GSDNet: Revisiting Incomplete Multimodality-Diffusion Emotion Recognition from the Perspective of Graph Spectrum

Yuntao Shou[1], **Jun Yao**[2], **Tao Meng**[1], **Wei Ai**[1], **Cen Chen**[3*], **Keqin Li**[4]
[1]College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, Hunan 410004, China
[2]Department of Computer Science, Anhui Normal University, Anhui 24100, China
[3]Future Technology Institute, South China University of Technology, Guangdong 510641, China
[4]Department of Computer Science, State University of New York, New Paltz, New York 12561, USA
shouyuntao@stu.xjtu.edu.cn, yj@ahnu.edu.cn, mengtao@hnu.edu.cn, aiwei@hnu.edu.cn, chencen@scut.edu.cn, lik@newpaltz.edu

Abstract

Multimodal emotion recognition in conversations (MERC) aims to infer the speaker's emotional state by analyzing utterance information from multiple sources (i.e., video, audio, and text). Compared with unimodality, a more robust utterance representation can be obtained by fusing complementary semantic information from different modalities. However, the modality missing problem severely limits the performance of MERC in practical scenarios. Recent work has achieved impressive performance on modality completion using graph neural networks and diffusion models, respectively. This inspires us to combine these two dimensions through the graph diffusion model to obtain more powerful modal recovery capabilities. Unfortunately, existing graph diffusion models may destroy the connectivity and local structure of the graph by directly adding Gaussian noise to the adjacency matrix, resulting in the generated graph data being unable to retain the semantic and topological information of the original graph. To this end, we propose a novel Graph Spectral Diffusion Network (GSDNet), which maps Gaussian noise to the graph spectral space of missing modalities and recovers the missing data according to its original distribution. Compared with previous graph diffusion methods, GSDNet only affects the eigenvalues of the adjacency matrix instead of destroying the adjacency matrix directly, which can maintain the global topological information and important spectral features during the diffusion process. Extensive experiments have demonstrated that GSDNet achieves state-of-the-art emotion recognition performance in various modality loss scenarios.

1 Introduction

Multimodal emotion recognition in conversations (MERC) aims to build an emotion recognition model with cross-domain understanding, reasoning, and learning capabilities by integrating video, audio, and text data [Li *et al.*, 2023], [Tsai *et al.*, 2019]. MERC research mainly focuses on how to effectively encode discriminative representations from different modalities and achieve more accurate information fusion and analysis [Ramesh *et al.*, 2021], [Ding *et al.*, 2023].

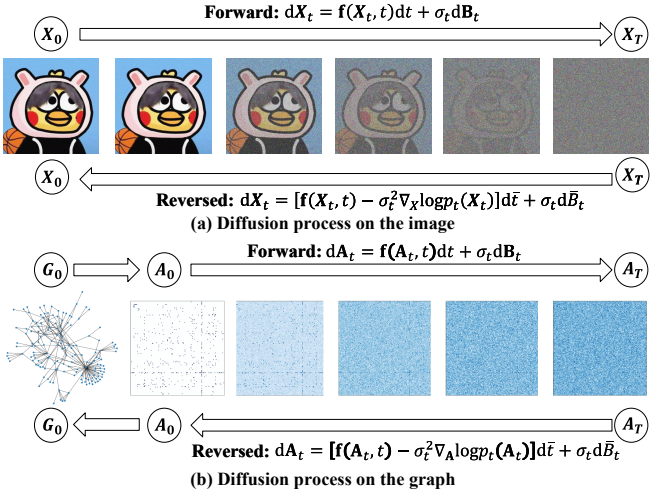


Figure 1: Illustration of the difference between images and graphs in the diffusion processes.

However, the modality missing problem is unavoidable in the real world, and it may severely degrade the performance of multimodal understanding models [Wang *et al.*, 2023a]. For example, the text modality may fail due to environmental noise interference, resulting in the inability to obtain valid text information. The acoustic modality may lose part of the sound information due to sensor failure. The visual modality may be affected by factors such as poor lighting conditions, object occlusion, or privacy protection requirements, resulting in the inability to obtain image or video data. Therefore, how to design and optimize a multimodal emotion recognition model for conversation that can cope with modality loss has become an important direction of current research [Wang *et al.*, 2024], [Zhang *et al.*, 2024].

*Corresponding author

Recently, graph neural networks (GCNs) and diffusion generative models (DGMs) have shown outstanding performance in multiple tasks such as vision, and language [Hu *et al.*, 2021], [Jo *et al.*, 2022]. The core of GCNs is message passing based on graph structure to establish complex dependencies. The core of DGMs is forward denoising and backward denoising to learn the original distribution of data. Considering the potential of GCNs and DGMs, some researchers have tried to apply GCNs and SGMs to modality completion. They alleviate the impact of modality loss on multimodal emotion recognition performance from different perspectives. For GCNs, [Lian *et al.*, 2023] applied graph completion networks to model semantic dependencies between different modal data to achieve modal recovery. For DGMs, [Wang *et al.*, 2024] proposed modal diffusion to learn the original distribution of data to achieve modal recovery. All of the above methods have shown significant results because they have strong modeling capabilities in their respective dimensions. Specifically, GCNs and DGMs respectively model the dependence and distribution between multi-modal data to achieve modal recovery. Generally, fusing complex semantic information between multi-modal features and capturing the original distribution of multi-modal data are crucial for emotion recognition performance in missing modalities. This inspires us to combine these two dimensions to obtain more powerful modal recovery capabilities.

However, unlike visual data, which has dense structural information, the structure of graph data is generally sparse, causing the data generated by the graph diffusion model to be unable to retain the topological information of the original graph. As shown in Fig. 1, the image perturbed by Gaussian noise still retains recognizable numerical patterns and local structural information in the early and middle stages of forward diffusion. For example, even if the details of the image are gradually covered by noise, its global contour and edge information can still be captured by the model to a certain extent. This enables the model to effectively restore the content of the image using this residual information during the reverse diffusion process. However, during the forward diffusion of graph data, the topological structure of the graph adjacency matrix is rapidly lost and a dense noise matrix is formed. Intuitively, the diffusion method of inserting Gaussian noise into the graph adjacency matrix seriously undermines the ability to learn the graph topology and feature representation. From a theoretical perspective, running diffusion over the entire space of the adjacency matrix will cause the signal-to-noise ratio (SNR) to drop rapidly and approach zero. Since the SNR is basically zero, the scoring network will not be able to effectively capture the gradient information of the original distribution during training.

