Unbiased Missing-modality Multimodal Learning

Ruiting Dai[1], Chenxi Li[1], Yandong Yan[2], Lisi Mo[1], Ke Qin[1,3], Tao He[1,3*]
[1]University of Electronic Science and Technology of China
[2]School of Computer Science, Peking University
[3]Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province
{rtdai,qinke,morris} @uestc.edu.cn, chenxi.li@std.uestc.edu.cn,
ai_yan@stu.pku.edu.cn, tao.he01@hotmail.com

Abstract

*Recovering missing modalities in multimodal learning has recently been approached using diffusion models to synthesize absent data conditioned on available modalities. However, existing methods often suffer from modality generation bias: while certain modalities are generated with high fidelity, others—such as video—remain challenging due to intrinsic modality gaps, leading to imbalanced training. To address this issue, we propose **MD^{2}N** (Multi-stage Duplex Diffusion Network), a novel framework for unbiased missing-modality recovery. MD^{2}N introduces a modality transfer module within a duplex diffusion architecture, enabling bidirectional generation between available and missing modalities through three stages: (1) global structure generation, (2) modality transfer, and (3) local cross-modal refinement. By training with duplex diffusion, both available and missing modalities generate each other in an intersecting manner, effectively achieving a balanced generation state. Extensive experiments demonstrate that MD^{2}N significantly outperforms existing state-of-the-art methods, achieving up to 4% improvement over IMDer on the CMU-MOSEI dataset. Project page: here.*

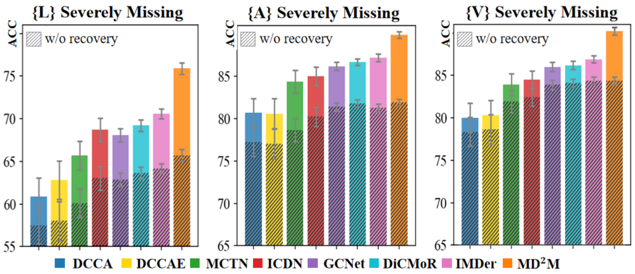

Figure 1. We evaluate state-of-the-art recovery-based incomplete multimodal learning models [3, 40, 50–52] on the MOSEI dataset under severe missing-modality scenarios, where **L**, **A**, and **V** denote text, acoustic, and visual modalities. Results show that most models perform unevenly across different missing-modality conditions, especially struggling when video data is missing. In contrast, our model achieves more balanced and robust performance across all missing scenarios.

1. Introduction

Multimodal learning leverages complementary information from heterogeneous data sources such as audio, images, and text [4, 7, 20, 56], achieving impressive success in modeling complex real-world phenomena. Its effectiveness is demonstrated across diverse applications, including visual question answering (VQA) [25] and affective computing [1], etc. Despite this progress, most existing multimodal approaches assume the availability of all modalities during both training and inference. This assumption rarely holds in real-world deployments, where data completeness is often compromised by practical issues such as sensor failures [32] or communication bottlenecks [6]. The resulting partial modality availability introduces a critical challenge: it results in a mismatch between training and testing conditions, fundamentally undermining the robustness and generalizability of multimodal learning systems.

Current missing modality learning approaches mainly use deep generative frameworks, such as diffusion models [30], to reconstruct absent modalities conditioned on the available ones. While these methods [26, 30] show promising reconstruction results, we observe a critical limitation: modality generation bias [47, 57]. Specifically, there are significant differences in the difficulty of generating different modalities. For example, text can often be synthesized reliably from visual inputs (e.g., images or videos), but generating high-quality images or videos from text remains more challenging, often resulting in substantial quality degradation (see Fig. 1 for empirical results). This bias leads to imbalanced training, where models perform well in

[*]Corresponding author.

some missing-modality scenarios but fail in others [14]. We argue that this limitation arises from the assumption that all modality generation tasks are of equal difficulty, ignoring the inherent differences between modalities and their generation complexities.

To address the challenge of biased missing modality recovery, we pose a question: *Can the recovery process be performed in an intersecting manner by integrating a further modality transfer process from missing to available modalities?* Motivated by this, we propose MD^{2}N (Multi-stage Duplex Diffusion Network), a novel framework that decomposes recovery into three sequential stages: (1) **Global structure generation** $(0, t_1])$: The model first reconstructs the coarse global structure of the target modality by leveraging cross-modal information, establishing a stable foundation for subsequent generation. (2) **Modality transfer** $(t_1, t_2])$: We introduce an intersecting transfer strategy that progressively integrates conditional information, enabling mutual knowledge flow between available and missing modalities, ensuring semantic alignment and learning of modality-invariant representations that capture shared characteristics across modalities. (3) **Local cross-modal refinement**$[t_2, T]$: In this stage, the model enhances local details to refine content quality, ensuring recovered data is both realistic and structurally coherent.

Our framework employs duplex diffusion models, allowing available and missing modalities to generate each other's data in an intersecting manner across all stages. Specifically, we adopt score-based diffusion models [43] as the generative backbone due to their ability to provide direct and flexible control over the reverse diffusion process, which is critical for accurately reconstructing complex modality structures. To further enhance generation stability, we introduce a time-step-based variance function that dynamically adjusts the noise variance at each diffusion step. This design effectively mitigates deviations and fluctuations caused by local perturbations, ensuring the generation process remains stable and coherent throughout. As a result, our method maintains global structural integrity while progressively enhancing fine-grained details, leading to high-quality and semantically aligned modality recovery outputs.

In summary, we make the following contributions:

- We empirically identify the issue of modality generation bias in recovery-based missing-modality models, which hinders the effectiveness of missing-modality recovery.
- We introduce a multi-stage diffusion process that decomposes the recovery task into three stages: global structure generation, modality transfer, and local detail refinement.
- We design a duplex cross-diffusion framework that simultaneously handles the diffusion processes for both available and missing modalities, facilitating the learning of modality-invariant knowledge.
- We conduct extensive experiments on several benchmark datasets, demonstrating that our model outperforms existing methods, achieving up to a 4% improvement over the SOTA models (e.g., IMDer) on the CMU-MOSEI dataset.

2. Related Works

Mulimodal Learning. Recent advancements in multimodal learning have led to significant breakthroughs across cross-modal generation [19, 61, 62], contextual learning [21, 45], and modality fusion [13, 15, 67]. For example, Huang *et al.* [21] introduced Multimodal Task Vectors (MTV), which compress multimodal exemplars into attention heads, circumventing context length constraints and enabling many-shot in-context learning. Zhou *et al.* [66] proposed a causal inference framework that leverages counterfactual reasoning and backdoor adjustment to mitigate modality prior-induced hallucinations, thereby enhancing the robustness of multimodal large language models. Yang *et al.* [58] developed ContextDIFF, a conditional diffusion model that propagates cross-modal context throughout the diffusion process, significantly improving text-guided visual synthesis and editing tasks.

In the domain of multimodal sentiment analysis, similar approaches have been widely adopted. For instance, Zhu *et al.* [68] proposed a BERT and Faster R-CNN-based framework that employs co-attention mechanisms and gating strategies to effectively integrate textual and visual features, leading to improved sentiment classification accuracy. Despite these recent innovations, the inherent challenge of missing modalities remains a critical bottleneck in practical multimodal applications, underscoring the need for further research to address this limitation.

