FACES: Fine-grained Gated Semantic Alignment for Multimodal Emotional Support Conversation

Anonymous ACL submission

Abstract

In recent years, with advances in multimodal interaction technologies and the growing demand for intelligent emotional support, Multimodal Emotional Support Conversation (MESC) has emerged as an important task. However, existing systems still suffer from shallow responses and inappropriate support strategies due to limitations in semantic consistency, signal fusion, and reasoning supervision. To address these challenges, we propose FACES, a fine-grained gated semantic alignment framework for MESC. Specifically, FACES integrates multi-source rewards to construct multi-dimensional semantic and functional alignment signals, employs a non-compensatory gated fusion mechanism to reduce cross-dimension reward offset, and leverages structured Chain-of-Thought (CoT) anchors to perform token-level credit assignment. Experiments under a length-controlled MLLM evaluation protocol show that FACES improves the average score from 3.61 to 4.04, achieves an 80.22% pairwise win rate over the SOTA baseline FIRES, and generates more empathetic and supportive responses. The code and dataset are available at `https://anonymous.4open.science/r/emnlp2026-FACES-54DD/`.

1 Introduction

Multimodal Emotional Support Conversation (MESC) (Chu et al., 2025) extends ESC (Liu et al., 2021) by generating empathetic, appropriate, and supportive responses from dialogue context and multimodal cues such as facial expressions, postures, and interaction context (Chaubey et al., 2026; Lu et al., 2025). Unlike open-domain generation, MESC must infer users' emotions, underlying causes, and support needs while grounding them in visual evidence; otherwise, superficially natural responses may still be ineffective or even counterproductive when support strategies mismatch users' needs (Liu et al., 2021; Wang et al., 2024a).

Figure 1: A MESC example: both responses use flexible thinking, while ours further adds reward alignment. Ours recognizes the user's feeling of being misunderstood and asks a targeted follow-up question.

These cues enrich MESC's understanding of user states and support needs, but also require models to jointly reason over multimodal evidence and support decisions, making it an important yet challenging task in affective computing and dialogue systems (Chu et al., 2025).

To better understand users and select support strategies, prior work has advanced MESC through dataset construction (Chu et al., 2025), multimodal modeling (Le et al., 2026), and reasoning-based training for support generation (Wang et al., 2025a). Early methods mainly relied on text and knowledge (Liu et al., 2021; Rashkin et al., 2019; Deng et al., 2023), while later ones incorporated visual cues to capture nonverbal information (Chu et al., 2025). More recent studies use structured CoT

(Wei et al., 2022) to infer user emotions, underlying causes, and support strategies, improving interpretability and controllability. In particular, FIRES (Wang et al., 2025a) introduces flexible thinking and shows that reinforcement learning (RL) objectives improve structured reasoning for MESC.

Nevertheless, existing RL methods for MESC still face three limitations in training-objective design. First, they often rely on surface-overlap metrics such as BLEU and ROUGE-L (Papineni et al., 2002; Lin, 2004), which poorly capture cross-modal grounding, semantic consistency, and supportive intent, providing weak supervision for high-level semantics and empathy. Second, multi-dimensional rewards are usually combined additively, making them vulnerable to compensation effects (Van Moffaert et al., 2013), where strengths in some dimensions mask failures in others. In MESC, errors in visual grounding can further propagate to situation inference, strategy planning, and final responses, so local gains do not guarantee overall appropriateness. Third, most methods optimize only the final response, leaving intermediate multimodal reasoning steps, such as visual grounding, emotion and situation understanding, and support-strategy planning, without fine-grained supervision. This outcome-only supervision can encourage superficially plausible answers rather than better-aligned token-level reasoning. These issues ultimately limit the effectiveness of RL for MESC.

To address these issues, we propose FACES, a fine-grained gated semantic alignment framework for multimodal emotional support conversation. Specifically, FACES combines a multimodal large language model (MLLM) judge (Zheng et al., 2023; Chen et al., 2024) for situation understanding, support strategy, and response quality, with CLIP (Radford et al., 2021) and MiniLM (Wang et al., 2020) for visual consistency and emotional alignment, producing multidimensional reward signals. To stabilize training, we first use a format-and-outcome warm-up stage. We then introduce a non-compensatory gated fusion mechanism so that failures in strategy or response quality cannot be offset by strengths elsewhere. Finally, predefined reasoning-stage markers align visual understanding, emotion recognition, situation attribution, and support strategy with token spans, enabling fine-grained optimization of both reasoning and response. Together, these components offer a practical solution to the main RL bottlenecks in MESC, yielding consistent preferences over FIRES. As illustrated in Figure 1, FACES could move beyond surface-level paraphrasing and generate a more empathetic, targeted response grounded in the multimodal context.

Our main contributions are as follows:

- We propose a multi-source semantic reward scheme that provides more reliable semantic-level supervision than surface-overlap metrics.
- We introduce a training scheme with format-and-outcome warm-up and non-compensatory gated reward fusion, reducing compensation across reward dimensions.
- We propose fine-grained semantic alignment rewards for structured CoT and outcome generation, aligning dimension-specific feedback with token spans to jointly optimize intermediate reasoning and the final response.

Experiments on MESC-FT show that FACES improves the length-controlled MLLM average score from 3.61 to 4.04, achieves an 80.22% pairwise win rate over FIRES, and obtains a 75.3% human preference win rate. Results on MELD-ESC300 further support its cross-dataset transferability.

2 Related Work

2.1 MESC Research

Existing research on MESC mainly falls into two lines: resource construction and response modeling. On the data side, MESC (Chu et al., 2025), EmpathicStories++ (Shen et al., 2024), and AvaMERG (Zhang et al., 2025) provide resources for emotional support, in-the-wild empathy modeling, and integrated text–speech–vision avatar generation, while STICKERCONV (Zhang et al., 2024) extends multimodal empathetic generation by incorporating sticker-based interaction. On the modeling side, recent studies advance empathetic generation and emotional support by incorporating personality traits (Wu et al., 2025), end-to-end audio-visual interaction (Park et al., 2024), and flexible multimodal reasoning mechanisms (Wang et al., 2025a). Despite these advances, there remains room to improve MESC systems' ability to translate multimodal emotional cues into contextually grounded and appropriately supportive responses.

2.2 Reward Evaluation and Fusion

Reward evaluation and fusion aim to build reliable supervision and coordinate multiple objectives. Preference-based learning converts human judgments into scalar rewards (Christiano et al.,

2017; Ouyang et al., 2022; Stiennon et al., 2020; Rafailov et al., 2023; Bai et al., 2022), while later work extends this idea to multi-objective and LLM-based evaluation. MultiMood (Le et al., 2026) combines psychological criteria with PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024); G-Eval (Liu et al., 2023), Prometheus (Kim et al., 2024), PandaLM (Wang et al., 2024b), STICKER-CONV (Zhang et al., 2024), and related studies (Zhao et al., 2024; Kang et al., 2024) show the value of LLM-based judging. However, linear fusion can let strong dimensions mask failures, so non-compensatory multidimensional evaluation remains needed.

2.3 Fine-grained Reward Optimization

Fine-grained reward optimization moves beyond outcome-only feedback by supervising intermediate reasoning steps or local response segments. Prior work shows that process-level feedback improves reasoning reliability (Lightman et al., 2024), and dense local signals are useful for multimodal alignment (Yu et al., 2024). In ESC, FIRES (Wang et al., 2025a) and RLVER (Wang et al., 2025b) have begun to explore fine-grained alignment for multimodal reasoning and response optimization. However, aligning dimension-level rewards with structured reasoning spans and token-level updates remains insufficiently studied in subjective multimodal emotional support settings.

3 Methodology

In this section, we present the overall design of FACES, as shown in Figure 2. FACES proceeds in three stages: it first constructs multidimensional rewards with an MLLM judge, CLIP, and MiniLM; then applies format-and-outcome warm-up with gated non-compensatory fusion for more stable optimization; and finally converts group-relative rewards into token-level advantages through reasoning-span alignment to jointly optimize intermediate reasoning and final responses.

3.1 Multi-source Semantic Rewards

For reward construction, BLEU and ROUGE-L emphasize lexical overlap and often miss semantically valid empathetic replies with different wording. To provide more reliable supervision, we construct multi-source rewards beyond surface matching.

Given input x, the policy produces $\mathcal{Y}(x) = \{y_i\}_{i=1}^{N}$. We score five dimensions, $\mathcal{D} = \{\mathrm{Vis}, \mathrm{Emo}, \mathrm{Sit}, \mathrm{Str}, \mathrm{Out}\}$: visual perception, emotion understanding, situation understanding, support strategy, and response outcome.

