ISSN:2790-1688 Volume-15-(2025)

Generating Missing Modalities: A Conditional Diffusion and Transformer Approach for Emotion Recognition

He Yu*, Chuwen Zhang

Stonybrook institution in Anhui University, Anhui University, Hefei, China.

R32214050@stu.ahu.edu.cn

Abstract. Multimodal models have significantly advanced traditional emotion recognition by utilizing information from text, audio, and visual modalities. Many studies have pushed the boundaries of this field. However, the absence of modalities remains a major challenge, hindering the model's ability to capture and integrate cross-modal interactions effectively. Besides, conventional modality completion approaches often fail to preserve fine-grained details. To break through these limitations, we propose a novel modality completion framework based on Conditional Diffusion and Transformer (CDTP). By incorporating three types of prompts and conditions, CDTP enables more detailed representations within and across modalities. Experiments and ablation studies demonstrate that our method substantially enhances emotion recognition performance and exhibits strong robustness in scenarios with missing modalities. The source code will be publicly available at https://github.com/cwzhang689/DPT.

Keywords: Diffusion Model, Transformer, Generative Model, Incomplete Modality, Multimodal Emotion Classification.

1. Introduction

In recent years, multimodal emotion recognition has continued to develop. However, most multimodal datasets are incomplete, necessitating data imputation. Current mainstream multimodal data imputation methods primarily include: Modality Imputation, which addresses missingness by filling in the absent modalities, such as replacing the missing modality with zero values or random values, or copying data from similar samples. However, this may introduce noise, cause information loss, reduce model efficiency, and increase the risk of overfitting [1, 2]. Alternatively, generative models can be used to synthesize missing modality data, such as Auto-Encoders [3], Generative Adversarial Networks (GANs), or Diffusion Models [2]. But this can lead to model generation inaccuracies, increased computational complexity, and potential generation biases [4]. Another approach is Representation Learning, which introduces constraints to align representations of different modalities in the semantic space, helping models train effectively even when modalities are missing [4]. Or, representations for the missing modality can be generated via small generative models [5]. The drawbacks are the risk of overfitting due to excessive reliance on constraints, and the fact that the generated modality representations may not fully reflect the true characteristics of the missing modality, thereby affecting training effectiveness. However, the paper "Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition" [6] proposed a novel multimodal Transformer framework that uses prompt learning to handle missing modality problems. This improved the performance of multimodal sentiment analysis and emotion recognition models in scenarios with missing modalities. Nevertheless, the module design in this method for generating missing modalities is relatively simple and insufficient to fully model the complex conditional dependencies between modalities [6]. Given the outstanding performance of Diffusion models as generative modules in recent years, inspired by this, this paper adopts a Diffusion model to replace the generation module. Our main contributions are as follows :

- We propose a multimodal missing data completion method based on the conditional diffusion model, which combines the relevant achievements of previous prompt learning, and effectively completes the missing modalities.
- The parameter quantities for condition diffusion and hints are based on the number of missing modalities, which is beneficial for reducing the required computing resources.

- The method we proposed outperforms the existing benchmarks in all aspects. At the same time, we have identified the optimal parameters for the model's performance.

2. Related Work

2.1 Multimodal Emotion Recognition

Multimodal Emotion Recognition (MER) aims to enhance the accuracy of emotion recognition by fusing information from multiple modalities (such as speech, image, text, etc.). Traditional emotion recognition methods primarily relied on single-modality information, for example, speech-based emotion recognition [7] or facial expression-based emotion recognition [8]. However, these methods face numerous challenges in practical applications, such as insufficient information from a single modality. To address this, recent research has focused on how to effectively fuse multimodal information to improve emotion recognition performance. Current MER methods can generally be categorized into three strategies: early fusion, late fusion, and intermediate fusion. Early fusion methods typically concatenate features from all modalities and input them directly into the model for training [9]. Intermediate fusion handles synergistic information between different modalities by designing cross-modal attention mechanisms or shared semantic spaces [10]. Late fusion methods process each modality separately and then fuse their outputs for decision-making [11]. In recent years, Transformer-based multimodal fusion methods [12] have gradually become mainstream research. They efficiently capture dependencies between modalities through self- attention mechanisms, thereby significantly improving the accuracy of emotion recognition [12]. However, all the above methods assume the dataset is complete. Our method can handle cases where modalities are missing.