To overcome these problems, we propose a novel Graph Spectral Diffusion Network (GSDNet) to strictly restrict the diffusion of Gaussian noise to the spectral space of the adjacency matrix. Specifically, we perform eigendecomposition on the adjacency matrix, decomposing it into eigenvalue matrix and eigenvector matrix and adding Gaussian noise to the eigenvalues without interfering with the eigenvectors. On the one hand, by operating the eigenvalues in the spectral space, the direct destruction of the local structure of the adjacency matrix can be effectively avoided, ensuring the generated graph still conforms to the global semantics. On the other hand, this constrained diffusion process can more naturally capture and preserve the topological information of the graph, ensuring that the generated graph has consistent spectral features. Overall, our contributions are as:

- We design a novel modality completion model, the Graph Spectral Diffusion Network (GSDNet), which can simultaneously model the dependencies between multimodal features and the distribution of original data to obtain powerful modality recovery capabilities.
- We strictly limit the diffusion of Gaussian noise in the spectral space of the adjacency matrix to avoid the destruction of the graph structure and ensure that the graph data generated by GSDNet still conforms to the global semantics.
- We conduct extensive experiments on multiple real-world datasets to demonstrate that our GSDNet outperforms state-of-the-art methods for conversational emotion recognition in incomplete multimodal scenarios.

2 Related Work

2.1 Incomplete Multimodal Learning

In practical applications, missing modalities are an inevitable problem. To address this challenge, an effective approach is to find a low-dimensional subspace that can be shared by all modalities, in which the correlation between different modalities is maximized. However, strategies based on shared low-dimensional subspaces often ignore the complementarity between heterogeneous modalities. To overcome this shortcoming, another more effective approach is to restore the missing modality through the existing modality. This process not only requires inferring the content of the missing modality based on the features of the known modality but also ensures that the restored modality can work together with other modalities. Existing modality restoration methods can be divided into several types, including zero-based restoration [Parthasarathy and Sundaram, 2020], average-based restoration [Zhang *et al.*, 2020], and deep learning-based restoration [Pham *et al.*, 2019]. Since zero-filling and average-based restoration methods do not use any supervised information, the data they restore often have a significant gap with the original data. In contrast, deep learning-based methods, with their powerful feature learning capabilities, can more accurately estimate the missing modality. For example, Tran et al. [Tran *et al.*, 2017] used a cascaded residual autoencoder to restore the missing modality, and the network's residual learning mechanism made the restoration effect more accurate. In addition, some researchers have proposed deep learning methods based on cross-modal restoration strategies, using cycle consistency loss to ensure the matching degree between the restored modality and the original modality [Zhao *et al.*, 2021]. Other studies use graph neural networks (GNNs) to solve the modality restoration problem. For example, Lian et al. [Lian *et al.*, 2023] introduced a graph neural network framework and combined the relationship between nodes and edges to enhance the correlation between modalities.

2.2 Score-based Generative Models

Score-based generative models (SGMs) estimate the probability distribution of data by parameterizing the score function [Song and Ermon, 2019], [Song and Ermon, 2020], [Song *et al.*,]. Specifically, SGMs model the scoring network $s(x;\theta)$ through learnable parameters θ, thereby training the model to estimate $\nabla_x \log p(x)$. Unlike likelihood-based generative models (e.g., regularized flows [Kingma and Dhariwal, 2018]), score-based generative models do not require regularization of the generation process. Specifically, in likelihood-based methods, model training usually relies on maximizing the likelihood function, which means that regularization terms need to be introduced to prevent overfitting and ensure model stability. In contrast, SGMs estimate the gradient of the data distribution by optimizing the score function. This approach usually does not require explicit regularization and reduces the complexity of model training. In addition, SGMs only need to focus on estimating the gradient of the data distribution by learning the score function, avoiding the complexity of accurately modeling the entire distribution process.

3 Preliminary Information

3.1 Score-based Generative Models

SGMs are efficient generative models that can generate high-quality data and model complex data distribution. Given an input $\mathbf{X} \in \mathbb{R}^d$ and a complicated data distribution $\mathcal{D}$, a forward noising process can be obtained through a stochastic differential equation (SDE) as follows:

$$\mathbf{X}_0 \sim \mathcal{D}, \mathrm{d}\mathbf{X}_t = \mathbf{f}(\mathbf{X}_t, t)\mathrm{d}t + \sigma_t \mathrm{d}\mathbf{B}_t,\ t \in [0,1] \tag{1}$$

where σ_t the diffusion coefficient, $\mathbf{B}$ represents the Brownian motion.

Assuming that p_t is a probability density function, the reverse denoising process can be established through the reversed time SDE as follows:

$$\begin{aligned} \mathrm{d}\bar{\mathbf{X}}_t &= (\mathbf{f}(\bar{\mathbf{X}}_t, t) - \sigma_t^2 \nabla \log p_t(\bar{\mathbf{X}}_t))\mathrm{d}\bar{t} + \sigma_t \mathrm{d}\bar{\mathbf{B}}_t \\ \bar{\mathbf{X}}_1 &\sim \mathbf{X}_1,\ t \in [0,1] \end{aligned} \tag{2}$$

where $d\bar{t} = -dt$ is the negative infinitesimal time step, $\bar{\mathbf{B}}$ is the reversed time Brownian motion.

In the reverse denoising process, the scoring network $s(\mathbf{X}(t), t; \theta)$ provides gradient information for the current noise sample $\mathbf{X}_t$, indicating how to adjust the value of the sample so as to gradually restore the original data distribution. During the training process, the scoring network is to minimize the gap between the model estimated score function and the true score function through the score matching loss function as follows:

$$\mathcal{L}_{\mathrm{s}} = \mathbb{E}_{t\sim\mathcal{U}(0,T)}[\mathbb{E}_{\mathbf{X}_t|\mathbf{X}_0}[\|s(\mathbf{X}(t), t; \theta) - \nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|X_0)\|_2^2] \tag{3}$$

where $\mathcal{U}(0,T)$ is a uniform distribution over $[0,T]$. Given a well-trained scoring network s_{θ^*}, we can generate realistic data by solving the learned reverse-time SDE as follows:

$$\begin{aligned} \mathrm{d}\hat{\mathbf{X}}_t &= (\mathbf{f}(\hat{\mathbf{X}}_t, t) - \sigma_t^2 s_{\theta^*}(\hat{\mathbf{X}}_t))\mathrm{d}\bar{t} + \sigma_t \mathrm{d}\bar{\mathbf{B}}_t \\ \hat{\mathbf{X}}_1 &\sim \pi,\ t \in [0,1] \end{aligned} \tag{4}$$

where π is the prior information of the data.

