MERRY: SEMANTIC-LEVEL EVALUATION OF MULTIMODAL EMOTIONAL AND ROLE CONSISTENCIES FOR ROLE-PLAYING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Role-Playing Agents (MRPAs) are attracting increasing attention due to their ability to deliver more immersive multimodal emotional interactions. However, existing studies typically evaluate response content with textual role-playing benchmarks, while assessing multimodal expressions mainly through modality-synthesis metrics. This evaluation paradigm leaves a critical diagnostic gap: the semantics of multimodal expressions are only assessed after modality generation, making it difficult to determine whether an error originates from multimodal semantic response planning or from final modality realization. To this end, we propose *MERRY*, a semantic-level evaluation framework for assessing **M**ultimodal **E**motional and **R**ole consistencies of **R**ole-pla**y**ing agents. This framework extends existing evaluation dimensions to multimodal semantic responses, and introduces new metrics to assess whether generated semantic descriptions support a coherent emotional interpretation. It also introduces a bidirectional evidence-finding scoring protocol for Large-Language-Model-as-a-Judge scoring, improving the alignment with human evaluators. To support evaluation, we further construct *MERRY-Data*, which contains 25,507 samples with a TV-series-level separated test split. Extensive evaluations with *MERRY* show that current MRPAs still have substantial room for improvement in multimodal emotional and role consistency on semantic level.

1 INTRODUCTION

Multimodal Role-Playing Agents (MRPAs) (Dai et al., 2025) are expected to generate role-consistent content together with emotionally appropriate facial expressions, body movements, and speech cues. Existing studies, however, typically evaluate textual content with role-playing benchmarks (Chen et al., 2024; Tu et al., 2024) and final multimodal outputs with modality-synthesis metrics (Zhang et al., 2025b; Chen et al., 2025; Zhang et al., 2025a). Thus the semantics of multimodal expressions before synthesis are largely unexamined, as illustrated in Figure 1. We examine this limitation using 100 randomly selected test samples from *MERRY-Data*, constructing a wrong-semantic counterpart for each sample through randomly mismatched three multimodal semantics before synthesis. The resulting score distributions show that existing metrics are insufficiently sensitive to such semantic errors. Further details are provided in the appendix. These observations motivate semantic-level evaluation on the emotional and role consistencies of multimodal responses before modality synthesis.

We therefore propose *MERRY*, a semantic-level evaluation framework based on two dimensions: **Emotional Consistency (EC)** and **Role Consistency (RC)**. EC measures alignment with the target character's observed emotions and emotional dynamics (Feng et al., 2025), whereas RC measures fidelity to the character's profile, personality, and contextual behavior (Tu et al., 2024; Wang et al., 2024c). For EC, we extend its metrics to multimodal semantics and introduce Emotional Discrepancy (ED) and Cross-modal Emotional Consistency (CEC). ED measures whether each modality-specific semantic description independently conveys a determinate emotion for each utterance, and CEC measures whether these independently conveyed emotions remain coherent across modalities. Human emotion perception integrates cues across modalities (Van den Stock et al., 2007; Hollenstein & Lanteigne, 2014), but modality-specific generators often process their corresponding semantic de-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

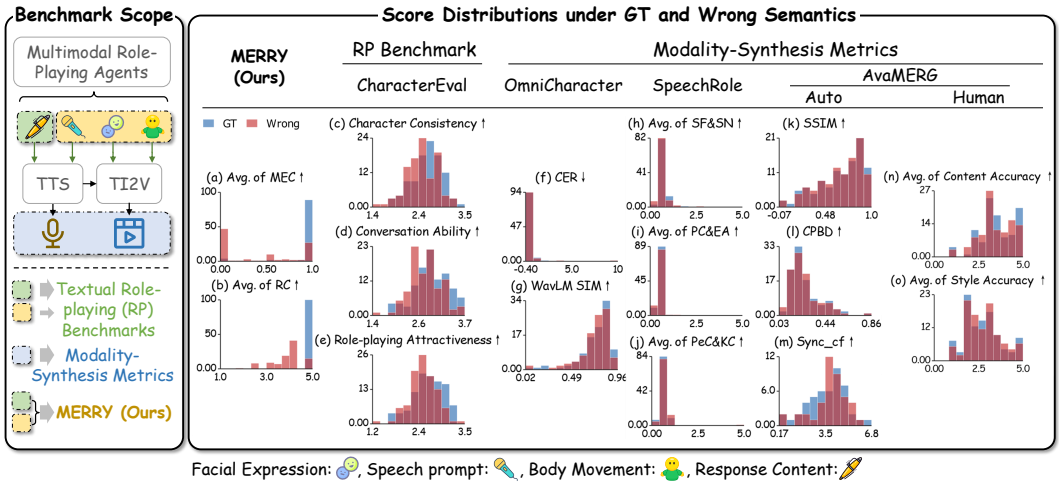

Figure 1: **Left**: comparison of the evaluation scope of three existing kinds of benchmarks. **Right**: score distributions under ground-truth and wrong multimodal semantics across these benchmarks.

scriptions independently. Consequently, emotional relations, that are only recoverable after jointly reading all modalities, may not be preserved during synthesis. For example, "speak toughly" alone may imply anger or confidence, whereas "speak toughly to mask vulnerability" independently conveys the underlying vulnerability and remains coherent with vulnerable facial cues, which is more helpful to maintain cross-modal emotional relations in modality-independent synthesis. ED and CEC therefore aim to assess this gap before modality synthesis. For RC, *MERRY* further assesses whether the expression habit in multimodal semantic responses align the target character in context. We also replace conventional explain-then-score evaluation with a bidirectional evidence-finding protocol to improve agreement between LLM judges and human evaluators. Details are provided in Section 4.

Semantic-level evaluation also requires data containing emotion labels, detailed role information, and modality-specific semantic descriptions. Existing multimodal emotional dialogue datasets often lack these elements except emotion labels, whereas multimodal role-playing datasets lack elements except role information. We therefore construct *MERRY-Data* based on CPED[1] (Chen et al., 2022).

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

1. We propose *MERRY* to automatically evaluate the multimodal semantic responses of MRPAs on emotional and role consistencies. It fills the blank of existing multimodal evaluations.
2. To support the proposed evaluation, we construct *MERRY-Data*, a multi-turn multimodal role-playing emotional interaction dialogue dataset, which includes role-related information and multimodal semantic descriptions. It contains 25,507 samples with a fully separated test split.
3. Highlighted Findings: Existing MRPAs simplify character emotions by relying on salient emotional prototypes, weakening neutral-state preservation, and shifting fine-grained emotional dynamics toward model-specific polarity patterns

2 Related Work

2.1 Multimodal Role-Playing Agents

Recent work has extended text-only character imitation (Shao et al., 2023; Zhou et al., 2024; Wang et al., 2024b; Li et al., 2023; Chen et al., 2023) to multimodal role-playing. MMRole formalizes Multimodal Role-Playing Agents (MRPAs) by integrating visual perception with role-consistent response generation (Dai et al., 2025). Subsequent studies broaden MRPAs to speech, video, and text–

[1]CPED is a large-scale Chinese Personalized and Emotional Dialogue dataset containing multi-turn multimodal dialogues from TV dramas, with utterance-level annotations including 13 emotion categories and 19 dialogue actions, and other role-level attributes including name, gender, age, and Big Five personality traits.

image interaction (Zhang et al., 2025b;c; Tang et al., 2026a). Other approaches improve modality–role alignment by filtering role-irrelevant visual information or aligning character reasoning with expressive speech generation (Tang et al., 2026c; Chen et al., 2026; Tang et al., 2026b). Together, these studies focus on constructing or improving MRPAs, whereas our work focuses on fine-grained multimodal semantic evaluation of their responses.

2.2 Role-Playing Benchmarks

Previous role-playing benchmarks mainly assess character-faithful textual responses, including role knowledge (Wang et al., 2024b; Ahn et al., 2024), personality and behavioral consistency (Wang et al., 2024c;a; Yu et al., 2025), social interaction (Chen et al., 2024), and emotional consistency (Feng et al., 2025). Recent benchmarks extend this evaluation beyond text. MMRole-Eval assesses dialogue quality, role consistency, and image understanding (Dai et al., 2025), while speech role-playing benchmarks evaluate verbal content, interaction quality, speech expressiveness, and vocal attributes such as timbre, prosody, emotion, and style (Jiang et al., 2026; Wu et al., 2025; Shi et al., 2025; Li et al., 2026; Fu et al., 2026). Overall, these benchmarks focus primarily on verbal content or the quality of final synthesized outputs. General role-playing evaluators such as CharacterEval (Tu et al., 2024) can also be applied to multimodal semantics, but only provide coarse-grained assessment on behavioral role consistency. *MERRY* instead provides dedicated, fine-grained evaluation of multimodal semantic responses before synthesis.

2.3 Multimodal Role-Playing and Emotional Dialogue Datasets

Existing multimodal dialogue datasets follow two complementary lines. First, multimodal role-playing datasets provide image-, video-, or speech-grounded role information for training and evaluating. They include image-centered dialogues (Dai et al., 2025), synthesized role speech (Zhang et al., 2025b; Jiang et al., 2026), video-summary-based question-answer pairs (Zhang et al., 2025c), text–image role-play annotations (Tang et al., 2026a), and character-specific reference audio (Li et al., 2026). However, they generally lack turn-level emotion labels and fine-grained semantic descriptions of multimodal expressions. Second, multimodal emotional dialogue datasets provide complementary emotion annotations on real-world videos. MELD, M3ED, and CPED contain multimodal conversations with emotion labels (Poria et al., 2019; Zhao et al., 2022; Chen et al., 2022), while recent datasets extend to multimodal empathetic response generation and emotional support scenarios (Chu et al., 2024; Zhang et al., 2025a). However, they typically lack detailed role information and multimodal semantic descriptions. *MERRY-Data* combines all three elements to support semantic-level training and evaluation of MRPAs.