2.2 Multimodal Data Imputation Methods

In multimodal emotion recognition tasks, missing modalities are a prevalent problem, especially in practical applications where certain modalities are often absent due to device failure or environmental interference. To address this issue, researchers have proposed various data imputation methods to restore missing modality information and ensure model performance. The earliest multimodal data imputation methods included value-based filling methods, which use zero values, random values, or copy values from other samples to impute missing modalities [13]. However, these methods are simple and inefficient, not only unable to effectively capture complex relationships between modalities, but also introduce bias. In recent years, generative models have been widely applied to data imputation tasks. Auto-Encoders [3] and Generative Adversarial Networks (GANs)[14] became co mmon generative models, where auto-encoders model data through compression and decompression processes, and GANs generate missing modality data via the adversarial process between generator and discriminator [15]. However, these methods suffer from problems like mode collapse and training instability. Diffusion Models, as an emerging generative model, have recently demonstrated outstanding performance in fields such as image generation and speech synthesis [16]. Consequently, an increasing number of studies have begun exploring the application of diffusion models to multimodal data imputation tasks [17]. But using diffusion models solely for missing modality generation lacks feature exchange between the existing modalities [16]. In contrast, our method uses conditional diffusion for missing modality imputation, where existing modalities guide the generation of the missing one, thus addressing this shortcoming.

2.3 Prompt Learning

Prompt learning involves generating prompts to utilize pre-trained models for different downstream tasks. It was initially widely used in Natural Language Processing (NLP). Recently, it has gradually developed in the multimodal domain. Guo et al. first extended the concept of soft prompts to sequential image-text sets, thereby transforming large language models into multimodal systems [6]. Subsequently, Barnum et al. proposed combining visual and text prompts at different stages of the Transformer, enabling synergy between vision-language modalities [5]. However, these

methods do not recover missing data. In contrast, our method leverages prompt learning combined with conditional diffusion, effectively generating missing modalities using the existing ones, further enhancing model performance.

Table 1. Quantitative results under six possible missing modality cases. For example, " a " means audio modality is available while video and text are missing. "Avg." means the average performance of the six possible cases. Bold: best result. Underline: second best result. We report the average result of five different random seeds.

Dataset	Method	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{a,t\}$		$\{v,t\}$		Avg.	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MOSI	LB	48.32	55.81	49.09	55.20	79.27	79.22	50.07	57.12	78.67	79.25	79.86	79.96	64.21	67.76
	MS	49.17	55.34	49.87	56.12	78.06	78.28	51.12	57.01	79.32	79.65	80.32	80.38	64.64	67.80
	MD	48.79	55.74	49.66	55.60	79.36	80.01	52.33	56.84	79.59	79.86	80.51	80.43	65.04	68.08
	MCTN	51.32	56.12	54.27	56.33	79.63	79.78	56.79	57.84	78.96	79.17	80.45	80.65	66.90	68.32
	MMIN	59.16	60.12	61.01	61.98	80.10	80.16	63.79	64.08	80.50	80.33	80.46	80.63	70.84	71.22
	MPMM	57.26	59.35	58.63	59.12	79.81	80.10	60.54	61.33	79.89	79.84	80.74	80.93	69.48	70.11
	MPLMM	62.71	63.65	63.12	63.74	80.12	80.31	65.02	65.41	80.76	81.09	81.12	81.19	72.14	72.57
	Ours	62.91	63.95	63.32	64.04	80.22	80.41	65.22	65.61	80.96	81.29	81.32	81.39	72.49	72.92
IEMOCAP	LB	46.35	46.21	48.07	47.58	56.06	55.28	58.12	57.89	72.18	72.25	65.63	65.28	57.74	57.42
	MS	47.65	47.52	47.68	47.36	59.27	59.22	57.48	56.60	72.30	72.18	66.81	66.93	58.53	58.30
	MD	48.22	48.09	48.26	47.98	61.26	61.28	58.08	57.96	72.40	72.31	67.08	68.22	59.22	59.31
	MCTN	51.62†	-	45.73†	-	63.78†	-	55.84†	-	69.46†	-	68.34†	-	59.19†	-
	MMIN	59.00†	-	51.60†	-	68.02†	-	65.43†	-	75.14†	-	73.61†	-	65.47†	-
	MPMM	58.69	57.66	55.18	55.36	68.39	68.08	63.68	63.47	74.90	74.98	73.80	72.67	65.77	65.37
	MPLMM	59.77	59.71	57.61	56.98	69.23	69.28	67.26	67.37	75.98	75.44	74.68	74.51	67.42	67.22
	Ours	59.97	59.91	57.81	57.18	69.53	69.58	67.46	67.57	76.18	75.64	74.88	74.71	67.75	67.55
MOSEI	LB	66.21	68.69	66.45	69.10	77.96	78.32	67.30	69.62	78.13	78.63	77.86	78.16	72.32	73.83
	MS	62.74	67.06	64.16	68.17	77.28	77.76	67.11	69.51	78.34	78.80	78.08	78.62	71.29	73.36
	MD	65.76	68.18	66.57	69.35	77.30	77.94	67.21	69.48	78.74	78.97	78.11	78.71	72.28	73.82
	MCTN	66.19	68.58	66.70	69.01	78.32	78.41	68.10	69.34	79.11	79.14	78.65	78.64	72.85	73.94
	MMIN	67.11	68.67	67.01	69.31	78.67	78.71	68.17	69.74	79.94	79.96	79.32	79.29	73.37	74.39
	MPMM	66.94	68.74	67.21	69.27	78.21	78.30	68.11	69.79	79.41	79.47	79.63	79.71	73.25	74.17
	MPLMM	67.33	68.71	67.29	69.40	79.12	79.17	68.21	69.91	80.45	80.43	80.11	80.13	73.75	74.68
	Ours	67.53	68.91	67.49	69.60	79.32	79.37	68.41	70.11	80.65	80.63	80.31	80.33	74.09	74.98

