How to Crack the Black Box of Emotion Generation? Interpretable Dual-Path Function Vectors for Token-Level Control

Anonymous authors
Paper under double-blind review

Abstract

Existing empathetic dialogue generation methods face limitations in the fine-grained control and cross-modal transfer of emotion expression, while generally lacking interpretability into the models' internal emotion generation mechanisms. We propose an interpretable dual-path emotional functional vector framework, integrated with a token-level adaptive injection strategy, to achieve transparent and controllable empathetic dialogue generation. Specifically, we design a mean difference pathway to extract semantic emotional features and a singular value decomposition pathway to extract parameter-structural emotion features, achieving dual-path fusion via adaptive weights. The token-level strategy dynamically assigns regulation weights based on an attention mechanism, matching the differential contributions of different words in emotional transmission. Additionally, we construct a text-to-speech cross-modal evaluation system to verify the consistency of emotional information across multimodal scenarios. Experimental results on the ED and ESConv datasets demonstrate that the proposed method significantly outperforms all existing state-of-the-art baselines. When using Llama-3.1 as the backbone model, we achieve a 18.22% improvement in Acc2 and a 6.02% improvement in F_Bert. Our work provides an interpretable and efficient solution for controllable emotion generation in LLMs.

1 Introduction

Empathetic dialogue generation aims to enable machines to understand users' emotion states and generate emotionally resonant responses, holding significant application value in fields such as intelligent customer service Chen et al. (2019); Leocádio et al. (2024), emotion companionship Loveys et al. (2022); Mitchell & Jeon (2025), and mental health support Ma et al. (2024); Li et al. (2025). Existing research primarily enhances the empathetic quality of responses by incorporating commonsense knowledge Huangfu et al. (2025); Cai et al. (2023) or optimizing generation strategies Wang et al. (2025a); Li et al. (2020), achieving remarkable progress in dialogue fluency and emotion alignment. While these methods generally suffer from a critical limitation: insufficient focus on the interpretability of models' internal emotional decision-making mechanisms. Operating as black boxes, they fail to clearly explain the rationale and process behind activating specific emotion features particular contexts. This restricts our ability to deeply understand and finely control the model's emotion expression.

In the field of interpretability research for large language models (LLMs), a variety of techniques have been proposed and applied, such as linear probe Arditi et al. (2024); Di Palma et al. (2025) and sparse autoencoder Lieberum et al. (2024); Rimsky et al. (2024). These methods primarily focus on tracing influential features backward from model outputs or extracting superficial correlations through structural decomposition of the network, and have played critical roles in domains like model safety Gu et al. (2025) and parameter editing Zhang et al. (2024). Nevertheless, they remain largely centered on correlational analyses between a model's external performance and input data, overlooking the exploration and application of the inherent functional vectors embedded within the model itself. As a result, they have not yet delved into the core vector mechanisms underlying emotion generation from the essential dimensions of the representation layer and intrinsic feature space.

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

We rethink the interpretability and emotion controllability of LLMs in text generation tasks from the perspective of emotion vectors. Based on a series of exploratory experiments (see Section 3 for details), we find that emotion vector extraction solely at the representation level is susceptible to corpus noise, while relying merely on parameter structures makes it difficult to capture fine-grained emotional semantic differences. Motivated by these observations, we propose *a dual-path functional vector framework for token-level emotion control to facilitate more effective empathetic dialogue generation.* Specifically, we design a synergistic and complementary dual-path architecture that integrates a mean difference (MD) optimization pathway at the representation layer and a singular value decomposition (SVD) enhancement pathway at the parameter layer. This dual-path architecture enables precise localization of emotion functional vectors, ensuring the interpretability of emotion vectors in both semantics and structure. Furthermore, we observe that traditional global vector injection methods lack differentiation in regulating non-emotional words and emotional key words, leading to insufficient emotional expression. Thus, we design an attention-based token-level adaptive injection strategy, which dynamically computes alignment weights between each token and the emotion vector, enabling fine-grained and interpretable fusion and regulation of emotional vectors.

Additionlly, existing text generation evaluation metrics Wang et al. (2025b); Li et al. (2022) are limited in real-world multimodal interactions. Since human emotion communication inherently integrates text, speech, and other multimodal signals, text-only metrics cannot fully assess emotional authenticity and persuasiveness, resulting in evaluation blind spots. To address this issue, we construct a text-to-speech (TTS) cross-modal emotion transfer pipeline and multi-dimensional evaluation system. It generates emotionally consistent speech, establishes cross-modal criteria, and provides interpretable, multi-dimensional validation of emotion expression in real-world scenarios.