Incomplete Multimodal Learning. Early research on missing modality recovery primarily focused on two approaches: removing incomplete samples [38] and recovering missing data [23, 55]. For example, FitRec [38] utilizes multimodal data during both training and prediction but fails to handle incomplete samples, leading to data depletion and model overfitting. Traditional imputation methods, which typically generate the absent modalities [7, 34, 39], encounter significant limitations when consecutive features are missing. Recently, deep learning-based methods, such as autoencoders [17, 18, 28, 37, 42] and Generative Adversarial Networks (GANs) [7, 39], have been applied to restore missing modalities in incomplete multimodal learning scenarios. However, these generative approaches often introduce additional noise, particularly when the number of modalities is large and sample completeness is low [11, 49]. More recently, diffusion models have been employed for modality recovery [12, 16, 41, 52]. For instance, IMDER [52] utilizes a diffusion model to restore missing modalities from Gaussian noise. Nevertheless, these methods consistently overlook a critical issue: modality generation bias, which hinders the quality and reliability of recovered data.

Diffusion Probabilistic Models. Diffusion models have achieved remarkable progress across various domains, including image restoration, 3D generation, and multimodal learning. For image restoration, Chan *et al.* [8] introduced the Dynamic Regulation Diffusion Anchoring mechanism (DRDA) to mitigate artifacts and color biases in low-light enhancement, while Li *et al.* [29] employed decoupled probabilistic modeling with uncertainty-guided attention to achieve high-quality reconstruction of complex textures. In 3D generation, Liu *et al.* [33] utilized 3D point cloud diffusion for modernizing traditional cultural elements, and Jo *et al.* [24] addressed semantic loss in text generation via statistical manifold mapping. In multimodal learning, diffusion models have been primarily applied to missing modality recovery. For example, Kebaili *et al.* [27] proposed an adaptive multimodal missing data completion framework that integrates an image-frequency fusion network (IFFN) with diffusion models to significantly improve medical image segmentation accuracy.

3. The Proposed Method

In this section, we detail our proposed method, covering the diffusion model preliminaries (Sec. 3.1), overall framework (Sec. 3.2), multimodal feature extractor (Sec. 3.3), duplex multi-stage diffused network (Sec. 3.3), and multimodal fusion (Sec. 3.4) for downstream tasks.

3.1. Preliminaries

Variance-preserving Diffusion Models (VPDM). In this work, we adopt the variance-preserving diffusion model [48] as the generator because it was demonstrated that the noise at each step is controlled to ensure stable and effective missing-modality data recovery. Specifically, VPDM is discretized the variance-preserving stochastic differential equation (VP-SDE) [44] with the Euler-Maruyama technique [35] and incorporates a time-step variance function $\beta(t)$ for dynamic noise adjustment as:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_{t-1}, t = 1, 2, \cdots, T, \quad (1)$$

where ϵ_{t-1} represents independent Gaussian noise.
Forward Processing. Following the standard Stochastic Differential Equation (SDE) [2, 43], the VP-SDE uses a variance-preserving mechanism to perturb the data **x**:

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\mathbf{w}, t \in [0, T], \quad (2)$$

where $\beta(t)$ is a time-dependent variance function controlling noise intensity and **w** denotes a Wiener process [64].
Reverse Processing. As [5, 36, 48], the reverse-time VP-SDE for sample generation during the recovery process is:

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)(\mathbf{x}\,\mathrm{d}t - \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\,\mathrm{d}t) + \sqrt{\beta(t)}\,\mathrm{d}\bar{\mathbf{w}} \quad (3)$$

For discrete time steps $t \in 1, 2, \cdots, T$, the reverse process can be represented as an iterative update as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t + \beta_t s_\theta(\mathbf{x}_t, t)) + \sqrt{\beta_t}\epsilon_t \quad (4)$$

where $s_\theta(\mathbf{x}_t, t)$ serves as an approximation of $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ using the score network s_θ [36, 44, 54].
Score-matching Loss Function. To optimize the score-matching loss and the score network s_θ, we leverage the transition kernel $p_{0t}(\mathbf{x}(t)|\mathbf{x}(0))$ in the VP-SDE. This kernel approximates the conditional distribution of the state $\mathbf{x}(t)$ at any time t, given the initial state $\mathbf{x}(0)$. Specifically, the transition kernel follows a Gaussian distribution, with both its mean and covariance determined by the initial state and the cumulative noise function over time. The equation is as:

$$p_{0t}(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}(\mathbf{x}(t); \mathbf{x}(0)\varphi(t), \mathbf{I}(1-\varphi(t)), \quad (5)$$

where $\varphi(t) = e^{-\frac{1}{2}\int_0^t \beta(s)\,\mathrm{d}s}$ is a decay factor that controls the diffusion process up to time t. The term $\mathbf{x}(0)\varphi(t)$ represents the gradual decay of the signal over time, and the covariance $\mathbf{I} - \mathbf{I}\varphi(t)$ reflects the cumulative noise introduced during the diffusion process.

Using this, the optimization of the score-matching loss is formulated as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}(0,T)} \left\| \frac{\epsilon}{\sqrt{\lambda(t)}} + s_\theta(\mathbf{x}(t), t) \right\|_2^2, \quad (6)$$

where $\mathcal{U}(0, T)$ is a uniform distribution over the time interval $[0, T]$, and $\lambda(t) = \mathbf{I} - \mathbf{I}e^{-\frac{1}{2}\int_0^t \beta(s)\,\mathrm{d}s}$ is a weighting function to balance the loss at different time steps.

3.2. Overall Framework

The overall framework, illustrated in Fig. 2(a), consists of three core components: Multimodal Feature Extractor, Multi-stage Duplex Diffusion Network, and Multimodal Fusion module. In the feature extractor, three independent encoders, $\mathcal{E}_K$ for $K \in \{L, V, A\}$, are utilized to extract features from the text (L), vision (V), and acoustic (A) modalities, respectively. The Multi-stage Duplex Diffusion Network adopts a cross-diffusion architecture to generate both available and missing modality data through three sequential stages: (1) Global Structure Generation, where cross-modal information is integrated to establish a coherent global structural framework; (2) Modality Transfer, which is progressively introduced to prevent premature modality dominance and ensure semantic alignment across modalities; and (3) Local Detail Refinement, which enhances fine-grained features to improve the quality and authenticity of the generated samples. Finally, the Multimodal Fusion module consolidates the recovered and available modality representations and utilizes a multimodal Transformer for downstream prediction tasks.