3. Method

As shown in Figure 1, our proposed method comprises three crucial components: the conditional diffusion model, cross-attention, and prompt learning.

We use MulT (Tsai et al., 2019) as the backbone, which introduced the Crossmodal Transformer for modeling unaligned data. We employ three types of different prompts in MPLMM: generative prompts, missing-signal prompts, and missing-type prompts. In our proposed method, we employ a conditional diffusion model. The generative prompts assist the available modalities in generating representations for the missing modalities.

In the first stage, we utilize a conditional diffusion model learning text, audio, and visual features, completing missing modalities. For the second stage, through cross-attention and self-attention, it captures cross-modality features.

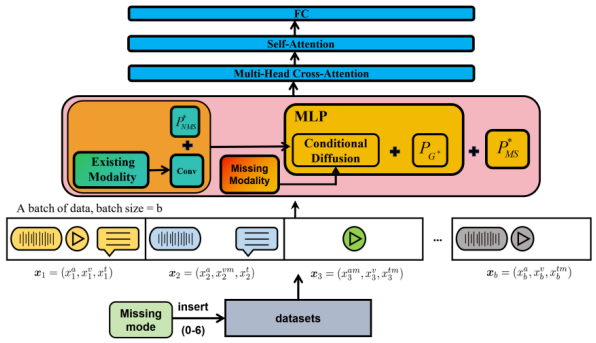

Fig. 1. Multimodal learning framework designed to handle missing modalities by leveraging conditional diffusion mechanisms.

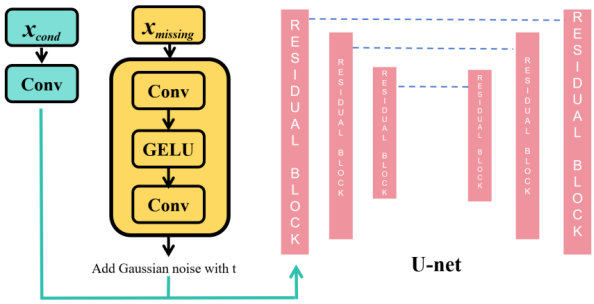

Fig. 2. The conditional diffusion backbone designed for reconstructing missing modalities.

3.1 Unmasked and Masked Modalities

For a multimodal dataset D with 3 modalities (e.g., text, audio, and vision), we use $x = (x^t, x^a, x^v)$ to represent a group of features without missing modalities; relatively, using x^{tm}, x^{am}, x^{vm} to indicate missing modalities.

We use a missing-mode signal m to classify 7 cases of three modalities. That is, it defines the mapping between the numerical missing-mode index and the corresponding set of missing modalities.

$$\text{missing\_mode}(m) = \begin{cases} \{\text{Text}\}, & m = 0, \\ \{\text{Audio}\}, & m = 1, \\ \{\text{Vision}\}, & m = 2, \\ \{\text{Text, Audio}\}, & m = 3, \\ \{\text{Text, Vision}\}, & m = 4, \\ \{\text{Audio , Vision}\}, & m = 5, \\ \{\text{G}\}, & m = 6, \end{cases}$$

Through a Uniform distribution to determine the missing mode, masking parts of the data to dice modalities:

$$m = \begin{cases} 6, & \text{if full\_data is true or } p > p_{\text{drop}}, \\ \text{Uniform}\{0,1,2,3,4,5\}, & \text{if } p \le p_{\text{drop}}. \end{cases}$$

where

$p \sim \text{Uniform}(0,1)$, p_{drop} is the drop rate of modalities.

