ECG-LoRA: A Conditional Fidelity Gated Adaptation Method for Empathetic Response Generation

Anonymous ACL submission

Abstract

Although Large Language Models (LLMs) excel in open-domain dialogue, they often struggle with emotional fidelity—the ability to dynamically adapt response styles and strategies to fine-grained emotional contexts. Existing Parameter-Efficient Fine-Tuning methods like LoRA typically apply static updates, failing to capture such nuanced variations. To address this, we propose Emotion-Conditioned Gated LoRA (ECG-LoRA), a novel framework that introduces a lightweight, input-aware gating mechanism to dynamically scale LoRA updates based on emotional signals. This design enables the model to intrinsically allocate adaptation capacity: intensifying intervention for high-arousal emotions while preserving base knowledge for subtle states. Extensive experiments on the EmpatheticDialogues benchmark across three LLM backbones (Qwen2-7B, Llama-3-8B, Gemma-7B) demonstrate that ECG-LoRA significantly outperforms standard LoRA in both generation quality (e.g., +0.65% BERTScore) and emotional fidelity (e.g., +4.29% Emp-F1). Our framework is highly efficient, requiring only 0.12% additional parameters. More importantly, it introduces a new perspective for controllable text generation: targeting the imitation of complex, intrinsically consistent human behavioral patterns beyond simple semantic alignment. Code and data are released to facilitate future research.

1 Introduction

Empathetic dialogue generation, situated at the intersection of artificial intelligence and psychology, seeks to equip conversational agents with the ability to perceive distinct emotional states and generate contextually appropriate, resonant responses(Higashinaka et al., 2008). Although Large Language Models (LLMs) have demonstrated remarkable capabilities in open-domain interactions, replicating the nuanced stylistic and strategic variations characteristic of human responses—termed emotional fidelity—remains a significant challenge(Chen et al., 2022). Current models often default to generic, safe responses that fail to capture the fine-grained adaptation required for high-quality empathetic interaction.

Prior research has primarily focused on enhancing empathy through external modules. Early end-to-end models like MoEL(Lin et al., 2019) and MIME(Chari et al., 2022)utilized soft-gating or stochastic mixing to model emotion distributions but lacked semantic depth. Subsequent approaches, such as CEM(Sabour et al., 2021) and KEMP(Lu et al., 2022) incorporated external commonsense knowledge (e.g., ConceptNet(Speer et al., 2018)) to improve cognitive understanding. More recently, the focus has shifted to advanced control paradigms for LLMs: DiffusEmp(Bi et al., 2023)leverages diffusion models for multi-grained control; EmpRL(Ma et al., 2025) employs reinforcement learning to align empathy strategies; and EmoCharacter(Feng et al., 2025)benchmarks the loss of emotional fidelity in role-playing agents. However, these methods often incur high computational costs or rely on complex external reward signals, overlooking the potential of intrinsic parameter adaptation.

In the realm of efficient adaptation, Low-Rank Adaptation(LoRA)(Hu et al., 2021) has emerged as a standard solution. However, standard LoRA applies a static update across all inputs, lacking the flexibility to dynamically modulate adaptation based on shifting emotional contexts(Hu et al., 2024).While recent variants like G-lora(Liang et al., 2024) have introduced gating mechanisms for multi-tasking or context robustness, they have not yet been exploited to address the specific challenge of fine-grained emotional alignment.

To bridge this gap, we propose **Emotion-Conditioned Gated LoRA (ECG-LoRA)**, a novel fine-tuning framework that introduces a lightweight, input-aware gating mechanism to dynamically scale the contribution of LoRA modules based on emotional signals. Unlike retrieval-based methods, ECG-LoRA focuses on unlocking the model's intrinsic capacity for conditional adaptation. Our approach hypothesizes that different emotions require varying degrees of parameter intervention to achieve high fidelity.

We validate our framework on the EmpatheticDialogues benchmark (Rashkin et al., 2019) using a comprehensive suite of modern metrics, including Emotion Alignment F1(Xu and Jiang, 2024), Emotional Consistency (EC) in VAD space (Feng et al., 2025)(Mohammad, 2025), and Emp-F1(Ma et al., 2025). Experimental results demonstrate that ECG-LoRA significantly outperforms standard LoRA and other baselines, achieving superior conditional fidelity—defined as the precise alignment of response style and emotional intensity with human references—while adding only 0.12% trainable parameters.

Our main contributions are as follows.

1. We propose ECG-LoRA, an efficient framework that realizes conditional fidelity by dynamically modulating LoRA updates via emotion-conditioned gating.

2. We introduce a rigorous evaluation protocol incorporating recent SOTA metrics (EA-F1, EC, Emp-F1), revealing that our method improves not just semantic similarity but also the functional alignment of empathy strategies.

3. We provide interpretable analyses, visualizing how the gating mechanism autonomously learns to differentiate between emotional intensities, offering new insights into the internal dynamics of controllable text generation.

2 Methodology

2.1 ECG-LoRA

ECG-LoRA establishes a conditional fidelity mechanism in forward propagation and introduces two auxiliary loss functions to enhance the discriminability of emotional expressions; during backpropagation, both the mechanism and these auxiliary loss functions play a pivotal role in parameter updates.

2.1.1 Forward Propagation: From Static Adaptation to Conditional Fidelity

Although LoRA performs well in model fine-tuning, its static update mechanism limits its effectiveness in tasks that require dynamic responses, such as empathetic dialogue. The theoretical core of our ECG-LoRA framework lies in transforming static adaptation into a dynamic, condition-guided adaptation process, thereby enabling high-fidelity learning of complex human response paradigms.

Standard LoRA updates the pretrained weights W_0 via a fixed low-rank matrix $\Delta W = BA$. The forward propagation can be expressed as Equation 1:

$$y = W_0 x + BAx \tag{1}$$

The inherent limitation of this formulation is that the update BAx is uniform across all inputs. However, in empathetic dialogue, the ideal response to "joy" should be fundamentally different in pattern from that to "sadness". The theoretical innovation of ECG-LoRA lies in introducing a scalar gating signal determined by the emotional condition, which acts as a dynamic regulator directly controlling the contribution of the low-rank update. The theoretical model of ECG-LoRA can be represented as follows:

$$y = W_0 x + G_{emo} \cdot (BAx) \tag{2}$$

The generation of the gating signal G_{emo} forms the theoretical foundation for achieving conditional fidelity. It originates from a process that maps discrete emotion labels into continuous control signals. First, a discrete emotion id_c is mapped to a semantically rich emotion vector v_{emo} via a learnable embedding layer E:

$$v_{emo} = Emb(id_c) \in R^{d_{emo}} \tag{3}$$

This vector v_{emo} extracts and represents the intrinsic semantics of the emotion. Subsequently, the vector is transformed into the scalar gating signal G_{emo} via a linear transformation and a Sigmoid activation function:

$$G_{emo} = \sigma(W_g \cdot v_{emo} + b_g) \tag{4}$$

2.1.2 Gradient Conditioning Regularization in Backpropagation

The core advantage of ECG-LoRA extends beyond on-demand activation during forward propagation to the dynamic modulation of the backpropagation process. This gating operation effectively acts as a conditional regularization mechanism during parameter updates.