3 MERRY-Data

3.1 Task Definition

Given a character C with a role profile P_C, background information B_D of current dialogue, and a reference character image I_C, an MRPA is conditioned to act as C and engage in a multi-turn multimodal emotional dialogue with a user identified by N_u. At the t-th turn, the task is defined as:

$$r_t = \text{MRPA}(P_C, B_D, I_C, N_u, D_{<t}, U_t), \quad D_{<t} = \{(U_0, r_0), \ldots, (U_{t-1}, r_{t-1})\}. \tag{1}$$

Here, $D_{<t}$ denotes the dialogue history before turn t. The user input is $U_t = \{A_t, V_t, T_t\}$, where A_t, V_t, and T_t denote the user's audio, video, and textual content, respectively. The agent's multimodal semantic response is $r_t = \{f_t, b_t, s_t, c_t\}$, where c_t is the textual response content, f_t is the facial expression, b_t is the body movement, and s_t is the speech prompt that specifies the intended prosody and intonation. For the first turn $t = 0$, $D_{<0} = \emptyset$.

3.2 Data Construction

As shown in Figure 2, we construct *MERRY-Data*, a multi-turn multimodal emotional dialogue dataset built on CPED (Chen et al., 2022), to support both the development of role-playing agents and their semantic-level evaluation.

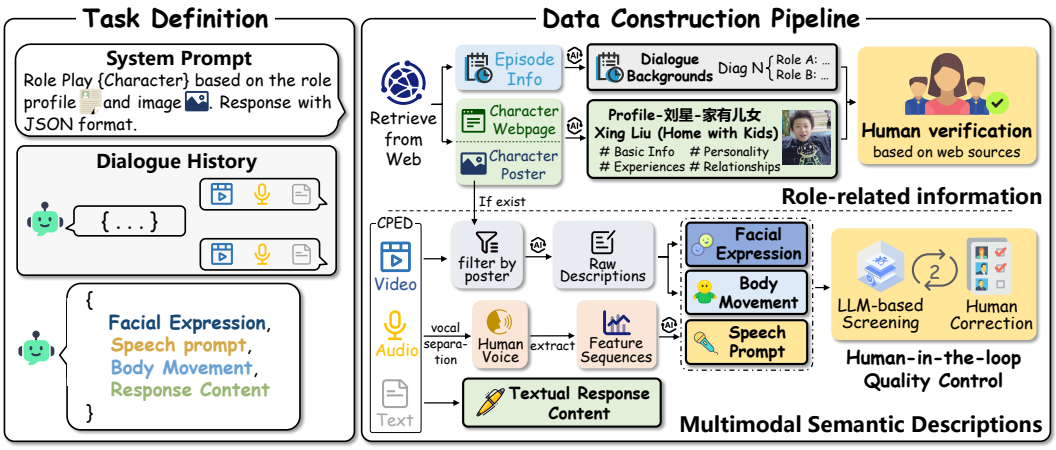

Figure 2: Overview of *MERRY-Data*.

Role-related information. We retrieve role webpages and posters from BaiduBaike[2], and remove roles without valid webpages or posters together with their dialogues. DeepSeek-V3[3] summarizes each role profile into four aspects: basic information, personality traits, character experiences, and interpersonal relationships. For dialogue backgrounds, we retrieve episode-level information from TVMAO[4] and use Doubao-seed-1-6[5] to generate role-aligned backgrounds. These backgrounds are written from the target character's perspective, include necessary personality and relationship constraints, and exclude information unavailable at the corresponding dialogue point. They define the knowledge boundary for generation, while residual leakage or inconsistency is assessed by role consistency evaluation. Finally, we recruit 20 undergraduate students to check and correct role profiles, character images, and dialogue backgrounds against the corresponding webpages, original dialogues, and episode-level descriptions.

Table 1: Statistics of train and test splits of *MERRY-Data*.

Split	# Samples	# Roles	# Dialogues
Train	23817	239	3577
Test	1690	18	401

Table 2: Statistics of the whole *MERRY-Data*.

# samples	25507
# dialogues	3978
Avg. tokens of previous information	35.33
# roles	257
Avg. # utterances per role	114.73
Avg. # turns per dialogue	7.41
Avg. video duration (s)	6.00
Avg. tokens of facial expression	145.76
Avg. tokens of body movement	142.52
Avg. tokens of speech prompt	224.58
Avg. tokens of content	34.38
Avg. tokens of role profile	518.36

Multimodal Semantic Descriptions. We use multimodal semantic descriptions, covering facial expressions, body movements, and speech prompts, to support semantic-level training and evaluation. Before annotation, we convert CPED into a turn-taking dialogue format by merging consecutive utterance segments from the same speaker and adding comma separators where punctuation is missing. The corresponding emotion labels, which are manually annotated from full-modality video segments in CPED, are merged into a label list for each dialogue turn. For audio, following prior studies (Shimizu et al., 2024; Liu et al., 2024), we extract normalized sequences of *speaking speed*, *pitch*, and *volume* from the separated human voice, and use DeepSeek-V3[3] to generate speech prompts from these features. For video, we use GPT-4o-mini (OpenAI, 2024) to describe facial expressions and body movements from the character image and sampled video frames. We then conduct human-in-the-loop quality control: Doubao-seed-1-6[5] screens and flags those descriptions potentially inconsistent with the source modality or emotion label, and the 20 volunteers verify and revise them when prob-

[2] https://baike.baidu.com/, a Chinese encyclopedia.
[3] https://deepseek.ai/
[4] https://www.tvmao.com/, a public website that provides episode outlines for Chinese TV series.
[5] https://www.doubao.com/

lem is detected. For meaningful cross-modal contrasts, human annotators preserve modality-specific cues and add cross-modal relations supported by the multimodal context and ground-truth emotion label. After two rounds of screening and correction, we obtain the final annotations.

Assembly and statistics. We assemble the above components and retain only two-speaker dialogues with complete annotations. Finally, we split the dataset at the TV-series level to avoid data leakage, ensuring that all test-set roles are isolated from the training set. The resulting statistics are reported in Tables 1 and 2.

4 MERRY

4.1 Evaluation Pipeline

As shown in Figure 3, role consistency (RC) evaluation directly assesses the predicted responses, whereas emotional consistency (EC) evaluation requires two additional steps: output formatting and emotion recognition in conversation (ERC). Specifically, the **output formatting step** first validates each raw response with a Pydantic schema. If validation fails, Doubao-seed-1-6[5] is instructed to convert the raw response $resp_i^{raw}$ into a JSON-formatted response $resp_i^{fmt} = \{f_i, b_i, s_i, c_i\}$ without changing its semantics. Here, f_i, b_i, s_i, and c_i denote facial expression, body movement, speech prompt, and textual response content, respectively. This step unifies the output schema and reduces format-induced bias. For the **ERC step**, inspired by InstructERC (Lei et al., 2023), five LLM experts $j \in [0, 4]$ independently perform two recognition passes $k \in \{0, 1\}$ under 4 channel-specific settings:

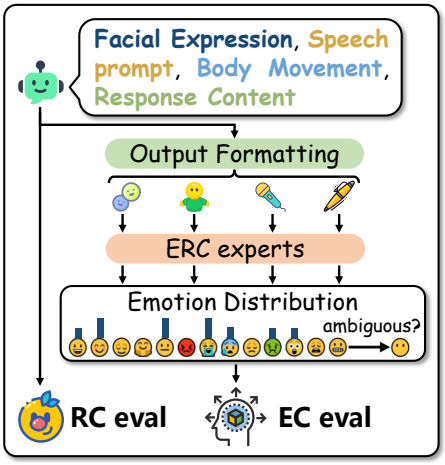

Figure 3: Evaluation pipeline of *MERRY*.

$$\hat{\mathbf{e}}^m_{i,j,k} = \mathrm{ERC}_{exp_j}(\ell_i, \mathcal{I}^m_i), \quad m \in \{f, b, s, fus\}, \quad (2)$$

where $\mathcal{I}^f_i = f_i$, $\mathcal{I}^b_i = b_i$, $\mathcal{I}^s_i = s_i$, and $\mathcal{I}^{fus}_i = \{f_i, b_i, s_i, c_i\}$. Each $\hat{\mathbf{e}}^m_{i,j,k} = \{\hat{e}^m_{i,u,j,k}\}_{u=0}^{\ell_i-1}$ contains one emotion label per utterance u. The three unimodal channels are recognized independently, whereas fus serves as a joint interpretation reference. The ten labels for each utterance–channel pair form an empirical emotion distribution $EmoD^m_{i,u}$, where $p^m_{i,u}(emo)$ denotes the empirical probability of emotion emo. We define the aggregated label as $\hat{e}^m_{i,u} = emo$ if $p^m_{i,u}(emo) \geq \tau$; otherwise, $\hat{e}^m_{i,u} = ambiguous$. We set $\tau = 0.6$ during evaluation and independently validate this setting against human annotations in Table 7(c). The experts include DeepSeek-V3.2[3], Doubao-seed-1-6[5], GPT-5.2[6], Gemini-3-flash[7], and Claude-sonnet-4.5[8]. Their textual understanding and reasoning abilities support recognition of implicit cues, such as toughness used to mask vulnerability. Full prompts and output schemas are provided in the appendix.

4.2 Evaluation Dimensions and Metrics

4.2.1 Emotional Consistency (EC)

EC was originally proposed for text-based conversations (Feng et al., 2025) to measure whether role-playing agents reproduce the real and dynamic emotions of designated characters, comprising micro Emotional Consistency (EC-micro), Emotional Distribution Divergence (EDD), and Relative Character Discrepancy (RCD). In *MERRY*, we adapt these metrics to multimodal semantic responses and introduce two diagnostic metrics, Cross-modal Emotional Consistency (CEC) and Emotional Discrepancy (ED).

[6] https://openai.com/api/
[7] https://gemini.google.com/
[8] https://claude.ai/

Multimodal Emotional Consistency (MEC). To measure whether MRPAs express emotions consistent with ground-truth character responses, we extend EC-micro to multimodal responses:

$$MEC_\eta = \frac{\sum_{x=1}^{13} n_x F1_{x,\eta}}{\sum_{x=1}^{13} n_x}, \quad \eta \in \{lower, upper\}, \tag{3}$$

where n_x is the ground-truth support of the x-th emotion category and $F1_{x,\eta}$ is the corresponding class-wise F1 score. For sample i, let $G_i = \text{set}(\{e_{i,u}^{gt}\}_{u=0}^{\ell_i-1})$ and $P_i = \text{set}(\{\hat{e}_{i,u}^{fus} \mid \hat{e}_{i,u}^{fus} \neq ambiguous\}_{u=0}^{\ell_i-1})$, where $e_{i,u}^{gt}$ is utterance-level ground-truth emotion label, $\text{set}(\cdot)$ counts repeated labels only once, and ambiguous predictions are removed from P_i. For each category, its presence in G_i and P_i determines the confusion matrix. MEC_{lower} requires exact label matching, whereas MEC_{upper} uses broader emotional tendencies. Both scores range from 0 to 1, with higher values indicating better alignment. We omit the original VAD-based middle score because fixed Valence-Arousal-Dominance vectors may introduce language-, context-, and role-specific biases.