3.2 Encoders

3.2.1 Positional embeddings.

To utilize the order of the sequence and temporal dependencies, we add "fixed positional encodings" to the modality representations. Through sine and cosine functions of different frequencies, the model injects both relative and absolute positional information into the modality representations. The formulas are computed as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (1)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (2)$$

where L represents the sequence length, $\text{pos} \in \{0,1,2,\ldots,L-1\}$ is the position index of the input token, d_{model} is the embedding dimension (e.g. 512), and $i \in \{0,1,\ldots,\lfloor d_{\text{model}}/2 \rfloor - 1\}$ is the dimension index. We use sin for odd dimensions and cos for even dimensions. Besides, we chose the fixed version instead of learned positional embeddings because it allows the model to extrapolate to sequence lengths longer than the ones encountered during training. After this initial encoding, every modality could contain the temporal features of the input, beneficially addressing both aligned and unaligned cases.

3.2.2 Conv1D Encoder.

Mapping from $d_{l,\text{orig}}$ to the target dimension d_l without bias is equivalent to a linear layer.

Input shape: $(\text{batch}, d_{l,\text{orig}}, L) \rightarrow$ Output shape: (batch, d_l, L). d_l is the number of tokens in the sequence, L is the length of the token sequence.

3.2.3 Conditional Encoder of Diffusion.

We design a lightweight conditional encoder to project modality features into the unified latent space. Formally, given an input sequence $x \in \mathbb{R}^{L\times d}$, the transformation is defined as

$$h = \text{Conv1D}_2\left(\text{GELU}\left(\text{Conv1D}_1(x)\right)\right),$$

where Conv1D_1 and Conv1D_2 are one-dimensional convolutional layers, and $\text{GELU}(\cdot)$ denotes the Gaussian Error Linear Unit activation.

3.3 Three kinds of Prompts

We introduce the three kinds of prompts from Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. They are generative prompts,

3.3.1 Generative Prompts.

For generative prompts, it guides the model to recover missing information. Generative prompts was denoted as

$$P_G = \left(P_{G_t}, P_{G_a}, P_{G_v}\right) \in \mathbb{R}^{3\times d_p\times \ell_p}$$

where P_{G_t}, P_{G_a} and P_{G_v} represent the generative prompts for the audio, video, and text modalities; d_p and ℓ_p represent the dimension and length of the prompts, respectively.

3.3.2 Missing-signal Prompts.

To indicate whether one modality is existential or generating, we use missing-signal prompts to better recognize. It will function at the corresponding Transformer. For each modality, there are two missing-signal prompts: P^*_{MS} means a modality is missing, and P^*_{NMS} means a modality is not missing.

3.3.3 Missing-type Prompts.

Missing-type prompts are essential because they encode the joint pattern of missing modalities, allowing the model to capture combinational correlations that missing-signal prompts alone cannot represent.

They prevent parameter explosion by generating combination-aware prompts through a projection mechanism instead of storing $2^M - 1$ separate prompts, if there are M modalities. They enable the Transformer to calibrate crossmodal trust and remain robust to distribution shifts by explicitly conditioning on the exact missing-modality configuration.

3.4 Completion Mechanism

To simplify the reasoning process of missing modalities cases, we mainly take the one-modality missing and two-modalities missing into consideration. There are unique examples to illustrate our method.

Given that

$$\mathbf{x} = (x^t, x^{am}, x^v)$$

the data only lacks one modality. For non-missing modality,

$$x^t_{\text{proj}} = \left(\text{Conv1D}_l((x^t)^\top)\right)^\top$$

$$x^t_{\text{final}} = x^t_{\text{proj}} + P^t_{\text{NMS}} \in \mathbb{R}^{L_t\times d'}$$

Similarly, x^t_{final} is also calculated as mentioned.

For the missing modality, we represent x^{am} using the existing x^t and x^v regard to the following equation:

$$\tilde{x}_a = [P_{Ga}; \mathrm{CondDiff}_{a\leftarrow t}(x_t); \mathrm{CondDiff}_{a\leftarrow v}(x_v)] \in \mathbb{R}^{(p+2*L'_a)\times d'}$$

where $P_{Ga} \in \mathbb{R}^{p\times d'}$ is generative prompt for audio modality.

$$x^a = \mathrm{MLP}(\tilde{x}_a) + P^a_{\mathrm{MS}} \in \mathbb{R}^{L_a\times d'}$$

Considering missing two modalities

$$\mathbf{x} = (x^t, x^{am}, x^{vm})$$

there are a few differences in the missing modalities dealing.