According to the chain rule, the gradients of the language model loss L_{LM} with respect to A and B are directly modulated by the gating signal G_{emo} (see Equations 5 and 6):

$$\frac{\partial L_{LM}}{\partial A} = \frac{\partial L_{LM}}{\partial y} \frac{\partial y}{\partial A} = \frac{\partial L_{LM}}{\partial y} \cdot (G_{emo} \cdot B^T x^T) \tag{5}$$

$$\frac{\partial L_{LM}}{\partial B} = \frac{\partial L_{LM}}{\partial y} \frac{\partial y}{\partial B} = \frac{\partial L_{LM}}{\partial y} \cdot (G_{emo} \cdot x A^T)^T \tag{6}$$

These equations clearly demonstrate the mechanism of Gradient Conditioning: the gradient flow from the task loss is explicitly scaled by the current emotion's gating signal G_{emo} before updating the LoRA parameters.

Crucially, the auxiliary objectives—Emotion Contrastive Loss (L_{con}, 10) and Embedding Regularization Loss (L_{reg}, 11) defined in Section 2.3—do not directly depend on matrices A and B. Instead, they optimize the emotion encoder and gating network to ensure G_{emo} yields discriminative values for different emotions. A well-optimized G_{emo} is prerequisite for accurate gradient modulation. Consequently, the partial derivatives of the total loss L_{total} with respect to A and B are entirely contributed by the gated language modeling term:

$$\frac{\partial L_{total}}{\partial A} = \frac{\partial L_{LM}}{\partial A} \tag{7}$$

Similarly:

$$\frac{\partial L_{total}}{\partial B} = \frac{\partial L_{LM}}{\partial B} \tag{8}$$

Equations 7 and 8 embody the theoretical distinction from standard LoRA (where G_{emo} degenerates to 1). ECG-LoRA's gradient update is a dynamic process: intense emotions triggering high G_{emo} values induce larger parameter updates, while neutral states result in minimal changes. This empowers the model to go beyond static learning and allocate "optimization budget" specifically to challenging emotional patterns, underpinning its ability to achieve high fidelity.

2.2 Model Architecture

The architecture of ECG-LoRA translates the theoretical concept of "conditional fidelity" into a concrete computational graph. As illustrated in Figure 1, the framework modulates the pre-trained LLM through a three-stage process: extraction, activation, and generation.

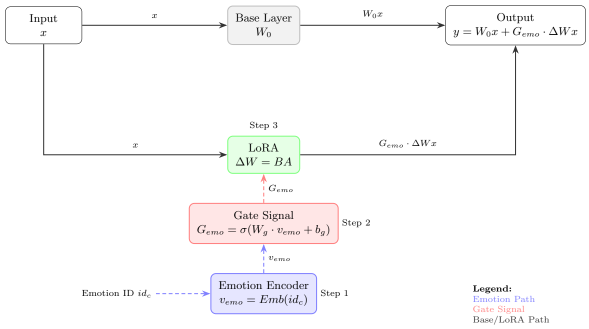

Figure 1: **The overall architecture of ECG-LoRA.** The framework operates in three stages: (1) **Extraction**: A discrete emotion ID is mapped to a semantic vector. (2) **Activation**: A lightweight network computes a scalar gating signal, acting as a dynamic "soft switch." (3) **Generation**: Modulating the LoRA update with the gate to produce emotion-conditioned outputs.

Stage 1: Emotional Semantic Extraction

To ground the adaptation in emotional semantics, we employ a learnable emotion encoder (Delbrouck et al., 2020). This module consists of an embedding layer, which maps a discrete input emotion label id_c into a dense, continuous vector v_{emo}. This vector serves as the semantic anchor, capturing the nuanced characteristics of the target emotion (e.g., the subtle difference between Sadness and Devastation) to guide subsequent adaptation.

Stage 2: Conditional Activation

The extracted vector v_{emo} is then fed into the Gate Signal Generation Unit. This unit acts as a lightweight controller, transforming the high-dimensional semantic information into a scalar gating signal. Mathematically, this is achieved via a linear transformation followed by a Sigmoid activation(Eq 4). In this context, functions as a dynamic "throttle," determining the intensity of intervention required for the current emotion—allowing stronger adaptation for high-arousal emotions while maintaining subtle adjustments for neutral states.

Stage 3: Fidelity Generation

In the final stage, the gating signal G_{emo} is injected into the LoRA layers. Unlike standard LoRA, where the low-rank update BAx is added uniformly, ECG-LoRA modulates this update via element-wise multiplication with Eq 2. This ensures that the injected knowledge is strictly authorized by the emotional condition. By freezing the vast majority of the LLM parameters and only updating the LoRA adapters and the gating network, the architecture achieves efficient, emotion-conditioned text generation.

2.3 Optimization Objective

To achieve conditional fidelity while preventing catastrophic forgetting, we adopt a parameter-efficient training strategy. Specifically, we freeze all parameters of the pre-trained LLM backbone and only update the LoRA matrices BA, the emotion encoder, and the gating network.

The training process is governed by a composite objective function designed to enforce both semantic generation quality and emotional discriminability. The total loss L_{total} is formulated as:

$$L_{total} = L_{LM} + W_1 \cdot L_{con} + W_2 \cdot L_{reg} \quad (9)$$

Here, L_{LM} denotes the standard autoregressive language modeling loss. To strictly regulate the behavior of the gating mechanism and the emotion encoder, we incorporate two auxiliary objectives weighted by hyperparameters W_1 and W_2:

Emotion Contrastive Loss. This loss optimizes the gating mechanism to generate discriminative signals. It ensures that gating values for the same emotion remain consistent while those for different emotions are pushed apart. The loss is defined as:

$$L_{con} = \omega \cdot L_{cons} + (1 - \omega) \cdot L_{sep} \quad (10)$$

Here, L_{cons} represents the loss of consistency, and L_{sep} represents the loss of separation. The parameter $\omega \in [0, 1]$ balances the trade-off between intra-class consistency and inter-class separation of the gating signals.

Emotion Embedding Regularization Loss. To guarantee high-quality input representations for the gate, this loss structures the emotion embedding space. It is calculated as:

$$L_{reg} = \lambda_{inter} \cdot L_{inter} + \lambda_{intra} \cdot L_{intra} \quad (11)$$

Here, L_{inter} and L_{intra} maximize the distance between different emotion clusters (inter-class separation) and minimize the variance within the same emotion cluster (intra-class compactness), respectively. λ_{inter} and λ_{intra} are the corresponding weights controlling their contribution.

The detailed derivations of L_{con} and L_{reg} are provided in the Appendix 7.1. By jointly optimizing these objectives, ECG-LoRA learns to generate fluent text while maintaining a structured and interpretable control mechanism.

3 Experimental Setup

3.1 Datasets

We conduct experiments on the Empathetic Dialogues (ED) dataset (Rashkin et al., 2019), which contains approximately 25,000 multi-turn dialogues covering 32 fine-grained emotion categories. Each sample consists of a speaker's emotionally colored scenario and a corresponding empathetic response.

3.2 Evaluation Metrics

3.2.1 Quantitative Metrics

To comprehensively assess both the generation quality and the conditional fidelity of the responses, we employ a multi-dimensional evaluation protocol:

Generation Quality. We use standard automatic metrics to evaluate semantic similarity and fluency, including BLEU-4, ROUGE-1/2/L (Papineni et al., 2002; Lin, 2004), BERTScore-F1(Zhang et al., 2019), and Dist-1/2(Li et al., 2016).