We conduct experiments on two widely used empathetic dialogue generation datasets, EmpatheticDialogues (ED) Rashkin et al. (2019) and Emotional Support Conversation (ESConv) Liu et al. (2021). Experimental results demonstrate that our model significantly outperforms all state-of-the-art (SoTA) baselines across all evaluation metrics in both text and speech dimensions. Specifically, by integrating our method with Llama-3.1 as the base backbone model, we achieve a remarkable 18.22% improvement in emotion accuracy (Acc2) and a 6.02% gain in semantic matching (F_Bert). Furthermore, we perform extensive experiments to verify the effectiveness of the dual-path functional vector extraction and token-level adaptive injection strategy.

Our main contributions are summarized as follows:

- We propose a dual-path framework for extracting emotion function vectors across the representation and parameter layers, coupled with a token-level adaptive regulation strategy, to achieve controllable empathetic dialogue generation.
- We construct a text-to-speech cross-modal evaluation system to address the limitation of text-only assessment in capturing emotion authenticity, thereby establishing a more comprehensive measurement standard.
- Our extensive experimentation on the ED and ESConv datasets demonstrate that our model achieves SoTA performance.

2 Related Works

2.1 Empathetic Dialogue Generation

Empathetic dialogue generation aims to enable models to accurately understand users' emotion states and generate emotionally resonant responses. Existing research mainly focuses on two core directions: enhancing emotion understanding Cai et al. (2023); Li et al. (2024); Xie et al. (2025) and optimizing generation strategies Lin et al. (2020); Yan et al. (2024); Wang et al. (2025b). In terms of emotion understanding, the primary focus is on the explicit modeling of emotional features. Rashkin et al. (2019) construct EmpatheticDialogues, the first large-scale emotion dialogue dataset. Wang et al. (2025a) propose a commonsense knowledge framework to mine implicit needs and improve empathy quality. Regarding the optimization of generation strategies, parameter-efficient fine-tuning and knowledge enhancement have become the mainstream technical paths. Sabour et al. (2022) introduce a commonsense knowledge graph to enhance the model's ability to reason. Huangfu et al.

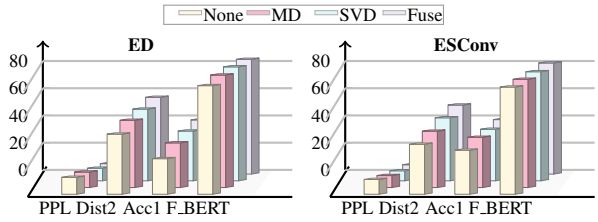


Figure 1: Performance comparison of empathetic dialogue generation under different function vector extraction methods on ED and ESConv datasets.

(2025) present a novel framework based on contrastive learning, context-sensitive entities, and social commonsense.

2.2 Functional Vectors

Functional vectors are a vectorization method for interpreting and intervening in the internal representations of neural networks, which have garnered widespread attention in the fields of explainable artificial intelligence and controlled text generation Hendel et al. (2023); Fu et al. (2025). Early researches primarily focuse on feature disentanglement and editing. Bau et al. (2019) achieve the editing of image attributes through vector operations in the latent space of GANs. Dathathri et al. (2020) propose controlling text attributes by updating hidden layer states via gradients. Subsequently, researchers begin exploring lightweight functional vector injection methods that do not require gradient updates. Turner et al. (2024) intervene model inference via concept activation difference vectors. Todd et al. (2024) verify functional vector mechanisms in autoregressive Transformers. Jiang et al. (2025) combine functional vectors with regularization to mitigate catastrophic forgetting. However, existing methods only use single extraction pipelines and crude injection strategies, neglecting varied emotional contributions across tokens.

3 Observation and Key Intuition

Functional vector techniques Todd et al. (2024); Zhang et al. (2024) provide a novel pathway toward transparent and controllable model behavior by intervening in the internal representations of neural networks. While their effectiveness in emotion generation tasks remains unclear. Inspired by the pursuit of interpretability and controllability in emotion analysis, we pose a key research question: *can we directly locate and regulate the functional units within the model that are associated with emotion generation?* We conduct a series of analyses aimed at developing a robust emotional vector extraction method, thereby enabling interpretable and fine-grained control over generated content.

Given that the essence of emotion generation lies in characterizing differences in emotion tendencies at the semantic level, *we mine emotion-related vectors from the representation layer by comparing semantic representations of emotion corpora.* Specifically, we design a representation-level mean difference (MD) extraction scheme. We construct a three-level balanced emotion corpus (D_{+core}, D_{+ext}, D_{-}), calculate the emotion functional vector as the mean difference of hidden layer representations between positive and negative emotion corpora, and inject it globally using a uniform coefficient. In Figure 1, experimental results show that this scheme achieves initial success. Compared to the baseline without vector injection, the emotion accuracy (Acc1) is improved by 6.66%, and the text naturalness (PPL) is also significantly better than the baseline. However, we find that this scheme only focuses on semantic-level emotion correlations, and the inherent structural features related to emotional generation in the model parameters are not fully utilized, which may lead to over-reliance on emotion corpora.