$$x_a = \mathrm{MLP}([P_{Ga}; \mathrm{CondDiff}_{a\leftarrow t}(x_t)]) + P^a_{\mathrm{MS}} \in \mathbb{R}^{L_a\times d'}$$
$$x_v = \mathrm{MLP}([P_{Gv}; \mathrm{CondDiff}_{v\leftarrow t}(x_t)]) + P^v_{\mathrm{MS}} \in \mathbb{R}^{L_v\times d'}$$
$$x_t = \mathrm{Conv1D}(x_t^\top)^\top + P^t_{\mathrm{NMS}} \in \mathbb{R}^{L_t\times d'}$$

Conditional Encoding and Noising. To incorporate conditional information for missing-modality completion, we employ a DiffusionConditionalLayer, consisting of a Conv1D-GELU-Conv1D stack. Given a conditional modality input $x_{\mathrm{cond}} \in \mathbb{R}^{L\times d}$, the conditional representation is computed as

$$x_{\mathrm{tgt}} = \mathrm{Conv1D}_2\left(\mathrm{GELU}\left(\mathrm{Conv1D}_1(x_{\mathrm{cond}})\right)\right).$$

To simulate the diffusion process, we add Gaussian noise to x_{tgt} at time step t :

$$\tilde{x}_{\mathrm{tgt}} = \sqrt{\alpha_t}x_{\mathrm{tgt}} + \sqrt{1-\alpha_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, I).$$

Finally, the noisy target representation $\tilde{x}_{\mathrm{tgt}}$ is concatenated with the available modality features, forming the input z_0 to the U-Net backbone:

$$x_{\mathrm{cond}} = \mathrm{Conv1D}_1(x)$$

here, x means a missing modality:

$$z_0 = \mathrm{cat}\left(x_{\mathrm{cond}}, \tilde{x}_{\mathrm{tgt}}\right)$$

For U-Net Backbone, we adopt a U-Net with residual blocks and temporal position encodings as backbone to parameterize the denoising function within the diffusion process. Conditional features are injected via Cross-Attention/ FiLM Layers. This allows the imputer to leverage available modalities while denoising the missing ones. Overall, if a modality is missing, the DDIM is used to generate the feature of that modality, which is then sent back to the MulT module for fusion. If all modalities are present, this module is bypassed. Formally, given an input z_0 (the concatenation of observed modalities and the noisy conditional representation) and the diffusion time step t, the U-Net can be written as

$$h = \mathrm{UNet}(z_0, t) = \mathrm{Dec}\left(\mathrm{Bottleneck}\left(\mathrm{Enc}(z_0, t)\right)\right),$$

where $\mathrm{Enc}(\cdot)$ denotes the encoder (down-sampling path), $\mathrm{Bottleneck}(\cdot)$ represents the central residual block stack, and $\mathrm{Dec}(\cdot)$ is the decoder (up-sampling path).

Each stage of the encoder and decoder is built from residual blocks. A residual block takes the form

$$\mathrm{ResBlock}(x) = x + F(x;\theta),$$

where $F(\cdot;\theta)$ is a sequence of operations (e.g., Conv1D/Conv2D, normalization, and non-linear activation).

Moreover, U-Net employs skip connections between the encoder and the decoder to preserve spatial/temporal information across scales. Specifically, the decoder feature at layer l is computed as:

$$h^{(l)}_{\mathrm{dec}} = \mathrm{DecBlock}\left(\left[h^{(l)}_{\mathrm{enc}}, h^{(l+1)}_{\mathrm{dec}}\right]\right),$$

where $[\cdot,\cdot]$ denotes concatenation along the feature dimension, $h^{(l)}_{\mathrm{enc}}$ is the encoder output at layer l, and $h^{(l+1)}_{\mathrm{dec}}$ is the decoder feature from the next deeper layer.

This encoder-decoder structure with residual and skip connections allows the U-Net to capture both global context and fine-grained details, which is crucial for robust multimodal generation under missing-modality conditions.

Notation. Let $\{\alpha_t\}_{t=1}^{T}$ be a noise schedule with $\beta_t = 1 - \alpha_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. At step t, the missing-modality state is $x_t \in \mathbb{R}^{L \times d}$ and the condition (from observed modalities after the conditional encoder) is $c \in \mathbb{R}^{L \times d_c}$. We feed the U-Net with concatenated features

$$\varepsilon_\theta(x_t, c, t) = \mathrm{UNet}([c, x_t], t),$$

and the U-Net directly predicts the noise ε_θ.

Forward diffusion (training data).