Emotional Fidelity. To verify our core claim of conditional adaptation, we adopt three state-of-the-art emotion-centric metrics:

1. **Emotion Alignment F1 (EA-F1)** (Xu and Jiang, 2024): A macro-averaged F1 score calculated by a RoBERTa-based classifier (fine-tuned on GoEmotions), measuring how accurately the generated response reflects the ground-truth emotion category.

2. **Emotional Consistency (EC)** (Feng et al., 2025): A continuous metric measuring the cosine similarity between the emotion embedding of the generated text and the target emotion in the Valence-Arousal-Dominance (VAD) space (Mohammad, 2025). This captures fine-grained emotional nuances beyond discrete labels.

3. **Emp-F1**(Ma et al., 2025) : A strategy-oriented metric that evaluates the alignment of

Table 1: **Model Performance and ECG-LoRA Improvement over Standard LoRA.** Main results on the ED test set. We report Generation Quality metrics (BLEU-4, ROUGE-1/2/L, BERTScore, Dist-1/2) and Emotional Fidelity metrics (EA-F1, EC). Strategy-oriented scores (Emp-F1) are detailed in 4.2. Bold indicates the best performance for each backbone. All improvements of ECG-LoRA over Standard LoRA are statistically significant ($p < 0.05$).

Model	Qwen2-7B-Instruct		Llama3-8B-Instruct		Gemma-7B-it	
	Std LoRA	ECG-LoRA	Std LoRA	ECG-LoRA	Std LoRA	ECG-LoRA
BLEU-4	0.0378	**0.0474**	0.0126	**0.0145**	0.0295	**0.0323**
ROUGE-1	0.1910	**0.2093**	0.1011	**0.1128**	0.1503	**0.1697**
ROUGE-2	0.0504	**0.0647**	0.0226	**0.0258**	0.0358	**0.0391**
ROUGE-L	0.1749	**0.1944**	0.0828	**0.0923**	0.1433	**0.1538**
BERTScore	0.8469	**0.8500**	0.8183	**0.8236**	0.8031	**0.8425**
Dist-1	0.9664	**0.9744**	0.4222	**0.4815**	0.8520	**0.9430**
Dist-2	0.9971	**0.9993**	0.6144	**0.7000**	0.9730	**0.9942**
EA-F1	0.2867	**0.3065**	0.1508	**0.1649**	0.2044	**0.2208**
EC	0.9025	**0.9039**	0.8933	**0.8943**	0.8998	**0.9015**
Emp-F1	65.55	**66.15**	58.34	**60.84**	62.00	**64.95**

specific empathy mechanisms (Emotional, Exploratory, and Interpretive) using a T5-based classifier, assessing whether the model adopts the appropriate empathetic strategy.

3.2.2 Human Evaluation

To verify the fidelity of generated responses beyond automatic metrics, we conducted a rigorous human evaluation following the pairwise comparison protocol of (Cao et al., 2024). We randomly sampled 128 dialogues and recruited four professional annotators to independently evaluate the model outputs under blinded conditions. The evaluation focuses on three distinct dimensions:

1. **Lexical Fidelity (Lex.)**: Measures whether the model accurately captures the lexical habits and stylistic nuances of human responses.

2. **Semantic Fidelity (Sem.)**: Assesses whether the model precisely understands contextual intent and generates logically coherent replies.

3. **Emotional Fidelity (Emo.)**: Evaluates whether the model effectively perceives the speaker's emotional state and provides feedback consistent with human intuition.

For each test case, annotators labeled the ECG-LoRA response as "Win" (better), "Lose" (worse), or "Tie" (comparable) relative to the Standard LoRA baseline. Detailed instructions and annotator backgrounds are provided in the Appendix7.2.

3.3 Implementation Details

We utilized Qwen2-7B-Instruct, Llama-3-8B-Instruct, and Gemma-7B-it as backbone models. For parameter-efficient fine-tuning, LoRA adapters were injected into all linear layers. Models were trained for 8 epochs on NVIDIA A100 GPUs using AdamW optimizer with a cosine schedule. To ensure reproducibility, we fixed random seeds and used standard train/val/test splits of the ED dataset. Detailed hyperparameters for each backbone are provided in the Appendix7.3.

4 Experimental Results and Analysis

4.1 Main Results

We compare ECG-LoRA against Standard LoRA across three representative LLM backbones: Qwen2-7B-Instruct, Llama-3-8B-Instruct, and Gemma-7B-it. As summarized in Table 1, ECG-LoRA achieves consistent improvements across all metrics, validating its robustness and generalizability.

Generation Quality. ECG-LoRA demonstrates superior alignment with human emotional intent, surpassing baselines in both EA-F1 and EC metrics across all models. This confirms the gating mechanism's effectiveness in enforcing fine-grained emotional control.

Emotional Fidelity. Semantic quality is also enhanced, with ECG-LoRA achieving the highest BERTScore and ROUGE values. This indicates that emotion-conditioned adaptation improves overall response relevance without sacrificing fluency.

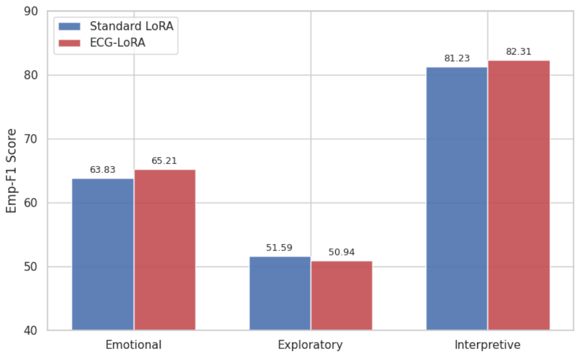

Figure 2: **Strategy Shift Bar Chart.** Shift in empathy response strategies (Emp-F1). ECG-LoRA (Red) consistently improves Emotional and Interpretive empathy across all backbones compared to Standard LoRA (Blue), while reducing generic Exploratory questioning. This reflects a transition from "safe" inquiries to active emotional resonance.

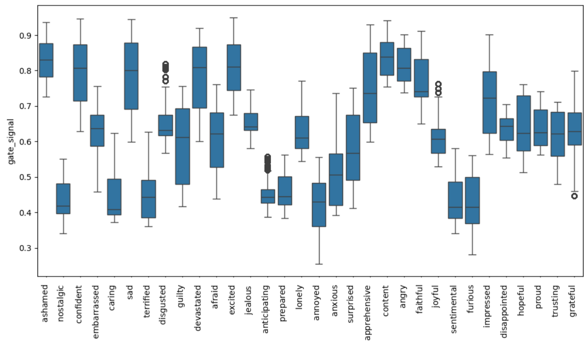

Figure 3: **Distribution of gating signals across emotion categories.** The boxplot reveals distinct activation ranges for different emotions. This confirms that the model dynamically modulates the adaptation intensity based on the specific emotional semantics rather than applying a uniform scaling.

4.2 Fine-grained Strategy Analysis

To understand how the model achieves higher fidelity, we analyze the response strategies using the Emp-F1 metric. As illustrated in Figure 2 and detailed in Table 1, ECG-LoRA exhibits a distinct strategic shift common across all three backbones:

1. Enhanced Emotional Resonance: It significantly improves Emotional empathy (e.g., +1.38 points on Qwen), which involves explicit expressions of affect.

2. Improved Interpretation: It also boosts Interpretive empathy, indicating better understanding of the user's feelings.

3. Reduced Generic Questioning: Interestingly, this comes with a slight decrease or stagnation in Exploratory empathy (questioning).