To address the limitation of the representation-level MD path's over-reliance on emotional corpora, we ask: *do the structural features of model parameters contain key information for emotion generation?* If emotion functional vectors can be mined from the parameter level, it may be possible to reduce dependence on corpus quality while complementing semantic information from the representation layer. Based on this hypothesis, we design a parameter-level singular value decomposition (SVD) extraction scheme. We perform SVD on the weight matrices of the MLP layers in the large

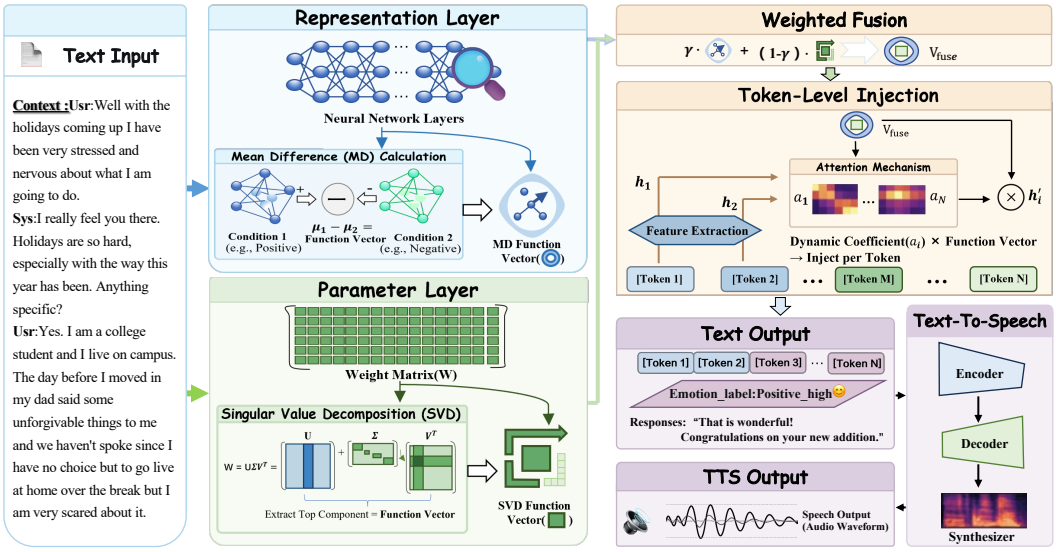


Figure 2: The overall architecture of our framework.

language model, identify sentiment-critical vectors in the parameter space through a sentiment-utility co-quantification mechanism, and inject them with the same globally unified coefficient. As shown in Figure 1, the results validate the effectiveness of this approach, with Acc1 improved by 10.41% compared to the baseline.

These two explorations demonstrate that the representation-level MD path excels at capturing semantic emotional orientation, while the parameter-level SVD path is adept at mining structural emotional features. Based on this, we propose our core hypothesis:

Combining semantic emotion information from the representation layer with structural emotion information from the parameter layer, we construct a dual-path fused functional vector extraction framework. This overcomes inherent limitations of single-path methods and achieves threefold optimization in emotion precision, scenario adaptability, and utility retention.

To verify this hypothesis, we perform weighted fusion of the two schemes and inject the combined vector using a global uniform coefficient. The results in Figure 1 show that the fused vector achieves significantly better performance than either single vector, validating the effectiveness of dual-path fusion. We formalize the emotion functional vector fusion and the token-level regulation into two theorems, thereby providing theoretical support for the dual-path fusion framework and the fine-grained regulation strategy (For detailed proof, see Appendix **??**).

Theorem 1. *Emotion Functional Vector Semantic-Structural Complementarity Theorem.* If the representation-level emotional vector V_{MD} satisfies emotion semantic distinctiveness and the parameter-level emotion vector V_{SVD} satisfies utility compatibility, then there exists a unique optimal fusion weight $\gamma \in [0, 1]$ such that the fused vector $V_{fuse} = \gamma V_{MD} + (1-\gamma)V_{SVD}$ outperforms either individual vector in terms of emotion regulation performance.

Theorem 2. *Token-Level Heterogeneity of Emotion Contribution Theorem.* In empathetic dialogue text, the contribution of different token types to emotional expression exhibits a hierarchical distribution property: the emotional contribution of emotional core words is greater than that of emotional modifiers and function words, and this distribution remains stable across different emotion types.

4 METHODOLOGY

To address functional vectors' single-path extraction and inadequate emotion expression precision, we propose a dual-path framework for token-level emotion control. As illustrated in Figure 2, we extract emotion function vectors via an MD-SVD dual-path mechanism. Leveraging the token-

level heterogeneity of emotion contribution theorem, we design an adaptive weight strategy for fine-grained emotion vector injection. Finally, we extend to cross-modal TTS processing and construct a full-link evaluation system.