3.2 Score-based Graph Generative Models

In the graph generation model, given a graph $\mathbf{G}(\mathbf{X}, \mathbf{A}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times n}$ with n nodes, where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the feature vectors of each node with dimension d , and $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the connection relationship between nodes. The goal of the graph generation model is to learn the underlying data distribution of the graph, A standard Graph Diffusion Score-based Model [Jo *et al.*, 2022], [Luo *et al.*, 2023] gradually generates a perturbation graph through a diffusion process and learns the generation process of the graph through a scoring network. Specifically, given each graph sample $(\mathbf{X}, \mathbf{A})$, a forward noising process is obtained through an SDE as follows:

$$\begin{cases} \mathrm{d}\mathbf{X}_t = \mathbf{f}^X(\mathbf{X}_t, t)\mathrm{d}t + \sigma_{X,t}\mathrm{d}\mathbf{B}_t^X \\ \mathrm{d}\mathbf{A}_t = \mathbf{f}^A(\mathbf{A}_t, t)\mathrm{d}t + \sigma_{A,t}\mathrm{d}\mathbf{B}_t^A \end{cases} \tag{5}$$

where $\sigma_{X,t}$, and $\sigma_{A,t}$ the diffusion coefficient, $\mathbf{B}_t^A$, and $\mathbf{B}_t^X$ represent the Brownian motion.

Assuming that p_t is a probability density function, the reverse denoising process can be established through the reversed time SDE as follows:

$$\begin{cases} \mathrm{d}\bar{\mathbf{X}}_t = \left(\mathbf{f}^X(\bar{\mathbf{X}}_t, t) - \sigma_{X,t}^2 \nabla_{\mathbf{X}} \log p_t(\bar{\mathbf{X}}_t, \bar{\mathbf{A}}_t)\right)\mathrm{d}\bar{t} + \sigma_{X,t}\mathrm{d}\bar{\mathbf{B}}_t^X \\ \mathrm{d}\bar{\mathbf{A}}_t = \left(\mathbf{f}^A(\bar{\mathbf{A}}_t, t) - \sigma_{A,t}^2 \nabla_{\mathbf{A}} \log p_t(\bar{\mathbf{X}}_t, \bar{\mathbf{A}}_t)\right)\mathrm{d}\bar{t} + \sigma_{A,t}\mathrm{d}\bar{\mathbf{B}}_t^A \end{cases} \tag{6}$$

where $\bar{\mathbf{B}}_t^A$, and $\bar{\mathbf{B}}_t^X$ represent the reversed time Brownian motion.

Given $\mathbf{G}_0$, the joint probability distribution of $\mathbf{X}_t$ and $\mathbf{A}_t$ can be simplified to the product of two simpler distributions [Jo *et al.*, 2022], so that the objective function of denoising score matching can be simplified in form as follows:

$$\begin{aligned} \mathcal{L}_s^\theta &= \mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{\mathbf{G}_t|\mathbf{G}}\|s_{\boldsymbol{\theta}} - \nabla \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0)\|^2 \\ \mathcal{L}_s^\phi &= \mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{\mathbf{G}_t|\mathbf{G}}\|s_{\boldsymbol{\phi}} - \nabla \log p_{t|0}(\mathbf{A}_t|\mathbf{A}_0)\|^2 \end{aligned} \tag{7}$$

4 Problem Definition

Assume that the set $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M\}$ represents the M modalities, where $\mathbf{X}_k$ represents the input of the k-th modality. We introduce a binary indicator $\alpha \in \{0,1\}$ to identify the availability status of each modality. If the k-th modality is missing, let $\alpha_k = 0$; conversely, if the k-th modality is available, let $\alpha_k = 1$. We can define a set of missing modalities $\mathcal{I}_m = \{k|\alpha_k = 0\}$. In this incomplete modality scenario, the goal is to recover these unobserved modalities to make up for the missing information. The process of modal recovery usually needs to rely on the existing observed modal information $\mathcal{I}_o = \{k|\alpha_k = 1\}$, and complete it by modeling the correlation between modalities.

Our main idea is to recover the missing emotion modality $\mathcal{I}_m$ from its latent distribution space conditioned on the observed modality $\mathcal{I}_{\mathrm{o}}$. We use the observed modality $\mathcal{I}_{\mathrm{o}}$ as a semantic condition to guide the generation of the missing modality, ensuring that the recovered modality data is consistent and relevant to the real data. Formally, we denote the data distribution of the missing modality as $p(\mathbf{X}_m)$ and the data distribution of the available modality as $p(\mathbf{X}_{\mathcal{I}_{\mathrm{o}}})$. Our ultimate goal is to sample the missing modality data from the conditional distribution $p(\mathbf{X}_m|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}})$. Inspired by graph completion networks [Lian *et al.*, 2023] and diffusion modality