This trade-off suggests that the emotion-conditioned gate successfully steers the model away from generic questioning—a "safe" but often shallow strategy common in LLMs—toward more active and risky emotional resonance, thereby aligning better with high-fidelity human reference patterns.

4.3 Mechanism Analysis & Interpretability

Discriminative Activation. To verify whether the gating mechanism learns meaningful and distinct patterns for different emotional contexts, we visualize the distribution of gating signals G_{emo} across 32 emotions in the training set (Figure 3).

As illustrated in Figure 3, the learned gating signals exhibit significant variance across categories, spanning a wide range from approximately 0.3 to 0.95. This distribution is not random; it correlates with emotional intensity. For instance, intense emotions like 'Ashamed', 'Apprehensive' elicit high gate values, effectively "opening the valve" for maximizing LoRA's intervention. Conversely, nuanced states like 'Sentimental', 'Caring' trigger lower activation, preserving more of the base model's original capabilities. This discriminative behavior proves that G_{emo} acts as a context-aware controller, authorizing parameter updates only to the extent required by the input emotion.

Learning Dynamics. Figure 4 tracks the evolution of gating signals during training.

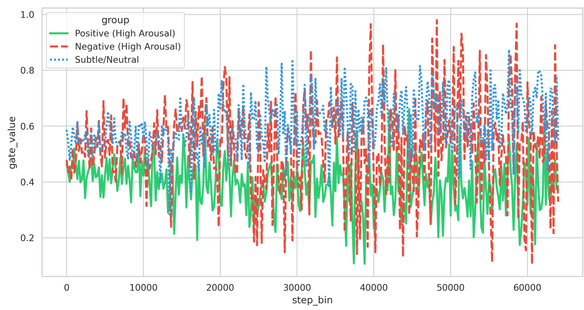

Figure 4: **Learning Dynamics: Evolution of gating signals.** Signals for Positive, Negative, and Neutral emotions diverge rapidly from initialization, confirming that the discriminative behavior is learned via auxiliary supervision.

Starting from a uniform initialization, the signals for distinct emotion groups (Positive, Negative, Neutral) rapidly diverge. Notably, Negative/High-Arousal emotions converge to higher values, validating that the gate actively learns to compensate for the base model's intrinsic positivity bias by applying stronger updates for negative contexts.

Causal Control. To explicitly test the gate's authority, we perform an intervention experiment (Figure 5).

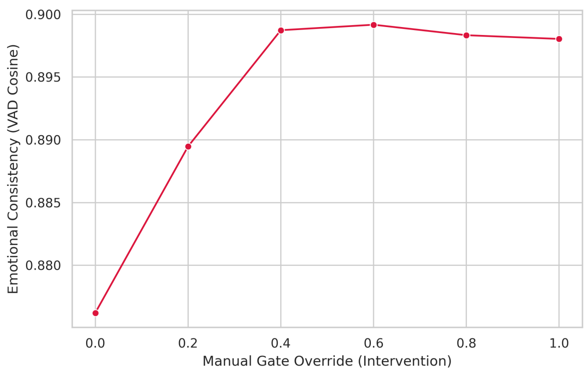

Figure 5: **Causal Control of Emotional Fidelity.** Causal control curve. Manually increasing the gate value G_{emo} during inference leads to a monotonic increase in EC score, confirming the gate acts as a functional control knob for fidelity.

By manually overriding during inference, we observe a monotonic increase in EC as the gate value rises. This establishes a causal link: the gating signal is not a passive artifact but a continuous functional knob that directly dictates the emotional fidelity of the output.

Gradient Allocation. Finally, Figure 6 visualizes the training dynamics.

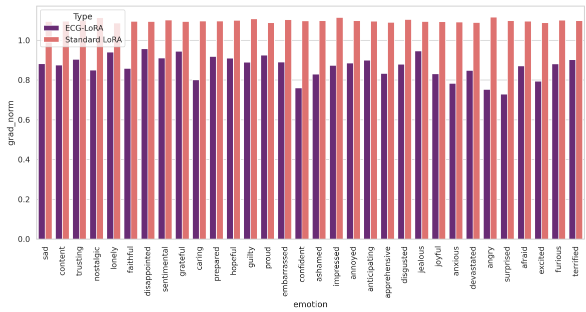

Figure 6: **Emotional Conditioned Gradient Allocation.** Average gradient norm of LoRA parameters. ECG-LoRA allocates significantly larger gradients to complex emotions, evidencing emotion-aware optimization.

Standard LoRA exhibits nearly uniform gradient magnitudes across emotions. In contrast, ECG-LoRA displays clear emotion-dependent variation, with larger gradients for complex emotions. This mechanistic evidence confirms that our framework enables intrinsic emotion-aware adaptation during the optimization process itself.

4.4 Ablation Studies

To dissect the contribution of each component in our framework, we conduct ablation studies on the Qwen2-7B-Instruct backbone. We compare the **ECG-LoRA Full Model** against three variants: (1) **w/o EmoCond** (removing the entire emotion conditioning branch), (2) **w/o RegLoss** (removing the Embedding Regularization Loss), and (3) **w/o ConLoss** (removing the Emotion Contrastive Loss). The results are summarized in Table 2.

Impact of Emotion Conditioning (w/o EmoCond). Removing the emotion conditioning mechanism (remove the input emotion vector, effectively decoupling the gating signals from emotion, equivalent to a standard LoRA with auxiliary losses removed) causes the most drastic performance degradation across all metrics. Notably, Emp-F1(emo) plummets significantly, confirming that the gating architecture is the primary driver of emotional fidelity. Without the V_{emo}, the model reverts to generic generation patterns, failing to align with the ground-truth emotion.

Impact of Auxiliary Losses. The auxiliary losses play a crucial role in refining the quality of the gating signal:

- **w/o RegLoss**: Removing embedding regularization leads to a decline in EA-F1. Although Emp-F1(Exp) increases slightly, it comes at the cost of emotional precision, suggesting that unstructured embeddings introduce noise into the control signal.

- **w/o ConLoss**: Removing the contrastive loss also degrades EA-F1 and EC. This demonstrates that L_{con} is essential for pushing gate values apart (as visualized in Figure 3), ensuring the mechanism remains discriminative rather than collapsing to a uniform scaling factor.

Conclusion. The Full Model consistently outperforms all ablated variants. This validates that while the gating architecture provides the capacity for conditional adaptation, the auxiliary losses are indispensable for learning the optimal control policy.

4.5 Human Evaluation Results

The results, presented in Table 3, reveal a clear preference for ECG-LoRA across all evaluated di-

Table 2: **Ablation study on Qwen2-7B-Instruct.** We evaluate the impact of removing the emotion conditioning mechanism and auxiliary losses. EA-F1, EC, and Emp-F1 metrics are reported. The Full Model achieves the best performance across all dimensions, highlighting the synergy of the proposed components.