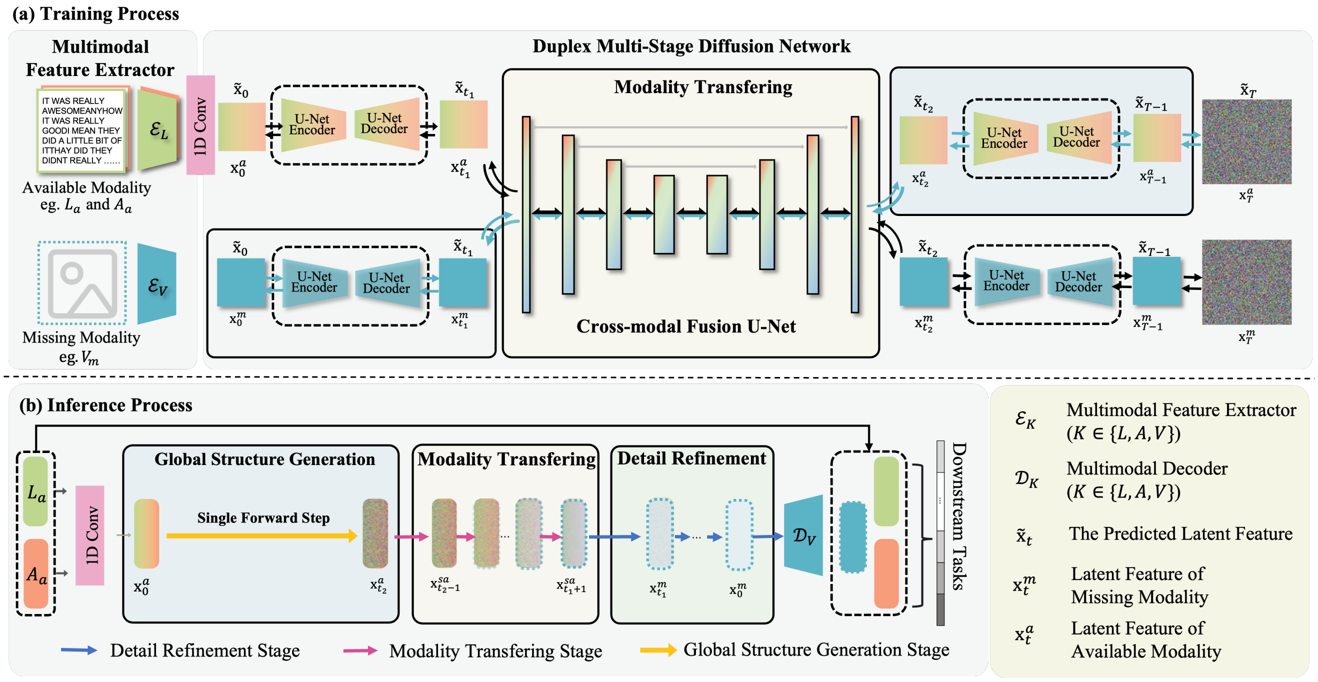
(a) Training Process
Multimodal Feature Extractor
IT WAS REALLY AWESOMEANYHOW IT WAS REALLY GOODI MEAN THEY DID A LITTLE BIT OF ITTHAY DID THEY DIDNT REALLY ......
$\mathcal{E}_L$
Available Modality eg. L_a and A_a
$\mathcal{E}_V$
Missing Modality eg. V_m
1D Conv
Duplex Multi-Stage Diffusion Network
$\tilde{x}_0$
x_0^a
U-Net Encoder
U-Net Decoder
$\tilde{x}_{t_1}$
$x_{t_1}^a$
x_0^m
$x_{t_1}^m$
Modality Transfering
Cross-modal Fusion U-Net
$\tilde{x}_{t_2}$
$x_{t_2}^a$
$\tilde{x}_{T-1}$
x_{T-1}^a
$\tilde{x}_T$
x_T^a
$x_{t_2}^m$
x_{T-1}^m
x_T^m
(b) Inference Process
L_a
A_a
1D Conv
Global Structure Generation
Single Forward Step
x_0^a
$x_{t_2}^a$
Modality Transfering
$x_{t_2-1}^{sa}$
$x_{t_1+1}^{sa}$
Detail Refinement
$x_{t_1}^m$
x_0^m
$\mathcal{D}_V$
Downstream Tasks
Detail Refinement Stage
Modality Transfering Stage
Global Structure Generation Stage
$\mathcal{E}_K$ Multimodal Feature Extractor ($K \in \{L, A, V\}$)
$\mathcal{D}_K$ Multimodal Decoder ($K \in \{L, A, V\}$)
$\tilde{x}_t$ The Predicted Latent Feature
x_t^m Latent Feature of Missing Modality
x_t^a Latent Feature of Available Modality

Figure 2. Overview of the proposed framework, which comprises three key components: multimodal feature extraction, a multi-stage duplex diffusion network, and multimodal fusion. Independent encoders first extract features from text (L), vision (V), and acoustic (A) modalities. The diffusion network then recovers missing modalities through three stages—global structure generation establishes a consistent cross-modal foundation, modality transferring achieves semantic alignment, and local detail refinement enhances fine-grained features. Finally, a multimodal Transformer fuses the refined representations for downstream prediction.

3.3. MD^{2}N: Multi-stage Duplex Diffusion Network

Motivation. Existing recovery-based missing modality learning models [51, 52] predominantly utilize conditional generative networks to directly reconstruct missing modality data. However, these approaches often fail to account for the inherent complexities and disparities among modalities, leading to substantial generation biases. We hypothesize that such biases arise from the modality gap, which reflects differences in data characteristics and structural representations across modalities. To address this limitation, we introduce a cross-modality transfer step within the generation process, specifically during the time interval $t \in (t_1, t_2]$ of the overall diffusion process $(0, T]$. This step is designed to preserve modality-invariant knowledge while effectively transferring modality-variant information. To achieve this, we propose Multi-stage Duplex Diffusion Networks (MD^{2}N), where two diffusion models collaborate by reconstructing each other's data through a cross-modality generation module. This collaborative mechanism enables each modality to contribute its unique information, resulting in more accurate, semantically aligned, and unbiased reconstruction of missing modalities.

Multimodal Feature Extractor. As illustrated in Fig. 2(a), the multimodal feature extractors process data from text (L), vision (V), and acoustic (A) modalities. The extraction network, denoted as $\mathcal{E}_k$ for $k \in \{L, V, A\}$, leverages modality-specific pre-trained encoders: BERT [10] for textual data, Facet [22] for visual inputs, and COVAREP [9] for audio. These encoders transform the raw modality data into a latent space. For each input $\mathbf{x}$, the encoders yield representations $\mathcal{X} = \{\mathbf{x}^k\}$, where $\mathbf{x}^k \in \mathbb{R}^{L \times D}$ (with L denoting the sequence length and D the latent dimensionality). Following [51, 52], each sample is assumed to have at least one available modality feature. Notably, during training, all modalities are complete; missing modality data occur only during the inference [40, 52, 53, 65].

