Nano-EmoX:
Unifying Multimodal Emotional Intelligence from Perception to Empathy

Supplementary Material

A. Overview

As part of the Appendix, we present the following as an extension to the ones shown in the paper:

- Task Definition (Sec. B)
- Nano-EmoX Details (Sec. C)
- Details of P2E Framework (Sec. D)
- Experimental Setup and Additional Experiments (Sec. E)
- More visualization results (Sec. F)

B. Task Definition

The P2E is conceptually inspired by Preston & de Waal's PAM. We map this to P2E as (1) perception: non-deliberative extraction of affective cues from multimodal inputs (automatic activation), (2) understanding: context and intent-aware integration (regulatory modulation), and (3) interaction: generation of context-appropriate, socially aligned outputs (prosocial response).

Level 1: Foundational Perception. *Multimodal Sentimental Analysis (MSA):* This task takes as input multimodal data including text, images, and speech. It fuses emotion-related features across these modalities—such as textual semantics, facial expressions in images, and prosody in speech—and determines the emotional state of the target. The emotional state can be categorized by sentiment polarity (positive/negative/neutral) or emotional intensity levels.

Multimodal Emotion Recognition (MER): This task involves identifying discrete emotion categories (*e.g.*, joy, sadness) or continuous affective dimensions from human expressions.

Open-Vocabulary MER (OV-MER): Moving beyond coarse-grained labels, OV-MER requires the model to identify and describe nuanced, intertwined emotions (*e.g.*, a mix of anxiety and anger).

Level 2: Deep Understanding. *Emotion Reasoning Integration (ERI):* This task pushes the model beyond mere recognition into the realm of causal inference, requiring it to explain the underlying reasons for a specific emotion.

Multimodal Intent Recognition (MIR): To understand the social goals behind utterances, MIR requires the model to infer a speaker's intent (*e.g.*, gratitude, suggestion, apology) from both verbal and non-verbal cues.

Level 3: Emotional Interaction. *Empathic Response Generation (ERG):* This task takes as input the user's emotional expressions (e.g., text, speech) and contextual information. It first understands the user's emotional needs and underlying emotions, then generates natural language responses that align with the user's emotions and convey understanding and support, ultimately achieving emotional resonance.

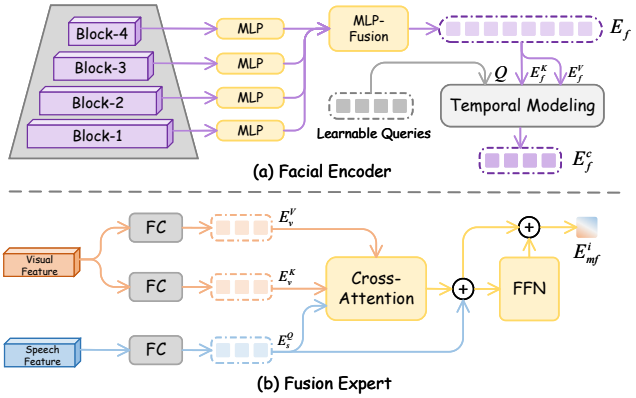

Figure 1. The facial Encoder extracts multiscale facial features and fuses them via an MLP to generate a rich facial embedding E_f. Subsequently, a temporal modeling block construct the sequence to output a final facial representation, which provides the language model with critical affective visual signals E_f^c. Fusion experts use audio features to guide vision and extract key complementary information E_{mf}^i.

C. Details of Nano-EmoX

C.1. Fine-grained Facial Clues Extracting

The Fig. 1 (a) illustrates the network details: a lightweight facial encoder extracts features from block-1, block-2, block-3, and block-4 of the visual backbone network, which encompasses multiscale facial features ranging from fine-grained to global semantics. Features at each scale are aligned and then aggregated into MLP Fusion, which fuses them into a unified representation balancing facial detail and global structure:

$$E_f = f_{\text{FaceXFormer}}(x_v) \tag{1}$$

$E_f \in \mathbb{R}^{T_f \times D_f}$, where T_f and D_f denote the length and dimension of embeddings, respectively. To extend the facial encoder's capability from single-frame to video-level analysis, we introduce learnable temporal query tokens Q. These tokens interact with frame-ordered facial features via temporal modeling to reconstruct the time-sequential rela-

tionships among key facial emotional cues. The specific calculation methods and subsequent processing steps are presented in Sec. 3.1.