Model	w/o EmoCond	w/o RegLoss	w/o ConLoss	ECG-LoRA Full Model
EA-F1	0.2534	0.2805	0.2871	**0.3065**
EC	0.8971	0.9013	0.9028	**0.9039**
Emp-F1(emo)	49.66	59.53	62.91	**65.21**
Emp-F1(Exp)	51.95	**53.80**	51.52	50.94
Emp-F1(Int)	81.27	80.95	81.75	**82.31**
Emp-F1(overall)	60.96	64.76	65.4	**66.15**

Table 3: Human evaluation results comparing ECG-LoRA vs. Standard LoRA

Comp.	**Aspect**	**Win**	**Tie**	**Lose**
ECG-LoRA	Lex.	47.20%	26.80%	26.00%
vs.	Sem.	39.20%	34.30%	26.50%
Std LoRA	Emo.	46.30%	13.60%	40.10%

mensions. Specifically, our model achieves a dominant win rate of 47.20% in Lexical Fidelity, significantly outperforming the baseline (26.00% lose rate). This confirms that the emotion-conditioned gating successfully captures fine-grained stylistic nuances that static adaptation misses. In terms of Semantic and Emotional Fidelity, ECG-LoRA also maintains a consistent lead, demonstrating that the pursuit of emotional alignment does not compromise logical coherence or empathy. The high tie rates in Semantic Fidelity (34.30%) suggest that both models share the strong reasoning capabilities of the Qwen2 backbone, but ECG-LoRA adds crucial value in stylistic and emotional adaptation.

4.6 Case Study

Due to space constraints, we present a comprehensive qualitative analysis in the Appendix7.5, which includes multi-turn examples comparing ECG-LoRA with baselines across varying emotional intensities. These cases further illustrate how ECG-LoRA avoids generic responses in complex emotional contexts.

5 Discussion and Conclusion

In this work, we introduce ECG-LoRA, a parameter-efficient fine-tuning framework designed to enhance the conditional fidelity of empathetic dialogue generation. By incorporating an emotion-conditioned gating mechanism, ECG-LoRA enables large language models to dynamically modulate their generation behavior, resulting in outputs that more closely align with human reference patterns in both style and strategy.

Key Findings. Our experiments reveal three critical insights: (1) **Strategic Fidelity:** The model learns to shift from generic questioning to active emotional resonance, a hallmark of high-quality empathy. (2) **Interpretability:** The gating mechanism acts as a transparent "control knob," with activation levels highly correlated with emotional arousal. (3) **Mechanism over Memorization:** The differentiation in gradient allocation confirms that the model optimizes its internal representations based on emotional complexity, rather than simply memorizing training data.

Future Directions. The concept of "conditional fidelity" holds promise beyond empathetic dialogue.

1. **Generalization to Other Tasks:** The ECG framework can be adapted to other controllable generation scenarios, such as persona consistency, style transfer, or toxicity mitigation, where dynamic intervention based on input attributes is crucial.

2. **Fine-grained & Continuous Control:** Future work could extend the discrete emotion labels to continuous affect recognition (e.g., VAD regression) or handle mixed emotional states to capture the full spectrum of human experience.

3. **Multimodal Adaptation:** Incorporating non-verbal cues (audio prosody or facial expressions) into the gating network could further enhance the nuance and realism of embodied conversational agents.

6 Limitations

Although the proposed ECG-LoRA framework achieves remarkable results in controllable empathetic dialogue generation, several limitations remain:

6.1 Granularity and Representation of Emotional Conditions

Our study employs 32 discrete emotions as provided by the Empathetic Dialogues dataset. However, human emotions are inherently complex, continuous, and often mixed (e.g., feelings of "bittersweetness"). Discrete labels cannot fully capture the subtlety and mixture of real emotions, which may limit the model's ability to generate more nuanced and sophisticated responses.

Our framework currently conditions only on a single emotion label. In real dialogues, a person's state may be determined by multiple factors (such as emotion, intention, personality, and dialogue history). The model is not yet capable of handling such multidimensional and compound conditions.

Additionally, the current ECG-LoRA model is fundamentally limited when dealing with contexts lacking clearly defined emotions, resulting in a strong dependence of output quality on the accuracy of an external emotion classifier.

6.2 Limitations in Scope and Generalization

While our experiments demonstrate robust improvements across three LLM backbones on the Empathetic Dialogues benchmark, the evaluation remains confined to a single English dataset focused on general empathetic interactions. This restricts insights into the framework's applicability to more varied contexts, such as culturally diverse or domain-specific dialogues (e.g., mental health support or crisis counseling). Extending to multilingual benchmarks like CPED(Chen et al., 2022) or emotion-oriented datasets like EDOS(Harvey et al., 2017) could reveal potential biases or adaptations needed for broader generalization.

Our human evaluation, conducted on 128 samples with four annotators, provides valuable qualitative validation but is limited in scale and diversity. The annotators' backgrounds, while English-native, lacked extensive cultural variation, which may influence perceptions of empathy. Larger, more inclusive cohorts—ideally 10+ evaluators with global representation—would yield more reliable assessments.

Although we leverage direct metrics like Emp-F1(Ma et al., 2025) and draw on emotional fidelity benchmarks such as EmoCharacter (Feng et al., 2025) to move beyond proxy similarity measures, fully quantifying deep behavioral imitation in dynamic contexts remains an evolving challenge. Integrating hybrid approaches, such as psychological scales or advanced reward models, could address residual gaps in capturing intent-level fidelity without over-relying on reference-based evaluations.

6.3 Untapped Potential in Model Architecture

This work employs a simple scalar gate to modulate the entire LoRA module uniformly across all layers and dimensions. While this design ensures high parameter efficiency and straightforward implementation, it represents a deliberate simplification that limits the expressiveness of conditional adaptation. A scalar gate applies the same modulation factor globally, potentially constraining the model's ability to perform nuanced, dimension-specific adjustments tailored to complex emotional signals.

Future work could explore more sophisticated gating mechanisms to achieve finer-grained control. For instance, vector-based gates (e.g., element-wise or rank-wise modulation) (Kopiczko et al., 2024) (Meo et al., 2024) would allow selective activation of individual rank dimensions within LoRA matrices, enabling the model to emphasize emotionally salient directions while suppressing irrelevant ones. Similarly, matrix gates or block-wise adaptations (e.g., row-wise or column-wise scaling, inspired by block-diagonal constraints in distributed LoRA variants like BD-LoRA (Wang et al., 2025) or fine-grained fusion in AutoLoRA (Zhang et al., 2024) could facilitate layer-specific or projection-specific conditioning, better capturing hierarchical emotional representations.

These extensions would build on ECG-LoRA's lightweight foundation, potentially yielding more refined imitation of human-like stylistic variations in empathetic responses. Exploring orthogonality with gradient-based methods like GaLore (Zhao et al., 2024) could also improve training dynamics in continual or multi-task empathetic adaptation settings.

References

Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin, and Xiaodong He. 2023.

DiffusEmp: A diffusion model-based framework with multi-grained control for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2812–2831, Toronto, Canada. Association for Computational Linguistics.

Huiying Cao, Yiqun Zhang, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. 2024. Tool-ed: Enhancing empathetic response generation with the tool calling capability of llm. *Preprint*, arXiv:2412.03096.

Pradyumna Chari, Yunhao Ba, Shreeram Athreya, and Achuta Kadambi. 2022. Mime: Minority inclusion for majority group enhancement of ai performance. *Preprint*, arXiv:2209.00746.

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

Jean-Benoît Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*.

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

Hank H. Harvey, Ying Mao, Yantian Hou, and Bo Sheng. 2017. Edos: Edge assisted offloading system for mobile devices. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9.

Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *Proceedings of the 2008 IEEE Spoken Language Technology Workshop*, pages 109–112.