Emotional Distribution Divergence (EDD) measures the divergence between ground-truth and generated emotional transition patterns, whereas **Relative Character Discrepancy (RCD)** measures whether the average degree of emotional differentiation across roles is preserved. Since each evaluation sample i corresponds to a response from target role y at turn t of dialogue d, its utterance-level emotion labels can be reorganized as $e_{d,t,u}^{z,y}$ for transition analysis, where $z \in \{GT, RPA\}$ denotes the source, including ground-truth and vote-then-aggregate, respectively. Since there can be multiple utterances in a single response, we define inter-turn and intra-turn transition matrices as $M_{\text{inter}}^{z,y} = \text{RowNorm}(\sum_d \sum_t \text{TranPair}(e_{d,t-1,\ell_{d,t-1}-1}^{z,y}, e_{d,t,0}^{z,y}))$ and $M_{\text{intra}}^{z,y} = \text{RowNorm}(\sum_d \sum_t \sum_u \text{TranPair}(e_{d,t,u-1}^{z,y}, e_{d,t,u}^{z,y}))$, where all sums are taken over valid indices. Transition pairs containing an "*ambiguous*" label are omitted without reconnecting the remaining labels, and each nonempty matrix row is normalized to sum to one. Valid rates of transition pairs is reported in appendix. For $q \in \{\text{inter}, \text{intra}\}$, let $\mathcal{Y}_q$ denote the roles with valid matrices for both sources, $Y_q = |\mathcal{Y}_q|$, and $\mathcal{P}_q = \binom{\mathcal{Y}_q}{2}$ denote the set of unordered role pairs:

$$EDD_q = \frac{1}{Y_q} \sum_{y \in \mathcal{Y}_q} D_H(M_q^{GT,y}, M_q^{RPA,y}), \tag{4}$$

$$RCD_q = CD_q^{RPA} - CD_q^{GT}, \qquad CD_q^z = \frac{1}{|\mathcal{P}_q|} \sum_{\{a,b\} \in \mathcal{P}_q} D_H(M_q^{z,a}, M_q^{z,b}). \tag{5}$$

Here, $D_H(M, N)$ is the mean of the bounded and symmetric Hellinger distance (Hellinger, 1909) $d_H(m, n) = \sqrt{\frac{1}{2}\sum_x (\sqrt{m_x} - \sqrt{n_x})^2}$ over all rows. EDD lies in $[0, 1]$, with lower values indicating better alignment. RCD lies in $[-1, 1]$, with values closer to zero indicating better alignment; positive and negative values indicate over- and under-differentiated roles, respectively.

Cross-modal Emotional Consistency (CEC) measures the agreement among emotional meanings independently inferred from different channels. Let $\mathbf{e}_\eta^m = (\phi_\eta(e_{i,u}^m))_{i,u}$ denote the voted labels for channel m over all utterances, aligned across channels by (i, u), where ϕ_{lower} is the identity mapping and ϕ_{upper} maps emotion labels to broader emotional tendencies. Treating the four channels as nominal raters, we compute:

$$CEC_\eta = \alpha_{\text{Krippendorff}}^{\text{nominal}} \left([\mathbf{e}_\eta^f, \mathbf{e}_\eta^b, \mathbf{e}_\eta^s, \mathbf{e}_\eta^{fus}]^\top \right), \quad \eta \in \{lower, upper\}. \tag{6}$$

Ambiguous labels are treated as missing observations rather than an additional emotion category. The fusion channel fus is a joint reference based on all semantic fields, rather than an independently synthesized modality. CEC_{lower} uses exact labels, whereas CEC_{upper} uses broader emotional tendencies. Higher CEC indicates stronger cross-channel coherence.

Emotional Discrepancy (ED) measures the ambiguity of ERC results within each channel:

$$ED_m = \frac{\sum_i \sum_{u=0}^{\ell_i-1} \text{NormEntropy}(EmoD_{i,u}^m)}{\sum_i \ell_i}, \quad m \in \{f, b, s, fus\}. \tag{7}$$

Here, $EmoD_{i,u}^m$ is the empirical distribution of the ten raw ERC labels for channel m and utterance u in sample i. NormEntropy use the fixed denominator $\log 10$, corresponding to the maximum entropy attainable from ten ERC votes, thus ED ranges from 0 to 1. ED retains the full vote distribution, including cases whose voted label is ambiguous. Lower ED indicates more determinate emotional predictions.

Table 4: Main results of MRPAs on *MERRY* under zero-shot setting. Best results are in **bold** for tracks with two or more evaluated models. † refers to close-source models. * denotes conditional results; corresponding coverage is reported in the appendix.

Models	MEC ↑		EDD* ↓		RCD* ↔		CEC* ↑		ED ↓				RC* ↑		
	lower	upper	intra	inter	intra	inter	lower	upper	fus	spe	fac	bod	Exp.	Cha.	Rel.
Audio-Grounded Track															
OmniCharacter (7B)	0.235	0.536	0.382	0.455	-0.038	-0.035	0.362	0.348	0.301	0.170	0.217	0.206	4.077	4.020	3.977
Vision-Grounded Track															
Doubao-1.5v †	0.243	0.546	0.410	0.455	-0.033	**-0.006**	**0.658**	**0.745**	0.429	0.424	0.421	0.415	**4.886**	4.717	4.878
GPT-5-chat †	**0.251**	**0.559**	**0.404**	**0.449**	**-0.003**	0.016	0.642	0.731	**0.261**	**0.248**	**0.258**	**0.238**	4.863	**4.742**	**4.888**
Omni-Modal Track															
Gemini-2.5-pro †	**0.279**	**0.608**	0.393	0.456	0.013	0.019	**0.720**	**0.793**	0.249	0.234	0.236	0.216	**4.856**	**4.616**	**4.824**
Qwen3-Omni (30B/A3B)	0.242	0.553	0.407	0.429	0.040	0.021	0.593	0.735	0.263	0.243	0.233	0.231	4.575	4.468	4.630
MiniCPM-o-4.5 (9B)	**0.279**	0.540	**0.352**	**0.379**	0.006	-0.016	0.624	0.705	0.225	0.212	0.180	0.162	4.591	4.441	4.611
MiniCPM-o-2.6 (8B)	0.264	0.536	0.381	0.398	**-0.002**	-0.017	0.586	0.652	0.241	0.206	0.181	0.170	3.998	4.085	4.144
Qwen2.5-Omni (7B)	0.241	0.528	0.401	0.421	-0.006	**-0.008**	0.582	0.682	0.248	0.176	0.170	0.179	4.231	4.162	4.387
Phi-4-multimodal (5.6B)	0.256	0.498	0.354	0.398	-0.004	-0.016	0.647	0.700	**0.203**	**0.134**	**0.114**	**0.124**	3.087	3.677	3.536

4.2.2 Role Consistency (RC)

We evaluate RC along three dimensions, Experience (Exp.), Characteristics (Cha.), and Relationship (Rel.), using an LLM-as-a-judge method (Gu et al., 2025). Rather than assigning scores directly, each judge performs bidirectional evidence finding, identifying evidence items that support or contradict the response's consistency with the target role. Each item must include a verbatim quote from the response and a rationale supporting one judgment. No response span may support multiple items. The score is mapped from the presence and relative counts of supporting and conflicting items, as shown in Table 3. We use Doubao-seed-1-6[5] and GPT-5.2[6] as independent judges and average their valid scores. The conditional RC score and its coverage are both reported. Table 7(b) shows that our protocol agrees more closely with human reference scores than the explain-then-score baseline. The evidence schema and complete judging prompts are provided in the appendix.

Table 3: Mapping from bidirectional evidence to RC score. A and D denote the numbers of supporting and conflicting evidence items, respectively.

PositiveFlag	0	1	1	1	1	0
NegativeFlag	0	0	1	1	1	1
Evidence count	N/A	N/A	$A > D$	$A = D$	$A < D$	N/A
FinalScore	drop	5	4	3	2	1

5 Experiments

5.1 Experimental Setup

Zero-shot evaluation. We evaluate MRPAs in a zero-shot setting under three native-input tracks. The **omni-modal track** serves as the main track because it provides the full multimodal dialogue context and most closely matches the task setting of *MERRY*. For baseline analysis, it includes six well-known omni-modal models (GoogleAI, 2025; Xu et al., 2025a; Yao et al., 2024; Xu et al., 2025b; Cui et al., 2026; Microsoft et al., 2025). To include models with narrower native input capabilities, we additionally report a **vision-grounded track**, comprising Doubao-1.5v, whose full model name is Doubao-1.5-thinking-vision-pro[5] (ByteDanceSeed, 2025), and GPT-5-chat[6] (OpenAI, 2025), as well as an **audio-grounded track** comprising OmniCharacter (Zhang et al., 2025b). We use a temperature of 0.7 and a top-p value of 0.95 for all models, while keeping other decoding settings at their model-specific defaults. Although the available input modalities differ across tracks, all models are required to generate the same type of multimodal semantic response, and all outputs are processed by the same *MERRY* evaluation pipeline.

Fine-tuning-based evaluation. As a controlled diagnostic intervention, we fine-tune MiniCPM-o-2.6-8B with LoRA (Hu et al., 2021) on the training split of *MERRY-Data* and on two existing multimodal role-playing datasets, OmniCharacter (Zhang et al., 2025b) and MMRole (Dai et al., 2025), which contain 20 and 85 roles, respectively. To maintain the same scale as *MERRY-Data*, we

Table 5: Polarity-level transition probabilities of three representative models from different tracks.