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(\sqrt{\alpha_t} x_{t-1}, \beta_t \mathbf{I}\right), \tag{3}$$

$$q(x_t \mid x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I}\right) \Leftrightarrow x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, \varepsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

Training objective. The U-Net predicts the added noise given (x_t, c, t) :

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \varepsilon} \| \varepsilon - \varepsilon_\theta(x_t, c, t) \|_2^2. \tag{5}$$

DDPM sampling. With $\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$, define the mean

$$\mu_\theta(x_t, c, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta(x_t, c, t) \right). \tag{6}$$

Then sample

$$p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}\left(\mu_\theta(x_t, c, t), \tilde{\beta}_t \mathbf{I}\right), \quad x_{t-1} \sim p_\theta(\cdot \mid x_t, c), \tag{7}$$

and iterate $t = T, T-1, \ldots, 1$ with $x_T \sim \mathcal{N}(0, \mathbf{I})$.

DDIM sampling. First estimate the clean sample

$$\hat{x}_0(x_t, c, t) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\varepsilon_\theta(x_t, c, t)}{\sqrt{\bar{\alpha}_t}}. \tag{8}$$

Choose $\sigma_t \in \left[0, \sqrt{1 - \bar{\alpha}_{t-1}} - \sqrt{1 - \bar{\alpha}_t}\right]$. The DDIM update is

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0(x_t, c, t) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \underbrace{\frac{x_t - \sqrt{\bar{\alpha}_t} \hat{x}_0(x_t, c, t)}{\sqrt{1 - \bar{\alpha}_t}}}_{= \varepsilon_\theta(x_t, c, t)} + \sigma_t \xi, \ \xi \sim \mathcal{N}(0, \mathbf{I}). \tag{9}$$

(9)

Setting $\sigma_t = 0$ yields the deterministic DDIM sampler.

3.5 Multi-Head Cross-Attention and Self-Attention

Due to better capture of interacting details among different modalities, we introduce a multihead cross-attention mechanism. Given that text (t), audio (a), and visual (v) modalities, we learn complementary representations across text (t).

$$x_l \in \mathbb{R}^{L_l \times d'} x_a \in \mathbb{R}^{L_a \times d'} x_v \in \mathbb{R}^{L_v \times d'}$$

$$\mathrm{Trans}_{qk}(Q, K, V) = \mathrm{Transformer}_{\mathrm{Encoder}}(Q, K, V)$$

where Q (Query) comes from current modality m_q (e.g., text), K, V come from other modalities m_k (e.g., audio and version).

For example, text is a current modality when audio modality is key and value:

$$\mathrm{Attention}^{(t \leftarrow a)} = \mathrm{Transformer}^{ta}_{\mathrm{Encoder}}(Q = x_t, K = x_a, V = x_a) \in \mathbb{R}^{L_t \times d_t}$$

To generalize,

$$h_t^{(a)} = \mathrm{Trans}_{ta}(x_l, x_a, x_a) \in \mathbb{R}^{L_t \times d_t}$$

$$h_t^{(v)} = \mathrm{Trans}_{tv}(x_l, x_v, x_v) \in \mathbb{R}^{L_t \times d_t}$$

$$h_t = \text{cat}\left(h_t^{(a)}, h_t^{(v)}\right) \in \mathbb{R}^{L_t \times 2d_t}$$

Similarly,

$$h_a = \text{cat}\left(\text{Trans}_{al}(x_a, x_l, x_l), \text{Trans}_{av}(x_a, x_v, x_v)\right) \in \mathbb{R}^{L_a \times 2d_a}$$
$$h_v = \text{cat}\left(\text{Trans}_{vl}(x_v, x_l, x_l), \text{Trans}_{va}(x_v, x_a, x_a)\right) \in \mathbb{R}^{L_v \times 2d_v}$$

After that, we need to combine missing-type prompts. For

$$\mathbf{x} = (x^t, x^a, x^{vm})$$

we introduce **a missing-type projection matrix:**

$$\mathbf{M}_P = \mathbf{M}_t \cdot P_{NMS}^t + \mathbf{M}_a \cdot P_{NMS}^a + \mathbf{M}_v \cdot P_{MS}^v$$

where $\cdot$ is the matrix multiplication, $\mathbf{M}_t, \mathbf{M}_a, \mathbf{M}_v \in \mathbb{R}^{d_p \times \ell_p}$ and $\mathbf{M}_P \in \mathbb{R}^{d_p \times d_p}$. Then, we can get the **missing-type prompts** P'_{MT} as follows:

$$P'_{MT} = P_{MT} \cdot \mathbf{M}_P$$

where P_{MT} represents the original missing-type prompts, P'_{MT} represents the projected missing-type prompts and $P_{MT}, P'_{MT} \in \mathbb{R}^{3 \times \ell_p \times d_p}$.