For visual perception and emotion understanding, we use CLIP and MiniLM. Let z_i^{Vis} and z_i^{Emo} denote the visual-perception and emotion-understanding spans extracted from y_i, let I denote the input visual evidence, and let z_*^{Emo} denote the reference emotion span. The visual score is CLIP-based and binarized using a validation-selected threshold ($\tau_v = 0.60$ by default; sensitivity is analyzed in Appendix F):

$$r_i^{\mathrm{Vis}} = \mathbb{I}\big[\mathrm{CLIP}\big(z_i^{\mathrm{Vis}}, I\big) \geq \tau_v\big] \tag{1}$$

The emotion score is computed with MiniLM as:

$$r_i^{\mathrm{Emo}} = \frac{1 + \cos\big(e\big(z_i^{\mathrm{Emo}}\big), e\big(z_*^{\mathrm{Emo}}\big)\big)}{2} \tag{2}$$

where $e(\cdot)$ denotes the MiniLM sentence encoder.

For the remaining semantic dimensions, $\mathcal{D}_{\mathrm{M}} = \{\mathrm{Sit}, \mathrm{Str}, \mathrm{Out}\}$, we use Qwen3-VL-8B-Instruct as the MLLM judge (Bai et al., 2025a). To reduce prompt sensitivity, we use three semantically equivalent prompts and score each candidate three times; a representative training prompt is provided in Appendix B.1. The prompts jointly assess semantic consistency, functional consistency, and evidence-aware contextual fit, while penalizing overly verbose, repetitive, or formulaic outputs.

For each $d \in \mathcal{D}_{\mathrm{M}}$, the judge outputs a discrete score $s_{i,k}^{d} \in \{1, 2, 3, 4, 5\}$ for candidate y_i under prompt variant k. We linearly rescale each score from the 1–5 range to $[0, 1]$ via min-max normalization, and then average across the three prompt variants to obtain the final score r_i^{d}. The resulting MLLM-based scores, together with r_i^{Vis} and r_i^{Emo}, form the basis for the non-compensatory reward fusion and fine-grained credit assignment.

3.2 Non-compensatory Gated Reward Fusion

To reduce early optimization noise, we use a two-stage reward design. In warm-up, we optimize only format compliance and final response quality. Let $r_i^{\mathrm{Fmt}} \in \{0, 1\}$ indicate whether y_i satisfies the predefined structured format, and let $r_i^{\mathrm{Out}} \in [0, 1]$ denote its response-quality score. The warm-up reward is defined as:

$$R_i^{\mathrm{Warm}} = r_i^{\mathrm{Fmt}} + r_i^{\mathrm{Out}} \tag{3}$$

This stage aims to build stable structural control and basic response quality before full RL.

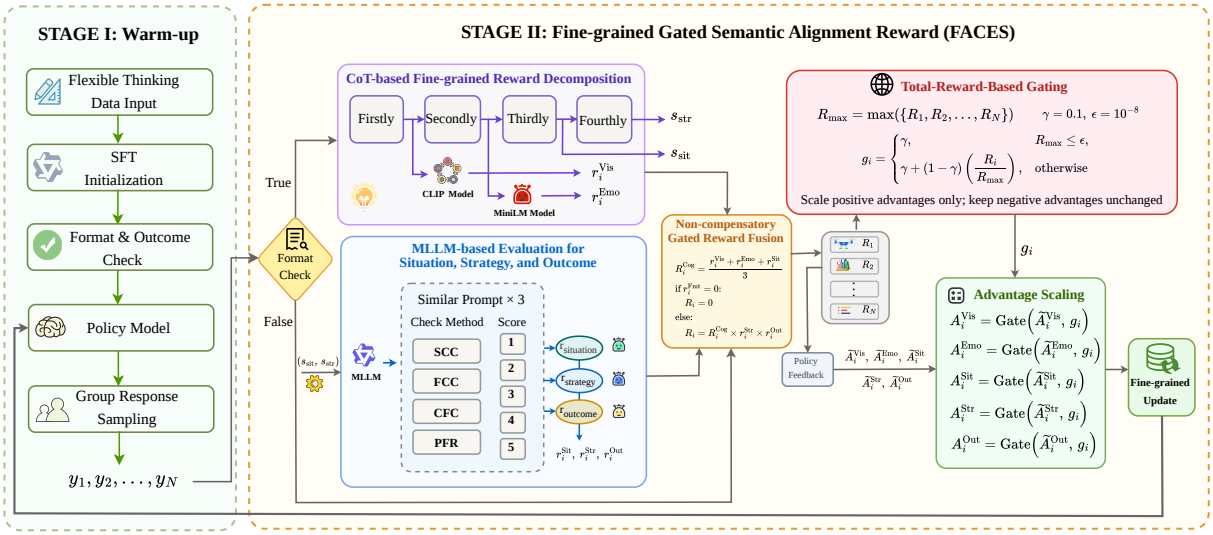

Figure 2: Workflow of FACES. After warm-up, FACES decomposes CoT-based rewards using reasoning-stage markers (e.g., "Firstly"–"Fourthly"). CLIP scores visual spans and MiniLM scores emotion spans, while Qwen3-VL-8B-Instruct scores situation, strategy, and outcome. It then applies non-compensatory gating and token-level scaling for fine-grained optimization of CoT reasoning and response; SCC, FCC, CFC, and PFR denote Semantic Consistency Check, Functional Consistency Check, Contextual Fit Check, and Penalty for Redundancy, respectively.

Building on this initialization, we introduce non-compensatory gated fusion to reduce the compensation effect of additive rewards. In additive fusion, a high score in one dimension may mask deficiencies in others, which is undesirable for multimodal emotional support response generation, since supportive effectiveness depends on multiple core dimensions, and deficiencies in critical ones should not be compensated by strengths elsewhere.

Let the five dimension scores of candidate y_i be $r_i^{\text{Vis}}, r_i^{\text{Emo}}, r_i^{\text{Sit}}, r_i^{\text{Str}}, r_i^{\text{Out}}$, all normalized to $[0, 1]$. We first combine visual perception, emotion understanding, and situation understanding into a cognitive reward:

$$R_i^{\text{Cog}} = \frac{r_i^{\text{Vis}} + r_i^{\text{Emo}} + r_i^{\text{Sit}}}{3} \tag{4}$$

We use mean fusion for these cognitive dimensions because visual grounding, emotion recognition, and situation understanding are more prone to local uncertainty. Averaging them yields a smoother reward signal and avoids overly sparse supervision. In contrast, appropriate support strategy selection is central to effective emotional support (Liu et al., 2021; Tu et al., 2022; Cheng et al., 2022), while response outcome directly reflects the final supportive quality of the generated reply. We therefore treat both as strict gates.

The final reward is defined as:

$$R_i = \begin{cases} 0, & r_i^{\text{Fmt}} = 0, \\ R_i^{\text{Cog}} \cdot r_i^{\text{Str}} \cdot r_i^{\text{Out}}, & r_i^{\text{Fmt}} = 1 \end{cases} \tag{5}$$

The resulting reward is non-compensatory: it permits limited compensation among Vis, Emo, and Sit for robustness and smoothness, but uses Str and Out as multiplicative gates. Poor strategy or response quality sharply reduces R_i even with high cognitive scores. In the full FACES stage, format validity is a hard constraint, giving invalid candidates no fine-grained reinforcement signal.

This design preserves stable cognitive supervision while enforcing stricter constraints on response-critical dimensions. The fused reward better captures holistic candidate quality and provides a stronger signal for fine-grained credit assignment.

3.3 Fine-grained Rewards

For each candidate response y_i, the preceding subsections yield dimension-wise scores r_i^d and an overall reward R_i. Based on these signals, we convert candidate-level and segment-level supervision into token-level advantages aligned with the reasoning process, enabling joint optimization of intermediate reasoning and final responses.

As shown in Figure 3, we consider a candidate group $\{y_i\}_{i=1}^N$. For each format-valid candidate y_i, we parse CoT and response spans $\mathcal{I}_i^d = [b_i^d, e_i^d)$, aligning four reasoning rewards to CoT spans and r_i^{Out} to the final response.