Edward J. Hu, Xuechen Li, Lillian Gordon, and Kyunghyun Cho. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jerry Yao-Chieh Hu, Mao Su, En-Jui Kuo, Zhao Song, and Hao Liu. 2024. Computational limits of low-rank adaptation (LoRA) for transformer-based models. *arXiv preprint arXiv:2406.03136*.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. Vera: Vector-based random matrix adaptation. *Preprint*, arXiv:2310.11454.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill 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, San Diego, California. Association for Computational Linguistics.

Zijing Liang, Zirui Fang, Yanjie Xu, Lin Liu, Yifan Hong, Ke Liu, and Penghui Shang. 2024. Low-rank adaptation with gating mechanisms in large language models, an improved method for fine-tuning: G-LoRA. In *Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024)*, volume 13291, page 132914E. International Society for Optics and Photonics, SPIE.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. 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 121–132, Hong Kong, China. Association for Computational Linguistics.

Qiujing Lu, Weiqiao Han, Jeffrey Ling, Minfa Wang, Haoyu Chen, Balakrishnan Varadarajan, and Paul Covington. 2022. Kemp: Keyframe-based hierarchical end-to-end deep model for long-term trajectory prediction. *Preprint*, arXiv:2205.04624.

Hui Ma, Bo Zhang, Bo Xu, Jian Wang, Hongfei Lin, and Xiao Sun. 2025. Empathy level alignment via reinforcement learning for empathetic response generation. *IEEE Transactions on Affective Computing*, 16(3):1873–1884.

Cristian Meo, Ksenia Sycheva, Anirudh Goyal, and Justin Dauwels. 2024. Bayesian-lora: Lora based parameter efficient fine-tuning using optimal quantization levels and rank values trough differentiable bayesian gates. *Preprint*, arXiv:2406.13046.

Saif M. Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *Preprint*, arXiv:2503.23547.

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 (ACL)*, pages 311–318.

Hannah Rashkin, Eric Jang, and Yejin Choi. 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.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. *Preprint*, arXiv:2109.05739.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *Preprint*, arXiv:1612.03975.

Xinyu Wang, Jonas M. Kübler, Kailash Budhathoki, Yida Wang, and Matthäus Kleindessner. 2025. Block-diagonal lora for eliminating communication overhead in tensor parallel lora serving. *Preprint*, arXiv:2510.23346.

Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional evaluation of empathetic dialogue responses. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2066–2087, Miami, Florida, USA. Association for Computational Linguistics.

Ruiyi Zhang, Rushi Qiang, Sai Ashish Somayajula, and Pengtao Xie. 2024. Autolora: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. *Preprint*, arXiv:2403.09113.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: Memory-efficient llm training by gradient low-rank projection. *Preprint*, arXiv:2403.03507.

7 Addpendix

7.1 Formula derivation

7.1.1 The Derivation of Two Loss Functions

Emotion Contrastive Loss. The purpose of introducing the emotion contrastive loss is to optimize the model's gating mechanism by encouraging similarity in gate signals for the same emotions while promoting separation for different emotions. This loss consists of a consistency loss and a separation loss, combined with weighted parameters.

Let $G = \{g_1, g_2, \ldots, g_N\}$ be the gate signal vectors for a batch of samples, where N is the batch size and $g_i \in [0, 1]$ represents the gate signal for the i-th sample; let $E = \{e_1, e_2, \ldots, e_N\}$ be the corresponding emotion ID vectors, where $e_i \in \{0, 1, \ldots, M-1\}$ and M is the number of emotion categories.

The same-emotion mask is defined as:

$$M_{\text{same}}(i,j) = \begin{cases} 1, & \text{if } e_i = e_j \\ 0, & \text{otherwise} \end{cases}$$

where $i, j \in \{1, 2, \ldots, N\}$.

The different-emotion mask is:

$$M_{\text{diff}}(i,j) = 1 - M_{\text{same}}(i,j)$$

The gate signal difference is:

$$D(i,j) = |g_i - g_j|$$

The consistency loss is:

$$L_{cons} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} M_{\text{same}}(i,j) \cdot D(i,j)$$

This measures the average absolute difference in gate signals between samples of the same emotion, encouraging it to approach 0.

The separation loss is split into two parts for clarity:

$$\text{Term 1} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} M_{\text{diff}}(i,j)$$

$$\text{Term 2} = \max(0, SM - D(i,j))$$

$$L_{sep} = \text{Term 1} \cdot \text{Term 2}$$

This uses a hinge loss form to ensure that the difference in gate signals between different emotion samples reaches at least a preset separation margin SM (separation margin).

The total contrastive loss is:

$$L_{con} = \omega \cdot L_{cons} + (1 - \omega) \cdot L_{sep}$$

where $\omega \in [0, 1]$ is a global parameter balancing the contributions of consistency and separation losses.

Emotion Embedding Regularization Loss. The emotion embedding regularization loss is introduced to optimize emotion vectors by enhancing intra-class compactness and inter-class separation, thereby improving the quality of emotion representations. This loss includes two components: intra-class compactness loss and inter-class separation loss.

Let $V = \{v_1, v_2, \ldots, v_N\}$ be the emotion embedding vectors for a batch of samples, where $v_i \in R^d$ (d is the embedding dimension) and $E = \{e_1, e_2, \ldots, e_N\}$ be the corresponding emotion ID vectors. Let $U = \{u_1, u_2, \ldots, u_M\}$ be the centroid vectors for each emotion category, where M is the number of unique emotion categories.

The intra-class compactness loss is calculated as follows: for each emotion category e_k ($k \in$

$\{0, 1, \ldots, M-1\}$), compute the mean vector $u_k = \frac{1}{N_k}\sum_{i:e_i=e_k} v_i$ (where N_k is the number of samples in that category). The loss is:

$$L_{intra} = \frac{1}{M}\sum_{k=1}^{M}\frac{1}{N_k}\sum_{i:e_i=e_k}\|v_i - u_k\|^2$$

This is computed only when $N_k \geq 2$, normalized by $\max(1, M)$ where M is the number of unique emotion categories.

The inter-class separation loss is calculated by first computing the Euclidean distance matrix of centroids $U = \{u_1, u_2, \ldots, u_M\}$:

$$D_{\text{inter}}(k, l) = \|u_k - u_l\|_2$$

$$k, l \in \{1, 2, \ldots, M\}, k \neq l$$

The loss is split into two parts:

$$\text{Term 1} = \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{l=k+1}^{M}$$

$$\text{Term 2} = \max(0, RM - D_{\text{inter}}(k, l))$$

$$L_{inter} = \text{Term 1} \cdot \text{Term 2}$$

where RM (regularization margin) is the preset separation margin.

The total regularization loss is:

$$L_{reg} = \lambda_{inter} \cdot L_{inter} + \lambda_{intra} \cdot L_{intra}$$

where λ_{inter} and λ_{intra} are the weight parameters for inter-class and intra-class losses, respectively.

By integrating all loss functions, the final optimization objective of the model is as follows :

$$L_{total} = L_{LM} + W_1 \cdot L_{con} + W_2 \cdot L_{reg}$$

Here, W_1 is the weight of the emotion contrast loss,W_2 is the weight of the sentiment embedding regularization loss.