For the attention mechanism, it's easy to show that

$$\text{AttnWeights} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + P'_{MT}\right)$$

4. Experiment

4.1 Datasets and Evaluation Metrics

To simulate real-world scenarios, we selected three widely used multimodal emotion and emotion recognition datasets to evaluate our proposed method. Among them, CMU-MOSEI (Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph) and CMUMOSI (MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos) are considered as high-resource datasets, while IEMOCAP (IEMOCAP: Interactive emotional dyadic motion capture database) is regarded as a low-resource dataset.

CMU-MOSI is a widely used benchmark dataset for multimodal sentiment analysis, covering three modalities: text, speech and vision. This dataset was collected from 93 English YouTube videos. Each passage was manually labeled with an emotion score ranging from -3 to +3, which was used to reflect the emotional intensity from strongly negative to strongly positive.

CMU-MOSE is a large-scale expansion of MOSI, featuring over 65 hours of video content, covering more than 1,000 speakers and over 250 topics, with a broader range of themes. Compared with MOSI, MOSEI not only has a larger scale and richer corpus sources, but also provides multidimensional annotation information. In terms of emotion

IEMOCAP consists of five pairs of two-person dialogues, involving 10 actors and lasting approximately 12 hours. The data includes text, audio and visual modalities, and provides multiple types of emotion labels (such as happy, sad, angry, neutral). In experiments, tasks are usually organized in a binary classification manner and evaluated by average accuracy and weighted F1 score.

In terms of evaluation indicators, we follow the Settings of existing studies: for MOSI and MOSEI, seven-classification accuracy (ACC-7), two-classification accuracy (ACC), F1 value, mean absolute error (MAE), and Pearson correlation coefficient (Corr) are adopted; For IEMOCAP, use the average accuracy and weighted F1. For CH-SIMS, ACC, F1, MAE and Corr are adopted.

4.2 Baselines

We compare our proposed method with the following baselines:

Lower Bound (LB) trains unimodal and partial-modality models using different modality combinations, serving as performance references when information is incomplete. **Modality Substitution (MS)** replaces missing modalities with default values or fixed placeholders, providing a simple imputation strategy. Modality **Dropout (MD)** trains models by randomly dropping modalities during the training stage to improve robustness against missing data. **MMIN (Zhao et al.,**

2021) predicts missing modality representations from the available ones, and is designed to handle arbitrary missing patterns. **MPLMM (Guo et al., 2024)** This method uses three prompts to generate missing modal features and promotes the learning of information within and between modalities.

4.3 Implementation Details

All models are implemented in PyTorch and trained on NVIDIA H200 GPUs with CUDA acceleration. Following prior works, we preprocess multimodal inputs into three feature streams: speech is processed using a pretrained acoustic feature extractor, video is processed using frame-level visual embeddings extracted from a Visual Transformer backbone, and text is processed using a pretrained BERT encoder.

For model configuration, we project each modality into a shared space of 256 dimensions and employ a 6-layer Transformer encoder with 8 attention heads. Dropout with rate 0.1 is applied to attention, feedforward, and residual connections.

Training is conducted for 50 epochs with a batch size of 32. We use the Adam optimizer with an initial learning rate of 1e-4.

To simulate missing-modality conditions, we follow MPLMM and apply a stochastic drop mechanism with a rate of 0.2 during training. For evaluation, the best model is selected based on validation performance.

