说不吃 不吃了啊 吃吧 快趁热吃 看着没看着都成这样儿了 赶快趁热赶紧吃 没事没事 来 来 再拿一块去 多好呀
情感标签：放松，快乐
得分：4

# 示例四
文本：我知道你担心，但你得相信我，我不会让你生气的。我们好好沟通，一切都会解决的。
情感标签：其他负面情绪
得分：2
```

以上示例仅供参考，现在开始打分。

文本：{content}
情感标签：{gt_emotion_label}
得分：

Given a piece of text and an emotional label, you need to identify the emotion conveyed in the text and evaluate its consistency with the given emotional label. Only output an integer between 0 and 5 (inclusive) without explanation, where 0 means completely inconsistent and 5 means completely consistent.

Examples:
```
# Example 1
Text: Sister Andy, you've been abroad for too long. You have no idea how terrifying the internet is now. You didn't see the short video I sent you just now. Look at what they're saying. And now the thread starter has only posted your photo, and those people in the thread are all scolding you. I think this post is targeting you, you know? I'm off work now. I'll go home first and continue working. You come back quickly too. Let's discuss together.
Emotional Label: depress
Score: 5

# Example 2
Text: This is a birthday surprise for you, Robin. Today is your 30th birthday, right? Do you want to open it and see what's inside?
Emotional Label: neutral
Score: 3

# Example 3
Text: It's written right here. It's sold to Uncle Wang. Mommy, I want to eat snacks. What do you think? Here you go. You said you wouldn't eat, but come on. Eat it while it's hot. Look, look, it's like this. Hurry up and eat it while it's still hot. It's okay. Come on, come on. Take another piece. It's so good.
Emotional Label: relax, happy
Score: 4

# Example 4
Text: I know you're worried, but you have to trust me. I won't make you angry. Let's communicate well and everything will be solved.
Emotional Label: negative-other
Score: 2
```

The above examples are for reference only. Now start scoring.

Text: {content}
Emotional Label: {gt_emotion_label}
Score:

Figure 7: The template of user prompt for TEC.

给定一个文本对，分别是说话内容和语气语调，你需要评价其情感的一致性。要求仅评价情感上的一致性，请忽略语义上或风格上的一致性。要求在输出时只输出0~5之间的整数，不需要给出解释，其中0表示完全不一致，5表示完全一致。

示例：
```
# 示例一
待评价文本对：
  - 我知道你担心，但你得相信我，我不会让你生气的。我们好好沟通，一切都会解决的。
  - 戴娜的声音有些低沉，带着一丝无奈
得分：3

# 示例二
待评价文本对：
  - 你这是什么态度？我是在帮你，你却这样对我？
  - 严厉的语气
得分：4

# 示例三
待评价文本对：
  - 求你接一个案子
  - 以中性的语气，语速略快，音量偏高且起伏较大，音调先低后高再回落，整体略显波动，传达出急切而诚恳的请求。
得分：5

# 示例四
待评价文本对：
  - 你从小到大哪件事情不是我逼出来的，你爸妈让你过来跟着我，就是为了让我来引导你，这么多不良青年我都收拾了，你我还教不会，除非你跟我说你一点都不喜欢宛瑜，要是你这么说了，我就去告诉她
  - 以中性语气为主，语速先快后缓再平稳，音量正常且略有起伏，音调整体偏高且波动较大，情绪由中性渐转为厌恶，最后放松。
得分：4
```

以上示例仅供参考，现在开始打分。

待评价文本对：
  - {content}
  - {speech_prompt}
得分：

Given a text pair consisting of spoken content and tone, you need to evaluate the emotional consistency between them. Focus only on emotional consistency, ignoring semantic or stylistic consistency. Output only an integer between 0 and 5 (inclusive) without explanation, where 0 means completely inconsistent and 5 means completely consistent.

Examples:
```
# Example 1
Text pair to evaluate:
• I know you're worried, but you have to trust me. I won't make you angry. Let's communicate well and everything will be solved.
• Dina's voice was a bit low, tinged with a hint of helplessness
Score: 3

# Example 2
Text pair to evaluate:
• What's your attitude? I'm trying to help you, but you treat me like this?
• A stern tone
Score: 4

# Example 3
• Please take on a case
• With a neutral tone, speak slightly faster, with a relatively high volume and significant fluctuations. The pitch starts low, then rises and finally drops back, showing an overall somewhat fluctuating pattern, conveying an urgent and sincere request.
Score: 5