4.1 Dual-path Emotion Functional Vector Extraction

Mean Difference Optimization Path. We design a mean difference extraction pathway with emotion semantic enhancement to accurately capture hierarchical differences in emotion semantics and enhance the emotion orientation specificity of functional vectors.

To this end, we first construct a three-level hierarchical emotion corpus, which is tailored to disentangle core emotional signals from contextual interference while preserving realistic interaction characteristics. Specifically, the corpus comprises: (i) A target emotion core corpus D_{+core}, containing only texts with pure target emotion expressions (free of redundant contextual noise) to anchor the fundamental emotional orientation. (ii) A target emotion extended corpus D_{+ext}, consisting of daily dialogue texts embedded with the target emotion to retain natural interactive contexts. (iii) A non-target emotion corpus D_{-}, encompassing texts with emotions opposite to or irrelevant to the target emotion to establish a clear emotional contrast. To mitigate sample distribution bias that could skew the learning process, we ensure the balance of corpus sizes as $|D_{+core}| + |D_{+ext}| = |D_{-}|$.

Building on this hierarchical corpus, we then perform hierarchical representation extraction: we feed the three corpus types into the base LLM separately, extract token-level hidden representations from the residual connection layer (a choice motivated by its ability to preserve fine-grained semantic information), and compute the average hidden representations for each corpus, denoted as X_{+core} (for D_{+core}), X_{+ext} (for D_{+ext}), and X_{-} (for D_{-}).

We further calculate a semantically enhanced positive emotion representation to integrate core emotion purity and contextual adaptability. Given the greater importance of core emotion signals in defining orientation specificity while the need to retain contextual adaptability for real-world applicability, we introduce hierarchical weights $\lambda_{core} = 0.7$ and $\lambda_{ext} = 0.3$ to balance these two aspects. The enhanced positive representation is thus computed as $X_{+enhance} = \lambda_{core} \times X_{+core} + \lambda_{ext} \times X_{+ext}$. Finally, the emotion functional vector is defined as:

$$V_{MD} = X_{+enhance} - X_{-} \tag{1}$$

This design leverages the contrast between enhanced positive emotions and non-target emotions to amplify emotional discriminability, enabling the functional vector to retain distinct emotion orientation while adapting to realistic conversational scenarios.

Singular Value Decomposition Augmentation Path. To synergistically optimize emotion feature preservation and utility protection, we design an emotion-utility co-quantified parameter-level SVD extraction pathway.

First, we decompose the model's parameter structure to isolate core feature directions. Specifically, we target the weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$ of the LLM's MLP layer—chosen for its role in modeling complex semantic and functional relationships—and perform SVD, yielding $W = U\Sigma V^T$. Here, $U = [u_1, u_2, ..., u_k]$ denotes the left singular vector matrix, $V = [v_1, v_2, ..., v_k]$ the right singular vector matrix, and Σ the diagonal matrix of singular values. Crucially, these singular vectors encode the dominant feature directions in the parameter space, making them ideal candidates for emotion-specific refinement.

Next, we construct dual-dimensional projection matrices to quantify emotion relevance and utility preservation separately. For the emotional dimension, we leverage a weakly supervised emotion calibration corpus D_{emo}, which includes texts annotated with weak, moderate, and strong emotional intensities, this design ensures the capture of graded emotion nuances rather than binary distinctions. From D_{emo}, we extract input activations $X_{in} \in \mathbb{R}^{d_{in} \times n}$ and solve for the low-rank matrix $\widehat{W}_{emo}$ (critical to emotion expression) by minimizing the Frobenius norm of output activation changes:

$$\widehat{W}_{emo} = \arg \min_{\text{rank} \leq r} \left\| W X_{in} - \widehat{W}_{emo} X_{in} \right\|_F^2 \tag{2}$$

We then perform SVD on the output activations of $\widehat{W}_{emo}$, take the top r left singular vectors to form U_{emo}, and derive the emotion projection matrix:

$$\Pi_{emo} = U_{emo} U_{emo}^T \tag{3}$$

This matrix effectively isolates the low-rank subspace for emotion expression. For the utility dimension, we select a general corpus D_{util} covering daily dialogue, knowledge QA, logical reasoning, scenarios reflecting the model's core utility requirements. Following the same procedure, we construct the utility projection matrix $\Pi_{util} = U_{util}U_{util}^T$, where U_{util} denotes the top r left singular vectors of activations from D_{util}. This matrix anchors the subspace critical to preserving general language capabilities.

Table 1: Experimental results of empathetic dialogue generation on text-level evaluation metrics for the ED and ESConv datasets. The downward arrow (↓) indicates that a lower value of the metric corresponds to better performance, while all other metrics indicate better performance with higher values.