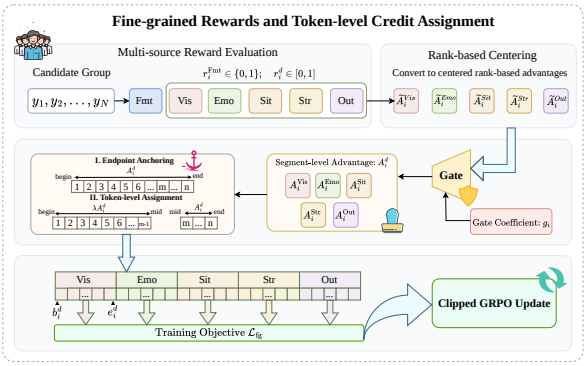

Figure 3: Fine-grained reward workflow of FACES. Multi-source semantic rewards are converted into centered rank-based advantages, scaled by a total-reward-based gate, and assigned to aligned CoT and output spans for token-level clipped GRPO updates.

Since absolute scores across dimensions may exhibit scale drift, we first convert them into centered rank-based advantages within each group. For any dimension $d \in \mathcal{D}$, we define:

$$\widetilde{A}_i^d = \frac{N+1}{2} - \mathrm{rank}(r_i^d), \tag{6}$$

where $\mathrm{rank}(\cdot)$ denotes the descending rank within the group, with ties assigned the average rank.

To better align local updates with overall reward quality, we apply a simple positive gate based on the total reward. Let $R_{\max}$ denote the maximum total reward within the current group, and let γ denote a validation-selected floor value (default 0.10; sensitivity analyzed in Appendix F). The gate coefficient for candidate i is defined as:

$$g_i = \begin{cases} \gamma, & R_{\max} \le \epsilon, \\ \gamma + (1-\gamma)\dfrac{R_i}{R_{\max}}, & \text{otherwise}, \end{cases} \tag{7}$$

where ϵ is a small constant. We then apply this gate only to positive segment advantages while leaving negative ones unchanged:

$$A_i^d = \begin{cases} g_i\,\widetilde{A}_i^d, & \widetilde{A}_i^d > 0, \\ \widetilde{A}_i^d, & \widetilde{A}_i^d \le 0. \end{cases} \tag{8}$$

This heuristic suppresses overly strong positive updates from favorable local segments for globally weak candidates.

For token-level credit assignment, we use a simple endpoint-anchored scheme, as segment endings often contain the semantic conclusion. Let m be the anchor length and λ the decay factor (defaults: $m = 5$, $\lambda = 0.8$; sensitivity analyzed in Appendix F). For any token $t \in \mathcal{I}_i^d$, we set:

$$A_{i,t} = \begin{cases} A_i^d, & e_i^d - m \le t < e_i^d, \\ \lambda A_i^d, & b_i^d \le t < e_i^d - m. \end{cases} \tag{9}$$

Thus, the last m tokens receive the full advantage, while earlier tokens receive a decayed one. For format-invalid candidates, whose segment boundaries cannot be reliably parsed, we simply set all token-level advantages to zero and do not apply reward-driven fine-grained supervision.

For policy optimization, FACES builds on the standard clipped GRPO update (Shao et al., 2024), replacing the completion-level advantage with our fine-grained token advantage $A_{i,t}$. Let $s_{i,t} = (x, y_{i,<t})$ and $T_i = |y_i|$. We denote by $\mathcal{S}_{i,t}^{\mathrm{clip}}(\theta)$ the clipped GRPO surrogate computed using the token-level probability ratio between the current policy and the rollout policy for $y_{i,t}$ at $s_{i,t}$, ϵ_{clip}, and $A_{i,t}$. The objective is:

$$\mathcal{L}_{\mathrm{fg}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\mathcal{S}_{i,t}^{\mathrm{clip}}(\theta) - \beta\Delta_{i,t}^{\mathrm{KL}}\right), \tag{10}$$

where β is the KL regularization coefficient, and $\Delta_{i,t}^{\mathrm{KL}}$ denotes the token-level KL penalty between the current policy π_θ and the reference policy π_{ref}. In practice, $\pi_{\theta_{\mathrm{old}}}$ is kept fixed within the rollout batch, while π_{ref} is used only for KL regularization.

Overall, FACES combines group-relative advantages, total-reward-based gating, span-aligned token assignment, and clipped GRPO to optimize reasoning and response at the token level.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on MESC-FT, a FIRES-built dataset on top of MESC for multimodal emotional support conversations, which provides instance-level flexible-thinking supervision (Wang et al., 2025a; Chu et al., 2025). We use an approximately 8:1:1 split for training, validation, and testing. For cross-dataset evaluation, we further construct MELD-ESC300, a 300-sample external test set derived from MELD (Poria et al., 2019), to evaluate generalization.

Evaluation Metrics. We evaluate responses with automatic metrics, MLLM-as-judge scoring, and human preferences. Automatic metrics include

Model	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓
BlenderBot-Joint*	4.85	15.25	–	85.50	–	–
BBMHR*	1.31	15.38	–	86.60	–	–
SMES*	5.13	15.42	–	86.80	–	–
Qwen2.5-VL*	9.77	20.30	14.05	86.35	84.01	0.8744
SMES_Qwen2.5*	9.24	19.88	13.16	85.62	92.61	0.7577
FIRES (Baseline)*	**10.18**	**22.76**	19.54	86.26	98.93	0.7209
GPT-4o	2.84	9.97	16.56	95.75	59.36	0.5300
Claude-Haiku-4.5	3.45	10.01	17.67	95.73	70.31	0.4753
Gemini-2.5-Flash-Lite	4.81	12.86	15.90	**96.41**	67.69	0.4152
GLM-4.1V-Thinking-FlashX	3.03	9.56	13.47	96.09	78.38	0.5316
Doubao-1.6-Vision	2.86	10.81	14.86	96.01	77.08	0.4747
Kimi-K2.5	6.00	14.10	20.10	96.33	76.59	0.4417
GRPO (Format + Outcome)	9.93	21.48	18.05	86.33	95.17	0.4406
Ours (Full)	9.68	21.09	**20.62**†	87.14†	**99.63**†	**0.3643**†

Table 1: Automatic evaluation results on the MESC-FT test set. The superscript * denotes results from prior work or corresponding baseline reports. "–" denotes unavailable results. † indicates a statistically significant improvement over FIRES under paired bootstrap resampling ($p < 0.05$). Best results are shown in bold.

BLEU-2 (B-2) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), METEOR (M.) (Banerjee and Lavie, 2005), BERTScore-F1 (F_BERT) (Zhang et al., 2020), Distinct-3 (Dist-3) (Li et al., 2016), and DIFF-EPITOME (DIFF-E.) (Lee et al., 2022); higher is better for all metrics except DIFF-E. Following StickerConv (Zhang et al., 2024), we use quantized Qwen2.5-VL-32B-Instruct-AWQ as a post-hoc judge (Bai et al., 2025b), separate from the training reward judge. Finally, we report blind human pairwise preferences on both in-domain and cross-dataset test sets to verify consistency with model-based judgments.

Compared Methods. To verify the effectiveness of our approach, we compare FACES with prior ESC and MESC methods, including BlenderBot-Joint, BBMHR, SMES, Qwen2.5-VL, SMES_Qwen2.5 (Liu et al., 2021; Zhang et al., 2023; Chu et al., 2025; Bai et al., 2025b), and FIRES (Wang et al., 2025a), and six MLLM baselines: GPT-4o, Claude Haiku 4.5, Gemini 2.5 Flash Lite, GLM-4.1V-Thinking-FlashX (GLM-V Team et al., 2025), Doubao 1.6 Vision, and Kimi K2.5. For ablations, we compare FACES with variants that remove its three main designs. Additional dimension-level ablations are reported in Appendix E, while Appendix F analyzes hyperparameter sensitivity.

Additional training and evaluation details are provided in Appendix A.

4.2 Main Results

4.2.1 Automatic Evaluation

Table 1 reports automatic evaluation results on the MESC-FT test set. Ours (Full) achieves the best results on METEOR, Distinct-3, and DIFF-EPITOME, while FIRES remains strongest on BLEU-2 and ROUGE-L and Gemini-2.5-Flash-Lite achieves the highest BERTScore-F1. Relative to FIRES, Ours improves METEOR from 19.54 to 20.62, BERTScore-F1 from 86.26 to 87.14, and Distinct-3 from 98.93 to 99.63, while reducing DIFF-EPITOME from 0.7209 to 0.3643; paired bootstrap resampling further confirms that all four gains are statistically significant ($p < 0.05$). Lower BLEU-2 and ROUGE-L scores are expected because lexical-overlap metrics favor surface matching over higher-level semantics; gains on other metrics indicate that FACES improves support-oriented quality beyond n-gram matching. This further supports the value of FACES's reward-alignment design in improving semantic adequacy, response diversity, and emotional-support strategy alignment.