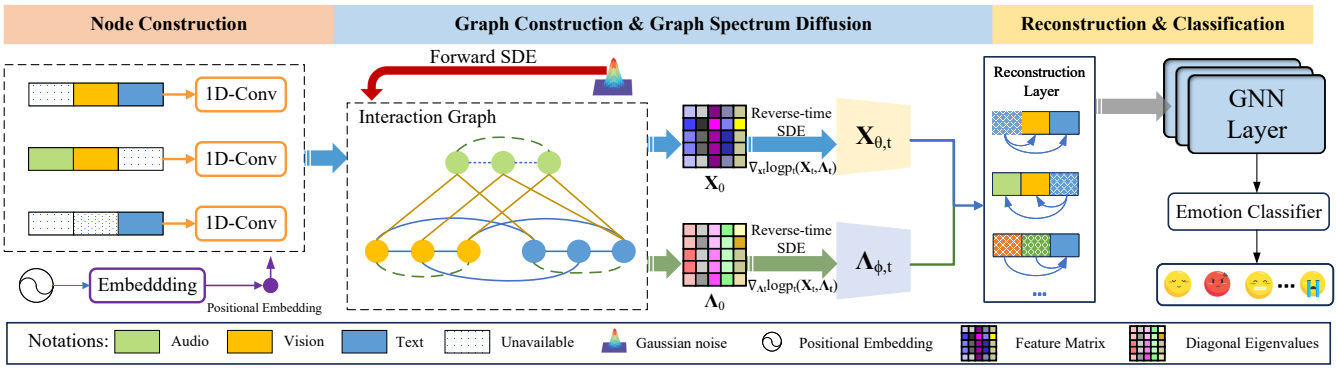


Figure 2: The framework of GSDNet. Given incomplete input data, GSDNet encodes shallow features through 1D-Conv and combines position embedding information. In the missing modal graph diffusion network, we sample from the prior noise distribution and add it to the node features and diagonal eigenvalues, and then solve the inverse time diffusion through the score model to denoise the features to generate new samples. Finally, the reconstructed features are used as complete data to predict the emotion label.

generation [Wang *et al.*, 2024], we combine the advantages of GNNs and diffusion models to simultaneously model the complementary semantic information between modalities and reconstruct high-quality missing modality features.

However, directly adding Gaussian noise to the adjacency matrix may seriously destroy the local structure of the graph, resulting in an unreasonable generated adjacency matrix. To overcome these problems, we propose to strictly restrict the diffusion of Gaussian noise to the spectral space of the adjacency matrix. On the one hand, by operating the eigenvalues in the spectral space, the direct destruction of the local structure of the adjacency matrix can be effectively avoided, ensuring that the generated graph still conforms to the global semantics. On the other hand, this constrained diffusion process can more naturally capture and preserve the topological information of the graph, ensuring that the generated graph has good connectivity and consistent spectral features.

Specifically, we consider a multi-step diffusion model to gradually construct the conditional distribution by perturbing $\mathbf{X}_m$ and $\mathbf{\Lambda}_m$, where $\mathbf{A}_m = \mathbf{U}_m\mathbf{\Lambda}_m\mathbf{U}_m$, $\mathbf{U}$ are the eigenvectors and $\mathbf{\Lambda}$ is the diagonal eigenvalues. In the t-th step, the conditional transfer distribution of the modal features and the adjacency matrix can be expressed as $p_t(\mathbf{X}_m(t)|\mathbf{X}_{\mathcal{I}_o}(0))$ and $p_t(\mathbf{\Lambda}_m(t)|\mathbf{\Lambda}_{\mathcal{I}_o}(0))$ and can be approximated as follows:

$$
\begin{aligned}
&p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0)) \\
&= \int p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t), \mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0)) p_t(\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0)) \mathrm{d}\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t) \\
&\approx \int p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)) p_t(\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0)) \mathrm{d}\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t) \\
&p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0)) \\
&= \int p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t), \mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0)) p_t(\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0)) \mathrm{d}\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t) \\
&\approx \int p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)) p_t(\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0)) \mathrm{d}\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)
\end{aligned}
\tag{8}
$$

According to the score-based diffusion model [Jo *et al.*, 2022], we calculated the conditional transition probability score $p_t(\mathbf{X}_m(t)|\mathbf{X}_{\mathcal{I}_o}(0))$ and $p_t(\mathbf{\Lambda}_m(t)|\mathbf{\Lambda}_{\mathcal{I}_o}(0))$ as follows:

$$
\begin{aligned}
&\nabla_{\mathbf{X}_{\mathrm{m}}} \log p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0)) \\
&\approx \nabla_{\mathbf{X}_{\mathrm{m}}} \log \mathbb{E}_{p_t(\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0))}[p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t))] \\
&\approx \nabla_{\mathbf{X}_{\mathrm{m}}} \log p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)) \\
&= \nabla_{\mathbf{X}_{\mathrm{m}}} \log p_t([\mathbf{X}_{\mathrm{m}}(t); \mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)]) \\
&\nabla_{\mathbf{\Lambda}_{\mathrm{m}}} \log p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0)) \\
&\approx \nabla_{\mathbf{\Lambda}_{\mathrm{m}}} \log \mathbb{E}_{p_t(\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0))}[p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t))] \\
&\approx \nabla_{\mathbf{\Lambda}_{\mathrm{m}}} \log p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)) \\
&= \nabla_{\mathbf{\Lambda}_{\mathrm{m}}} \log p_t([\mathbf{\Lambda}_{\mathrm{m}}(t); \mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)])
\end{aligned}
\tag{9}
$$

where $\mathbf{X}_{\mathcal{I}_o}(t)$ and $\mathbf{\Lambda}_{\mathcal{I}_o}(t)$ is a random sample from $p_t(\mathbf{X}_{\mathcal{I}_o}(t))|\mathbf{X}_{\mathcal{I}_o}(0))$ and $p_t(\mathbf{\Lambda}_{\mathcal{I}_o}(t))|\mathbf{\Lambda}_{\mathcal{I}_o}(0))$, respectively. Eq. 9 is held because:

$$
\begin{aligned}
&\nabla_{\mathbf{X}_m} \log p_t([\mathbf{X}_{\mathrm{m}}(t); \mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)]) \\
&= \nabla_{\mathbf{X}_m} \log p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)) + \nabla_{\mathbf{X}_m} \log p_t(\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)) \\
&= \nabla_{\mathbf{X}_m} \log p_t(\mathbf{X}_{\mathrm{m}}(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(t)) \\
&\nabla_{\mathbf{\Lambda}_m} \log p_t([\mathbf{\Lambda}_{\mathrm{m}}(t); \mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)]) \\
&= \nabla_{\mathbf{\Lambda}_m} \log p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)) + \nabla_{\mathbf{A}_m} \log p_t(\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t)) \\
&= \nabla_{\mathbf{\Lambda}_m} \log p_t(\mathbf{\Lambda}_{\mathrm{m}}(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(t))
\end{aligned}
\tag{10}
$$