Forward Process. As illustrated in Fig. 2, the duplex diffusion networks dilute both the available data $\mathbf{x}^a$ and the missing modality data $\mathbf{x}^m$ into Gaussian noise. Importantly, these two diffusion processes are intersectantly connected via the modality transferring process. We denote the output at each time step as $\tilde{\mathbf{x}}_t$, highlighted in deep blue in Fig. 2. The entire forward process is divided into three time stages:

(1) Stage 1: Global Structure Generation ($t \in (0, t_1]$). In this stage, noise is injected into the features $\mathbf{x}^m(t)$ and $\mathbf{x}^a(t)$. For example, the forward process for $\mathbf{x}^m(t)$, based on Eq. (1), is formulated as:

$$\tilde{\mathbf{x}}_t^m = \sqrt{1 - \beta_t}\mathbf{x}_{t-1}^m + \sqrt{\beta_t}\epsilon_{t-1}. \qquad (7)$$

Similarly, $\tilde{\mathbf{x}}_t^a$ can be obtained through the same process.

(2) Stage 2: Modality Transfer ($t \in (t_1, t_2]$). In this stage, the latent features $\mathbf{x}_t^m$ and $\mathbf{x}_t^a$ undergo cross-time-step transferring until t_2. Taking the transfer from $\mathbf{x}^a \to \mathbf{x}^m$ as an example, the forward process for $\tilde{\mathbf{x}}_t^m$ is written as:

$$\tilde{\mathbf{x}}_t^m = \sqrt{1-\beta_t}\tilde{\mathbf{x}}_{t-1}^m + \sqrt{\beta_t}\epsilon_{t-1} - \Phi(\mathbf{x}_0^m - \mathbf{x}_0^a), \tag{8}$$

where Φ is a coefficient derived by aligning the model with Eq. (1), with its detailed derivation provided in the supplementary materials. The reverse transfer from $\mathbf{x}^m \to \mathbf{x}^a$ follows an analogous formulation.

(3) Stage 3: Local Cross-Modal Refinement ($t \in (t_2, T]$). In the final stage, the injected noise disrupts the latent features $\mathbf{x}_t^a$ and $\mathbf{x}_t^m$ into Gaussian noise. For simplicity, we use $*$ to denote either superscript a or m. The forward process in this stage is expressed as:

$$\tilde{\mathbf{x}}_t^* = \sqrt{1-\beta_t}\mathbf{x}_{t-1}^* + \sqrt{\beta_t}\epsilon_{t-1}. \tag{9}$$

Reverse Process. Based on the time-dependent score model s_θ, we construct the corresponding reverse-time SDE and numerically simulate it to generate samples from p_0. Starting from samples $\mathbf{x}(T) \sim p_T$, the process is reversed to obtain $\mathbf{x}(0) \sim p_0$.

As shown in Eq. (7) and Eq. (9), for the time stages $(0, t_1]$ and $(t_2, T]$, the forward latent $\tilde{\mathbf{x}}_t^*$ aligns with the discretized VP-SDE formulation. Therefore, both stages share the same parameterized reverse iteration, formulated as:

$$\tilde{\mathbf{x}}_{t-1}^* = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t^* + \beta_t s_\theta^{(*)}(\tilde{\mathbf{x}}_t^*, y, t)\right) + \sqrt{\beta_t}\epsilon_t, \tag{10}$$

where $s_\theta^{(*)}(\tilde{\mathbf{x}}_t^*, y, t)$ is the prediction network estimating $\mathbf{x}_0^*$ from the noisy latent feature $\tilde{\mathbf{x}}_t^*$, and y denotes the generated other modalities. Similarly, for the stage $t \in (t_2, T]$, the reverse process is written as:

$$\tilde{\mathbf{x}}_{t-1}^* = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t^* + \beta_t s_\theta^{(*)}(\tilde{\mathbf{x}}_t^*, y, t)\right) + \sqrt{\beta_t}\epsilon_t. \tag{11}$$

Following Eq. (3), the modality transferring process from $\mathbf{x}^a$ to $\mathbf{x}^m$ in $t \in (t_1, t_2]$ is expressed as:

$$\tilde{\mathbf{x}}_{t-1}^a = \frac{1}{\sqrt{1-\beta_t}}(\tilde{\mathbf{x}}_t^a - \Phi(\mathbf{x}_0^m - \mathbf{x}_0^a) + \beta_t s_\theta^{(m)}(\tilde{\mathbf{x}}_t^m, y, t)) + \sqrt{\beta_t}\epsilon_t. \tag{12}$$

where Φ denotes the transfer coefficient derived in the supplementary materials. Conversely, the process from $\mathbf{x}^m$ to $\mathbf{x}^a$ in $t \in (t_1, t_2]$ is formulated as:

$$\tilde{\mathbf{x}}_{t-1}^m = \frac{1}{\sqrt{1-\beta_t}}(\tilde{\mathbf{x}}_t^m - \Phi(\mathbf{x}_0^a - \mathbf{x}_0^m) + \beta_t s_\theta^{(a)}(\tilde{\mathbf{x}}_t^a, y, t)) + \sqrt{\beta_t}\epsilon_t. \tag{13}$$

Diffusion Optimization Objective. As shown in Fig. 2(a), the optimization objectives for $s_\theta^{(a)}(\tilde{\mathbf{x}}_t^a, y, t)$ and $s_\theta^{(m)}(\tilde{\mathbf{x}}_t^m, y, t)$ follow the formulation in Eq. (6), and are denoted as $\mathcal{L}_a$ and $\mathcal{L}_m$, respectively. The overall optimization objective $\mathcal{L}_{\text{score}}$ is defined as:

$$\mathcal{L}_{\text{score}} = \mathcal{L}_a + \mathcal{L}_m, \tag{14}$$

where each $\mathcal{L}_*$ is computed as:

$$\mathcal{L}_* = \mathbb{E}_{\mathbf{x}^*,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}(t_1,T)}\left\| s_\theta^{(*)}(\tilde{\mathbf{x}}_t^*, y, t) + \frac{\epsilon}{\sqrt{\lambda(t)}}\right\|_2^2, \tag{15}$$

and $*$ denotes either m or a. Here, $\lambda(t)$ is the noise scaling factor defined in Eq. (6), and $\mathcal{U}(t_1, T)$ denotes the uniform distribution over the diffusion time interval.

Besides, we deploy a decoder to reconstruct the generated data. Thus, we leverage reconstruction loss to optimize the generated data by:

$$\mathcal{L}_{\text{dec}} = \|\hat{\mathbf{x}}^m - \mathbf{x}^m\|_2^2 + \|\hat{\mathbf{x}}^a - \mathbf{x}^a\|_2^2. \tag{16}$$

By combining $\mathcal{L}_{\text{score}}$ and $\mathcal{L}_{\text{dec}}$, the objective of our multimodal recovery diffused network is $\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{score}} + \mathcal{L}_{\text{dec}}$.