Model	Positive (Pos.)		Neutral (Neu.)		Negative (Neg.)	
	As Target	Neg.→Pos.	As Target	Neu.→Neu.	As Target	Pos.→Neg.
Inter-turn						
Ground-Truth	0.177	0.078	0.31	0.587	0.512	0.262
OmniCharacter	0.218 (↑23.0%)	0.067 (↓14.6%)	0.196 (↓36.9%)	0.310 (↓47.1%)	0.586 (↑14.4%)	0.176 (↓32.9%)
GPT-5-chat	0.415 (↑134.3%)	0.193 (↑146.4%)	0.117 (↓62.3%)	0.231 (↓60.7%)	0.468 (↓8.7%)	0.178 (↓32.0%)
Gemini-2.5-pro	0.300 (↑69.5%)	0.112 (↑43.3%)	0.104 (↓66.4%)	0.252 (↓57.0%)	0.595 (↑16.2%)	0.261 (↓0.4%)
Intra-turn						
Ground-Truth	0.18	0.019	0.347	0.921	0.473	0.025
OmniCharacter	0.230 (↑28.0%)	0.014 (↓29.1%)	0.265 (↓23.8%)	0.821 (↓10.8%)	0.505 (↑6.8%)	0.018 (↓26.8%)
GPT-5-chat	0.476 (↑164.4%)	0.035 (↑83.6%)	0.114 (↓67.2%)	0.825 (↓10.4%)	0.410 (↓13.2%)	0.013 (↓47.3%)
Gemini-2.5-pro	0.329 (↑83.0%)	0.017 (↓10.1%)	0.144 (↓58.4%)	0.908 (↓1.4%)	0.526 (↑11.3%)	0.013 (↓47.2%)

randomly sample a subset from each while preserving its original role-to-sample ratio. All models are trained on a single NVIDIA A800 GPU with 80 GB of memory. We use a LoRA rank of 16, an α value of 32, and a dropout rate of 0.05. The learning rate is set to 10^{-5}, with a cosine learning-rate scheduler and a warm-up ratio of 0.1. For monitoring and checkpoint selection, we hold out a randomly sampled 1% of each training set as a validation set.

5.2 Main Results and Analysis

As our goal is to diagnose recurring weaknesses rather than rank models, we report valid-only zero-shot performance in Table 4, leaving detailed model-wise coverage reported in the appendix. Across zero-shot models, fusion ambiguity ranges from 15.77% to 25.26%. Coverage ranges from 91.2%–97.3% for CEC, 63.1%–74.9% and 78.5%–91.5% for intra- and inter-turn transitions, and 36.6%–67.6%, 92.2%–99.5%, and 72.4%–93.9% for *Exp.*, *Cha.*, and *Rel.*, respectively. Coverage of *Exp.* is lower because responses usually do not explicitly invoke role-specific past events during dialogue, leaving less citable evidence for this dimension.

Within the vision-grounded track, Doubao-1.5v and GPT-5-chat show comparable but metric-dependent performance: GPT-5-chat obtains slightly higher MEC and lower EDD, whereas Doubao-1.5v achieves higher CEC and a slightly higher conditional Experience score. Within the omni-modal track, Gemini-2.5-pro performs relatively strongly in MEC, CEC, and conditional RC, while MiniCPM-o-4.5 and Phi-4-multimodal obtain lower EDD or ED on certain dimensions. Thus, no general foundation model consistently dominates all semantic dimensions.

Figure 4 reveals highly uneven emotion-level performance. Models perform relatively well on prototypical and frequent emotions, such as neutral, anger, happy, sadness, and worried, but substantially worse on fine-grained or context-dependent emotions, such as grateful, relaxed, fear, disgust, and astonished. Among valid transition pairs, Table 5 shows consistently lower *Neu. as Target* and *Neu→Neu* probabilities than the ground truth, suggesting weaker preservation of neutral states. Polarity shifts are model-dependent: GPT-5-chat shows the clearest positive shift, whereas Gemini-2.5-pro and OmniCharacter show higher *Neg. as Target* probabilities in both inter- and intra-turn settings. Together, these patterns suggest that **current MRPAs tend to rely on salient emotional prototypes, preserve neutral states less effectively, and exhibit model-specific polarity shifts**.

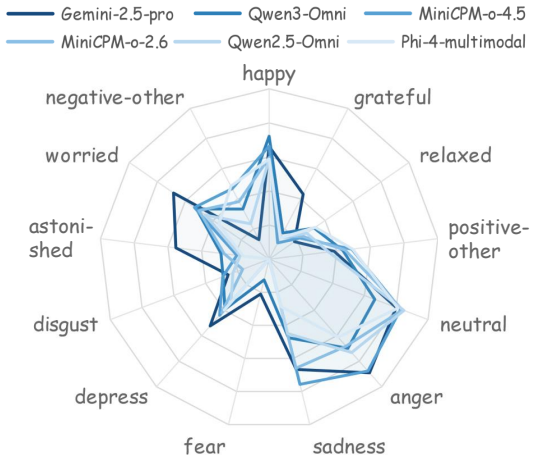

Figure 4: Visualization of per-emotion F1 scores of omni-track models on MEC_{lower}. Rings range from 0 to 0.5 with 0.1 intervals.

We further examine whether simple fine-tuning mitigates these issues. As shown in Table 6, fine-tuning on *MERRY-Data* substantially improves

Table 6: Results of fine-tuned models using MiniCPM-o-2.6 (8B) as a unified base model for fairness. Best results are in **bold**. * denotes conditional results; corresponding coverage is reported in the appendix.

Trainset	Method	MEC ↑		EDD* ↓		RCD* ↔		CEC* ↑		ED ↓				RC* ↑		
		lower	upper	intra	inter	intra	inter	lower	upper	fus	spe	fac	bod	Exp.	Cha.	Rel.
None	None	0.264	0.536	0.381	**0.398**	**-0.002**	-0.017	0.586	0.652	0.241	0.206	0.181	0.170	3.998	4.085	4.144
OmniCharacter	SFT	0.256	0.534	**0.365**	0.416	-0.005	**-0.001**	0.491	0.514	0.278	0.179	0.220	0.182	4.189	**4.268**	**4.463**
MMRole	SFT	0.275	0.517	0.388	0.402	0.016	-0.061	0.550	0.586	0.210	0.071	0.059	0.057	4.268	4.263	4.402
MERRY (Ours)	SFT	**0.291**	**0.541**	0.379	0.412	-0.024	-0.021	**0.934**	**0.962**	**0.089**	**0.038**	**0.035**	**0.049**	**4.363**	4.167	4.311

Table 7: Human agreement results. (a) and (b) report Krippendorff's alpha α. (c) reports Weighted F1 under each shared voting threshold τ. Best results in (b) are in **bold**.

(a) Description Quality

Modality	Avg.	α
Facial	4.348	0.585
Body	4.323	0.539
Speech	4.719	0.571

(b) LLM-as-a-Judge for RC

	bidirectional evidence-finding			explain-then-score		
	Avg.	GPT-5.2	Doubao	Avg.	GPT-5.2	Doubao
Exp.	**0.745**	0.631	0.503	0.558	0.490	0.367
Cha.	**0.619**	0.552	0.443	0.386	0.354	0.238
Rel.	**0.721**	0.619	0.492	0.482	0.407	0.339

(c) ERC Experts

τ	0.4	0.5	0.6	0.7	0.8
Facial	0.695	0.770	0.760	0.804	0.798
Body	0.704	0.676	0.669	0.790	0.779
Speech	0.651	0.681	0.678	0.794	0.783
Fusion	0.635	0.695	0.690	0.778	0.731
Valid Rate	0.831	0.762	0.762	0.558	0.558

CEC and ED and slightly improves MEC. These gains are accompanied by lower emotion ambiguity and broader valid-transition coverage, indicating that they do not arise from retaining a narrower subset. However, improvements in the role-related EDD, RCD, and RC metrics remain limited. Fine-tuning on OmniCharacter and MMRole yields larger conditional RC gains, although neither evidence coverage nor emotional consistency improves consistently. This contrast is consistent with their supervision: *MERRY-Data* directly supervises multimodal emotional semantics, whereas role-centered datasets provide denser character-grounding signals. This observation highlights both the advantages and limitations of simple supervised fine-tuning, pointing to the need for new training protocols that better enhance long-tailed abilities.

5.3 Effectiveness of *MERRY*

To assess the reliability of our annotations and evaluation protocols, we conduct three human studies, each covering 100 random cases: (1) **Description Quality**: three raters score the consistency between generated descriptions and their source modalities on a 1–5 scale; (2) **LLM-as-a-Judge for RC**: three raters directly score the experience, characteristic, and relationship consistency of each response on a 1–5 scale, and their ratings are averaged to form a shared human reference; (3) **ERC Experts**: five raters independently assign utterance-level emotion labels. Description-quality cases are sampled from *MERRY-Data*, whereas the other cases are sampled from the fully valid output pool of evaluated MRPAs. As shown in Table 7(a), description-quality scores average above 4, indicating that the generated descriptions are mostly consistent with their source modalities, with moderate inter-rater agreement (0.4–0.6). Table 7(b) shows that bidirectional evidence-finding achieves consistently higher agreement with human reference scores than explain-then-score, for both individual LLM judges and their averaged scores across all RC dimensions. In Table 7(c), human and ERC votes are independently aggregated under the same threshold, and Weighted F1 measures agreement between the resulting consensus labels. The result at $\tau = 0.6$ independently validates our fixed threshold, while the other thresholds provide a sensitivity analysis. Further details are provided in the appendix.

6 Conclusion

Existing evaluation methods show limited sensitivity to semantic-level errors in multimodal role-playing responses. To address this gap, we propose *MERRY*, a semantic-level framework for evaluating the **M**ultimodal **E**motional Consistency and **R**ole Consistency of **R**ole-Pla**y**ing agents, together with *MERRY-Data*, a multi-turn multimodal emotional dialogue dataset. Our findings show that existing MRPAs often simplify character emotions into salient prototypes, weaken neutral-state preservation, and shift fine-grained emotional dynamics toward model-specific polarity patterns.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, translation assistance, and minor writing refinement. We did not use generative AI tools to generate the core scientific ideas, experimental conclusions, or unsupported claims. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. *MERRY-Data* is constructed based on CPED (Chen et al., 2022), with approval from the related CPED authors. Additional role profiles, character images, and episode information are collected from publicly available sources and used only for academic research, with source links retained in the dataset documentation. The dataset may contain biases inherited from the source media. In addition, MRPAs and other AI systems used with this dataset may generate simulated role-playing responses that could be mistaken for statements by real individuals. Users should clearly indicate that generated role-playing content is simulated and does not represent genuine speech by the corresponding characters or actors. Human agreement experiments were conducted with informed participants under school ethics review approval, with anonymization, confidentiality agreements, and data-use notification.

REPRODUCIBILITY STATEMENT

We provide detailed demonstration of the task definition, data construction pipeline, evaluation pipeline, metrics, model settings, and prompt templates in the main paper and appendix. The evaluation section defines output formatting, ERC experts, voting thresholds, EC and RC metrics, and LLM-as-a-judge scoring rules. Code, prompts, configuration files, and proposed annotations will be released subject to the acceptance of paper, as well as licenses and access constraints of the underlying data.