Finally, we derive the score-matching objective as follows:

$$
\begin{aligned}
\mathcal{L}_s^{\theta} &= \mathbb{E}_{\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}, \mathbf{X}_m, t\sim\mathcal{U}(0,T)} \mathbb{E}_{\mathbf{G}_t|\mathbf{G}} \| s_{\boldsymbol{\theta}} - \nabla \log p_{t|0}(\mathbf{X}_m(t)|\mathbf{X}_{\mathcal{I}_{\mathrm{o}}}(0)) \|^2 \\
\mathcal{L}_s^{\phi} &= \mathbb{E}_{\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}, \mathbf{\Lambda}_m, t\sim\mathcal{U}(0,T)} \mathbb{E}_{\mathbf{G}_t|\mathbf{G}} \| s_{\boldsymbol{\phi}} - \nabla \log p_{t|0}(\mathbf{\Lambda}_m(t)|\mathbf{\Lambda}_{\mathcal{I}_{\mathrm{o}}}(0)) \|^2
\end{aligned}
\tag{11}
$$

5 The Proposed Method

5.1 Modality Encoder

Since the original features of text, audio, and video modalities usually have significant dimensional differences, directly using the original features of the modalities to recover missing modalities may lead to difficulties in semantic alignment or even introduce noise.To ensure that the unimodal sequence representations of the three modalities can be mapped to in the same feature space, we input these modalities into a one-dimensional convolutional layer to achieve feature alignment:

$$
\mathbf{X}'_{\mathrm{m}} = \mathrm{Conv1D}\left(\mathbf{X}_{\mathrm{m}}, l_m\right) \in \mathbb{R}^{N\times d},\ m \in \{t, a, v\} \tag{12}
$$

where l_m represents the size of the one-dimensional convolution kernel corresponding to the m-th modality. N represents the number of utterances in the conversation. d represents the dimension of the common feature space.

To make full use of the position and order information in the sequence, we introduced position embedding when processing the sequence after convolution:

$$\begin{aligned}\mathbf{PE}_{(pos,2i)} &= \sin\left(\frac{pos}{10000^{2i/d}}\right)\\ \mathbf{PE}_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{2i/d}}\right)\end{aligned} \tag{13}$$

where pos represents the index of the sequence, and dimension i represents the index of the feature dimension.

Overall, we feed position embeddings into a convolutional sequence as follows:

$$\mathbf{X}'_{\mathrm{m}} = \mathbf{X}'_{\mathrm{m}} + \mathbf{PE} \tag{14}$$

5.2 Missing Modality Graph Spectral Diffusion Network

We train two scoring networks s_θ and s_ϕ to model the distribution of missing modalities $m \in \mathcal{I}_m$, respectively. Similar to score-based diffusion [Jo *et al.*, 2022], the corresponding inverse-time SDE can be derived as follows:

$$\begin{aligned}\mathrm{d}\mathbf{X}'_m &= (\mathbf{f}(\mathbf{X}'_m, t) - \sigma^2_{X,t} s_\theta(\mathbf{X}'_m, \mathbf{X}'_{\mathcal{I}_\mathrm{o}}, t; \theta_m))\mathrm{d}\bar{t} + \sigma_{X,t}\mathrm{d}\bar{\mathbf{B}}^X_t\\ \mathrm{d}\mathbf{\Lambda}'_m &= (\mathbf{f}(\mathbf{\Lambda}'_m, t) - \sigma^2_{\Lambda,t} s_\phi(\mathbf{\Lambda}'_m, \mathbf{\Lambda}'_{\mathcal{I}_\mathrm{o}}, t; \theta_m))\mathrm{d}\bar{t} + \sigma_{\Lambda,t}\mathrm{d}\bar{\mathbf{B}}^\Lambda_t\end{aligned} \tag{15}$$

The core of Eq. 15 is to guide the inverse diffusion process through the trained score network, gradually transforming random noise into missing modal data consistent with the true distribution. We assume that the language modality $\mathbf{X}_l$, the visual modality $\mathbf{X}_v$, and the corresponding diagonal eigenvalues $\mathbf{\Lambda}_l$ and $\mathbf{\Lambda}_v$ are observed, while the acoustic modality $\mathbf{X}_a$ and the corresponding diagonal eigenvalues $\mathbf{\Lambda}_a$ are missing. Our goal is to model the missing acoustic modality $\mathbf{X}_a$ and $\mathbf{\Lambda}_a$, and recover their data through the score network s_θ and s_ϕ conditioned on $\mathbf{X}_l$, $\mathbf{X}_v$, $\mathbf{\Lambda}_l$ and $\mathbf{\Lambda}_v$ as follows:

$$\begin{aligned}\mathbf{X}'_a(t-\Delta t) &= \mathbf{X}'_a(t) - \mathbf{f}(\mathbf{X}'_m, t)\\ &\quad + \sigma^2_{X,t} s_\theta(\mathbf{X}'_a(t), [\mathbf{X}'_l; \mathbf{X}'_v](t), t; \theta_a)\Delta t\\ &\quad + \sigma_{X,t}\sqrt{\Delta t}\epsilon_{X,t}\\ \mathbf{\Lambda}'_a(t-\Delta t) &= \mathbf{\Lambda}'_a(t) - \mathbf{f}(\mathbf{\Lambda}'_m, t)\\ &\quad + \sigma^2_{\Lambda,t} s_\phi(\mathbf{\Lambda}'_a(t), [\mathbf{\Lambda}'_l; \mathbf{\Lambda}'_v](t), t; \theta_a)\Delta t + \sigma_{\Lambda,t}\sqrt{\Delta t}\epsilon_t\end{aligned} \tag{16}$$

where Δt is a discrete time step size and $\epsilon_t \sim \mathcal{N}(0, I)$. After enough iterations, we can gradually guide the noise data to approach the target distribution and finally obtain the restored acoustic modal data $\mathbf{X}'_a$ and the corresponding diagonal eigenvalues $\mathbf{\Lambda}'_a$. To generate more refined acoustic modalities, we input the restored acoustic data xa into a specially designed acoustic modal reconstruction module $\mathcal{D}_X$ and diagonal eigenvalues reconstruction module $\mathcal{D}_\Lambda$ to obtain the final reconstructed acoustic modal $\hat{\mathbf{X}}_a = \mathcal{D}_X(\mathbf{X}'_a)$ and diagonal eigenvalues $\hat{\mathbf{\Lambda}}_a = \mathcal{D}_\Lambda(\mathbf{\Lambda}'_a)$. We define a reconstruction loss function $\mathcal{L}_{rec}$ to measure the difference between the reconstructed data and the original target modal data under any missing patterns as follows:

$$\mathcal{L}_{\text{rec}} = \sum_{i\in\mathcal{I}_\mathrm{m}} \left\|\hat{\mathbf{X}}_i - \mathbf{X}_i\right\|_2^2 + \left\|\hat{\mathbf{\Lambda}}_i - \mathbf{\Lambda}_i\right\|_2^2 \tag{17}$$