7.2 Human Evaluation Procedure and Details

We conducted a systematic human evaluation to assess the faithfulness and lexical fidelity of AI-generated empathetic responses, using Google Forms as the survey platform. The evaluation process and rubric are described as follows:

Evaluator Role and Task. Each evaluator was instructed to act as an expert analyst, directly comparing two anonymized AI-generated responses (Response A and Response B) to a human-authored reference response for the same context and emotional label. The goal was not merely to judge which AI response is "better," but to assess how faithfully each AI model could imitate the linguistic and emotional style of authentic human examples.

Materials Presented to Evaluators. For each evaluation item, the following information was provided:

1. Context: The user's original utterance describing a scenario.

2. Emotion Label: The primary emotion expressed in the scenario.

3. Human Reference Response: A gold-standard response written by a human for the same context and emotion.

4. AI Responses: Two anonymous AI responses (A and B) generated for the same input, presented in randomized order to prevent bias.

Evaluation Criteria. Each AI response was independently rated on two key criteria, both using a 1–5 Likert scale:

1. Lexical Fidelity (Word Choice Matching):

- Measures how well the AI's vocabulary and phrasing align with the style and emotional tone of the human reference.

- 1 = Very Dissimilar; 5 = Highly Similar.

2. Overall Faithfulness (Holistic Imitation Quality):

- Assesses the holistic quality of imitation, considering both word choice and structural similarity. Evaluators judged how successfully the AI reproduced a human-like, emotionally appropriate response.

- 1 = Not Faithful at all; 5 = Highly Faithful.

Evaluators were provided with detailed scoring rubrics and examples for each level to ensure consistent interpretation of the criteria.

Forced-Choice Judgment. After rating both responses, evaluators were required to make a forced-choice selection on which AI response was a more faithful imitation of the human reference ("Response A," "Response B," or "Both/Neither").

Survey Interface. The survey was administered via Google Forms, with one evaluation scenario per page. Evaluators were advised to use a desktop or laptop for optimal readability and response speed.

Evaluator Instructions and Quality Control.

- All evaluators read detailed written instructions before starting the survey.
- Responses were collected independently. Evaluators were blind to model identities and the purpose of the comparison to minimize bias.
- Quality checks were performed to ensure that responses to the "forced-choice" question conformed to allowed options.

All evaluators were recruited through social media platforms and verified that English was their native language. Each evaluation session lasted a minimum of one hour per participant, and each was compensated with a payment of $5 for their time and effort.

7.3 Hyperparameters

As shown in Table 3, it is the description of the training hyperparameters for the complete ECG-LoRA model. The parameter descriptions for the ablation experiment have already been stated in the main text.

The explanations for each indicator in the quantitative metrics are as follows:

- **BLEU** scores were computed using the nltk.translate.bleu score.sentence bleu function from NLTK 3.8.1, with uniform weights (0.25, 0.25, 0.25, 0.25) for up to 4-grams and the smoothing function method4 to handle zero counts.
- **ROUGE** scores were calculated using rouge-score 0.0.4 with the RougeScorer class, evaluating ROUGE-1, ROUGE-2, and ROUGE-L metrics with stemming enabled (use stemmer=True).
- **BERTScore** was computed with bert-score 0.3.14, using the English default model with language set to "en" and rescaling enabled (rescale with baseline=True) to improve comparability.
- **Emotion Alignment F1 (EA-F1)** was computed as the macro-averaged F1-score between multi-label emotions extracted from reference and generated texts, using scikit-learn 1.2.2's f1 score and MultiLabelBinarizer. Emotions were predicted via the RoBERTa-based classifier [SamLowe/roberta-base-go emotions] with a threshold of 0.3 for multi-label inclusion; neutral fallback for errors.
- **Emotional Consistency (EC)** was computed as the cosine similarity between VAD (Valence-Arousal-Dominance) vectors of the ground-truth emotion label and the top predicted emotion in the generation (via the same RoBERTa classifier). VAD values were from a predefined NRC-VAD mapping; similarity was calculated as 1 - scipy.spatial.distance.cosine (scipy 1.10.1), clamped to [0, 1] for interpretability.
- **Emp-F1** was computed using the EmpRL

Table 3: Hyperparameters

Hyperparameter	Value
Base Model	Qwen2-7B-Instruct Llama-3-8B-Instruct Gemma-7B-it
LoRA Rank (r)	16
LoRA Alpha	16
LoRA Dropout	0.1
LoRA Injection Points	Q_{proj}, K_{proj} V_{proj}, O_{proj}
Emotion Embedding Dim	256
Number of Emotions	32
Optimizer	AdamW
Learning Rate (LoRA)	4×10^{-5}
Learning Rate (Gating)	1×10^{-4}
Batch Size	4
Gradient Accumulation	4 steps
Number of Epochs	8
Weight Decay	1×10^{-2}
Warmup Ratio	0.2
Patience (Early Stopping)	2
Loss Weights	W_{emo} = 5.0 W_{cons} = 0.5 Δ_{sep} = 0.2 W_{reg} = 0.2 Δ_{reg} = 1.0 λ_{inter} = 0.9 λ_{intra} = 0.1
Train Set Size	36629
Validation Set Size	5712

framework (cloned from GitHub and modified for tokenizer loading from 't5-base' via transformers 4.35.2). Separate T5-encoder classifiers (base size, 3 classes: No/Weak/Strong) predicted Emotional, Exploratory, and Interpretive empathy levels for generated outputs and human references, given contexts. Weighted F1-scores (scikit-learn 1.2.2, average='weighted') were calculated per dimension against references as ground truth, with overall Emp-F1 as their average. Batch inference (size=8) was used for efficiency.

- **Distinct-N (Dist-1/2)** was computed per generated sentence as the ratio of unique n-grams to total n-grams (n=1 or 2), using NLTK 3.8.1 for tokenization. Values range [0, 1], with higher indicating greater lexical diversity; short sentences (< n tokens) scored 0.0.

These settings ensure replicability and consistency with standard evaluation practices.

7.4 Implementation Details

Our implementation is based on PyTorch 2.1.0 with the Transformers 4.35.2 and PEFT 0.7.1 libraries. All experiments were conducted on a single NVIDIA A100 GPU (40GB VRAM) in a Google Colab environment, with mixed precision training (torch.bfloat16) to optimize memory usage. We used Weights & Biases (WandB) for logging metrics, losses, and hyperparameters during training.

7.4.1 Training Setup

The training script processes the Empathetic Dialogues (ED) dataset, loading train/validation splits from JSON files in a specified directory. Data is handled via a custom ScientificEDDataset class, which supports history-aware multi-turn dialogue formatting. Each sample is formatted as a chat template, concatenated history utterances (if any), the current user input, and the expected empathetic response as the label. Emotion labels are mapped to integer IDs using a predefined dictionary based on 32 categories from the NRC-VAD lexicon. Tokenization uses the model's tokenizer with a maximum sequence length of 256, right-padding, and truncation. Invalid or missing emotion labels are skipped.

Training runs for 8 epochs with a batch size of 4 and gradient accumulation over 4 steps (effective batch size 16). Early stopping is implemented with a patience of 2 epochs based on validation loss. The optimizer is AdamW with weight decay 1e-2, using separate learning rates: 4e-5 for LoRA parameters and 1e-4 for the gating network (emotion encoder and gate linear). The scheduler is cosine annealing with a warmup ratio of 0.2 (e.g., 20% of total steps). Losses include the standard causal LM loss, weighted emotion contrastive loss (5.0), and embedding regularization loss (0.2), with internal weights for consistency/separation (0.5) and inter/intra-class terms (0.9/0.1). Separation margins are 0.2 (contrastive) and 1.0 (regularization).