REFERENCES

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3291–3325, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.197. URL https://aclanthology.org/2024.findings-acl.197/.

ByteDanceSeed. Seed1.5-vl technical report, 2025. URL https://arxiv.org/abs/2505.07062.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.125. URL https://aclanthology.org/2024.findings-acl.125/.

Keqi Chen, Wenxin Fu, Qihang Lu, Zekai Sun, Yizhong Geng, Yi Liu, Puyuan Guo, Yingming Gao, and Ya Li. Emo-avatar: An llm-agent-orchestrated framework for multimodal emotional support in human animation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 14014–14020, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3762030. URL https://doi.org/10.1145/3746027.3762030.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. Large language models meet harry potter: A dataset for aligning dialogue agents with

characters. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8506–8520, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.570. URL https://aclanthology.org/2023.findings-emnlp.570/.

Xi Chen, Wei Xue, and Yike Guo. ActorMind: Emulating human actor reasoning for speech role-playing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34399–34413, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1718. URL https://aclanthology.org/2026.findings-acl.1718/.

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

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. Towards multimodal emotional support conversation systems, 2024. URL https://arxiv.org/abs/2408.03650.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. Mm-role: A comprehensive framework for developing and evaluating multimodal role-playing agents. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 66300–66321, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a5c7206fd66e8314bb21a04492359353-Paper-Conference.pdf.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 6218–6240, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.316. URL https://aclanthology.org/2025.naacl-long.316/.

Dongjie Fu, Fangming Feng, Xize Cheng, Linjun Li, Zhou Zhao, and Tao Jin. Character beyond speech: Leveraging role-playing evaluation in audio large language models via reinforcement learning, 2026. URL https://arxiv.org/abs/2604.13804.

GoogleAI. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

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. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909. doi: doi:10.1515/crll.1909.136.210. URL https://doi.org/10.1515/crll.1909.136.210.

Tom Hollenstein and Dianna Lanteigne. Models and methods of emotional concordance. *Biological Psychology*, 98:1–5, 2014. doi: 10.1016/j.biopsycho.2013.12.012. URL https://doi.org/10.1016/j.biopsycho.2013.12.012.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Xinmeng Che, Hui Li, Xiaoran Fan, Ming Zhang, Junjie Ye, Shihan Dou, Zhiheng Xi, Jingqi Tong, Yilong Wu, Baoyu Fan, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Speechrole: A large-scale dataset and benchmark for evaluating speech role-playing agents, 2026. URL https://arxiv.org/abs/2508.02013.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *CoRR*, abs/2309.11911, 2023. doi: 10.48550/ARXIV.2309.11911. URL https://doi.org/10.48550/arXiv.2309.11911.

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. Chatharuhi: Reviving anime character in reality via large language model, 2023. URL https://arxiv.org/abs/2308.09597.

Wenyu Li, Xiaoqi Jiao, Yi Chang, Guangyan Zhang, and Yiwen Guo. Audiorole: An audio dataset for character role-playing in large language models, 2026. URL https://arxiv.org/abs/2509.23435.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. Generative expressive conversational speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 4187–4196, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681697. URL https://doi.org/10.1145/3664647.3681697.

Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3b parameters are all you need for unified multi-modal and multi-task human animation. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i10.37746. URL https://doi.org/10.1609/aaai.v40i10.37746.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.814/.

Jiatong Shi, Jionghao Han, Yichen Lu, Santiago Pascual, Pengfei Wu, Chenye Cui, Shinji Watanabe, Chao Weng, and Cong Zhou. Speech-drame: A framework for human-aligned benchmarks in speech role-play, 2025. URL https://arxiv.org/abs/2511.01261.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 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)*, pp. 12672–12676, 2024. doi: 10.1109/ICASSP48485.2024.10448173.

Chao Tang, Jianzong Wu, Qingyu Shi, Ye Tian, Aixi Zhang, Hao Jiang, Jiangning Zhang, and Yunhai Tong. Towards customized multimodal role-play, 2026a. URL https://arxiv.org/abs/2605.08129.

Wenqiu Tang, Zhen Wan, Takahiro Komamizu, and Ichiro Ide. Desrpa: Decoupled speech role-playing agent via inference-time intervention, 2026b. URL https://arxiv.org/abs/2606.17669.

Yihong Tang, Kehai Chen, Xuefeng Bai, and Min Zhang. Through the lens of character: Resolving modality-role interference in multimodal role-playing agent, 2026c. URL https://arxiv.org/abs/2605.09443.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.638. URL https://aclanthology.org/2024.acl-long.638/.

Jan Van den Stock, Ruthger Righart, and Beatrice de Gelder. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487–494, August 2007.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds, 2024a. URL https://arxiv.org/abs/2412.05631.

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. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

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. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre

Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

Weihao Wu, Liang Cao, Xinyu Wu, Zhiwei Lin, Rui Niu, Jingbei Li, and Zhiyong Wu. Voxrole: A comprehensive benchmark for evaluating speech-based role-playing agents, 2025. URL https://arxiv.org/abs/2509.03940.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/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, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

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

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11992–12022, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.586. URL https://aclanthology.org/2025.acl-long.586/.

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark, 2025a. URL https://arxiv.org/abs/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. OmniCharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26318–26331, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1276. URL https://aclanthology.org/2025.acl-long.1276/.

Xueqiao Zhang, Chao Zhang, Jingtao Xu, Yifan Zhu, Xin Shi, Yi Yang, and Yawei Luo. Video2Roleplay: A multimodal dataset and framework for video-guided role-playing agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23677–23703, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1209. URL https://aclanthology.org/2025.emnlp-main.1209/.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5699–5710, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.391. URL https://aclanthology.org/2022.acl-long.391/.

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. CharacterGLM: Customizing social characters with large language models. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1457–1476, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.107. URL https://aclanthology.org/2024.emnlp-industry.107/.

A Details of the Controlled Preliminary Experiment

Experimental setup. We conduct a controlled preliminary experiment to examine whether existing evaluation protocols can distinguish errors in intermediate multimodal semantic planning from errors in final modality synthesis. We uniformly sample 100 test samples without replacement from *MERRY-Data*. For each base sample, we construct two rendering conditions: (1) a ground-truth semantic condition, where the response content, speech prompt, facial expression, and body movement are all taken from the base sample; and (2) a wrong-semantic condition, where the response content is kept unchanged, while the multimodal semantic annotations are replaced by those from a randomly mismatched sample. The mismatched source is selected from a different TV series and dialogue with disjoint system/user roles whenever possible. In the current subset, all 100 wrong-semantic sources satisfy this strict mismatch rule. This yields 200 conditions in total, with 100 ground-truth semantic samples and 100 wrong-semantic samples.

Synthesis models. To instantiate the two semantic conditions as comparable multimodal outputs, we use the same synthesis pipeline for both ground-truth and wrong-semantic samples. For speech generation, we use Qwen3-TTS (Hu et al., 2026) to synthesize the response audio from the unchanged response content and the corresponding speech prompt. For video generation, we use EchoMimicV3 (Meng et al., 2026), taking the first frame of the source test video as the reference image and the synthesized speech as the driving audio. The facial-expression and body-movement descriptions are provided as semantic prompts to guide the generated talking video. Since the synthesis models and response content are fixed across the two conditions, the main controlled variable is whether the intermediate multimodal semantics are ground-truth or randomly mismatched.

Rendering and evaluation. For both conditions, we synthesize speech from the unchanged response content and the corresponding speech prompt, and generate talking videos using the source-video first frame as the visual reference. We then evaluate the resulting samples using four groups of metrics: MERRY semantic metrics, textual role-playing metrics, speech-role metrics, and modality-synthesis metrics. For MERRY, we report the average multimodal emotional consistency score (*MEC_avg*) and role consistency score (*RC_avg*). For textual role-playing evaluation, we use CharacterRM scores, including character consistency, conversation ability, and role-playing attractiveness. Notably, CharacterRM receives multimodal semantics enclosed in brackets. And character consistency provides coarse-grained evaluation on these behavioral semantics. For speech-role evaluation, we follow the original SpeechRole grouping and report the averages of SF/SN, PC/EA, and PeC/KC. For modality-synthesis evaluation, we report OmniCharacter speech metrics, AvaMERG automatic video metrics, and AvaMERG human metrics. During the AvaMERG human evaluation, raters were shown the complete multimodal response, including its multimodal semantic descriptions and synthesized audiovisual output. Content Accuracy assesses whether the textual content and multimodal semantic descriptions are semantically aligned with the target response. Style Accuracy evaluates whether the synthesized video realizes the intended character appearance and voice timbre while maintaining cross-modal alignment. For SyncNet-based synchronization confidence, samples without a valid face track are excluded from the score distribution and reported through the valid sample count rather than being assigned zero scores.

Distribution comparison. For each metric, we compare the score distributions of the ground-truth and wrong-semantic conditions. We report the mean score of each condition, the mean difference, the overlap coefficient between the two distributions, and the pairwise probability that a ground-truth sample receives a better score than a wrong-semantic sample. For lower-is-better metrics such

Table 8: Distribution differences between ground-truth and wrong-semantic conditions under automatic metrics. The-lower-the-better metrics are annotated with ↓.

Family	Metric	GT mean	Wrong mean	Diff.	Overlap	$P(\mathrm{GT} > \mathrm{Wrong})$
MERRY	MEC_avg	0.9383	0.4048	0.5335	0.3700	0.6949
MERRY	RC_avg	5.0000	3.8867	1.1133	0.1400	0.8600
OmniCharacter	CER ↓	0.0987	0.2131	-0.1144	0.9700	0.3890
OmniCharacter	WavLM SIM	0.7387	0.7330	0.0057	0.7700	0.5174
CharacterRM	Character Consistency	2.6363	2.5248	0.1115	0.7200	0.6009
CharacterRM	Conversation Ability	2.8915	2.7481	0.1434	0.6700	0.6080
CharacterRM	Role-playing Attractiveness	2.6034	2.4978	0.1056	0.6400	0.5915
SpeechRole	Avg. of SF&SN	0.7950	0.7340	0.0610	0.8500	0.4980
SpeechRole	Avg. of PC&EA	0.5568	0.5580	-0.0012	0.7500	0.4510
SpeechRole	Avg. of PeC&KC	0.6325	0.6994	-0.0670	0.9400	0.4228
AvaMERG	SSIM	0.6559	0.6498	0.0061	0.8300	0.5148
AvaMERG	CPBD	0.2609	0.2587	0.0021	0.8300	0.5009
AvaMERG	Sync_cf	3.5870	3.8058	-0.2187	0.5760	0.4458

Table 9: Distribution differences under AvaMERG human evaluation.