	ED						ESConv					
Method	**PPL ↓**	**Dist1**	**Dist2**	**Acc1**	**Acc2**	**F_BERT**	**PPL ↓**	**Dist1**	**Dist2**	**Acc1**	**Acc2**	**F_BERT**
EmpDG	37.43	0.46	1.99	18.66	23.09	68.74	30.64	2.03	5.37	20.07	19.73	66.14
CEM	36.33	0.62	2.39	19.25	23.63	69.52	28.82	2.12	6.53	21.63	21.32	67.92
DCKS	16.08	2.19	9.61	20.03	24.36	70.17	18.26	3.41	10.16	25.16	24.86	69.37
NEC	**12.89**	**3.15**	13.91	22.37	26.55	73.28	15.31	3.68	12.89	27.58	26.36	72.13
Sibyl	5.12	2.55	9.67	24.52	28.92	76.42	8.03	4.49	16.34	30.76	29.74	76.75
llama	12.41	15.39	44.15	26.13	33.01	79.78	10.85	8.52	36.68	32.31	31.79	78.72
Sibyl + Ours	3.95	7.26	19.33	31.56	41.24	80.63	5.88	7.65	38.76	39.48	37.43	82.21
Ours	**3.01**	**19.45**	**59.08**	**43.12**	**51.23**	**85.78**	**4.43**	**14.16**	**55.38**	**44.12**	**44.62**	**83.73**

Building on this, we perform collaborative scoring to select emotion-critical vectors while safeguarding utility. First, we quantify the emotion importance of each left singular vector u_i as $s_{emo,i} = u_i^T \Pi_{emo} u_i$, a metric that reflects how well u_i aligns with the emotional subspace and thus its contribution to emotion expression. We simultaneously compute its utility importance as $s_{util,i} = u_i^T \Pi_{util} u_i$, which measures the vector's retention in the utility subspace and its role in maintaining general performance. To balance these two objectives, we design a collaborative scoring function:

$$score_i = s_{emo,i} - \beta s_{util,i} \tag{4}$$

where the utility protection coefficient β suppresses vectors that would overly compromise utility while amplifying those with strong emotional contributions. Finally, we select the top-k singular vectors by $score_i$ to form the emotion-critical singular vector set $U_{emo-key} = [u_1, u_2, ..., u_k]$, paired with the corresponding right singular vector set $V_{emo-key} = [v_1, v_2, ..., v_k]$. We construct the emotion functional vector as $V_{SVD} = [U_{emo-key}, V_{emo-key}]$, fusing left and right singular vectors to fully preserve the parameter-level structural information underlying emotion expression.

Dual-path Adaptive Weighted Fusion. We adopt an adaptive weighted fusion strategy to combine V_{MD} and V_{SVD} into the final emotion functional vector V_{fuse}:

$$V_{fuse} = \gamma \times V_{MD} + (1 - \gamma) \times V_{SVD} \tag{5}$$

where γ is a dynamic fusion coefficient, adaptively adjusted based on the emotional accuracy on the validation set. This enables the synergistic enhancement of emotion semantics from the representation layer and emotion structure from the parameter layer.

4.2 Token-level Adaptive Injection Strategy

To address the one-size-fits-all limitation of global emotion vector injection, we design an attention-based token-level regulation strategy enabling refined emotion vector injection

We first extract the token hidden state matrix $H = [h_1, h_2, ..., h_n]$ from the model's target layer, where n denotes sequence length and $h_i \in \mathbb{R}^d$ represents the i-th token's hidden state vector. These hidden states encode fine-grained semantic information of individual tokens, laying the foundation for discriminative regulation. We then leverage the fused functional vector V_{fuse} as emotion query vector q_{emo}. To quantify alignment between each token and target emotional orientation, we adopt scaled dot-product attention to compute the adaptive regulation coefficient:

$$\alpha_i = \text{Softmax}\left(\frac{h_i \cdot q_{emo}^T}{\sqrt{d}}\right) \tag{6}$$

Table 2: Experimental results on speech-level and human-level evaluation metrics in ED and ESConv datasets.

		Speech-Level								Human-Level		
Dataset	Method	V	B-S	BLEU	TD ↓	UA	WA	F1	L2 ↓	Emp	Rel	Flu
ED	Sibyl	15.79	68.76	8.47	48.96	17.06	17.32	12.28	2.15	3.66	3.79	4.13
	Sibyl +Ours	19.91	73.59	14.88	45.67	20.99	22.23	18.12	1.89	4.37	4.28	4.32
	Llama	33.28	70.63	11.21	47.12	17.17	17.11	14.17	2.27	3.74	3.82	3.96
	Ours	**35.76**	**74.84**	**17.62**	**45.65**	**28.79**	**29.14**	**25.15**	**1.78**	**4.59**	**4.61**	**4.72**
ESConv	Sibyl	15.42	69.42	12.74	51.17	11.98	23.19	27.38	2.02	3.57	3.69	3.72
	Sibyl + Ours	22.25	73.83	18.15	48.71	18.42	30.26	33.21	1.54	4.03	4.17	4.21
	Llama	16.03	70.83	15.82	50.57	17.65	25.13	28.89	1.66	3.58	3.82	3.74
	Ours	**25.03**	**74.57**	**18.96**	**46.63**	**24.27**	**34.87**	**39.47**	**1.42**	**4.26**	**4.38**	**4.35**