Discussion. Building upon the duplex diffusion networks, we divide the inference recovery process into three distinct stages: global structure generation, modality transfer, and detail refinement. At inference (see Fig. 2(b)), since only the available modality x^a is observed, we utilise the generation direction from x^a to the missing modality x^m. (1) Global structure generation stage ($t \in [T, t_2)$): As illustrated in Fig. 2(c), this stage begins by sampling Gaussian noise $\tilde{z}_T$ and generating the coarse structural representation of the missing modality x^m by predicting its latent feature $\tilde{x}_t^a$. (2) Modality transfer stage ($t \in [t_2, t_1)$): Here, modality transformation is performed by gradually converting the noised latent feature $\tilde{\mathbf{x}}_{t_2}$ into $\tilde{\mathbf{x}}_{t_1}$ of the target missing modality via the conditional score function $s_\theta^{(m)}(\tilde{\mathbf{x}}_t^m, y, t)$. This process integrates cross-modal knowledge, enabling effective semantic transfer between modalities. (3) Detail refinement stage ($t \in [t_1, 0)$): Finally, fine-grained details of the generated missing modality are refined by predicting $\mathbf{x}_0^m$ from the noisy latent feature $\tilde{\mathbf{x}}_{t_1}^m$, enhancing the quality and realism of the reconstructed data.

3.4. Multimodal Fusion and Prediction

For any missing pattern, the set of recovered data is denoted as $\hat{\mathcal{X}}^{\text{miss}}$, while the available data is represented as $\mathcal{X}^{\text{ava}}$. These are combined to form the complete multimodal input for downstream fusion and prediction tasks.

We employ multimodal Transformers [46] to fuse the features from $\hat{\mathcal{X}}^{\text{miss}} \cup \mathcal{X}^{\text{ava}}$. The resulting fused representation is subsequently passed through multi-layer perceptrons

	Model	$\{\mathbf{L}\}$	$\{\mathbf{A}\}$	$\{\mathbf{V}\}$	$\{\mathbf{L}, \mathbf{A}\}$	$\{\mathbf{L}, \mathbf{V}\}$	$\{\mathbf{A}, \mathbf{V}\}$
CMU-MOSI	DCCA [3]	73.6 / 73.8 / 30.2	50.5 / 46.1 / 16.3	47.7 / 41.5 / 16.6	74.7 / 74.8 / 29.7	74.9 / 75.0 / 30.3	50.8 / 46.4 / 16.6
	DCCAE [50]	76.4 / 76.5 / 28.3	48.8 / 42.1 / 16.9	52.6 / 51.1 / 17.1	77.0 / 77.0 / 30.2	76.7 / 76.8 / 30.0	54.0 / 52.5 / 17.4
	MCTN [40]	79.1 / 79.2 / 41.0	56.1 / 54.5 / 16.5	55.0 / 54.4 / 16.3	81.0 / 81.0 / 43.2	81.1 / 81.2 / 42.1	57.5 / 57.4 / 16.8
	TransM [53]	80.1 / 80.0 / 41.2	55.2 / 55.0 / 15.2	55.8 / 55.8 / 16.2	82.2 / 82.3 / 43.9	82.1 / 82.1 / 42.0	58.1 / 58.0 / 17.2
	ICDN [63]	83.1 / 83.2 / 42.0	55.5 / 55.4 / 15.0	56.7 / 56.7 / 16.1	83.1 / 83.1 / 43.3	82.9 / 83.0 / 42.1	59.3 / 59.3 / 17.2
	MMIN [65]	83.8 / 83.8 / 41.6	55.3 / 51.5 / 15.5	57.0 / 54.0 / 15.5	84.0 / 84.0 / 42.3	83.8 / 83.9 / 42.0	60.4 / 58.5 / 19.5
	GCNet [31]	83.7 / 83.6 / 42.3	56.1 / 54.5 / 16.6	56.1 / 55.7 / 16.9	84.5 / 84.4 / 43.4	84.3 / 84.2 / 43.4	62.0 / 61.9 / 17.2
	DiCMoR [51]	84.5 / 84.4 / 44.3	60.5 / 60.8 / 20.9	62.2 / 60.2 / 20.9	85.5 / 85.5 / 44.6	85.5 / 85.4 / 45.2	64.0 / 63.5 / 21.9
	IMDer [52]	84.8 / 84.7 / 44.8	62.0 / 62.2 / 22.0	61.3 / 60.8 / 22.2	85.4 / 85.3 / 45.0	85.5 / 85.4 / 45.3	63.6 / 63.4 / 23.8
	MD^{2}N	**87.4 / 87.3 / 45.9**	**68.8 / 68.8 / 27.5**	**67.1 / 67.0 / 27.2**	**87.4 / 87.4 / 45.6**	**87.5 / 87.4 / 46.0**	**70.4 / 70.4 / 28.3**
CMU-MOSEI	DCCA [3]	78.5 / 78.7 / 46.7	62.0 / 50.2 / 41.1	61.9 / 55.7 / 41.3	79.5 / 79.2 / 46.7	80.3 / 79.7 / 46.6	63.4 / 56.9 / 41.5
	DCCAE [50]	79.7 / 79.5 / 47.0	61.4 / 53.8 / 40.9	61.1 / 57.2 / 40.1	80.0 / 80.0 / 47.4	80.4 / 80.4 / 47.1	62.7 / 59.2 / 41.6
	MCTN [40]	82.6 / 82.8 / 50.2	62.7 / 54.5 / 41.5	62.6 / 57.1 / 41.6	83.5 / 83.3 / 50.7	83.2 / 83.2 / 50.4	63.7 / 62.7 / 42.1
	TransM [53]	82.3 / 82.3 / 49.6	60.2 / 57.1 / 40.0	60.8 / 59.6 / 41.2	83.6 / 83.3 / 51.1	83.4 / 83.3 / 50.0	64.0 / 63.3 / 41.9
	ICDN [63]	82.7 / 83.2 / 50.0	58.5 / 58.4 / 40.1	61.7 / 61.2 / 40.9	84.0 / 83.9 / 51.4	83.7 / 83.7 / 50.9	63.3 / 60.7 / 40.6
	MMIN [65]	82.3 / 82.4 / 51.4	58.9 / 59.5 / 40.4	59.3 / 60.0 / 40.7	83.7 / 83.7 / 52.0	83.8 / 83.4 / 51.2	63.5 / 61.9 / 41.8
	GCNet [31]	83.0 / 83.2 / 51.2	60.2 / 60.3 / 41.1	61.9 / 61.6 / 41.7	84.3 / 84.4 / 51.3	84.3 / 84.4 / 51.1	64.1 / 57.2 / 42.0
	DiCMoR [51]	84.2 / 84.3 / 52.4	62.9 / 60.4 / 41.4	63.6 / 63.6 / 42.0	85.0 / 84.9 / 52.7	84.9 / 84.9 / 53.0	65.2 / 64.4 / 42.4
	IMDer [52]	84.5 / 84.5 / 52.5	63.8 / 60.6 / 41.7	63.9 / 63.6 / 42.6	85.1 / 85.1 / 53.1	85.0 / 85.0 / 53.1	64.9 / 63.5 / 42.8
	MD^{2}N	**88.4 / 88.3 / 55.2**	**69.7 / 69.7 / 43.5**	**70.1 / 70.0 / 44.7**	**88.5 / 88.4 / 56.5**	**88.3 / 88.4 / 56.1**	**71.2 / 70.6 / 44.6**

Table 1. Comparison with the state-of-the-arts on CMU-MOSI [59] and CMU-MOSEI [60] under fixed missing scenario. $\{K\}$ means modality $\{*\}$ is available ($* \in \{\mathbf{L}, \mathbf{A}, \mathbf{V}\}$). The values in each cell denote ACC$_2$/F$_1$/ACC$_7$. **Bold** is the best.