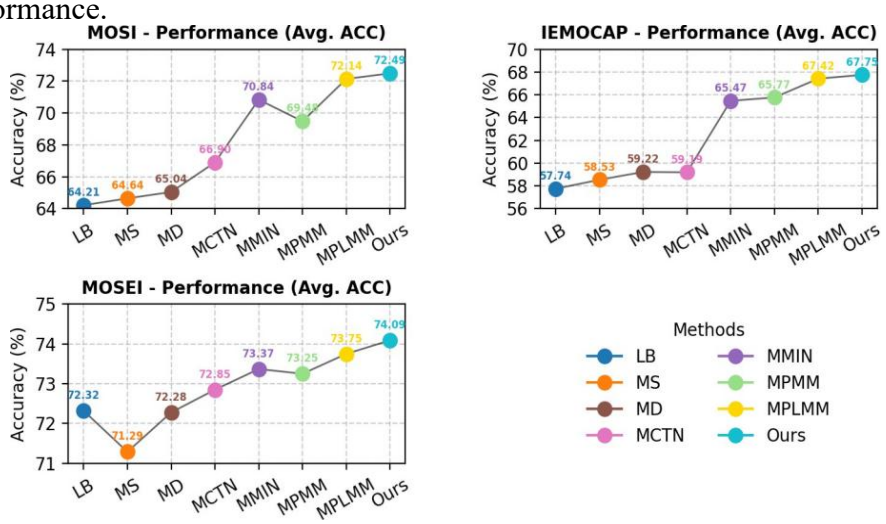

Fig. 3. The model performance on MOSI, IEMOCAP, and MOSEI datasets

4.4 Main Results

Table 1 shows quantitative results on three datasets. The baselines LB, MS, MD, MPMM, and MPLMM have the same backbone as our method. Comparing the baseline MS and MD, it is apparent that random discarding of data modalities during training improves the generalization ability of the model. Observing the performance of our proposed method in all datasets, it exceeds all baselines under all six missing modality cases. Hence, it reflects the effectiveness of our proposed method to deal with missing modalities. We implement different feature extraction approaches on three datasets. The results from Table 1 show that our model can adapt to extract features by different methods.

All models are implemented in PyTorch and trained on NVIDIA H200 GPUs with CUDA acceleration. Following prior works, we preprocess multimodal inputs into three feature streams: speech is processed using a pretrained acoustic feature extractor, video is processed using frame-level visual embeddings extracted from a Visual Transformer backbone, and text is processed using a pretrained BERT encoder.

For model configuration, we project each modality into a shared space of 256 dimensions and employ a 6-layer Transformer encoder with 8 attention heads. Dropout with a rate of 0.1 is applied to attention, feedforward, and residual connections. Training is conducted for 50 epochs with a batch size of 32. We use the Adam optimizer with an initial learning rate of 1e-4. To simulate missing-

modality conditions, we follow MPLMM and apply a stochastic drop mechanism with a rate of 0.2 during training. For evaluation, the best model is selected based on validation performance.

4.5 Generalization Ability

To further assess the generalization capacity of our approach, we evaluate it on multiple representative MSA/MER backbones. In particular, we conduct experiments on MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), and UniMSE (Hu et al., 2022), and report the results in Table 2. For each backbone, our generative prompts and module are introduced after the feature extraction stage. In the case of UniMSE, the missing-signal and missing-type prompts are injected into its multimodal fusion layers, while for MMIM and MISA, they are incorporated into their modality-specific en- coders as well as the fusion components. As shown in the table, our method consistently improves the robustness of different backbones against missing-modality scenarios. Moreover, even when the full set of modalities is available, the integration of our prompts yields further gains, demonstrating their effectiveness in enhancing both intra-modality and inter-modality representation learning.

5. Conclusions

In this paper, we proposed a multimodal modality completion method based on a diffusion model and a Transformer with prompt learning. By utilizing the diffusion model, our method greatly improved the modality completion ability, putting emotion recognition accuracy at a higher stage. At the same time, three kinds of prompts, which are generative prompts, missing-signal prompts, and missing-type prompts, guide the model in generating, distinguishing, and fusing three modalities. It has a more comprehensive understanding of the relationships within and between modalities. Through experiments and ablation studies, our method demonstrated effectiveness and robustness. In the future, it is promising to explore more Diffusion Transformers with better adaptability to complex scenarios and abundant cases in emotion recognition.

References

[1] Kaicheng Yang, Hua Xu, and Kai Gao. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia, pages 521–528, 2020.

[2] Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. arXiv preprint arXiv:2210.12798, 2022.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[4] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10760–10770, 2020.

[5] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In Proceedings of the 30th ACM international conference on multimedia, pages 1642–1651, 2022.

[6] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. arXiv preprint arXiv:2407.05374, 2024.

[7] Chawki Barhoumi and Yassine BenAyed. Real-time speech emotion recognition using deep learning and data augmentation. Artificial Intelligence Review, 58(2):49, 2024.

[8] David Dukić and Ana Sovic Krzic. Real-time facial expression recognition using deep learning with application in the active classroom environment. Electronics, 11(8):1240, 2022.

[9] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning. arXiv preprint arXiv:2011.07191, 2020.

[10] R Gnana Praveen and Jahangir Alam. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4803–4813, 2024.

[11] JohnPaul Quilingking Tomas, Ruth Anne S. Jamilla, Kim S. Lopo, and Cherisse E. Camba. Multimodal emotion detection model implementing late fusion of audio and lyrics in filipino music. In Proceedings of the 2020 3rd International Conference on Computing and Big Data, pages 78–84, 2020.

[12] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3507–3511. IEEE, 2020.

[13] Mike Van Ness and Madeleine Udell. In defense of zero imputation for tabular deep learning. In NeurIPS 2023 Second Table Representation Learning Workshop, 2023.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

[15] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. arXiv preprint arXiv:1902.09599, 2019.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[17] Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. Unified multi-modal image synthesis for missing modality imputation. IEEE Transactions on Medical Imaging, 44(1):4–18, 2024.