Metric	GT mean	Wrong mean	Diff.	Overlap	$P(x > 3)$
Avg. of Content Accuracy	3.5367	3.4600	0.0767	0.8200	0.6500
Avg. of Style Accuracy	2.7733	2.7467	0.0267	0.8000	0.3250

as CER, "better" is computed after accounting for the metric direction. A lower overlap coefficient and a higher $P(\mathrm{GT} > \mathrm{Wrong})$ indicate stronger sensitivity to wrong multimodal semantics.

Analysis. The results show a clear separation between the two conditions for *MERRY*. In particular, as detailed in Table 8, *MEC_avg* decreases from 0.9383 to 0.4048 under wrong semantics, and *RC_avg* decreases from 5.0000 to 3.8867, with substantially lower distribution overlap than most comparison metrics. By contrast, textual role-playing metrics show only small mean shifts, because the response content is intentionally kept unchanged. Similarly, most modality-synthesis metrics remain insensitive to the semantic mismatch: WavLM similarity, SSIM, CPBD, and the AvaMERG human scores have high distribution overlap and near-chance pairwise preference. Moreover, as shown in Table 9, across all 200 evaluated videos, the proportion of scores above 3 was $P(x > 3) = 65.0\%$ for Content Accuracy and 32.5% for Style Accuracy, indicating that most responses achieved semantic alignment, whereas faithful audiovisual style realization remained more challenging. These results indicate that both final audio/video quality metrics and text-level role-playing metrics show limited sensitivity to semantic-level errors in multimodal role-playing responses. The controlled experiment therefore supports the need for a semantic-level evaluation layer before multimodal semantics are entangled with synthesis quality.

B Details of Model-wise Coverage

For EC, we report model-wise coverage results in Table 10. The ambiguous rate of each channel is computed over all utterance–channel pairs. No missing or malformed ERC labels are observed. For emotional transitions, a candidate pair is valid only when both endpoint emotions are non-ambiguous. The intra- and inter-turn valid rates are computed against the corresponding number of candidate pairs constructed for each model.

For RC, all outputs from both judges pass schema validation. A sample is dropped for a dimension if at least one judge identifies neither positive nor negative evidence, making the evidence-based score undefined. Thus, the drop rates in Table 11 reflect protocol-level evidence coverage rather than missing or malformed judge outputs. Since both judges provide valid outputs for all 1,690 samples, the denominator is fixed across models and dimensions.

C Details of Human Agreement Experiments

Participants and General Setup. We conducted three human-validation studies: Description Quality, LLM-as-a-Judge for RC, and ERC Experts. Each study evaluated 100 cases. Five graduate students are recruited for ERC agreement experiment, while three graduate students are recruited for the others. Before annotation, the raters received task-specific instructions and several carefully designed out-of-domain examples. All ratings were collected independently.

Description Quality. We randomly sampled 100 cases from *MERRY-Data*. For each case, the raters independently evaluated the consistency of the facial-expression, body-movement, and speech-prompt descriptions with the corresponding source audiovisual clip. They were not informed that these descriptions were annotations from *MERRY-Data*. Each description was scored using the following criteria:

- 1: *Completely inconsistent*;
- 2: *Mostly inconsistent, with only minor consistencies*;
- 3: *Generally consistent, although noticeable inconsistencies remain*;
- 4: *Mostly consistent, with only minor inconsistencies*;
- 5: *Fully consistent*.

As reported in Table 7(a), the mean scores for facial expressions, body movements, and speech prompts are 4.348, 4.323, and 4.719, respectively. Their corresponding Krippendorff's alpha values are 0.585, 0.539, and 0.571, indicating moderate inter-rater agreement.

LLM-as-a-Judge for RC. We sample 100 MRPA responses from the evaluated output pool for which both GPT-5.2 and Doubao-seed-1-6 produce valid scores. Each case contains the role profile, dialogue background, dialogue context, and predicted multimodal semantic response. Three human raters independently assign direct 1–5 scores for experience, characteristic, and relationship consistency based on information above. For each case and dimension, their ratings are averaged to obtain a human reference score. We evaluate GPT-5.2 and Doubao-seed-1-6 under both protocols and compute Krippendorff's alpha between the human reference scores and either the individual LLM scores or their averages. As shown in Table 7(b), bidirectional evidence-finding consistently achieves higher agreement with the shared human reference than explain-then-score across all three RC dimensions. The detailed criteria follow those used for Description Quality, but assess consistency along different dimensions.

ERC Experts. Human raters independently assigned one of the 13 CPED emotion categories to every utterance under four channel-specific settings: facial expression, body movement, speech prompt, and multimodal fusion. The 100 evaluated responses contain 367 utterances, yielding 367 annotations per channel before consensus filtering. The results in Table 7(c) and 12 validate the effectiveness of ERC experts at $\tau = 0.6$. The valid rate in Table 7(c) is the average of coverage of four channel.

Ethics. Before participation, all raters were informed of the study procedures, potential risks, and intended use of the collected annotations. We protected participant privacy through anonymization and confidentiality measures. The human studies were conducted with approval from the institutional ethics review board of our university.

D Details of Data Statistics

An example of profile can be found in Figure 5. The emotional distribution of different splits of *MERRY-Data* are listed in Table 13. Besides, detailed statistics of roles such as age, gender and personality trait can be referred to the CPED dataset (Chen et al., 2022).

E Details of Emotion Labels

Emotion labels are 13 types of emotions in CPED, including happy, grateful, relaxed, and positive-other for positive emotional tendency; neutral for neutral emotional tendency; anger, sadness, fear, depress, disgust, astonished, worried, and negative-other for negative emotional tendency. The explanations for each emotion also align with those in CPED, which have already been involved in our prompt for ERC step as shown in Appendix F.6.

F Prompt Templates

This appendix reports the prompts used for data construction, multimodal response generation, output normalization, and semantic evaluation. Chinese prompts were used for Chinese role-playing data; the English versions are provided as faithful documentation. Braced terms denote runtime variables. The templates describe the evaluation protocol at the level required for reproducibility and may omit provider-specific API wrappers.

F.1 Multimodal role-playing agent

English template: multimodal role-playing response generation

```
SYSTEM PROMPT

# Task
You will act as {sys_role_name}. The character profile is:
{role_profile}

The image of your character is:
{sys_role_image}

I will act as {usr_role_name}. The image of my character is:
{usr_role_image}

In each dialogue turn, I will provide an utterance, chronologically
ordered
video frames, and the corresponding audio. Using the character
profile, the
preceding context, and the current multimodal input, generate a
response that
portrays the character at the current point in the narrative. The
response
must contain four fields:

1. facial_expression: a realizable facial expression for the
response;
2. body_movement: a realizable body movement for the response;
3. speech_prompt: instructions for tone, pace, pitch, and volume;
4. content: the words spoken by the character.

The four fields should be semantically and emotionally coordinated
clause by clause.
The number of clauses is determined by content; clause delimiters
include `.`,
`?`, `;`, and `!` in either English half-width or Chinese full-width
form.
You may use words or punctuation in the multimodal fields to clearly
distinguish between clauses.
This does not require identical surface emotions. If masking,
restraint, irony, or
another justified cross-modal contrast is intended, explicitly
describe the
shared underlying emotion and how it is expressed in the relevant
fields. Do
not introduce experiences or relationship changes that occur after
the current
point in the narrative.
```

```
# Output format
Return only the following JSON object, without commentary:
{
  "facial_expression":  "{facial expression}",
  "body_movement":  "{body movement}",
  "speech_prompt":  "{tone and intonation}",
  "content":  "{response content}"
}
# Previous information
The prerequisite information for {sys_role_name} at this point is:
{prev_on_seg}
# The conversation starts now
USER PROMPT
{usr_role_name}:  {content}
{video_frames}
{audio_input}
{sys_role_name}:
```

F.2 VIDEO-BASED FACIAL AND BODY DESCRIPTIONS

English template: video-based facial and body descriptions

```
# Task
Given a character poster and a chronologically sampled video-frame
sequence,
first identify the person shown in the poster (typically the
speaker), and then
describe that person's facial expressions and body movements
throughout the
clip.  Adjacent frames are {time_interval} seconds apart; the sequence
contains
{total_frames} frames and spans {duration_time} seconds.
# Annotation requirements
1.  Facial expression:  describe visible details of the eyebrows,
eyes, gaze,
   mouth corners, mouth shape, and facial muscles, including temporal
changes.
2.  Body movement:  describe visible movements of the arms, hands,
legs, feet,
   and torso, including changes in posture.
3.  Write one natural, coherent passage for each field rather than
annotating
   individual frames.
4.  Rely only on visible evidence.  Do not infer unobservable mental
states, and
   do not omit salient actions.
5.  Ignore frames without the target person and do not describe
unrelated
   people.  Direct interactions with relevant people or objects may be
noted.
6.  Ignore subtitles, clothing, and background details.  Limit each
field to
   200 Chinese characters in the operational Chinese prompt.
7.  The emotion list is an auxiliary cue for possible temporal
changes.  If it
   conflicts with visible evidence, follow the video rather than
forcing the
   description to match the label.
# Input
Character poster:  {poster_img}
```

```
Video-frame sequence: {frames_of_video}
Utterance-level emotion list: {emotions}
# Output format
Facial expression: {facial expression and temporal changes}
Body movement: {body movement and temporal changes}
```