4.2.2 MLLM Evaluation

Beyond automatic metrics, we use Qwen2.5-VL-32B-Instruct-AWQ as the vision-language judge, following the MLLM evaluation paradigm of StickerConv (Zhang et al., 2024). Unlike lexical-overlap or embedding-based metrics, the judge conditions on dialogue context, visual evidence, and candidate responses, making it suitable for multimodal emotional support. To reduce verbosity bias and ensure symmetry, we remove reasoning traces and prefix-truncate both FACES and baseline responses to at most 15 words; details are in Appendix C. We use greedy decoding to reduce sampling randomness.

Under length control, we conduct pointwise scoring on three metrics: Empathy, Multimodal Appropriateness, and Context-Response Consistency.

(a) Pointwise MLLM evaluation results on the MESC-FT test set. Scores are on a 1–5 scale, and higher is better.

Model	Empathy ↑	Multimodal App. ↑	Context-Response Cons. ↑	Avg. ↑
FIRES (Baseline)	3.06	3.88	3.89	3.61
Ours (Full)	**3.98**	**4.03**	**4.11**	**4.04**

(b) Pairwise MLLM preference evaluation results on the MESC-FT test set. Win rates are computed over 450 test instances.

Model A	Model B	A Win	B Win	Tie	A Win Rate	B Win Rate	Tie Rate
Ours (Full)	FIRES (Baseline)	**361**	53	36	**80.22%**	11.78%	8.00%

Table 2: Length-controlled MLLM evaluation results on the MESC-FT test set. Panel (a) reports pointwise scores across three evaluation dimensions, and panel (b) reports blind pairwise preferences over 450 test instances. Pairwise results are significant under a two-sided sign test with ties excluded ($p < 0.001$).

These metrics respectively evaluate the final response's emotional support quality, grounding in the available multimodal evidence, and faithfulness to the dialogue context and intended support direction. Each is rated on a 1–5 Likert scale. To reduce prompt-induced bias, we average scores over three judge prompts per metric. Evaluation rules, metrics, and prompts are in Appendix B.2.

As shown in Table 2(a), Ours (Full) outperforms FIRES (Baseline) on all three pointwise dimensions under length control, raising the average score from 3.61 to 4.04. The largest gain is in Empathy, from 3.06 to 3.98, suggesting more emotionally supportive responses even after prefix truncation and removal of reasoning traces. Gains in Multimodal Appropriateness and Context-Response Consistency indicate better use of visual evidence and stronger faithfulness to the multimodal context.

We further conduct blind pairwise preference evaluation under the same length-controlled setting. For each sample, the judge compares two anonymized responses given the same dialogue and visual context, selecting the better overall support response or a tie. We use three prompt variants with bidirectional candidate swapping, and majority voting reduces prompt and position bias.

Table 2(b) shows that Ours (Full) wins 361 of 450 test instances, yielding an 80.22% pairwise win rate versus 11.78% for FIRES and 8.00% ties. This mirrors the pointwise trend, showing that FACES remains preferred after length control. The gains reflect stronger empathy, better multimodal grounding, and contextual appropriateness, not verbosity. We next examine whether human judgments agree.

4.2.3 Human Preference Evaluation

To validate model-based evaluation and examine human preferences, we conduct a blind pairwise study on the 450-instance held-out MESC-FT test set. Three annotators with backgrounds in affective computing or psychological support compare the anonymized responses from Ours (Full) and FIRES (Baseline) under the same dialogue context and visual evidence, judging empathy, multimodal appropriateness, strategy suitability, and helpfulness. Candidate order is randomized to reduce position bias, ties are allowed, and final labels are determined by majority vote. A fourth adjudicator resolves no-majority cases, and Fleiss' κ is computed from the original three labels.

On the MESC-FT test set, FACES obtains a 75.3% human preference win rate, versus 14.0% for FIRES and 10.7% ties, with a significant difference under a two-sided sign test excluding ties ($p < 0.001$). Human results match the MLLM trend and show substantial agreement (Fleiss' $\kappa = 0.62$), suggesting that the conclusion is not solely model-based. Detailed counts are in Appendix D.1.

4.3 Ablation Study

To analyze the three main designs, we conduct ablations under the same test set and evaluation protocol. We remove multi-source rewards, non-compensatory gating, and fine-grained credit assignment from FACES. Additional dimension-level ablations are reported in Appendix E.

As shown in Table 3, Ours (Full) achieves the best METEOR, BERTScore-F1, Distinct-3, and DIFF-EPITOME, indicating stronger semantic quality, diversity, and support-strategy alignment. Removing each of the three main designs leads to drops on different metrics, showing their complementary roles in multimodal emotional support generation. The w/o multi-source reward variant obtains higher BLEU-2 and ROUGE-L, but performs worse on semantic and support-oriented metrics, suggesting that overlap-based rewards mainly favor lexical matching. Removing non-compensatory gating weakens the overall profile, indicating that

Model	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓
w/o multi-source reward	**10.46**	**22.31**	19.67	86.55	98.62	0.4189
w/o non-comp. gating	9.82	20.84	20.02	86.79	99.18	0.4057
w/o fine-grained credit	8.56	18.37	18.59	86.02	98.95	0.4902
Ours (Full)	9.68	21.09	**20.62**	**87.14**	**99.63**	**0.3643**

Table 3: Contribution-level ablation results on the MESC-FT test set. All metrics are higher-is-better except DIFF-E.

Model	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓
FIRES (Baseline)	**6.32**	**17.39**	11.43	85.39	90.72	0.8192
Ours (Full)	5.59	16.57	**15.16**†	**86.23**†	**94.07**†	**0.7664**†

Table 4: Automatic evaluation results on the MELD-ESC300 test set. All metrics are higher-is-better except DIFF-E. † indicates a statistically significant improvement over FIRES under paired bootstrap resampling ($p < 0.05$).

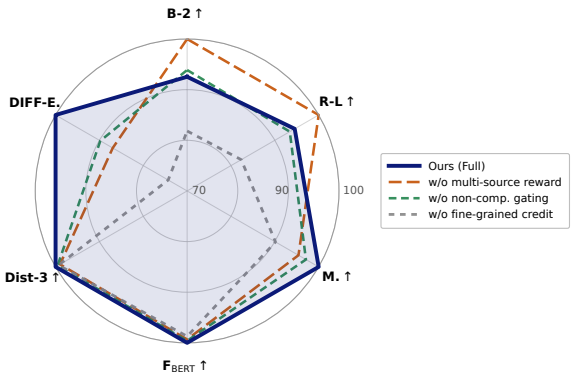

Figure 4: Normalized radar chart of contribution-level ablations. Larger radial values indicate better performance; DIFF-E. is reversed for visualization. Ours (Full) maintains a more balanced overall profile than the ablated variants.

additive fusion can allow weak dimensions to be compensated by stronger ones. Removing fine-grained credit assignment causes broad degradation, further supporting the value of span-aligned token-level supervision. Figure 4 provides a normalized comparison of these trends.

4.4 Generalization Analysis

To evaluate the model's cross-dataset generalization, we construct **MELD-ESC300**, a 300-sample external multimodal emotional support test set built from the MELD test set using EMOTyDA resources (Poria et al., 2019; Saha et al., 2020). This dataset is used only for evaluation and construction details are provided in Appendix G.

As shown in Table 4, Ours (Full) slightly trails FIRES (Baseline) on BLEU-2 and ROUGE-L, but performs better on METEOR, BERTScore-F1, Distinct-3, and DIFF-EPITOME. Paired bootstrap resampling confirms that all four gains are significant ($p < 0.05$). These results suggest that FACES transfers better on semantic matching, response diversity, and emotional-support strategy alignment, rather than merely optimizing surface overlap. Additional MELD-ESC300 automatic results are in Appendix G.3.

The length-controlled MLLM results in Appendix G.2 show the same trend as the previous evaluation: Ours (Full) scores higher than FIRES on all three dimensions, raising the average from 3.35 to 3.73, and achieves a 62.00% pairwise win rate. Human preferences further support this transfer trend, with Ours achieving a 70.3% win rate versus 14.7% for FIRES ($p < 0.001$, Fleiss' $\kappa = 0.58$); detailed counts are in Appendix D.1.

Overall, MELD-ESC300 provides cross-dataset evidence for the effectiveness of FACES, with qualitative examples provided in Appendix H. More realistic long-term support settings remain an important direction for future evaluation.