where d is the dimension of the token hidden state, dividing by $\sqrt{d}$ serves to mitigate the scaling effect of high-dimensional dot-product results, ensuring gradient stability during attention computation. Finally, we dynamically adjust each token's hidden state using the weighted functional vector, realizing refined regulation of emotional expression:

$$h_i' = h_i + \alpha_i \cdot V_{fuse} \tag{7}$$

This design inherently links regulation intensity to token-level emotional relevance: emotion-critical tokens receive higher α_i values and thus stronger emotional enhancement, while non-emotion tokens retain moderate regulation.

4.3 Text to Speech Cross Modal Extension

We construct an XTTS-based cross-modal extension framework to ensure faithful emotional transmission across text and speech. The emotionally refined text from the token-level adaptive regulation module serves as input to the XTTS model. Leveraging XTTS's inherent strengths in prosody modeling and multi-language adaptation, we enhance its emotional embedding layer to encode text-based emotional cues, mapping emotional intensity and tonal features to core acoustic attributes (pitch f_0, speech rate r, volume v) via:

$$[f_0, r, v] = \text{XTTS}(\text{Emb}_{emo}(T), \theta_{emo}) \tag{8}$$

where $\text{Emb}_{emo}(T)$ denotes the emotion embedding of input text T, and θ_{emo} represents the fine-tuned parameters for emotion alignment. Synchronous output of the emotion text and high-fidelity synthesized speech not only delivers an immersive empathetic interaction experience but also provides a complete carrier for cross-modal emotion evaluation.

5 Experiments

Table 3: Ablation results of empathetic dialogue generation on ED and ESConv datasets.

	Text–Level						Speech–Level							
	PPL ↓	Dist1	Dist2	Acc1	Acc2	F_BERT	V	B-S	BLEU	TD ↓	UA	WA	F1	L2 ↓
Ours(ED)	**3.01**	**19.45**	**59.08**	**43.12**	**51.23**	**85.78**	**35.76**	**74.84**	**17.62**	**45.65**	**28.79**	**29.14**	**25.15**	**1.78**
w/o Token	7.68	18.55	56.37	39.81	47.39	84.27	35.02	73.69	15.48	45.98	26.51	26.82	22.96	1.93
w/o MD	8.54	17.97	53.94	37.69	46.13	83.79	34.71	73.02	14.86	46.27	24.39	23.96	21.73	1.98
w/o SVD	10.36	17.25	50.37	34.72	43.86	82.66	34.13	72.53	13.29	46.49	21.65	21.73	19.65	2.06
Ours(ESConv)	**4.43**	**14.16**	**55.38**	**44.12**	**44.62**	**83.73**	**25.03**	**74.57**	**18.96**	**46.63**	**24.27**	**34.87**	**39.47**	**1.42**
w/o Token	6.86	12.11	50.69	40.03	39.87	81.63	22.49	73.36	17.32	47.33	22.68	31.66	36.55	1.48
w/o MD	7.59	10.79	47.24	38.59	37.06	80.74	20.75	72.54	16.71	48.55	20.84	30.45	34.92	1.55
w/o SVD	8.47	10.13	42.77	37.24	36.28	79.97	19.32	71.95	16.19	48.98	19.77	28.68	32.27	1.59

5.1 Experimental Setup

Settings. We adopt the source large language model Llama-3.1-8B-Instruct as the testing tool.

Datasets. We evaluate the effectiveness of our model on two widely used empathetic dialogue generation datasets, EmpatheticDialogues (ED) Rashkin et al. (2019) and Emotional Support Conversation (ESConv) Liu et al. (2021).

5.2 Baseline Systems

To verify the effectiveness of our model, we compare it with the following SoTA baselines.

- Classic baseline methods: EmpDG Li et al. (2020), CEM Sabour et al. (2022), DCKS Cai et al. (2023), NEC Huangfu et al. (2025), and Sibyl Wang et al. (2025a).
- Backbone model baseline: we use Llama-3.1 as the base backbone model, taking its performance as a reference to demonstrate the optimization effect of our method on a strong model foundation.
- Generalization validation of the method: we integrate our functional vector method into an existing SoTA empathetic dialogue generation model(Sibyl + Ours).