Therefore, the loss of the missing modality graph diffusion network is as follows:

$$\mathcal{L}_{\text{miss}} = \mathcal{L}_{\text{rec}} + \mathcal{L}^\theta_s + \mathcal{L}^\phi_s \tag{18}$$

5.3 Multimodal Fusion and Prediction

The recovered data and the observed available data are combined to obtain the complete multimodal data $\mathbf{H}$ and the adjacency matrix $\mathbf{A}$. To achieve the fusion of multimodal data, we use GCN to capture the complementary semantic information between the modalities as follows:

$$\mathbf{H}^{(l+1)} = \mathbf{ReLU}(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{19}$$

where $\mathbf{W}^{(l)}$ is the learnable weight matrix of the l-th layer.

To train the entire model, we combine the losses of the above reconstruction and prediction tasks into a joint optimization objective function as follows:

$$\mathcal{L}_{\text{total}} = \beta\mathcal{L}_{\text{miss}} + \mathcal{L}_{\text{pred}} \tag{20}$$

where β is a hyperparameter.

6 Experimental Database and Setup

6.1 Datasets

We conduct extensive experiments on two MERC datasets to conduct experiments, including CMU-MOSI [Zadeh *et al.*, 2016], and CMU-MOSEI [Zadeh *et al.*, 2018]. On the two datasets, we extract the lexical modality features via pre-trained RoBERTa-Large model [Liu *et al.*, 2019] and obtain a 1024-dimensional word embedding. For visual modality, each video frame was encoded via DenseNet model [Huang *et al.*, 2017] and obtain a 1024-dimensional visual feature. The acoustic modality was processed by wav2vec [Schneider *et al.*, 2019] to obtain the 512-dimensional acoustic features.

6.2 Baselines

We compare our proposed method GSDNet to the state-of-the-art incomplete learning methods, including MCTN, MMIN [Zhao *et al.*, 2021], GCNet [Lian *et al.*, 2023], DiCMoR [Wang *et al.*, 2023a], IMDer [Wang *et al.*, 2023b].

7 Results and Discussion

7.1 Comparison with the state-of-the-arts

Tables 1 and 2 lists the quantitative results of the different missing modalities and the random missing ratio on CMU-MOSI and CMU-CMSEI datasets, showing the performance of different methods under the missing modal. Specifically, GSDNet achieved the best results on the two datasets, verifying its superiority in dealing with modal missing. The performance improvement of GSDNet may be attributed to its ability to explicitly restore the missing modality, which not only helps to restore the lost information but also provides additional complementary information for MERC. In

Datasets	Available	MCTN	MMIN	GCNet	DiCMoR	IMDer	GSDNet (Ours)
CMU-MOSI	$\{l\}$	79.1/79.2/41.0	83.8/83.8/41.6	83.7/83.6/42.3	84.5/84.4/44.3	84.8/84.7/44.8	**86.4/86.6/45.7**
	$\{v\}$	55.0/54.4/16.3	57.0/54.0/15.5	56.1/55.7/16.9	62.2/60.2/20.9	61.3/60.8/22.2	**64.1/63.7/25.3**
	$\{a\}$	56.1/54.5/16.5	55.3/51.5/15.5	56.1/54.5/16.6	62.2/60.2/20.9	62.0/62.2/22.0	**64.4/64.1/24.6**
	$\{l, v\}$	81.1/81.2/42.1	83.8/83.9/42.0	84.3/84.2/43.4	85.5/85.4/45.2	85.5/85.4/45.3	**86.5/86.4/46.7**
	$\{l, a\}$	81.0/81.0/43.2	84.0/84.0/42.3	84.5/84.4/43.4	85.5/85.5/44.6	85.4/85.3/45.0	**86.7/86.6/46.8**
	$\{v, a\}$	57.5/57.4/16.8	60.4/58.5/19.5	62.0/61.9/17.2	64.0/63.5/21.9	63.6/63.4/23.8	**65.2/64.8/24.9**
	$\{l, v, a\}$	81.4/81.5/43.4	84.6/84.4/44.8	85.2/85.1/44.9	85.7/85.6/45.3	85.7/85.6/45.3	**87.7/87.3/46.8**
	Average	70.2/69.9/31.3	72.7/71.4/31.6	73.1/72.8/32.1	75.4/75.1/34.7	75.5/75.3/35.5	**77.3/77.1/37.3**
CMU-MOSEI	$\{l\}$	82.6/82.8/50.2	82.3/82.4/51.4	83.0/83.2/51.2	84.2/84.3/52.4	84.5/84.5/52.5	**86.6/86.1/55.3**
	$\{v\}$	62.6/57.1/41.6	59.3/60.0/40.7	61.9/61.6/41.7	63.6/63.6/42.0	63.9/63.6/42.6	**65.1/65.7/44.9**
	$\{a\}$	62.7/54.5/41.4	58.9/59.5/40.4	60.2/60.3/41.1	62.9/60.4/41.4	63.8/60.6/41.7	**64.6/64.2/43.1**
	$\{l, v\}$	83.2/83.2/50.4	83.8/83.4/51.2	84.3/84.4/51.1	84.9/84.9/53.0	85.0/85.0/53.1	**87.3/87.0/56.2**
	$\{l, a\}$	83.5/83.3/50.7	83.7/83.3/52.0	84.3/84.4/51.3	85.0/84.9/52.7	85.1/85.1/53.1	**86.2/86.4/55.5**
	$\{v, a\}$	63.7/62.7/42.1	63.5/61.9/41.8	64.1/57.2/42.0	65.2/64.4/42.4	64.9/63.5/42.8	**66.7/66.3/45.2**
	$\{l, v, a\}$	84.2/84.2/51.2	84.3/84.2/52.4	85.2/85.1/51.5	85.1/85.1/53.4	85.1/85.1/53.4	**87.3/87.2/54.9**
	Average	74.6/72.5/46.8	73.7/73.5/47.1	74.7/73.7/47.1	75.8/75.4/48.2	76.0/75.3/48.5	**77.7/77.6/50.7**

Table 1: The performance of different methods is shown under different missing modalities on the CMU-MOSI and CMU-MOSEI datasets. The values reported in each cell represent the ACC_2/F1/ACC_7. Bold indicates the best performance.