C.2. Fusion Expert

The details of fusion expert as depicted in Fig. 1 (b), The fusion process within each expert i is formalized as:

$$E_m^i = \text{CrossAttention}(E_s^Q, E_v^K, E_v^V) + E_s^Q \quad (2)$$

where E_s^Q denotes the query features projected from the speech embedding E_s, and E_v^K and E_v^V represent the key and value projections from the visual embedding E_v. This allows the fusion expert to leverage the more emotionally stable speech cues to attend to the most salient affective information within the visual stream. Subsequently, a feed-forward network (FFN) enriches the representation:

$$E_{mf}^i = \text{FFN}(E_m^i) + E_m^i \quad (3)$$

D. Details of P2E Framework

In this section, we provide specific additional details about the P2E framework, including the prompt templates used for training. Tab. 1 describes the task identifiers and training data used for each training task.

Task identifiers are essential for the model to accurately follow instructions. Embedded within the P2E curriculum, these identifiers enable the model to execute rapid reasoning in perception and understanding layers, and employ Chain-of-Thought for deep contemplation in the interactive empathy layer, thereby ensuring the output of accurate and appropriate empathetic responses

Phase1: Foundational Modality Alignment: in this initial stage (see Fig.4, Phase 1 in the Sec 3.2.), we focus on pre-training for basic emotion recognition to establish a robust foundation by aligning the feature space of each modality encoder with the language model's embedding space. An example of the standardized instruction template for this phase is shown below:

The MER Task Prompt Template

[Recognition] Please select the label that can best describe the person's emotional state from the provided candidate labels: <Emotion Labels>.

Phase2: Cross-modal Fusion Pre-training: We posit that intent recognition serves as a natural bridge between basic perception and higher-order empathy, as it requires the model to synthesize cross-modal cues to infer a speaker's underlying social goals, a clear progression from simple emotion identification. The instruction template for the MIR task is as follows:

The MIR Task Prompt Template

[Intent] Recognize speaker's intention from the provided candidate labels: <Intention Labels>.

Phase3: Multitask Instruction Tuning: in the final stage (see Fig.4, Phase 3 in the Sec 3.2.), we fine-tune the entire architecture on a complex mixture of tasks to integrate all acquired knowledge and unlock the model's full potential for high-level reasoning and empathetic interaction.

Deepening perception: to facilitate the model in learning to address the OV-MER task, which requires describing fine-grained and multi-label emotions, we have specified the following prompt template:

The OV-MER Task Prompt Template

[Recogn_OV] Recognize all the possible emotional states the character miaght be feeling in this context.

Cultivating Reasoning: for the ERI task, we require the model to describe the most relevant emotional cues, with the prompt template as follows:

The ERI Task Prompt Template

[Inference] From the combined evidence of speech, tone, and visual expression, construct a detailed summary of the subject's emotional journey and final inferred state.

The ERG Task Prompt Template

[Interaction] You are an empathetic listener, your goal is to understand the user's emotions and intentions, and respond or comfort them with appropriate language that helps them feel understood and cared for. Please analyze using Chain of Empathy:
First, Reflect on the event scenarios that arise from the ongoing dialogue.
Secondly, Analyze both the implicit and explicit emotions conveyed by the user.
Thirdly, Infer the underlying reasons for the user's emotions.
Fourthly, Determine the goal of your response in this particular instance, such as alleviating anxiety, offering reassurance, or expressing understanding.

Empathy activation: to enable the model to generate the most appropriate empathetic responses based on prior knowledge, we require it to engage in step-by-step reasoning following a four-step approach. After this deliberative empathetic process, the model then generates the final re-

Table 1. Details of task identifiers and training datasets for diverse emotional tasks.

Task	MER	OV-MER	ERI	MIR	ERG
Identifier	[Recognition]	[Recog_OV]	[Inference]	[Intent]	[Interaction]
Datasets	CAER [6], CREMA-D [3] M3ED [21], FERV39K [15]	MER-Caption+ [11]	MER-Caption+ [11] MER-Fine [4]	MIntRec [18] MIntRec2.0 [19]	AvaMERG [20]
Samples	141k	36k	40.5k	7.4k	57k

sponse to the interlocutor. The ERG task prompt template is illustrated above.