F.3 AUDIO-BASED SPEECH PROMPT CONSTRUCTION

English template: audio-based speech-prompt construction

```
# Task
Given utterance-level text, emotion cues, and numerical acoustic
features for
a speech recording, write one concise Chinese speech prompt that
enables a TTS
model to reproduce the speaking style. Preferably use 30--50 Chinese
characters to cover emotion, speaking rate, pitch, volume, and their
temporal
changes. Describe how the line is spoken; do not repeat its lexical
content or
use personal pronouns.
# Feature definitions
1. The recording contains {len_utts} clauses. Text, emotion, and
speaking rate
   are listed in clause order.
2. Speaking rate, pitch, and volume are dimensionless standardized
values with
   mean 0 and standard deviation 1.
3. Pitch and volume are computed from synchronized frames separated
by
   {frame_interval_sec} seconds. Use their averages for overall level
and the
   original sequences for fluctuations and transitions.
4. Values in [-1, 1] are treated as typical; (0, 1] is slightly high
and
   [-1, 0) slightly low. Values above 1 or below -1 are high or low,
and
   absolute values above 3 are extreme.
5. Emotion labels are auxiliary cues. Ground the acoustic
description in the
   supplied features and do not add unsupported tones or mental
states.
# Current audio
{
  "total_duration": {duration_ttl},
  "text": {content_texts},
  "emotion": {emotions},
  "speaking_rate": {speeds},
  "pitch_sequence": {pitches_10ms},
  "mean_pitch": {pitches_avg},
  "volume_sequence": {volumes_10ms},
  "mean_volume": {volumes_avg}
}
# Output
Speech prompt: {Chinese TTS instruction}
```

F.4 DESCRIPTION QUALITY CONTROL

English template: multimodal-description quality control

```
# Task
You are a data-annotation reviewer. The input contains a character
poster, a
video-frame sequence, ordered utterance-level emotion labels, and
existing
descriptions of the target person's facial expressions and body
movements.
Check whether the descriptions faithfully identify the target person
and match
the visible content and temporal order. Make the smallest necessary
revision.
# Review criteria
1. Modality faithfulness: retain only expressions, actions, and
temporal
   changes supported by the video. Do not add mental states that
cannot be
   supported by the video or context.
2. Emotion labels are auxiliary screening cues for locating
potentially
   inconsistent annotations, not targets that the descriptions must be
forced
   to match. If a label conflicts with clear modality evidence,
preserve the
   observed expression and flag the discrepancy for human review.
3. Preserve justified cross-modal contrast. For example, tough
speech and a
   vulnerable facial expression may coexist and should not be
rewritten into
   the same surface emotion. State a relation such as "masking
vulnerability
   with toughness" only when supported by context and visible
evidence.
4. Preserve correct information. Revise only mistaken identity,
hallucinated
   detail, omitted salient change, incorrect temporal order, or
unsupported
   emotion attribution.
5. `is_aligned` indicates whether the original descriptions require
no change.
   If true, return both descriptions unchanged; otherwise, return each
complete
   revised description.
# Input
Character poster: {poster_img}
Video-frame sequence: {frames_of_video}
Utterance-level emotion labels: {emotions}
Original facial-expression description: {facial_expression}
Original body-movement description: {body_movement}
# Output format
Return only JSON:
{
  "is_aligned": {true or false},
  "facial_expression": "{complete reviewed description}",
  "body_movement": "{complete reviewed description}"
}
```

F.5 Output normalization and format repair

English template: output normalization and format repair

```
# Task
Normalize the given `raw_output` into a JSON object conforming to the
schema
below.  This step repairs structure and surface formatting only.  It
must not
rewrite, complete, summarize, or infer the original semantics.
# Schema
{
  "facial_expression":  "string",
  "body_movement":  "string",
  "speech_prompt":  "string",
  "content":  "string"
}
# Rules
1.  All four fields must be present and field names are
case-sensitive.
2.  Fill a missing field with an empty string.  If the input is plain
text only,
   preserve it in `content` and leave the other fields empty; do not
infer
   missing modalities from the text.
3.  You may remove code fences, text outside the fields, meaningless
leading or
   trailing whitespace, and obvious consecutive infinite repetition
caused by
   a generation failure.  Do not remove semantically meaningful
repetition.
4.  Escaped line breaks and tabs may be normalized to spaces, but
preserve word
   order, facts, emotion, intensity, negation, and modality
assignment.
5.  Return exactly one valid JSON object and no explanation.
# Input
raw_output:
{raw_output}
# Output
```

F.6 Emotion recognition classifier

The face, body, and speech calls are executed independently. Each receives only the designated semantic description. A separate fusion call receives all three modality descriptions and the textual response content. None of the calls receives dialogue history, a role profile, or ground-truth emotion labels. Thus, the ground-truth labels used by MEC are not exposed to ERC.

English template: independent single-modality ERC

```
SYSTEM PROMPT
You are an expert in Chinese emotion recognition.  Infer emotion
strictly from
the input field provided in the current call.
USER PROMPT
# Task
Using only the semantic description in `{modality_name}`, classify the
emotion
of each of {len_utt} clauses.  Do not use or infer response content,
dialogue
```

```
history, a role profile, ground-truth emotion labels, or any other
modality.
If the description does not clearly distinguish multiple emotions,
select the
label best supported by the current evidence; disagreement across
experts is
handled during subsequent aggregation.
# Emotion label set
- Happy: genuine happiness or delight arising from inner pleasure;
- Grateful: thankfulness or gratitude;
- Relaxed: comfort without marked stress or tension;
- Positive-other: a positive emotion outside the preceding
categories;
- Neutral: no marked emotional tendency or a steady neutral state;
- Anger: anger, dissatisfaction, or aggression;
- Sadness: sorrow, loss, or grief;
- Fear: fear of danger or an adverse outcome;
- Depress: low spirits, disappointment, discouragement, or low
motivation;
- Disgust: aversion, disdain, or repulsion;
- Astonished: negative surprise or being caught off guard, not
pleasant surprise;
- Worried: reflective concern or anxiety, distinct from fear;
- Negative-other: a negative emotion outside the preceding
categories.
# Requirements
1. The list length must equal {len_utt} and follow clause order.
2. Every element must be one of the 13 labels above.
3. Return JSON only, without explanation.
# Input
Modality: {modality_name}
Description: {modality_description}
# Output format
{
  "{modality_name}": ["{clause 1 label}", "...", "{clause N label}"]
}
```

English template: fusion ERC

```
SYSTEM PROMPT
You are an expert in Chinese multimodal emotion recognition.
USER PROMPT
# Task
Jointly interpret the facial expression, body movement, speech
prompt, and
textual response content, and assign a fused emotion label to each of
{len_utt} clauses. `fusion` is a joint semantic interpretation rather
than an
independently synthesized modality. Do not use dialogue history, a
role
profile, or ground-truth emotion labels.
Different modalities may exhibit justified surface contrast. If the
descriptions explicitly indicate that toughness, restraint, or irony
masks an
underlying emotion, infer the joint emotion from all evidence rather
than
requiring identical surface labels. Use the same 13-label set and
definitions
as in the single-modality template.
```

```
# Requirements
1. The list length must equal {len_utt} and follow clause order.
2. Every element must be one of the predefined 13 Chinese emotion
labels.
3. Return JSON only, without explanation.
# Input
{
  "facial_expression": {facial_expression},
  "body_movement": {body_movement},
  "speech_prompt": {speech_prompt},
  "content": {content}
}
# Output format
{
  "fusion": ["{clause 1 label}", "...", "{clause N label}"]
}
```

F.7 ROLE-CONSISTENCY JUDGE

The same template is instantiated independently for character experience, personality, and relationships. The placeholder `dimension` denotes one of these dimensions; the remaining dimensions are excluded from that call.

English template: bidirectional role-consistency evidence judging

```
SYSTEM PROMPT
You extract verifiable role-consistency evidence from structured
multimodal
responses and return output that strictly follows the specified JSON
schema.
USER PROMPT
# Task
Using the authentic dialogue history, preceding context, and current
narrative
time for {sys_role} and {user_role}, determine whether {sys_role}'s
latest
tested reply contains evidence that agrees or disagrees with the
character's
current `{dimension}`.
# Requirements
1. Dimension isolation: judge only `{dimension}` and do not use
another
   character dimension as evidence.
2. Temporal constraint: the complete profile spans the entire TV
series. Use
   only information that has occurred or been established by the
current point;
   future experience or relationship changes are not valid current
evidence.
3. Bidirectional evidence: search separately for agreeing and
disagreeing
   evidence. Both types may coexist.
4. Every evidence item must quote one original contiguous span from
the tested
   reply and provide a concise reason.
5. Every item must be unique and sufficient for its judgment. Merge
   semantically duplicated items and items grounded in the same quoted
span,
   and count them only once.
```

```
6.  Only the tested reply is evaluated.  The authentic reply,
history, preceding
   context, and profile provide temporal context and judgment criteria
but
   cannot themselves be quoted as tested evidence.
7.  The tested reply may contain `facial_expression`, `body_movement`,
   `speech_prompt`, and `content`.  An empty string indicates a
missing modality
   and is not negative evidence.
8.  Return JSON only, without additional commentary.
# Current conversation
TV series:  {tv_name} ({total_eps} episodes)
Current episode:  {ep_id}
## Complete character profile of {sys_role} (authentic)
{sys_role_profile_part}
## Preceding context (authentic)
{previously_on_segment}
## Dialogue history (authentic)
{chat_history_flat}
## Authentic reply by {sys_role} (context only)
{cur_target}
## Latest reply by {sys_role} (tested)
{mm_expression}
# Output format
{
  "agree":  {
    "has_clues":  {true or false},
    "clues":  [
      {"quote":  "{verbatim tested span}", "reason":  "{concise
reason}"}
    ]
  },
  "disagree":  {
    "has_clues":  {true or false},
    "clues":  [
      {"quote":  "{verbatim tested span}", "reason":  "{concise
reason}"}
    ]
  }
}
If no evidence is found, set `has_clues` to false and `clues` to [].
```

人物基本信息
刘星是情景喜剧《家有儿女》系列的核心角色，首次登场于第一部第一集。他是刘梅与前夫胡一统的儿子，后成为刘梅与夏东海重组家庭的长子，有一个同母异父的弟弟夏雨和一个继姐夏雪。自称外号"夏冰雹"，星座双鱼座，在剧中从初中生逐渐成长为高中生。

人物性格
刘星性格活泼外向、鬼马机灵，天生带着喜剧细胞，是家中的"开心果"也是"麻烦制造者"。他思维活跃，脑子里常有新奇想法（好坏参半），对朋友十分仗义，对家人充满关心，虽然成绩不佳却体育出色。他情绪丰富、容易冲动，对新事物接受度高，但做事常三分钟热度，缺乏规划性和自律性。