# Example 4
• Which thing in your life from childhood to adulthood wasn't forced by me? Your parents asked you to come and follow me just to let me guide you. I've dealt with so many delinquent youths. Do you think I can't teach you? Unless you tell me that you don't like Wanyu at all. If you say so, I'll go and tell her.
• Mainly in a neutral tone, the speaking speed is fast at first, then slows down and finally becomes steady. The volume is normal with slight fluctuations. The pitch is overall high and varies significantly. The emotion gradually changes from neutral to disgusted, and finally relaxes.
Score: 4
```

The above examples are for reference only. Now start scoring.

Text: {content}
Emotional Label: {gt_emotion_label}
Score:

Figure 8: The template of user prompt for SEC. VEC and CEC have a similar form as SEC, with difference only on few-shot examples.

你需要结合对话历史，评价{sys_role}的最新一轮回复是否符合角色简介中的人物经历及其可能的发展，即衡量其经历幻觉程度。例如如果在最新一轮回复中出现了与人物经历相悖的，则打低分；如果最新一轮回复中出现了人物经历中的内容且内容正确，则打高分；如果在最新一轮回复中出现了人物经历中不存在的，则酌情打分。要求仅评价人物经历的一致性，忽略人物性格和人物关系的一致性。要求在输出时只输出0~5之间的整数，不需要给出解释，其中0表示完全不一致，5表示完全一致。 {sys_role}的角色简介： {sys_role_profile} {sys_role}与{user_role}的对话历史： {chat_history_flat_content_only} （最新一轮回复）{sys_role}：{content} 得分：	You need to evaluate whether {sys_role}'s latest response aligns with the character's experiences and possible developments in the role profile by incorporating the conversation history, i.e., measuring the degree of experience hallucination. For example, a low score will be given if the latest response contradicts the character's experiences; a high score will be given if the latest response includes content from the character's experiences and the content is correct; and a discretionary score will be given if the latest response contains content not present in the character's experiences. Only the consistency of the character's experiences should be evaluated, ignoring the consistency of the character's personality and relationships. Only an integer between 0 and 5 (inclusive) should be output without explanation, where 0 means completely inconsistent and 5 means completely consistent. {sys_role}'s role profile: {sys_role_profile} Conversation history between {sys_role} and {user_role}: {chat_history_flat_content_only} (Latest response) {sys_role}: {content} Score: (a)
你需要结合对话历史，评价{sys_role}的最新一轮回复是否符合角色简介中的人物性格，即衡量其性格偏离程度。例如如果在最新一轮回复中出现了与人物性格不一致的内容，则打低分，否则打高分。要求仅评价人物性格的一致性，忽略人物经历和人物关系的一致性。要求在输出时只输出0~5之间的整数，不需要给出解释，其中0表示完全不一致，5表示完全一致。 {sys_role}的角色简介： {sys_role_profile} {sys_role}与{user_role}的对话历史： {chat_history_flat_content_only} （最新一轮回复）{sys_role}：{content} 得分：	You need to evaluate whether {sys_role}'s latest response aligns with the character's personality in the role profile by incorporating the conversation history, i.e., measuring the degree of personality deviation. For example, a low score will be given if the latest response contains content inconsistent with the character's personality, and a high score otherwise. Only the consistency of the character's personality should be evaluated, ignoring the consistency of experiences and relationships. Output only an integer between 0 and 5 (inclusive) without explanation, where 0 means completely inconsistent and 5 means completely consistent. {sys_role}'s role profile: {sys_role_profile} Conversation history between {sys_role} and {user_role}: {chat_history_flat_content_only} (Latest response) {sys_role}: {content} Score: (b)
你需要结合对话历史，评价{sys_role}最新一轮回复是否符合角色简介中的人物关系，即衡量其关系混淆程度。例如如果在最新一轮回复中出现了与人物关系相悖的称呼、语气等，则打低分；如果最新一轮回复中出现了与人物关系相符的话语，则打高分；若人物关系中未明显提及{sys_role}与{user_role}的关系，可以从两人的共同朋友或亲戚中推理出其关系，否则当陌生人处理并酌情给分。要求仅评价人物关系的一致性，忽略人物经历和人物性格的一致性。要求在输出时只输出0~5之间的整数，不需要给出解释，其中0表示完全不一致，5表示完全一致。 {sys_role}的角色简介： {sys_role_profile} {user_role}的人物关系： {user_role_relationship} {sys_role}与{user_role}的对话历史： {chat_history_flat_content_only} （最新一轮回复）{sys_role}：{content} 得分：	You need to evaluate whether {sys_role}'s latest response aligns with the character relationships in the role profile by incorporating the conversation history, i.e., measuring the degree of relationship confusion. For example, a low score will be given if the latest response contains address terms, tone, etc., that contradict the character relationships; a high score will be given if the response includes discourse consistent with the character relationships; if the role profile does not explicitly mention the relationship between {sys_role} and {user_role}, their relationship can be inferred from mutual friends or relatives, otherwise, treat them as strangers and score appropriately. Only the consistency of character relationships should be evaluated, ignoring the consistency of experiences and personality. Output only an integer between 0 and 5 (inclusive) without explanation, where 0 means completely inconsistent and 5 means completely consistent. {sys_role}'s role profile: {sys_role_profile} {user_role}'s character relationships: {user_role_relationship} Conversation history between {sys_role} and {user_role}: {chat_history_flat_content_only} (Latest response) {sys_role}: {content} Score: (c)

Figure 9: The template of user prompt for (a)Experience, (Characteristics), (c)Relationships.