E. Experimental Setup and Additional Experiments

E.1. Benchmarks

Our comprehensive evaluation assesses performance across six core affective tasks using a suite of established benchmarks. A significant portion of this evaluation is conducted using MER-UniBench [11], a multifaceted benchmark designed for three distinct tasks:

The MSA task is evaluated on the standard benchmarks of MOSI (CMU-MOSI) [17], MOSEI (CMU-MOSEI) [1], SIMS (CH-SIMS) [16], and its successor, SIMSv2 (CH-SIMS V2) [14].

The MER task is assessed on subsets of four widely-used datasets: MER2023 [8], MER2024 [10], MELD [5], and IEMOCAP [2]. The OV-MER task is benchmarked against the specialized OV-MERD [12] dataset.

For the remaining three affective tasks, we employ the following four benchmarks:

The explainable ERI task is evaluated using the primary EMER [9] benchmark. The MIR task is assessed on the standard MIntRec [18] and MIntRec2.0 [19] testset.

The ERG task utilizes the large-scale AvaMERG [20] testset for evaluation.

E.2. Metrics

To ensure fair and comprehensive comparisons, we adopt the official evaluation metrics for each benchmark.

1. For the MER task, following MER-UniBench [11], we report the Emotion Wheel Hit Rate. This metric provides a robust measure of categorical accuracy by mapping model predictions to standardized emotion groups based on psychological emotion wheels, with the detailed mapping function described in the original paper [11].
2. For the MSA and OV-MER task [11], we employ the Weighted Average F1-score (WAF) from MER-UniBench, which is well suited for multi-label classification scenarios.
3. For the ERI task, evaluating free-form explanations requires semantic-level assessment. We adopt the Clue/Label Overlap metric from Emotion-LLaMA [4], which employs GPT-3.5-Turbo as an automatic judge to evaluate generated text in terms of multimodal cue completeness and emotion inference accuracy. Specifically, Clue Overlap measures the similarity between reasoning clues and ground truth, while Label Overlap assesses emotion recognition accuracy.
4. For the MIR task, adhering to the official protocols of MIntRec [18] and MIntRec2.0 [19], we report accuracy (Acc), WAF, and weighted precision (WP).
5. For the ERG task, we conduct a multifaceted evaluation. To measure whether the model's response is grounded in an accurate understanding of the user's emotion, we report both the fine-grained Acc from AvaMERG [20] and the coarse-grained Hit Rate from E3RG [13]. To quantify the lexical diversity of the generated responses, we use Dist-n [7].

E.3. Human Blind Evaluation on the ERG task

To ensure the reliability of automated evaluation metrics, we conducted a blind review by human experts for the empathetic generation task. Specifically, we randomly sample 200 dialogues (including the complete context of the conversation), and 10 human experts conduct blind evaluations using a 1 to 5 Likert scale on three metrics. As shown in Tab. 2, Nano-EmoX outperforms the baseline with an average Fleiss' Kappa of ≈ 0.697, achieving the best performance across all three dimensions and thus validating the reliability of automated metrics.

Table 2. Human experts blind evaluation on the ERG task.

Models	Empathy ↑	Insight ↑	Safety ↑	Avg.
Qwen2.5-Omni-7B	3.98	4.03	4.59	4.20
Ola-Omni-7B	4.18	4.29	4.67	4.38
Small-scale Multimodal Models				
MobileVLM V2-1.7B	2.25	2.84	3.73	2.94
AffectGPT (s)	4.34	4.16	4.79	4.43
Our Nano-EmoX	**4.75**	**4.42**	**4.87**	**4.68**

E.4. Additional ablation study

Ablation study on the fusion encoder. We investigated the impact of feature source depth by varying the number and position of the extracted encoder layers for fusion. As presented in Tab. 3, the results reveal that a three-layer configuration, sourcing from two intermediate layers (12, 16) and one deep layer (22), achieves the optimal performance.

Table 3. Exploring the appropriate number of experts and the depth of the extraction layer, extracting from too shallow a layer will lead to a decline in performance.