Datasets	Missing Rate	MCTN	MMIN	GCNet	DiCMoR	IMDer	GSDNet (Ours)
CMU-MOSI	0.0	81.4/81.5/43.4	84.6/84.4/44.8	85.2/85.1/44.9	85.7/85.6/45.3	85.7/85.6/45.3	**87.7/87.3/46.8**
	0.1	78.4/78.5/39.8	81.8/81.8/41.2	82.3/82.3/42.1	83.9/83.9/43.6	84.9/84.8/44.8	**87.1/86.5/46.2**
	0.2	75.6/75.7/38.5	79.0/79.1/38.9	79.4/79.5/40.0	83.9/83.9/43.6	83.5/83.4/44.3	**86.4/86.1/45.2**
	0.3	71.3/71.2/35.5	76.1/76.2/36.9	77.2/77.2/38.2	80.4/80.2/40.6	81.2/81.0/42.5	**85.2/85.0/44.3**
	0.4	68.0/67.6/32.9	71.7/71.6/34.9	74.3/74.4/36.6	77.9/77.7/37.6	78.6/78.5/39.7	**83.3/82.9/42.1**
	0.5	65.4/64.8/31.2	67.2/66.5/32.2	70.0/69.8/33.9	76.7/76.4/36.4	76.2/75.9/37.9	**81.2/81.1/40.6**
	0.6	63.8/62.5/29.7	64.9/64.0/29.1	67.7/66.7/29.8	73.3/73.0/32.7	74.7/74.0/35.8	**80.1/79.7/38.7**
	0.7	61.2/59.0/27.5	62.8/61.0/28.4	65.7/65.4/28.1	71.1/70.8/30.0	71.9/71.2/33.4	**77.6/77.3/35.6**
	Average	70.6/70.1/34.8	73.5/73.1/35.8	75.2/75.1/36.7	78.9/78.7/38.5	79.6/79.3/40.5	**83.6/83.2/42.3**
CMU-MOSEI	0.0	84.2/84.2/51.2	84.3/84.2/52.4	85.2/85.1/51.5	78.9/78.7/38.5	85.1/85.1/53.4	**87.3/87.2/54.9**
	0.1	81.8/81.6/49.8	81.9/81.3/50.6	82.3/82.1/51.2	78.9/78.7/38.5	84.8/84.6/53.1	**86.7/86.5/54.2**
	0.2	79.0/78.7/48.6	79.8/78.8/49.6	80.3/79.9/50.2	81.8/81.5/51.4	82.7/82.4/52.0	**85.3/85.1/53.5**
	0.3	76.9/76.2/47.4	77.2/75.5/48.1	77.5/76.8/49.2	79.8/79.3/50.3	81.3/80.7/51.3	**83.3/83.0/52.2**
	0.4	74.3/74.1/45.6	75.2/72.6/47.5	76.0/74.9/48.0	78.7/77.4/48.8	79.3/78.1/50.0	**81.4/81.2/51.4**
	0.5	73.6/72.6/45.1	73.9/70.7/46.7	74.9/73.2/46.7	77.7/75.8/47.7	79.0/77.4/49.2	**80.5/80.1/50.7**
	0.6	73.2/71.1/43.8	73.2/70.3/45.6	74.1/72.1/45.1	77.7/75.8/47.7	78.0/75.5/48.5	**79.4/79.1/49.4**
	0.7	72.7/70.5/43.6	73.1/69.5/44.8	73.2/70.4/44.5	75.4/72.2/46.2	77.3/74.6/47.6	**78.2/78.1/48.6**
	Average	77.0/76.1/46.9	77.3/75.4/48.2	77.9/76.8/48.3	79.9/78.6/49.6	80.9/79.8/50.6	**82.8/82.5/51.9**

Table 2: The performance of different methods is shown at different missing ratios on the CMU-MOSI and CMU-MOSEI datasets. The values reported in each cell represent the ACC_2/F1/ACC_7. Bold indicates the best performance.

addition, GSDNet has a significant advantage in maintaining consistency between the restored modality and the original modality. This distribution consistency ensures that the information fusion between different modalities is smoother and more accurate, further improving the overall performance of the model. Compared with other MERC methods, the performance degradation of GSDNet decreases as the modal missing rate increases. In practical applications, when the modal missing rate is high, most recovery-based models will experience significant performance degradation.

7.2 Ablation study

We conduct ablation experiments on the CMU-MOSI and CMU-MOSEI datasets. The results in Table 3 show that GSDNet consistently outperforms all variants. Removing the frequency diffusion degrades the performance, which highlights the role of frequency diffusion in capturing the distribution of multimodal data.

Methods	CMU-MOSI			CMU-MOSEI		
	ACC_2	F1	ACC_7	ACC_2	F1	ACC_7
GSDNet	75.7	70.6	35.3	78.1	77.4	47.4
GSDNet w/spectral	**83.6**	**83.2**	**42.3**	**82.8**	**82.5**	**51.9**

Table 3: Ablation study of graph spectral diffusion on GSDNet under average random missing ratios.