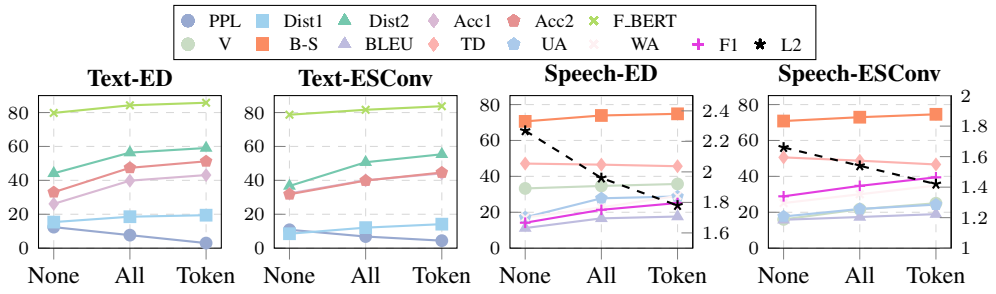


Figure 3: Comparison results of different injection strategies. For the Speech-ED and Speech-ESConv subplots, the L2 metric uses the right-hand y-axis scale, while all other metrics use the left-hand y-axis scale.

5.3 Main Results

Table 1 and Table 2 report text, speech and human evaluation results respectively. Our method surpasses all SOTA baselines on all metrics over two datasets, with three core conclusions as follows. First, consistent gains on all text and speech metrics prove our dual-path semantic emotion vector extraction, structured emotional parameters and token-level adaptive control can accurately capture subtle dialogue emotion intentions. Second, experiments based on Llama3 show our model lifts Acc2 by 18.22% and F_BERT by 6.02% with low PPL, verifying the effectiveness and compatibility of our functional vector framework for LLM emotional optimization. Furthermore, the combined "Sibyl + Ours" model outperforms original Sibyl, demonstrating strong cross-model generalizability of our approach.

5.4 Ablation Study

In Table 3, we design ablation experiments to verify the necessity of each core module in our framework on ED and ESConv datasets. Removing token-level adaptive injection degrades all metrics, as it balances emotional word adjustment and semantic integrity via dynamic token weights. When the MD optimization path is removed, the model exhibits an overall and significant performance decline. This fully confirms that the pathway effectively uncovers both explicit and implicit emotional cues through hierarchical corpora, even across diverse datasets. Eliminating the SVD enhancement path leads to notable performance drops. In the ED dataset, Acc2 and F_BERT fall by 7.37% and 3.12% in text metrics, while WA and F1 decrease by 7.41% and 5.50% in speech metrics. This shows that emotion-utility co-quantified SVD preserves core emotion functions without disrupting general utility. Its parameter emotional features complement MD semantic vectors to boost emotional generation.

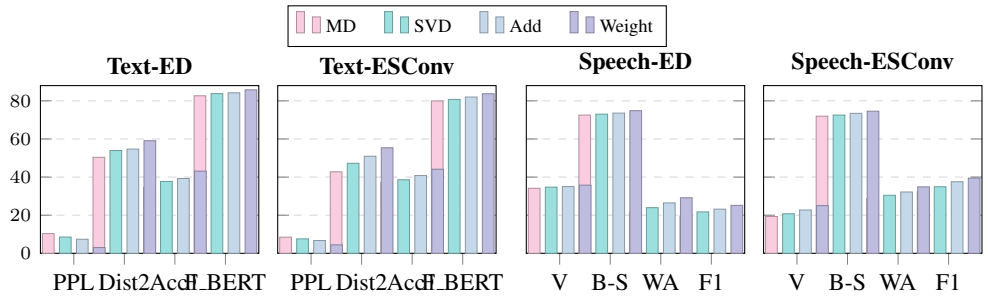


Figure 4: Comparison results of different functional vector fusion methods.

5.5 Impact of Functional Vector Injection Strategies

To investigate the token-level adaptive injection strategy, we design three comparative experiments: 1) No functional vector injection (None): Serving as the baseline without introducing any emotional functional vectors. 2) Uniform injection for all tokens (All): Adopting a global unified regulation coefficient to inject the dual-path fused functional vector (V_{fuse}) into all tokens indiscriminately. 3) Token-level adaptive injection (Token). Figure 3 proves our adaptive strategy outperforms the other two on all text and speech metrics. For text, it lifts Acc2 by 3.84% and F_BERT by 1.51% over uniform injection, with lower PPL and higher Dist scores for better fluency and diversity. For speech, WA rises by 6.45% and F1 by 3.78%, alongside smaller emotional embedding L2 distance for more accurate emotion delivery. The adaptive scheme dynamically tunes adjustment strength per token's emotional weight to balance emotion precision and natural semantics. Uniform injection still beats the baseline, confirming the validity of emotional functional vectors.