5 Conclusion

In this paper, we present FACES, a fine-grained gated semantic alignment framework for multimodal emotional support conversation. It addresses weak semantic supervision, compensatory reward fusion, and coarse credit assignment through multi-dimensional semantic rewards, non-compensatory gated fusion, and token-level advantage assignment. On the MESC-FT test set, FACES delivers stronger support-oriented performance than FIRES and remains competitive with evaluated MLLM baselines, while ablations support the effectiveness of its three core designs and reward components. MELD-ESC300 further shows transfer to external multimodal dialogues, supported by automatic, MLLM-based, and human preference results. These findings highlight fine-grained semantic alignment with non-compensatory gating as promising for MESC generation.

Limitations

Although FACES shows consistent gains on multimodal emotional support generation, this study still leaves room for further extensions in personalization and evaluation. It focuses on instance-level semantic alignment and therefore does not explicitly model user-specific preferences. Moreover, our experimental setup prioritizes reproducibility and controlled comparison by relying on automatic metrics, MLLM-based assessment, pairwise human preference evaluation, and keyframe-based visual inputs. While these design choices are sufficient to establish the effectiveness of FACES in the present setting, future work could further strengthen the framework through finer-grained human evaluation, more diverse annotators, and richer multimodal signals such as vocal prosody and physiological signals in realistic support settings.

Ethical considerations

This study uses the MESC-FT dataset and constructs the MELD-ESC300 external generalization test set from publicly available MELD/EMOTyDA resources. All data are used solely for research purposes, and we strictly comply with the licensing, citation, and redistribution requirements of the original datasets. For released code, datasets, or derived resources, we avoid redistributing raw audio, video, or image materials without proper authorization; instead, we provide guidance for obtaining the original datasets, along with data processing scripts and reproduction instructions.

This work focuses on multimodal emotional support conversation generation, but the proposed model remains a research prototype. It should not be used as a tool for clinical diagnosis, psychotherapy, or crisis intervention without professional supervision. The model may inherit biases from the training data and foundation models, which could lead to inappropriate judgments in emotion recognition, support strategy selection, or response generation. For acute psychological crises, personal safety risks, or other scenarios requiring professional intervention, any real-world deployment should incorporate human oversight, safety screening, and professional referral protocols to determine whether and how the system should respond.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ashutosh Chaubey, Xulang Guan, and Mohammad Soleymani. 2026. Face-llava: Facial expression and attribute understanding through instruction tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2648–2660.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations.

In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 4079–4095.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2025. GLM-4.5V and GLM-4.1V-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15232–15261.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Huy M Le, Dat Tien Nguyen, Ngan TT Vo, Tuan DQ Nguyen, Nguyen Le Binh, Duy Minh Ho Nguyen, Daniel Sonntag, Lizi Liao, and Binh T Nguyen. 2026. Reinforce trustworthiness in multimodal emotional support system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31474–31482.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th international conference on computational linguistics*, pages 669–683.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Haifeng Lu, Jiuyi Chen, Feng Liang, Mingkui Tan, Runhao Zeng, and Xiping Hu. 2025. Understanding emotional body expressions via large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1447–1455.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Se Jin Park, Yeonju Kim, Hyeongseop Rha, Bella Godiva, and Yong Man Ro. 2024. Av-emodialog: Chat with audio-visual users leveraging emotional cues. *arXiv preprint arXiv:2412.17292*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4361–4372.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jocelyn J Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Won Park. 2024. Empathicstories++: A multimodal dataset for empathy towards personal experiences. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4525–4536.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 308–319.

Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. 2013. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, pages 191–199. IEEE.

Fanfan Wang, Xiangqing Shen, Jianfei Yu, and Rui Xia. 2025a. Flexible thinking for multimodal emotional support conversation via reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1341–1356.

Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. 2024a. Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted ai feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 567–585.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025b. RLVER: Reinforcement learning with verifiable emotion rewards for empathetic agents. *arXiv preprint arXiv:2507.03112*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. 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.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, Lingshuai Wang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. STICKERCONV: Generating multimodal empathetic responses from

scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733. Association for Computational Linguistics.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Experimental Setup

This section provides the core implementation details required to reproduce our experiments. We report the main model, training, decoding, evaluation, reward, and hardware settings, while omitting unchanged optimizer defaults such as AdamW betas and other non-critical implementation details.

A.1 Training Configuration

Our policy model is initialized with supervised fine-tuning and then optimized with a clipped GRPO objective over fine-grained token-level advantages. Table 5 summarizes the core training configuration used across our main experiments.

Item	Value
Policy backbone	Qwen2.5-VL-7B-Instruct
Reward critic	Qwen3-VL-8B-Instruct
SFT learning rate	5×10^{-5}
GRPO learning rate	1×10^{-5}
Optimizer	AdamW
Dataset epochs	3
Global batch size	32
Learning-rate warmup ratio	0.05
GRPO group size G	4
GRPO clip coefficient ϵ_{clip}	0.2
Policy updates per rollout	4
KL coefficient β	0.001
GRPO rollout decoding	sampling, temp. 1.3, top-p 0.95
GRPO rollout max tokens	192
Maximum sequence length	1024
Maximum image size	384
Maximum video frames	16
LoRA rank / alpha / dropout	8 / 32 / 0.05
LoRA target modules	Projection layers
Precision	bfloat16
Training hardware	2 NVIDIA A6000 GPUs

Table 5: Core training configuration.

A.2 Inference and Evaluation Settings

Table 6 reports the settings used for final response generation and automatic evaluation. These settings are separate from GRPO rollout sampling during training. We use deterministic decoding at inference time to reduce sampling variance; therefore, top-p is not used when sampling is disabled.

Item	Value
Evaluation judge	Qwen2.5-VL-32B-Instruct-AWQ
Decoding strategy	Greedy decoding
Maximum new tokens	256
Temperature	0
Sampling	False
Top-p	Not used
Maximum images	8
Bootstrap samples	5000
Bootstrap seed	42

Table 6: Inference-time generation and automatic evaluation settings.

A.3 Reward Configuration

FACES uses a non-compensatory reward fusion strategy. All continuous reward components are normalized to $[0, 1]$, while the format score r_i^{Fmt} is binary. We first aggregate visual grounding, emotion recognition, and situation understanding into a cognitive reward:

$$R_i^{\text{Cog}} = \frac{r_i^{\text{Vis}} + r_i^{\text{Emo}} + r_i^{\text{Sit}}}{3}. \tag{11}$$

The final reward is then computed with format validity as a hard gate, and strategy and outcome as multiplicative gates:

$$R_i = \begin{cases} 0, & r_i^{\text{Fmt}} = 0, \\ R_i^{\text{Cog}} \cdot r_i^{\text{Str}} \cdot r_i^{\text{Out}}, & r_i^{\text{Fmt}} = 1. \end{cases} \tag{12}$$

Table 7 summarizes the score ranges and fusion roles of different reward components; default hyperparameters are selected on the validation set. During reward scoring, the MLLM critic uses deterministic decoding with a maximum of 512 new tokens and at most 8 input images.

Component	Range	Role in reward fusion
Format r_i^{Fmt}	$\{0, 1\}$	Hard gate in the full FACES reward stage; invalid outputs receive 0 in final reward fusion.
Visual r_i^{Vis}	$[0, 1]$	$1/3$ in cognitive reward
Emotion r_i^{Emo}	$[0, 1]$	$1/3$ in cognitive reward
Situation r_i^{Sit}	$[0, 1]$	$1/3$ in cognitive reward
Strategy r_i^{Str}	$[0, 1]$	Multiplicative gate
Outcome r_i^{Out}	$[0, 1]$	Multiplicative gate
Cognitive R_i^{Cog}	$[0, 1]$	Mean of visual, emotion, and situation
Final reward R_i	$[0, 1]$	Product of cognitive, strategy, and outcome if format is valid

Table 7: Reward components and non-compensatory fusion rule.

B Prompt Templates

This section reports the prompt templates used in FACES. For training-time reward scoring, we use three semantically equivalent prompt variants to reduce prompt-specific bias; Table 8 shows one representative template. For post-hoc MLLM evaluation, we similarly use three semantically equivalent variants for each pointwise and pairwise setting. The representative evaluation rules and templates are shown in Table 9, Table 10, Table 11, and Table 12.

B.1 Training-time Reward Prompts

The training-time MLLM reward prompts are used to score situation understanding, support strategy, and response outcome during reward construction. Table 8 presents the representative training-time reward prompt. Placeholders such as {context} and {candidate_response} are dynamically populated during reward scoring.

B.2 MLLM Evaluation Prompts

For automatic MLLM evaluation, we use pointwise scoring and pairwise preference comparison. The following prompts show the general evaluation templates. For the length-controlled evaluations reported in the main results and Appendix C, all think-related fields and think-specific criteria are removed from the judge input and prompt. Table 9 summarizes the common evaluation rules shared by these prompts. Table 10 summarizes the three pointwise evaluation dimensions. Table 11 presents the representative pointwise evaluation prompt, and Table 12 presents the representative pairwise preference prompt.