(MLPs) to produce the final predictions. The overall optimization objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \gamma \mathcal{L}_{\text{rec}}, \tag{17}$$

where $\mathcal{L}_{\text{task}}$ denotes the task-specific loss, implemented as cross-entropy loss in our experiments, and γ is a balancing coefficient that controls the relative importance of the reconstruction loss $\mathcal{L}_{\text{rec}}$. The entire optimization is conducted in an end-to-end manner. Detailed training configurations, including modality missing rate settings, are provided in the experimental section.

4. Experiments

In this section, we conduct extensive experiments on the missing modality multimodal learning and a suite of ablation studies.

4.1. Datasets and Implementation Details

Datasets. To verify the effectiveness of MD^{2}M, we conduct experiments on two multimodal sentiment analysis datasets: CMU-MOSI [59] and CMU-MOSEI [60]. Each sample is labeled with a sentiment score ranging from -3 (strongly negative) to +3 (strongly positive). We evaluate the performance using the following metrics: 7-class accuracy (ACC$_7$), binary accuracy (ACC$_2$), and F$_1$ score.

Baseline. We compare MD^{2}M with several state-of-the-art incomplete multimodal learning methods, including recovery-based methods (MCTN [40],TransM [53], ICDN [63],MMIN [65], GCNet [31], DiCMoR [51], IMDer [52]) and non-recovery methods (DCCA [3], DCCAE [50]).

Implementation Details. On the two datasets, we extract the text features via pre-trained BERT model text[10] and obtain a 768-dimensional hidden state as the word features. For visual modality, each video frame was encoded via Facet [22] to represent the presence of the total 35 facial action units [9]. The acoustic modality was processed by COVAREP [9] to obtain the 74-dimensional features. Each experiment was run five times, and the average results on the test set are reported, using PyTorch on an NVIDIA A800 GPU. We explore the effectiveness of various methods in two distinct scenarios: one where a specific modality is consistently missing, and another where the missing modality is randomly selected. For the fixed missing modality scenario, we systematically discard either one modality (*i.e.* $\{L, A\}$, $\{L, V\}$, $\{A, V\}$) or two modalities (*i.e.* $\{L\}$, $\{A\}$, $\{A\}$) throughout the evaluation. For the random missing senario, we define the missing rate $r_{\text{miss}} = (1 - \frac{\sum_{i=1}^{N} m_i}{N \times M}) \times 100\%$ to quantify the overall extent of missing modalities across the samples, where N denotes the total number of modalities,m_i represents the number of available modalities for i^{th} sample and M corresponds to the total number of modalities. In the case of three modalities, we select eight values of r_{miss} from the range $[0\%, 10\%, 20\%, \ldots, 70\%]$, where 70% represents the max approximate missing rate while ensuring that at least

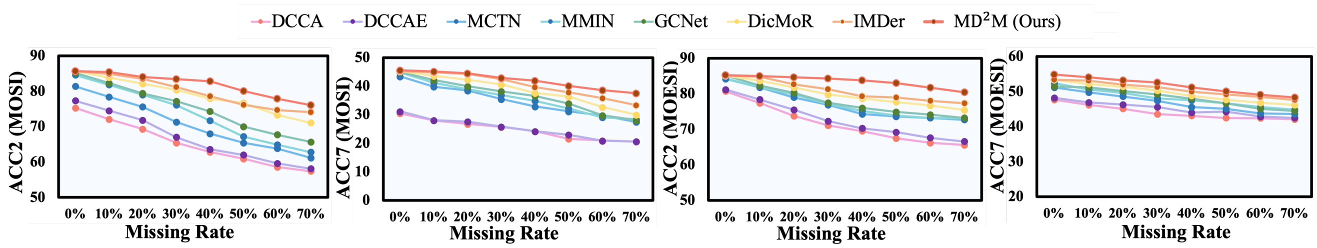

Figure 3. Comparison results on CMU-MOSI [59] and CMU-MOSEI [60] under randomly missing scenario.

one modality is available at any given time.

4.2. Comparison with the state-of-the-arts

Table 1 and Fig. 3 present the quantitative results of our model across both datasets. From the results, we could make the following key observations:

(1) **Effectiveness of recovery-based models**: Recovery-based methods, such as our MD^{2}N model, consistently outperform non-recovery approaches [3, 50]. This improvement is possibly attributed to their ability to utilize missing modality data more effectively.

(2) **State-of-the-art performance**: Among all recovery-based methods [31, 40, 53, 63, 65], our model achieves the best overall performance. We attribute this to its dual cross-diffusion structure, which dynamically controls noise and integrates cross-modal semantic information, preserving global consistency while enhancing local detail fidelity.

(3) **Robustness to missing patterns and rates**: As shown in Tab. 1 and Fig. 3, our model exhibits smaller performance degradation under increasing missing rates compared to other recovery methods, indicating its strong robustness across different missing patterns and levels of modality incompleteness.

4.3. Ablation Study

Impact of Different Configurations. To validate the effectiveness of different configurations, we ablate our MD^{2}N into three variants: (i) Base-model: Direct recovery from the available modality to the missing modality, similar to IMDer [52]. (ii) Adding VP-SDE Model: Using the VP-SDE technique for a score model, optimising time-step noise control to predict the missing modality $\mathbf{x}_0^m$. (iii) Adding with Multi-stage Model: Incorporating VP-SDE within a score model $s_\theta^{(m)}$ and adopts a multi-stage diffusion process without duplex modeling.

As shown in Tab. 2, we make the following observations: (i) vs. (ii): Optimising time-step noise control in the VP-SDE model significantly improves the accuracy of missing modality predictions compared to the base model. (ii) vs. (iii): Adding a multi-stage model outperforms the one only using the VP-SDE model, indicating that the multi-stage diffusion process better captures cross-modal interactions. (iii) vs. (iv): Our full model achieves the best performance across all metrics on both datasets, benefiting from the additional guidance of the duplex diffusion structure. This result demonstrates the effectiveness of our duplex training strategy in enhancing multimodal recovery and learning robust modality-invariant representations.

Dataset	Method Type	Results
CMU-MOSI	(i) Base-model	76.5 / 73.4 / 35.2
	(ii) + VP-SDE	76.7 / 76.6 / 35.3
	(iii) + Multi-stage	77.1 / 77.1 / 36.0
	(iv) **Ours**	**80.0 / 80.0 / 40.2**
CMU-MOSEI	(i) Base-model	79.0 / 77.3 / 49.3
	(ii) + VP-SDE	79.4 / 78.0 / 49.5
	(iii) + Multi-stage	79.9 / 78.4 / 49.8
	(iv) **Ours**	**82.6 / 82.6 / 50.3**

Table 2. Ablation study on various configurations. The average results for missing rates ranging from 30% to 70% are shown in the cells, representing ACC$_2$/F$_1$/ACC$_7$. **Bold** indicates the best performance.