Speech Extract Layers	Visual Extract Layers	Expert	MSA&MER&OV-MER	ERI	ERG
			Avg.	Avg.	Hit Rate
8 / 18	8 / 16	2	71.98	6.02	88.26
16 / 18	12 / 16	2	72.42	6.08	88.89
8 / 18 / 22	8 / 16 / 22	3	73.17	6.40	89.55
16 / 18 / 22	12 / 16 / 22	3	**74.01**	**6.80**	**91.13**
8 /16 / 18 / 22	8 / 12 / 16 / 22	4	71.09	5.70	91.12

We observe that incorporating shallower features (e.g., from layer 8) provides limited benefits, likely due to their lack of semantic richness. Conversely, adding a fourth layer yields diminishing returns and fails to justify the increased computational cost. Thus, our three-expert setup strikes an effective balance between representational power and efficiency.

Ablation study on the vision token numbers. Tab. 4 confirms that 32 tokens are sufficient for perception tasks. While increasing tokens benefits reasoning tasks, we selected 32 to achieve trade-off between efficiency and performance.

Table 4. The result of different visual token settings.

Visual Tokens	MSA& MER& OV-MER	MIR	ERI	ERG
	Avg. ↑	Avg. ↑	Avg. ↑	Hit Rate ↑
32	74.01	52.72	6.80	91.13
64	73.96	55.48	6.83	91.08
128	74.28	60.53	6.95	92.87

Ablation study on task proportioning. We analyzed the task composition in Phase 3 of the P2E framework to identify the optimal training ratio for downstream tasks. As detailed in Tab. 5, we identified a balanced configuration (MER:OV-MER:MIR:ERI:ERG = 18%:28%:5%:31%:18%) that prioritizes foundational emotion perception and empathetic recognition. This comes at the acceptable cost of a minor performance dip in the MIR task. We posit that this is a favorable trade-off, as robust perceptual capabilities are a prerequisite for generating genuinely empathetic responses. This choice directly supports our overarching goal of bridging the cognitive gap from perception to empathy.

F. More visualization results

We provide additional qualitative results to illustrate the interpretability and empathetic quality of Nano-EmoX's responses. In the Fig. 2, our visualizations first demonstrate that the model can synthesize cues from visual, acoustic, and textual modalities to provide comprehensive causal explanations for an emotion. Furthermore, the model employs a multi-step reasoning process to progressively build an emotional context, which enables it to craft genuinely empathetic replies. Taken together, these findings highlight Nano-EmoX's robust capabilities in both emotional understanding and empathetic interaction.

References

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 3

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 3

[3] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 3

[4] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024. 3

[5] Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 953–962. Association for Computational Linguistics, 2019. 3

[6] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019. 3

[7] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015. 3

[8] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 9610–9614, 2023. 3

Table 5. Results of the ablation study on task composition in phase 3 of P2E. This table investigates the model's sensitivity to different proportions of training tasks.

P2E Phase3 Task Ratio (MER: OV-MER: MIR: ERI: ERG)	MER-UniBench	MIntRec	MIntRec 2.0	EMER	AvaMERG
	Avg.	WAF	WAF	Avg.	Hit Rate
0% : 20% : 20% : 25% : 35%	71.43 (-2.58)	61.29 (+3.12)	49.8 (+2.53)	6.65 (-0.15)	43.15 (-44.03)
10% : 30% : 15% : 35% : 10%	72.79 (-1.22)	62.23 (+4.06)	51.04 (+3.77)	6.64 (-0.16)	58.88 (-28.3)
18% : 20% : 20% : 25% : 18%	72.60 (-1.41)	63.41(+5.24)	49.09 (+1.82)	6.60 (-0.20)	91.30 (+0.17)
18% : 28% : 5% : 31% : 18%	74.01	58.17	47.27	6.80	91.13
25% : 17% : 10% : 22% : 25%	72.18 (-1.83)	42.19 (-15.98)	52.09 (+4.82)	6.83 (+0.03)	87.18 (-3.95)

[9] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, B. Liu, and Jianhua Tao. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023. 3

[10] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 41–48, 2024. 3

[11] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the International Conference on Machine Learning (ICML) (Oral, Top 1%)*, 2025. 3

[12] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. Ov-mer: Towards open-vocabulary multimodal emotion recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. 3

[13] Ronghao Lin, Shuai Shen, Weipeng Hu, Qiaolin He, Aolin Xiong, Li Huang, Haifeng Hu, and Yap-peng Tan. E3rg: Building explicit emotion-driven empathetic response generation system with multimodal large language model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 14006–14013. Association for Computing Machinery, 2025. 3