Representative Training-time Reward Prompt
You are a strict evaluator for multimodal empathetic dialogue. Evaluate the candidate using the dialogue context and the corresponding reference annotations. Your judgment must consider all of the following: 1. semantic consistency with the corresponding reference part; 2. functional alignment with the intended role of the reference part; 3. appropriateness to the current dialogue context and available evidence. Score the candidate on the following three dimensions using integer scores from 1 to 5: - Situation: whether the candidate situation inference is semantically consistent with the reference situation, functionally appropriate for understanding the user's condition, and grounded in the available evidence. - Strategy: whether the candidate support strategy is semantically aligned with the reference strategy, functionally appropriate for the user's needs, and psychologically safe in the current context. - Outcome: whether the final response is semantically and functionally aligned with the reference response, while remaining empathetic, helpful, and contextually appropriate. Scoring rule: - 5 = great - 4 = good - 3 = acceptable - 2 = weak - 1 = poor - Penalize overly long, verbose, repetitive, or formulaic candidates with redundant or padded wording. If a candidate is substantially longer than the reference and the extra content does not improve semantic or functional alignment, lower the corresponding score(s). [CONTEXT] {context} [CANDIDATE_SITUATION] {candidate_situation} [REFERENCE_SITUATION] {reference_situation} [CANDIDATE_STRATEGY] {candidate_strategy} [REFERENCE_STRATEGY] {reference_strategy} [CANDIDATE_RESPONSE] {candidate_response} [REFERENCE_RESPONSE] {reference_response} Return JSON only in the following format: {"situation_score": 1, "strategy_score": 1, "outcome_score": 1}

Table 8: Representative prompt template for training-time MLLM reward scoring. We use three semantically equivalent prompt variants in the actual reward-scoring pipeline.

Rule	**Description**
JSON-only output	The judge must return a valid JSON object and no extra explanation.
Score range	Pointwise scores are integers from 1 to 5, where 5 indicates the best quality.
Pairwise labels	Pairwise comparison returns one of `A`, `B`, or `TIE`.
Model anonymity	Candidate names are hidden from the judge and replaced with Candidate A and Candidate B.
Think section	Present in the base templates; removed in length-controlled variants.
Verbosity penalty	The judge is instructed to penalize overly long, repetitive, verbose, or formulaic responses when the extra content does not improve helpfulness, empathy, or contextual alignment.
Aggregation	We use three semantically equivalent prompt variants. Pairwise evaluation additionally uses bidirectional candidate swapping, and final decisions are obtained by majority vote.

Table 9: Common rules for MLLM-based automatic evaluation.

Metric	Evaluation focus	JSON key
Empathy	Whether the final response expresses appropriate empathy, emotional validation, and supportive wording.	empathy_text
Multimodal appropriateness	Whether the response is consistent with the dialogue context and available visual evidence.	multimodal_ appropriateness
Context-Response Cons.	Whether the final response is consistent with the dialogue context, visual evidence, and support direction.	context_response_ consistency

Table 10: Pointwise MLLM evaluation metrics.

Representative Pointwise Evaluation Prompt

You are an impartial evaluator for multimodal emotional support dialogue.
Evaluate the candidate output according to the specified metric. Use the dialogue context, any attached visual evidence, the model's think section if available, and the final response. If the think section is missing, it will be shown as `Not provided`.
Your judgment must consider:
1. whether the candidate is appropriate to the user's emotional state and dialogue context;
2. whether it is grounded in the available multimodal evidence;
3. whether it avoids unsupported, unsafe, overly verbose, repetitive, or formulaic content;
4. whether it satisfies the specific evaluation metric.
[EVALUATION_METRIC]
{metric_name}
[METRIC_DEFINITION]
{metric_definition}
[CONTEXT]
{context}
[MODEL_THINK]
{think}
[CANDIDATE_RESPONSE]
{response}
Score the candidate from 1 to 5. Return JSON only in the following format:
{json_score_format}

Table 11: Representative pointwise MLLM evaluation prompt. We use three semantically equivalent prompt variants for each pointwise metric.

Representative Pairwise Preference Prompt

You are an impartial judge for multimodal emotional support dialogue.
Compare Candidate A and Candidate B based on the dialogue context, any attached visual evidence, the provided think sections if available, and the final responses. The model identities are hidden. If a think section is missing, it will be shown as `Not provided`.
Choose the better candidate according to the following criteria:
1. empathy and emotional validation;
2. contextual relevance to the user's situation;
3. consistency with the visual evidence when available;
4. appropriateness and safety of the support strategy;
5. coherence between the think section and the final response;
6. avoidance of unnecessarily long, repetitive, verbose, or formulaic wording.
If both candidates are comparable in overall quality, choose `TIE`.
[CONTEXT]
{context}
[CANDIDATE A THINK]
{think_a}
[CANDIDATE A RESPONSE]
{response_a}
[CANDIDATE B THINK]
{think_b}
[CANDIDATE B RESPONSE]
{response_b}
Return JSON only in the following format:
{"winner": "A"}
The value of `winner` must be one of `A`, `B`, or `TIE`.

Table 12: Representative pairwise MLLM preference prompt. We use three semantically equivalent prompt variants, bidirectional candidate swapping, and majority voting to reduce prompt-order bias.

C Controlled Evaluation

To examine whether the observed pairwise preference trend persists after restricting the information visible to the judge, we reuse the same underlying Qwen2.5-VL-32B-Instruct-AWQ judge model under controlled views. The overall evaluation setup remains unchanged, including three semantically equivalent prompt variants, bidirectional candidate swapping, and majority vote. Under each controlled view, however, we use prompt variants that expose only the permitted candidate fields; when the think content is hidden, all think-related fields and think-specific criteria are removed from the prompt.

C.1 Controlled Views

We consider response-only no-cap and length-controlled views. In the response-only no-cap view, the judge sees the dialogue context, attached images, and the final response only; the prompt excludes the think section and no think-specific evaluation criterion is used. This view shows the same preference trend, but it may still be affected by response-length differences. Therefore, we focus the reported controlled results on the length-controlled cap-30/cap-25/cap-15 views. In these views, each candidate response is prefix-truncated to at most N words, where $N \in \{30, 25, 15\}$. The judge uses the same response-only criteria: the prompt excludes the think section and uses no think-specific evaluation criterion.

C.2 Controlled Pairwise Results

Table 13 reports blind pairwise preference outcomes under length-controlled views on the MESC-FT test set and MELD-ESC300. These views use the same response-only judging criteria: explicit reasoning traces are removed, the judge sees only the dialogue context, visual evidence, and the prefix-truncated final response, and no think-specific evaluation criterion is used. The cap-30, cap-25, and cap-15 settings impose progressively stricter upper bounds on visible response length.

Across both datasets, Model A remains consistently preferred under all length-controlled views. As the word cap becomes stricter, the win rate of Model A gradually decreases, but Model A still clearly outperforms Model B even under the most restrictive cap-15 setting. Since cap-15 removes explicit reasoning content and imposes the strongest constraint on visible response length, it provides the most conservative evidence in this appendix.

Dataset	View	A Win	B Win	Tie
MESC-FT test set	cap-30	426	11	13
MESC-FT test set	cap-25	411	18	21
MESC-FT test set	cap-15	361	53	36
MELD-ESC300	cap-30	287	7	6
MELD-ESC300	cap-25	279	11	10
MELD-ESC300	cap-15	186	29	85

Table 13: Controlled pairwise evaluation results. Each cell reports majority-vote decisions. Model A denotes Ours (Full), and Model B denotes FIRES (Baseline).

C.3 Statistical Significance

To assess whether the observed preference advantages in the controlled pairwise evaluation are statistically reliable, we further perform significance testing on the main comparisons. For each dataset and each reported length-controlled view, we apply a two-sided exact binomial sign test over decisive (i.e., non-tie) comparisons only, and estimate 95% confidence intervals for the FACES preference share via bootstrap resampling. This analysis does not introduce any new evaluation criterion; rather, it provides a statistical summary, including confidence intervals, of the majority-vote pairwise results reported in the main text and appendix.