Methods	CMU-MOSI	CMU-MOSEI
MCTN	71.3 / 71.2 / 35.5	76.9 / 76.2 / 47.4
MCTN w/$s_{\theta_i}^{(m,i)}$	**75.7 / 75.6 / 38.4**	**79.0 / 79.1 / 48.5**

Table 3. Ablation study of multi-stage duplex diffusion network $s_{\theta_i}^{(m,i)}$ on MCTN[40] under 30% missing rate. The value in each cell denotes ACC$_2$/F$_1$/ACC$_7$. **Bold** is best.

Effects of Multi-stage Duplex Diffused Network. To further evaluate the generalizability and effectiveness of our proposed multi-stage duplex diffusion network (MD^{2}N), we integrate it into the MCTN [40]. As shown in Tab. 3, incorporating MD^{2}N into MCTN (denoted as MCTN w/$s_{\theta_i}^{(m,i)}$) consistently outperforms the original MCTN [40] on both datasets, achieving approximately a 2-point performance improvement. These results demonstrate that our proposed module can be seamlessly integrated into existing models, providing consistent gains and highlighting its broad applicability and effectiveness.

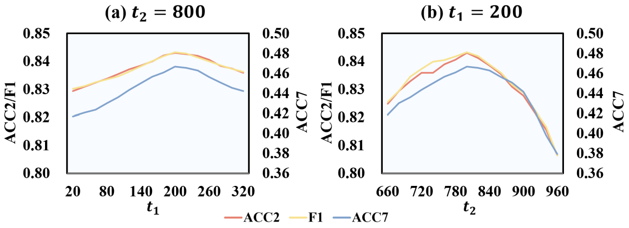

Figure 4. Ablation study on the effect of varying t_1 and t_2 under 30% missing rate on MOSEI. (a) shows the impact of changing t_1 with $t_2 = 800$, and (b) shows the effect of changing t_2 with $t_1 = 200$. Metrics: ACC$_2$, F$_1$, ACC$_7$.

Effects of the Multi-Stage Process Configuration. We investigate the impact of varying t_1 and t_2 on model performance. When discretizing the time into 1000 steps, the configuration $t_1 = 200$ and $t_2 = 800$ achieves the best overall results across all evaluation metrics, as illustrated in Fig. 4. Specifically, in experiments where t_2 is fixed at 800 and t_1 is varied (Fig. 4(a)), we observe that as t_1 decreases from 200 to 20, performance slightly declines. This degradation is attributed to the detailed refinement stage becoming too brief to sufficiently modify modality details. Conversely, when t_1 increases beyond an optimal point, the conditional fusion stage shortens, which also leads to a drop in performance. Furthermore, as shown in Fig. 4(b), when t_1 is fixed at 200 and t_2 is varied, a longer conditional fusion stage consistently yields better results. These findings highlight the importance of balancing the durations of both stages to maximize overall model effectiveness.

Effects of the Multimodal Feature Extractor and Decoder. To evaluate the contributions of the multimodal feature extractor $\mathcal{E}_K$ and the decoder $\mathcal{D}_K$, we conduct comparative experiments using four model variants. All configurations employ the full multi-stage duplex diffusion network, differing only in whether they include $\mathcal{E}_K$ and $\mathcal{D}_K$. The results in Tab. 4 lead to the following observations.

Using both the feature extractor and decoder consistently yields the strongest performance on both CMU-MOSI and CMU-MOSEI, providing an improvement of approximately 2–5 points across all evaluation metrics compared to configurations lacking either component. This demonstrates that $\mathcal{E}_K$ effectively extracts modality-specific information from text, image, and audio inputs, while $\mathcal{D}_K$ enables accurate reconstruction of the original data.

Datasets	$\mathcal{E}_K$	$\mathcal{D}_K$	Results
CMU-MOSI	✓	✓	**83.4 / 83.4 / 42.9**
	✗	✓	81.2 / 81.2 / 41.0
	✓	✗	80.4 / 80.2 / 40.4
	✗	✗	78.5 / 78.5 / 37.3
CMU-MOSEI	✓	✓	**84.3 / 84.4 / 52.6**
	✗	✓	82.4 / 82.2 / 50.2
	✓	✗	81.3 / 81.1 / 49.4
	✗	✗	79.4 / 79.6 / 47.3

Table 4. Ablation study on the effects of the multimodal feature extractor and decoder under 30% missing rate on MOSI ans MOSEI. The value in each cell denotes ACC$_2$/F$_1$/ACC$_7$. **Bold** is best.

Excluding the feature extractor leads to a moderate performance drop of about 2 points, indicating its crucial role in mapping multimodal signals into a unified latent space and supporting robust cross-modal representation learning. Without $\mathcal{E}_K$, the model struggles to capture complementary cues across modalities. Similarly, removing the decoder results in a comparable performance decrease of around 3 points, highlighting its importance for reconstructing the original data from intermediate representations. The decoder ensures reliable recovery by optimizing the reconstruction that combines scoring and decoding losses.

Finally, removing both modules causes the most substantial performance degradation, with a drop of roughly 5–7 points across datasets. This confirms that both components are critical for robust multimodal representation learning and for recovering missing modalities, especially under incomplete input conditions.

5. Conclusion

In this paper, we tackled the challenge of modality generation bias in multimodal learning for missing modality recovery. Existing diffusion-based approaches often struggle to balance generation quality across modalities due to inherent modality gaps. To address this issue, we proposed the Multi-Stage Duplex Diffusion Network (**MD^{2}N**), which introduces a modality transfer module to enable smooth and unbiased cross-modal generation. By leveraging a duplex diffusion framework consisting of three progressive stages—global structure generation, modality transfer, and local cross-modal refinement—our method facilitates mutual influence between available and missing modalities, resulting in a more balanced and effective recovery process. Extensive experiments demonstrate that MD^{2}N substantially outperforms state-of-the-art methods, confirming its effectiveness in reducing modality generation bias and enhancing multimodal learning under missing modalities.

In future work, we plan to extend our framework to handle scenarios involving more than three modalities simultaneously and explore adaptive stage configurations to automatically adjust the diffusion process based on input modality availability. Additionally, we aim to evaluate MD^{2}N in real-world downstream tasks such as multimodal sentiment analysis under partial observation and cross-modal retrieval with missing modality conditions to further validate its generalizability and practical impact.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (NSFC) (granted No. 62306064), the Central-Guided Local Science and Technology Development (granted No. 2023ZYD0165), and Sichuan Science and Technology Program (granted No. 2024ZDZX0011). We appreciate all the authors for their fruitful discussions. In addition, thanks are extended to anonymous reviewers for their insightful comments.