5.6 Impact of Functional Vector Fusion Methods

We investigate the impact of different functional vector fusion methods on model performance in Figure 4, comparing four schemes: 1) MD-only path (MD): Injecting only the semantic emotion vector from the representation layer. 2) SVD-only path (SVD): Injecting only the structural emotion vector from the parameter layer. 3) Direct concatenation fusion (Add): Injecting V_{MD} and V_{SVD} after simple concatenation. 4) Adaptive weight fusion (Weight). Adaptive weighted fusion delivers state-of-the-art results on all text and speech metrics, as it fully exploits the complementarity of dual-path vectors: the MD path's semantic emotion discrimination and the SVD path's structural emotion retention. Simple concatenation also surpasses both single-path baselines, proving representation-layer semantic features and parameter-layer structural features are indispensable for high-fidelity emotion modeling. Even without learnable adaptive weights, concatenation merges the MD path's explicit emotion semantics and context sensitivity and the SVD path's native structural emotion patterns and generalizability, overcoming incomplete emotion capture from single-path extraction.

6 Conclusion

In this paper, we propose a unified framework integrating interpretable dual-path emotion vectors and token-level adaptive injection, enabling fine-grained emotion control for empathetic dialogue generation. By synergistically optimizing emotion function vector extraction across the representation and parameter layers, combined with an attention-driven token-level adaptive injection mechanism, the framework effectively mitigates semantic distortion from traditional global injection. Additionally, we construct a text-to-speech cross-modal multi-dimensional evaluation system, providing a more comprehensive benchmark for emotion generation tasks. Experimental results on the ED and ESConv datasets demonstrate that our method significantly outperforms all existing baselines across all metrics, offering an efficient pathway for fine-grained emotional control in empathetic dialogue and controllable text generation.

AI USE STATEMENT

In this work, we use generative AI tools for polishing the English manuscript and assisting with writing LaTeX figure codes.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4502–4511, 2019.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7858–7873, 2023.

Cen Chen, Xiaolu Zhang, Sheng Ju, Chilin Fu, Caizhi Tang, Jun Zhou, and Xiaolong Li. Antprophet: an intention mining system behind alipay's intelligent customer service bot. In *IJCAI*, volume 8, pp. 6497–6499, 2019.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 6124–6142. Association for Computational Linguistics, 2025.

Shuhao Fu, Esther Goldberg, Ying Nian Wu, and Hongjing Lu. Multimodal function vectors for spatial relations. *arXiv preprint arXiv:2510.02528*, 2025.

Peijian Gu, Quan Wang, and Zhendong Mao. Improve safety training of large language models with safety-critical singular vectors localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4941–4954, 2025.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Yuanxiang Huangfu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. Non-emotion-centric empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 989–999, 2025.

Gangwei Jiang, Caigao Jiang, Zhaoyi Li, Siqiao Xue, Jun Zhou, Linqi Song, Defu Lian, and Ying Wei. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Diogo Leocádio, Leonel Guedes, José Oliveira, João Reis, and Nuno Melão. Customer service with ai-powered human-robot collaboration (hrc): A literature review. *Procedia Computer Science*, 232:1222–1232, 2024.

Bobo Li, Hao Fei, Fangfang Su, Fei Li, and Donghong Ji. Integrating discourse features and response assessment for advancing empathetic dialogue. *Information Processing & Management*, 61(5):103803, 2024.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th international conference on computational linguistics*, pp. 4454–4466, 2020.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 10993–11001, 2022.

Zhuoyang Li, Zihao Zhu, Xinning Gui, and Yuhan Luo. "this is human intelligence debugging artificial intelligence": Examining how people prompt gpt in seeking mental health support. *International Journal of Human-Computer Studies*, pp. 103555, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13622–13623, 2020.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 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)*, pp. 3469–3483, 2021.

Kate Loveys, Catherine Hiko, Mark Sagar, Xueyuan Zhang, and Elizabeth Broadbent. "i felt her company": A qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. *International Journal of Human-Computer Studies*, 160: 102771, 2022.

Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2024.

Jennifer J Mitchell and Myounghoon Jeon. Exploring emotional connections: A systematic literature review of attachment in human-robot interaction. *International Journal of Human–Computer Interaction*, pp. 1–22, 2025.

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

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Sahand Sabour, Chujie Zheng, and Minlie Huang. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11229–11237, 2022.

Eric Todd, Millicent Li, Arnab Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*. ICLR, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. 2024.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 123–140, 2025a.

Xu Wang, Bo Wang, Yihong Tang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. Ecc: Synergizing emotion, cause and commonsense for empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5475–5485, 2025b.

Jingran Xie, Shun Lei, Yue Yu, Yang Xiang, Hui Wang, Xixin Wu, and Zhiyong Wu. Leveraging chain of thought towards empathetic spoken dialogue without corresponding question-answering data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Haoqiu Yan, Yongxin Zhu, Kai Zheng, Bing Liu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Talk with human-like agents: Empathetic dialogue through perceptible acoustic reception and reaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15009–15022, 2024.

Shaolei Zhang, Tian Yu, and Yang Feng. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8908–8949, 2024.