7.3 Visualization of Embedding Space

Fig. 3 shows the distribution of the restored data and the original data in the feature space obtained by different restoration methods under the condition of fixed missing modalities. In order to compare these distributions more intuitively, we

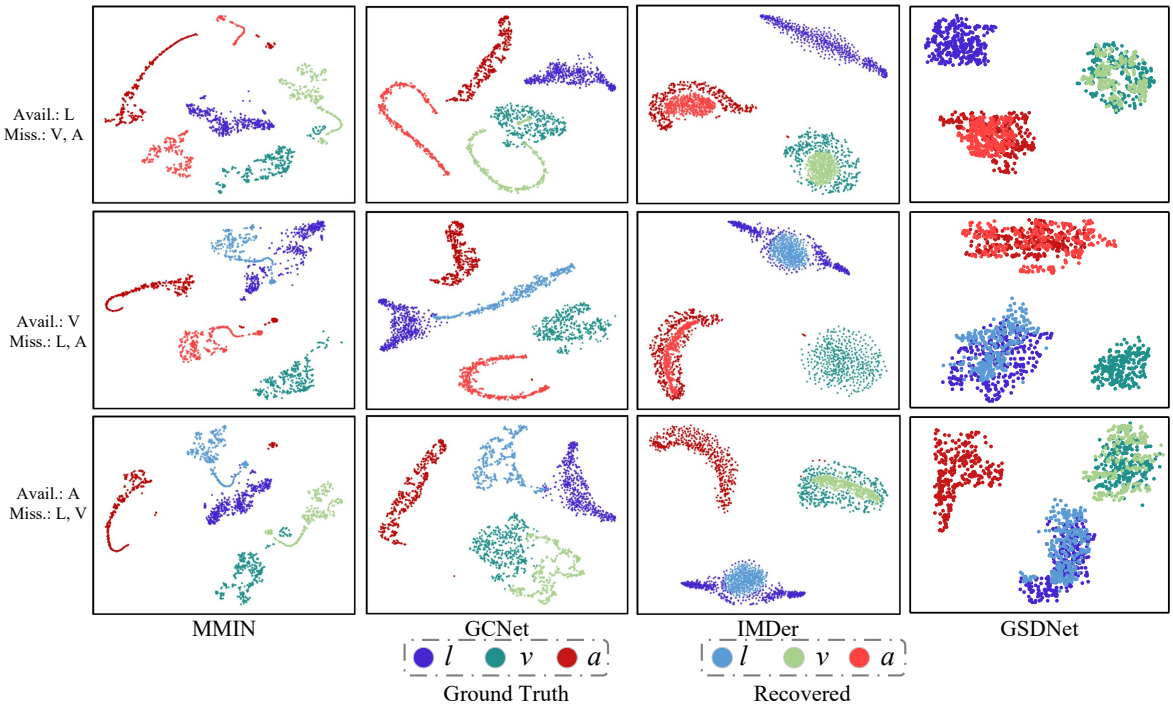
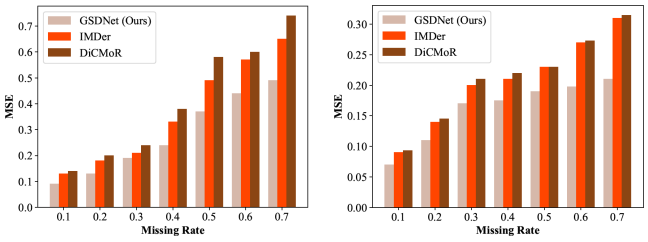


Figure 3: Visualization of restored modalities. Avail. indicates available.

use t-SNE dimensionality reduction technology on the CMU-MOSEI dataset to project the high-dimensional features into two-dimensional space for visualization. As can be seen from Fig. 3, the modal data restored by GSDNet is closest to the distribution of the original data, which shows that GSDNet can better maintain the original feature distribution of the data when restoring the missing modalities. In contrast, there is a clear difference between the distribution of the restored data of other methods and the original data, especially in some local areas, the degree of overlap of the distribution is low.



(a) CMU-MOSI (b) CMU-MOSEI

Figure 4: The comparison of interpolation performance under different missing rates shows the interpolation effects of various methods when dealing with different missing rates.

7.4 Imputation Performance

Fig. 4 shows the interpolation results of different methods under different missing rates. By comparing the performance of baseline models under different missing rates, our proposed GSDNet always outperforms other baseline methods in the CMU-MOSI and CMU-MOSEI datasets and all missing rate conditions. Specifically, GSDNet not only shows strong interpolation performance under low missing rates but also has more outstanding performance advantages under high missing rates. The experimental results show that speaker dependency and data distribution consistency play a vital role in data interpolation tasks. Most baseline methods often ignore the synergy of these dependencies, which limits their interpolation performance when dealing with missing data. In contrast, GSDNet can use the speakers relationship to perform more accurate interpolation while maintaining data distribution consistency through the graph diffusion model, so that GSDNet can always maintain relatively good performance under various missing rates.

8 Conclusions

In this paper, we introduce a novel GSDNet, which maps Gaussian noise to the graph spectral space of missing modalities and recover the missing data according to original distribution. GSDNet only affects the eigenvalues of the adjacency matrix instead of destroying the adjacency matrix directly, which can maintain the global topological information and important spectral features during the diffusion process. Extensive experiments have demonstrated that GSDNet achieves state-of-the-art emotion recognition performance in various modality loss scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62372478), the Research Foundation of Education Bureau of Hunan Province of China (Grant No. 22B0275), and the Hunan Provincial Natural Science Foundation Youth Project (Grant No. 2025JJ60420).

References

[Ding *et al.*, 2023] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Hu *et al.*, 2021] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, 2021.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[Jo *et al.*, 2022] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International conference on machine learning*, pages 10362–10383. PMLR, 2022.

[Kingma and Dhariwal, 2018] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31, 2018.

[Li *et al.*, 2023] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023.

[Lian *et al.*, 2023] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Luo *et al.*, 2023] Tianze Luo, Zhanfeng Mo, and Sinno Jialin Pan. Fast graph generation via spectral diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Parthasarathy and Sundaram, 2020] Srinivas Parthasarathy and Shiva Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, 2020.

[Pham *et al.*, 2019] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899, 2019.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. Pmlr, 2021.

[Schneider *et al.*, 2019] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of the Interspeech*, pages 3465–3469, 2019.

[Song and Ermon, 2019] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[Song and Ermon, 2020] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems*, 33:12438–12448, 2020.

[Song *et al.*,] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

[Tran *et al.*, 2017] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[Wang *et al.*, 2023a] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034, 2023.

[Wang *et al.*, 2023b] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[Wang *et al.*, 2024] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

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

[Zhang *et al.*, 2020] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2402–2415, 2020.

[Zhang *et al.*, 2024] Yunhua Zhang, Hazel Doughty, and Cees Snoek. Learning unseen modality interaction. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhao *et al.*, 2021] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021.