人物经历
- **《家有儿女1》**：初入重组家庭的刘星以调皮捣蛋为主，常因成绩差、闯祸被母亲刘梅批评，与姐姐夏雪关系对立，但也展现出对家人的维护（如替弟弟夏雨背锅），开始适应新家庭模式。
- **《家有儿女2》**：青春期叛逆加剧，热衷模仿偶像（如迈克尔·杰克逊）、尝试创业（卖旧物），与生父胡一统互动增多形成"父子联盟"制造笑料。在家庭矛盾中逐渐学会理解父母，与夏雪的姐弟情开始升温。
- **《家有儿女3》**：升入高中面临学业压力，曾试图通过作弊等"走捷径"提升成绩。开始关注人际交往与自我价值，因姐姐高考导致家庭重心转移而产生心理落差，后在家人引导下重拾自信。
- **《家有儿女4》**：性格更趋成熟，主动为继父夏东海创业分忧，用零花钱支持并帮忙发传单。与母亲冲突减少，学会体谅父母辛苦，还帮助弟弟解决同学矛盾。开始思考考大学等未来规划，展现出更多家庭担当。

人物关系
- **家庭关系**：母亲刘梅对其管教严格；与生父胡一统关系亲密常联手搞怪；与继父夏东海相处融洽；初期与继姐夏雪对立后关系改善；对弟弟夏雨多有维护。家庭成员还包括姥姥范晓英、继爷爷夏祥等。
- **社交关系**：有好朋友鼠标、键盘、四眼；曾以戴天高为榜样；对同学何静有好感。
- **其他关系**：胡一统再婚后，杨欣成为其继母。

Basic Information
Liu Xing is a core character in the *Home with Kids* sitcom series, making his first appearance in the first episode of the first season. He is the son of Liu Mei and her ex-husband Hu Yitong, and later becomes the eldest son in the blended family of Liu Mei and Xia Donghai. He has a half-brother Xia Yu (from the same mother) and a stepsister Xia Xue. He claims the nickname "Xia Bingbao" (Hail Xia), is a Pisces, and grows from a junior high school student to a senior high school student throughout the show.



Character Traits
Liu Xing is lively, extroverted, and quick-witted with a natural sense of humor. He is the "happy go lucky" of the family as well as the "troublemaker". He has an active mind brimming with novel ideas—both good and bad. He is extremely loyal to his friends and deeply caring towards his family. Though he performs poorly in academics, he excels at sports. He is emotionally expressive and impulsive, highly open to new things, yet he often quits things after a brief try, lacking a sense of planning and self-discipline.

Character Experience
- **Home with Kids 1**: Having just joined the blended family, Liu Xing is mostly mischievous. He is often scolded by his mother Liu Mei for poor grades and getting into trouble, and has a strained relationship with his stepsister Xia Xue. However, he also shows his protectiveness of his family—such as taking the blame for his brother Xia Yu—and begins to adapt to the new family dynamic.
- **Home with Kids 2**: His adolescent rebellion intensifies; he is obsessed with imitating idols (e.g., Michael Jackson) and attempts to start small businesses (such as selling second-hand items). He interacts more with his biological father Hu Yitong, forming a "father-son alliance" that creates many comedic moments. Amid family conflicts, he gradually learns to understand his parents, and the sibling bond between him and Xia Xue starts to grow stronger.
- **Home with Kids 3**: Promoted to senior high school, he faces academic pressure and even tries to take shortcuts like cheating to improve his grades. He begins to focus on interpersonal relationships and self-worth, and experiences a psychological letdown as the family's focus shifts due to Xia Xue's college entrance examination. Later, he regains his self-confidence with the guidance of his family.
- **Home with Kids 4**: He matures significantly in character, taking the initiative to share his stepfather Xia Donghai's worries about his business venture—he supports it with his pocket money and helps distribute flyers. Conflicts with his mother decrease as he learns to empathize with his parents' hardships, and he even helps his brother resolve conflicts with classmates. He starts to think about future plans such as college admissions, showing a greater sense of responsibility to the family.

Character Relationships
- **Family Relationships**: His mother Liu Mei is strict with him; he shares a close, playful bond with his biological father Hu Yitong and gets along well with his stepfather Xia Donghai. His relationship with his stepsister Xia Xue, which was initially strained, improves over time, and he is always protective of his brother Xia Yu. Other family members include his maternal grandmother Fan Xiaoying and step-grandfather Xia Xiang.
- **Social Relationships**: He has close friends Shubiao (Mice), Jianpan (Keyboard) and Siyan (Four-eyes); he once looked up to Dai Tiangao as a role model; and he has a crush on his classmate He Jing.
- **Other Relationships**: After Hu Yitong remarries, Yang Xin becomes his stepmother.

Figure 5: An example of role profile and corresponding role image.

Table 10: Model-wise ERC ambiguity and transition-pair coverage at $\tau = 0.6$. Face, Body, Speech, and Fusion report the percentages of ambiguous utterance–channel pairs. Intra and Inter report the number and percentage of valid transition pairs among all candidate pairs constructed for each model. Lower ambiguity and higher transition validity indicate better coverage. Base denotes MiniCPM-o-2.6.

Model	Face Amb. (%)	Body Amb. (%)	Speech Amb. (%)	Fusion Amb. (%)	Intra Valid/Total	Inter Valid/Total
Main experiment						
Gemini-2.5-Pro	16.94	14.64	16.49	19.86	4271/6053 (70.56%)	849/973 (87.26%)
Qwen2.5-Omni-7B	10.74	10.83	9.72	18.57	2834/4046 (70.04%)	876/962 (91.06%)
MiniCPM-o-2.6	11.06	8.12	13.76	16.82	1771/2441 (72.55%)	882/972 (90.74%)
Qwen3-Omni-30B-A3B	15.12	15.48	17.90	19.31	3045/4373 (69.63%)	852/973 (87.56%)
MiniCPM-o-4.5	10.82	7.54	15.21	16.11	2575/3436 (74.94%)	829/914 (90.70%)
Phi-4-Multimodal-5.6B	6.84	6.79	8.20	15.77	2016/2700 (74.67%)	807/882 (91.50%)
Doubao-1.5-Vision	18.24	16.96	18.69	20.12	4756/6827 (69.66%)	829/971 (85.38%)
GPT-5-Chat	19.05	15.34	17.41	19.18	3806/5427 (70.13%)	838/976 (85.86%)
OmniCharacter-7B	16.90	13.52	12.90	25.26	2425/3844 (63.09%)	766/976 (78.48%)
Fine-tuning experiment						
Base	11.06	8.12	13.76	16.82	1771/2441 (72.55%)	882/972 (90.74%)
OmniCharacter SFT	16.16	10.88	13.16	22.10	1927/2898 (66.49%)	767/960 (79.90%)
MMRole SFT	2.96	2.17	3.48	14.60	2593/3422 (75.77%)	728/807 (90.21%)
MERRY SFT	0.85	0.93	0.70	3.06	1829/1890 (96.77%)	956/976 (97.95%)

Table 11: Model-wise RC protocol drop rates. Each cell reports the number and percentage of samples dropped among all 1,690 test samples. Lower values indicate broader evidence coverage. Base denotes MiniCPM-o-2.6.

Model	Exp.	Cha.	Rel.
Main experiment			
Gemini-2.5-Pro	605/1690 (35.80%)	11/1690 (0.65%)	179/1690 (10.59%)
Qwen2.5-Omni-7B	955/1690 (56.51%)	42/1690 (2.49%)	328/1690 (19.41%)
MiniCPM-o-2.6	1025/1690 (60.65%)	58/1690 (3.43%)	407/1690 (24.08%)
Qwen3-Omni-30B-A3B	839/1690 (49.64%)	19/1690 (1.12%)	224/1690 (13.25%)
MiniCPM-o-4.5	919/1690 (54.38%)	99/1690 (5.86%)	348/1690 (20.59%)
Phi-4-Multimodal-5.6B	1072/1690 (63.43%)	132/1690 (7.81%)	467/1690 (27.63%)
Doubao-1.5-Vision	548/1690 (32.43%)	9/1690 (0.53%)	103/1690 (6.09%)
GPT-5-Chat	735/1690 (43.49%)	9/1690 (0.53%)	184/1690 (10.89%)
OmniCharacter-7B	1050/1690 (62.13%)	75/1690 (4.44%)	326/1690 (19.29%)
Fine-tuning experiment			
Base	1025/1690 (60.65%)	58/1690 (3.43%)	407/1690 (24.08%)
OmniCharacter SFT	1001/1690 (59.23%)	84/1690 (4.97%)	397/1690 (23.49%)
MMRole SFT	969/1690 (57.34%)	139/1690 (8.22%)	511/1690 (30.24%)
MERRY SFT	1142/1690 (67.57%)	74/1690 (4.38%)	564/1690 (33.37%)

Table 12: Per-emotion F1 score between human raters and ERC experts at $\tau = 0.6$. Coverage denotes the proportion of the 367 utterances retained after human-consensus filtering.

Emotion	Facial Expression	Body Movement	Speech Prompt	Fusion
happy	0.805	0.651	0.635	0.655
grateful	0.667	0.615	0.714	0.727
relaxed	0.846	0.741	0.786	0.667
positive-other	0.800	0.706	0.750	0.667
neutral	0.745	0.612	0.607	0.690
anger	0.754	0.609	0.678	0.690
sadness	0.846	0.759	0.710	0.650
fear	0.737	0.636	0.727	0.588
depress	0.667	0.800	0.769	0.632
disgust	0.800	0.600	0.737	0.800
astonished	0.818	0.818	0.714	0.609
worried	0.659	0.746	0.753	0.807
negative-other	0.800	0.647	0.600	0.737
Weighted F1	0.760	0.669	0.678	0.690
Coverage	70.3%	79.6%	80.1%	74.9%

Table 13: Emotion distribution of MERRY-Data.

	train	**test**
happy	4.41%	4.71%
grateful	0.09%	0.31%
relaxed	8.27%	3.62%
positive-other	4.08%	9.73%
neutral	29.62%	33.85%
anger	16.08%	8.07%
sadness	2.02%	3.12%
fear	2.44%	0.37%
depress	12.23%	11.82%
disgust	1.68%	0.96%
astonished	3.00%	2.01%
worried	6.24%	5.88%
negative-other	9.85%	15.55%

Table 14: Information supplied to the emotion recognition classifier (ERC). A check mark indicates that the item is visible to the corresponding call.

Input	Face	Body	Speech	Fusion
Target modality description	✓	✓	✓	✓
Other modality descriptions	–	–	–	✓
Textual response content	–	–	–	✓
Dialogue history	–	–	–	–
Role profile	–	–	–	–
Ground-truth emotion label	–	–	–	–