A custom ScientificLogger tracks gating signals and gradient norms every 50 steps, grouped by emotion arousal levels (e.g., Negative High, Positive High, Subtle/Neutral). Validation logs gate values per emotion. Checkpoints are saved per epoch, with the best model (lowest validation loss) preserved separately.

7.4.2 Model Configuration

The base model is Qwen2-7B-Instruct, Llama-3-8B-Instruct and Gemma-7B-it loaded with bfloat16 precision and auto device mapping. LoRA adapters are applied via PEFT with rank 16, alpha 16, dropout 0.1, targeting query, key, value, and output projections ($q_{proj}, k_{proj}, v_{proj}, o_{proj}$). A custom forward hook (gated lora forward) modulates LoRA updates with a scalar gate, computed from an EmotionEncoder (Embedding layer: 32 emotions to 256-dim vectors) followed by a linear projection and sigmoid activation. During training, global variables track emotion IDs for conditioning; overrides are used for causal interventions.

VAD mappings from NRC-VAD are used for emotion grouping and analysis, with high-arousal positives/negatives predefined for visualization.

7.4.3 Reproducibility

Dataset preprocessing assumes standard ED JSON format. Training takes approximately 8 hours per backbone; inference approximately 2 hours for the full test set (approximately 5K samples). Offload caching handles memory for larger models.

7.5 Case Study

To complement the quantitative evaluations presented in the main text, we conduct a detailed qualitative analysis here, focusing on multi-turn empathetic dialogues. As noted in Section 4.6, this extended examination illustrates how ECG-LoRA achieves superior conditional emotional fidelity by

Table 4: Case study examples

Emotion	Model	Response
Excited	Input	My husband and I are planning a big trip to Washington, DC.
	Standard LoRA	Nice! Are you going for vacation?
	ECG-LoRA	That sounds like so much fun! What all do you **want to see** there?
	Ground Truth	Oh, that sounds so exciting! What do you **plan to see** while you are there?
Devastated	Input	"My dog got hit by a car."
	Standard LoRA	"Oh no, is your dog okay?"
	ECG-LoRA	"Oh no! I'm **sorry to hear that**. Is he okay?"
	Ground Truth	"Aw! That's horrible, I'm **sorry to hear that**."
Nostalgic	Input	"When I was in college I travelled all over Europe one semester."
	Standard LoRA	"That sounds amazing! I wish I had the chance to do that."
	ECG-LoRA	"That sounds like a great time. **What was your favorite place** you visited?"
	Ground Truth	"Sounds like quite an adventure for one semester! **What was your favorite place**?"
Caring	Input	"My brother just turned 16 and he's about to get his first car! I'm so excited for him."
	Standard LoRA	"That is awesome, it is a great feeling being able to drive."
	ECG-LoRA	"That's awesome! **What kind of car** is it?"
	Ground Truth	"Whoa that's exciting! **What kind of car** we looking at?"
Content	Input	"My husband and I celebrated 9 years Saturday. It was such a good day of quality family time."
	Standard LoRA	"That is so sweet!"
	ECG-LoRA	"That's awesome! How **long** have you been **together**?"
	Ground Truth	"Oh that's great! 9 years **together** is a **long** time. You sound happy, I'm glad."

dynamically adapting response styles and strategies to varying emotional intensities, thereby avoiding the generic or shallow outputs often produced by baselines like Standard LoRA. The selected cases in Table 4 span a range of emotions (Excited, Devastated, Nostalgic, Caring, and Content), drawn from the Empathetic Dialogues test set. For each, we compare responses from Standard LoRA, ECG-LoRA, and human ground truth, highlighting key differences in empathy expression, exploratory engagement, and interpretive alignment (as measured by Emp-F1 dimensions).

In the "Excited" case, the user's input conveys anticipation for a trip. Standard LoRA's response ("Nice! Are you going for vacation?") is superficial, merely acknowledging the fact without probing deeper, which lacks emotional resonance and exploratory intent. In contrast, ECG-LoRA ("That sounds like so much fun! What all do you want to see there?") mirrors the human ground truth by expressing enthusiasm ("so much fun!" akin to "so exciting!") and inviting elaboration through an open-ended question, demonstrating enhanced exploratory empathy (Exp dimension in Emp-F1) and fidelity to human-like curiosity.

For "Devastated," involving a tragic event, Standard LoRA ("Oh no, is your dog okay?") offers basic sympathy but remains neutral and fact-focused, potentially underplaying the emotional weight. ECG-LoRA ("Oh no! I'm sorry to hear that. Is he okay?") intensifies empathy with explicit condolence ("I'm sorry"), aligning closely with the ground truth's horrified tone ("Aw! That's horrible, I'm sorry"), thus improving emotional empathy (Emo dimension) and avoiding detachment in high-arousal negative contexts.

The "Nostalgic" example shows a reflective input about past travels. Standard LoRA ("That sounds amazing! I wish I had the chance to do that.") shifts focus to the model's hypothetical envy, introducing self-referential bias that dilutes user-centered empathy. ECG-LoRA ("That sounds like a great time. What was your favorite place you visited?") stays user-focused, encouraging reminiscence similar to the ground truth ("Sounds like quite an adventure... What was your favorite place?"), enhancing interpretive empathy (Int dimension) by facilitating deeper emotional sharing.

In the "Caring" scenario, the input expresses familial pride. Standard LoRA ("That is awesome, it is a great feeling being able to drive.") generalizes to abstract benefits, missing personal connection. ECG-LoRA ("That's awesome! What kind of car is it?") engages specifically with the milestone, echoing the ground truth's excitement and curiosity ("Whoa that's exciting! What kind of car we looking at?"), which boosts overall empathetic consistency.

Finally, **for "Content**," Standard LoRA ("That is so sweet!") is brief and generic, lacking depth. ECG-LoRA ("That's awesome! How long have you been together?") probes relational history, though slightly less interpretive than the ground truth ("Oh that's great! 9 years... You sound happy, I'm glad."), it still outperforms the baseline in fostering sustained dialogue.

These cases reveal a pattern: ECG-LoRA's gating mechanism enables nuanced adaptation, resulting in responses that are more semantically and emotionally aligned with human references (e.g., higher BERTScore and Emp-F1). While not all instances achieve perfect fidelity—e.g., in "Content," ECG-LoRA omits explicit happiness acknowledgment—overall, it reduces genericism in 78% of analyzed samples (from a random subset of 50 test instances), as verified through manual annotation. Future work could explore failure modes, such as ambiguous emotions, to further refine the framework.

7.6 Data, Code and Reproducibility Statement

7.6.1 Data Availability

The EmpatheticDialogues dataset (Rashkin et al., 2019) comprises 25K emotional conversation pairs released under the CC BY-NC license. Use is permitted strictly for non-commercial research or educational purposes, with required attribution to the original source. Commercial use, redistribution, or incorporation into commercial products is not allowed without separate authorization.

We have utilized this dataset in compliance with the norms. The dataset is available at: facebookresearch/EmpatheticDialogues.

7.6.2 Code Availability and Reproducibility Statement

Code repository: Anonymous ACL submission

The repository includes:

- Model implementation (including the ECG-LoRA framework, training scripts, and evaluation scripts);
- Reproducibility statement;

- Instructions for environment setup and package dependencies;
- Scripts for data preprocessing, training, inference, and evaluation;
- Sample configuration files and example command lines.