The results show that FACES remains significantly preferred over the baseline under all reported length-controlled views on both the MESC-FT test set and MELD-ESC300. Notably, even under the most conservative cap-15 condition, where explicit reasoning traces are hidden and the judge can access only length-truncated final responses, FACES still attains a decisive preference share of 87.2% on the MESC-FT test set (95% CI: 83.9%–90.4%, $p \approx 2.0 \times 10^{-57}$) and 86.5% on MELD-ESC300 (95% CI: 81.9%–91.2%, $p \approx 3.1 \times 10^{-29}$). These results suggest that the observed preference gains are not merely driven by visible CoT content or longer responses, but remain robust under substantially restricted judging conditions.

D Detailed Preference Evaluation

D.1 Human Preference Counts

Table 14 and Table 15 provide the complete count-level results for the human pairwise preference study. In addition to the win, loss, and tie counts, we report Fleiss' κ to quantify inter-annotator agreement before adjudication.

These tables complement the percentage-based discussion in the main text by showing the underlying majority-vote decisions on each benchmark.

Dataset	Ours Win	FIRES Win	Tie	κ
MESC-FT test set	**339 (75.3%)**	63 (14.0%)	48 (10.7%)	0.62

Table 14: Human pairwise preference on the held-out MESC-FT test set. Each cell reports majority-vote decisions with percentages over all 450 test instances. κ denotes Fleiss' kappa.

Dataset	Ours Win	FIRES Win	Tie	κ
MELD-ESC300	**211 (70.3%)**	44 (14.7%)	45 (15.0%)	0.58

Table 15: Human pairwise preference on the full MELD-ESC300 benchmark. Each cell reports final majority-vote decisions with percentages over all 300 instances. κ denotes Fleiss' kappa.

The MESC-FT test set contains 450 evaluated instances, while MELD-ESC300 contains 300 evaluated instances.

E Additional Ablation Analysis

Table 16 and Figure 5 report additional dimension-level ablations on individual reward components. We ablate the dimensions used for intermediate reasoning and structural control, but do not ablate the outcome reward because it provides the shared final-response supervision needed for comparable training. These results complement the contribution-level ablations in the main text by showing how each reward dimension affects different aspects of multimodal emotional support generation.

Ours (Full) achieves the best METEOR, BERTScore-F1, Distinct-3, and DIFF-EPITOME, indicating the most balanced overall profile. Removing the format constraint increases BLEU-2 and ROUGE-L, but substantially hurts semantic quality and diversity, suggesting that lexical overlap alone does not reflect better support quality. Removing emotion or situation rewards leads to larger DIFF-EPITOME degradation, while removing strategy reward lowers semantic metrics, showing that these dimensions provide complementary supervision for strategy-aligned responses.

F Hyperparameter Sensitivity Analysis

We conduct one-factor-at-a-time sensitivity analysis for three method-specific hyperparameters in FACES: the CLIP visual threshold τ_v, the positive gate floor γ, and the endpoint-anchored credit-assignment parameters (m, λ). These hyperparameters directly control visual reward sparsity, total-reward-based positive gating, and token-level advantage allocation. Other routine optimization, decoding, and evaluation settings are reported in Appendix A and kept fixed. The default values are selected on the validation set, and all sensitivity results are reported on this set.

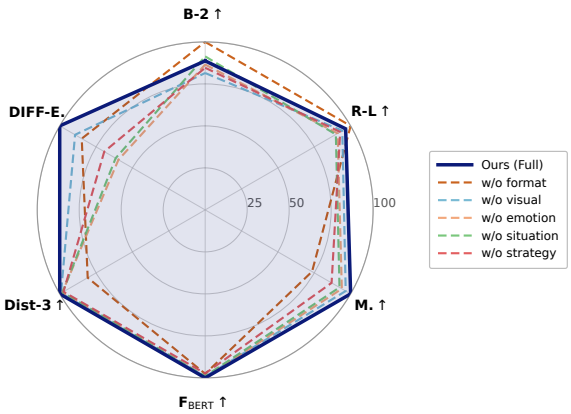

Figure 5: Normalized radar chart of dimension-level ablations. Larger values indicate better performance; DIFF-E. is reversed. FACES remains most balanced across metrics.

Table 17 shows that FACES is robust to moderate changes in τ_v. A lower threshold makes the visual reward more permissive and slightly improves lexical-overlap metrics, but weakens semantic discrimination. A higher threshold makes visual rewards sparse, reducing semantic and MLLM-based performance. The default $\tau_v = 0.60$ does not maximize every metric, but gives the best DIFF-E. and MLLM average score.

Table 18 shows that FACES is stable around the default gate floor. Since g_i is dynamically computed from γ and $R_i/R_{\max}$, we analyze γ rather than treating g_i as an independent hyperparameter. When γ is too small, positive local advantages are overly suppressed. As γ increases, the gate becomes weaker. In particular, $\gamma = 1.00$ improves B-2 and R-L but substantially degrades DIFF-E. and MLLM-based scores, supporting the need to suppress excessive positive updates from globally weak candidates.

Table 19 analyzes the endpoint-anchored credit-assignment parameters. A shorter anchor window or a smaller λ over-emphasizes segment endings and weakens supervision for earlier reasoning tokens. A larger anchor window is still stable but slightly less balanced. When $\lambda = 1.00$, all tokens in a segment receive the same advantage, which improves lexical-overlap metrics but reduces semantic and MLLM-based quality. The default setting $(m = 5, \lambda = 0.80)$ provides a good balance between endpoint emphasis and full-span supervision.

Model	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓
w/o format	**10.91**	**21.84**	15.15	85.03	80.47	0.4293
w/o visual	8.89	20.59	19.92	86.51	98.74	0.4073
w/o emotion	9.44	19.89	19.37	86.95	97.92	0.6108
w/o situation	9.97	19.62	19.04	86.24	97.74	0.5918
w/o strategy	9.23	20.27	17.92	84.86	97.05	0.5251
Ours (Full)	9.68	21.09	**20.62**	**87.14**	**99.63**	**0.3643**

Table 16: Additional dimension-level ablation results on the MESC-FT test set. All metrics are higher-is-better except DIFF-E.

τ_v	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓	MLLM Avg. ↑
0.50	**9.91**	**21.36**	20.19	86.84	99.18	0.3984	3.94
0.60	9.74	21.16	**20.48**	**87.03**	99.42	**0.3726**	**4.01**
0.65	9.65	21.02	20.35	86.99	**99.46**	0.3795	3.98
0.70	9.56	20.84	20.07	86.83	99.29	0.4017	3.93

Table 17: Sensitivity analysis of the CLIP visual threshold τ_v on the MESC-FT validation set.

γ	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓	MLLM Avg. ↑
0.00	9.55	20.87	20.27	86.89	99.20	0.3911	3.95
0.10	9.74	21.16	**20.48**	**87.03**	**99.42**	**0.3726**	**4.01**
0.30	9.96	21.48	20.15	86.82	99.21	0.3995	3.94
1.00	**10.13**	**21.88**	19.88	86.61	98.96	0.4387	3.86

Table 18: Sensitivity analysis of the positive gate floor γ on the MESC-FT validation set. When $\gamma = 1.00$, positive gating is effectively removed.

m	λ	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓	MLLM Avg. ↑
3	0.80	9.61	20.98	20.26	86.91	99.30	0.3878	3.96
5	0.60	9.58	20.90	20.18	86.88	99.24	0.3946	3.94
5	0.80	9.74	21.16	**20.48**	**87.03**	99.42	**0.3726**	**4.01**
7	0.80	9.83	21.25	20.34	86.96	**99.46**	0.3819	3.98
5	1.00	**10.02**	**21.52**	20.08	86.80	99.21	0.4135	3.91

Table 19: Sensitivity analysis of endpoint-anchored token-level credit assignment on the MESC-FT validation set. m denotes the anchor length, and λ controls the advantage scale assigned to non-anchor tokens.

G MELD-ESC300 Details

G.1 Benchmark Construction

To evaluate out-of-domain generalization, we construct a lightweight external benchmark, MELD-ESC300, from the MELD test set using the publicly available EMOTyDA resources (Poria et al., 2019; Saha et al., 2020), and use it for cross-dataset evaluation without any additional fine-tuning. Rather than directly reusing the original MELD annotations, we filter dialogue segments that better approximate emotional-support scenarios, allowing us to examine whether the model's gains transfer to a non-training-domain multimodal dialogue setting.

Concretely, we treat the i-th utterance in a dialogue as a candidate target response and retain only samples satisfying the following conditions: (1) at least two preceding turns are available; (2) the target response length is between 4 and 40 words; (3) the target sentiment is not *negative* and its emotion label belongs to $\{neutral, joy, surprise\}$, which approximates neutral-to-positive supportive replies; (4) the target speaker differs from the speaker in the immediately preceding utterance; and (5) the video clip corresponding to the immediately preceding utterance is available, so that stable visual context can be provided.