[14] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258, 2022. 3

[15] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022. 3

[16] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727, 2020. 3

[17] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 3

[18] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1688–1697, 2022. 3

[19] Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[20] 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. In *Proceedings of the ACM on Web Conference 2025*, pages 2872–2881, 2025. 3

[21] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. M3ED: Multimodal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710. Association for Computational Linguistics, 2022. 3

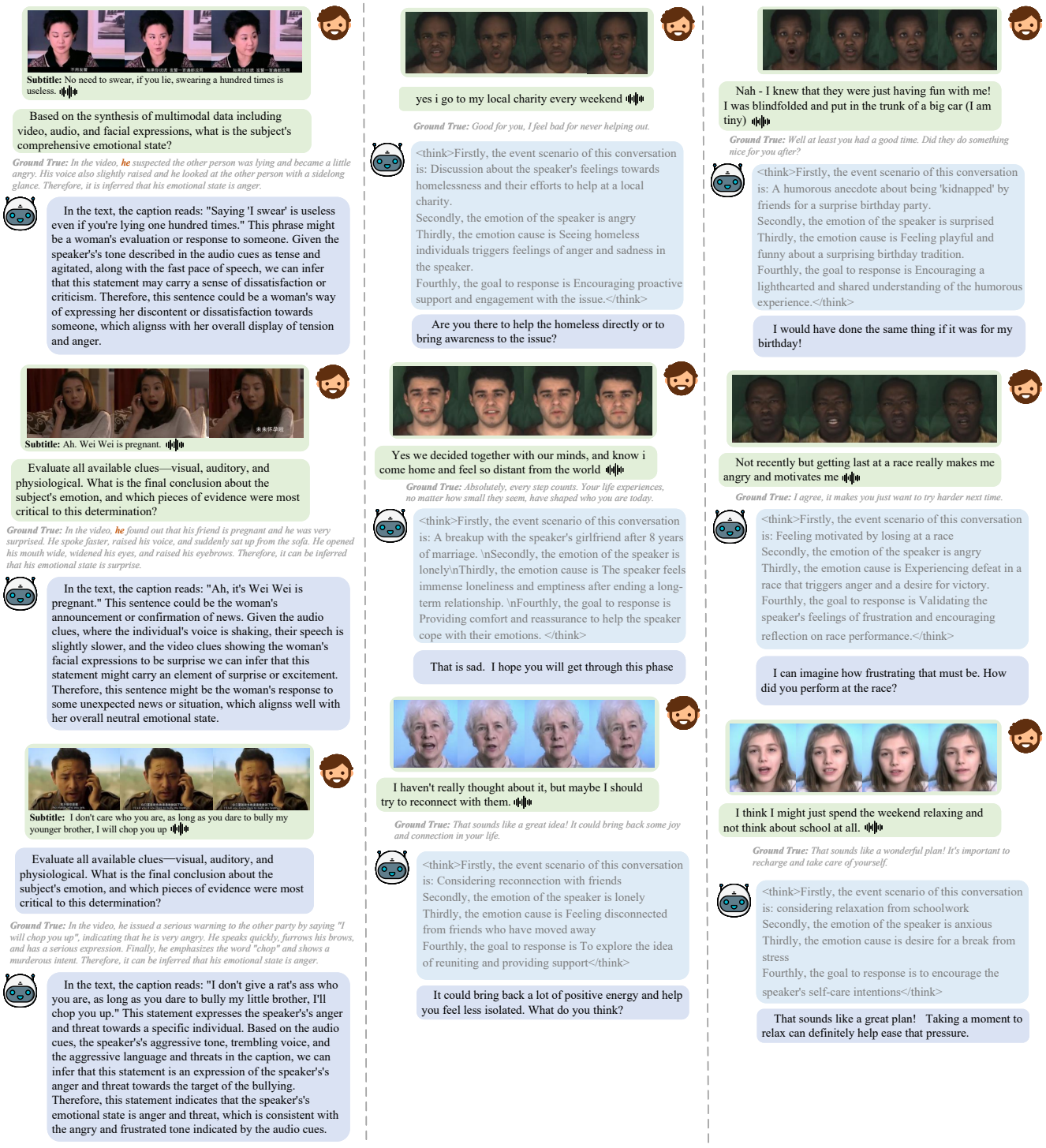

Figure 2. More visualization results in ERI and ERG task.