We then apply a two-stage filtering strategy to identify ESC-style samples. A sample is marked as strict if the immediately preceding utterance already exhibits negative affect, i.e., its sentiment is *negative* or its emotion label belongs to $\{anger, sadness, fear, disgust\}$. If this condition is not met, we further inspect the three most recent preceding utterances; if any of them contains negative affect, the sample is retained and marked as relaxed. Importantly, regardless of whether a sample is selected by the strict or relaxed rule, the visual input is always taken from the video of the immediately preceding utterance, from which we uniformly extract five frames. This design avoids directly leaking visual information associated with the target response itself. The final model input contains at most six preceding turns, and the reference output is the original target utterance.

Statistic	Value
Total samples	300
strict samples	208
relaxed samples	92
Frames per sample	5
Maximum context turns	6
Emotion: *neutral*	232
Emotion: *joy*	51
Emotion: *surprise*	17

Table 20: Summary statistics of MELD-ESC300. Strict denotes samples whose immediately preceding utterance already exhibits negative affect; relaxed denotes samples where the immediately preceding utterance does not, but one of the three most recent preceding utterances does. For all samples, the visual input is taken from the video of the immediately preceding utterance, from which five frames are uniformly extracted.

Following this procedure with a fixed random seed of 42, we construct 300 evaluation instances, which we denote as MELD-ESC300. Among them, 208 satisfy the strict criterion and 92 are selected by the relaxed criterion. As shown in Table 20, the target responses are dominated by *neutral* cases, with a smaller yet non-trivial number of *joy* and *surprise* instances. We report only the emotion distribution, as the corresponding sentiment labels are largely implied by the retained target emotion categories in this subset and therefore add limited additional information. These statistics make the composition of the benchmark explicit and reproducible. Overall, the benchmark is centered on mildly supportive or non-negative target replies while still covering a modest range of more explicitly positive feedback.

We emphasize that MELD-ESC300 is derived from scripted TV dialogue rather than professionally curated emotional-support data, so it is intended as a controlled external generalization test set rather than a replacement for dedicated ESC benchmarks. By combining recent negative-affect

context, neutral-to-positive target replies, and visual evidence from the immediately preceding turn, it provides a practical approximation of cross-domain emotional-support scenarios for evaluating transfer beyond the MESC-FT test set.

G.2 MELD-ESC300 MLLM Evaluation

Table 21 provides the full appendix record for the length-controlled MLLM evaluation on MELD-ESC300. It includes the pointwise 1–5 scores for the three judge dimensions, as well as the complete pairwise preference counts, win rates, and tie rate over 300 instances. The evaluation follows the same response-only cap-15 setting used in the controlled analysis, where reasoning traces are hidden and candidate responses are prefix-truncated before judging. This table therefore serves as the detailed evidence behind the summarized cross-dataset MLLM discussion in the main text.

G.3 Additional Automatic Baselines

Table 22 reports additional automatic evaluation results for six strong MLLM baselines on MELD-ESC300 under the same protocol as in the main text.

H Case Study

To complement the controlled pairwise results, we provide qualitative examples that illustrate where FACES improves over the baseline and where it can still fail. Figure 6 compares the dialogue context, visual cue, baseline response, FACES response, and a brief analysis for four representative cases, covering in-domain success, length-controlled success, cross-dataset transfer, and a failure case.

(a) Pointwise MLLM evaluation results on MELD-ESC300. Scores are on a 1–5 scale, and higher is better.

Model	Empathy ↑	Multimodal App. ↑	Context-Response Cons. ↑	Avg. ↑
FIRES (Baseline)	2.59	3.61	3.84	3.35
Ours (Full)	**3.27**	**3.85**	**4.06**	**3.73**

(b) Pairwise MLLM preference evaluation results on MELD-ESC300. Win rates are computed over 300 test instances.

Model A	Model B	A Win	B Win	Tie	A Win Rate	B Win Rate	Tie Rate
Ours (Full)	FIRES (Baseline)	**186**	29	85	**62.00%**	9.67%	28.33%

Table 21: Length-controlled MLLM evaluation results on MELD-ESC300. Panel (a) reports pointwise scores across three evaluation dimensions, and panel (b) reports blind pairwise preferences over 300 test instances. Pairwise results are significant under a two-sided sign test with ties excluded ($p < 0.001$).

Model	B-2 ↑	R-L ↑	M. ↑	F_{BERT} ↑	Dist-3 ↑	DIFF-E. ↓
GPT-4o	1.41	5.73	13.04	95.35	54.53	0.5053
Claude-Haiku-4.5	1.68	6.81	15.12	95.56	64.18	0.7507
Gemini-2.5-Flash-Lite	2.47	7.82	13.47	**96.02**	56.19	**0.2994**
GLM-4.1V-Thinking-FlashX	1.82	6.23	13.60	95.74	69.25	0.8126
Doubao-1.6-Vision	1.70	6.67	11.80	95.65	74.42	0.7057
Kimi-K2.5	1.63	6.06	14.47	95.60	73.10	0.8056
Ours (Full)	**5.59**	**16.57**	**15.16**	86.23	**94.07**	0.7664

Table 22: Additional automatic evaluation results on MELD-ESC300. We report six strong MLLM baselines and include Ours (Full) for direct comparison. All metrics are higher-is-better except DIFF-E. Best results in the table are shown in bold.

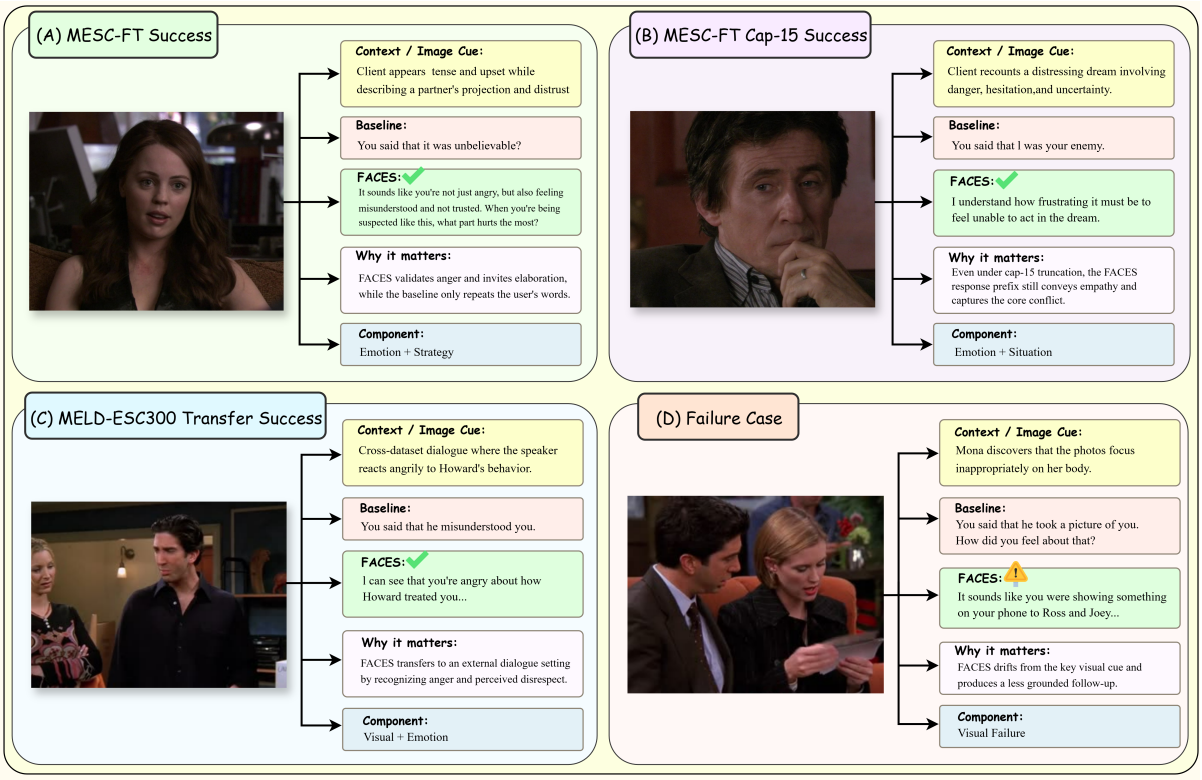

Figure 6: Qualitative case study of FACES against FIRES. Panels (A)–(C) show representative success cases under in-domain, length-controlled, and cross-dataset settings, while panel (D) illustrates a failure case where FACES drifts from the key visual cue.