References

[1] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01):73–79, 2021. 1

[2] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 3

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 1, 6, 7

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1

[5] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 3

[6] Filipe Betzel, Karen Khatamifard, Harini Suresh, David J Lilja, John Sartori, and Ulya Karpuzcu. Approximate communication: Techniques for reducing communication bottlenecks in large-scale parallel systems. *ACM Computing Surveys (CSUR)*, 51(1):1–32, 2018. 1

[7] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multimodality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018. 1, 2

[8] Cheuk-Yiu Chan, Wan-Chi Siu, Yuk-Hee Chan, and H Anthony Chan. Anlightendiff: Anchoring diffusion probabilistic model on low light image enhancement. *IEEE Transactions on Image Processing*, 2024. 3

[9] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014. 4, 6

[10] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 6

[11] Craig K Enders. *Applied missing data analysis*. Guilford Publications, 2022. 2

[12] Bin Fu, Fanghua Yu, Anran Liu, Zixuan Wang, Jie Wen, Junjun He, and Yu Qiao. Generate like experts: Multi-stage font generation by incorporating font transfer process into diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6892–6901, 2024. 2

[13] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 6–15, 2021. 2

[14] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8472–8480, 2024. 2

[15] Tao He, Yuan-Fang Li, Lianli Gao, Dongxiang Zhang, and Jingkuan Song. One network for multi-domains: domain adaptive hashing with intersectant generative adversarial networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2477–2483, 2019. 2

[16] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuan-Fang Li. Semantic compositional learning for low-shot scene graph generation. *arXiv preprint arXiv:2108.08600*, 2021. 2

[17] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Semisupervised network embedding with differentiable deep quantization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4791–4802, 2021. 2

[18] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. State-aware compositional learning toward unbiased training for scene graph generation. *IEEE Transactions on Image Processing*, 32:43–56, 2022. 2

[19] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Transferable and differentiable discrete network embedding for multi-domains with hierarchical knowledge distillation. *Information Sciences*, 629:520–532, 2023. 2

[20] Trong Nghia Hoang, Shenda Hong, Cao Xiao, Bryan Low, and Jimeng Sun. Aid: Active distillation machine to leverage pre-trained black-box models in private data settings. In *Proceedings of the Web Conference 2021*, pages 3569–3581, 2021. 1

[21] Brandon Huang, Chancharik Mitra, Leonid Karlinsky, Assaf Arbelle, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. *Advances in Neural Information Processing Systems*, 37:22124–22153, 2025. 2

[22] iMotions. Facial expression analysis. Website, 2017. https://imotions.com/products/imotions-lab/modules/fea-facial-expression-analysis/. 4, 6

[23] Jong-Hwan Jang, Junggu Choi, Hyun Woong Roh, Sang Joon Son, Chang Hyung Hong, Eun Young Kim, Tae Young Kim, Dukyong Yoon, et al. Deep learning approach for imputation of missing values in actigraphy data: Algorithm development study. *JMIR mHealth and uHealth*, 8(7):e16113, 2020. 2

[24] Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. *arXiv preprint arXiv:2502.11564*, 2025. 3

[25] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 1

[26] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. 1

[27] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. Amm-diff: Adaptive multi-modality diffusion network for missing modality imputation. *arXiv preprint arXiv:2501.12840*, 2025. 3

[28] Linchao Li, Bowen Du, Yonggang Wang, Lingqiao Qin, and Huachun Tan. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems*, 194:105592, 2020. 2

[29] Wenbo Li, Xin Yu, Kun Zhou, Yibing Song, Zhe Lin, and Jiaya Jia. Image inpainting via iteratively decoupled probabilistic modeling. *arXiv preprint arXiv:2212.02963*, 2022. 3

[30] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023. 1

[31] 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. 6, 7

[32] Gong-Xu Liu, Ling-Feng Shi, and Dong-Jin Xin. Data integrity monitoring method of digital sensors for internet-of-things applications. *IEEE Internet of Things Journal*, 7(5): 4575–4584, 2020. 1

[33] Yubo Liu, Han Li, Qiaoming Deng, and Kai Hu. Diffusion probabilistic model assisted 3d form finding and design latent space exploration: A case study for taihu stone spacial transformation. In *The International Conference on Computational Design and Robotic Fabrication*, pages 11–23. Springer, 2023. 3

[34] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2302–2310, 2021. 2

[35] Xuerong Mao. The truncated euler–maruyama method for stochastic differential equations. *Journal of Computational and Applied Mathematics*, 290:370–384, 2015. 3

[36] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[37] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, pages 689–696, 2011. 2

[38] Jianmo Ni, Larry Muhlstein, and Julian McAuley. Modeling heart rate and activity data for personalized fitness recommendation. In *The World Wide Web Conference*, pages 1343–1353, 2019. 2

[39] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multimodality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6839–6853, 2021. 2

[40] 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*, pages 6892–6899, 2019. 1, 4, 6, 7

[41] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 2

[42] Kaveh Samiee and Péter Kovács. Ecg decomposition using cascaded spline projection residual auto encoders. In *2023 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2023. 2

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2, 3

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[45] Yan Tai, Weichen Fan, Zhao Zhang, and Ziwei Liu. Link-context learning for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27176–27185, 2024. 2

[46] 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 conference. Association for computational linguistics. Meeting*, page 6558. NIH Public Access, 2019. 5

[47] Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia*, 2024. 1

[48] Jingjing Wang, Dan Zhang, and Feng Luo. Unified directly denoising for both variance preserving and variance exploding diffusion models. *arXiv preprint arXiv:2405.21059*, 2024. 3

[49] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020. 2

[50] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015. 1, 6, 7

[51] 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. 4, 6

[52] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023. 1, 2, 4, 6, 7

[53] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of the web conference 2020*, pages 2514–2520, 2020. 4, 6, 7

[54] Zeyu Wang, Jingyu Lin, Yifei Qian, Yi Huang, Shicen Tian, Bosong Chai, Juncan Deng, Lan Du, Cunjian Chen, Yufei Guo, et al. Diffx: Guide your layout to cross-modal generative modeling. *arXiv preprint arXiv:2407.15488*, 2024. 3

[55] Renjie Wu, Hu Wang, and Hsiang-Ting Chen. A comprehensive survey on deep multimodal learning with missing modality. *arXiv preprint arXiv:2409.07825*, 2024. 2

[56] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013. 1

[57] Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pages 464–481. Springer, 2024. 1

[58] Ling Yang, Zhilong Zhang, Zhaochen Yu, Jingwei Liu, Minkai Xu, Stefano Ermon, and CUI Bin. Cross-modal contextualized diffusion models for text-guided visual generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[59] 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. 6, 7

[60] 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. 6, 7

[61] Maciej Żelaszczyk and Jacek Mańdziuk. Text-to-image cross-modal generation: A systematic review. *arXiv preprint arXiv:2401.11631*, 2024. 2

[62] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[63] Qiongan Zhang, Lei Shi, Peiyu Liu, Zhenfang Zhu, and Liancheng Xu. Retracted article: Icdn: integrating consistency and difference networks by transformer for multimodal sentiment analysis. *Applied Intelligence*, 53(12): 16332–16345, 2023. 6, 7

[64] Zhengxin Zhang, Xiaosheng Si, Changhua Hu, and Yaguo Lei. Degradation data analysis and remaining useful life estimation: A review on wiener-process-based methods. *European Journal of Operational Research*, 271(3):775–796, 2018. 3

[65] 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. 4, 6, 7

[66] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*, 2024. 2

[67] Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019. 2

[68] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. Multimodal sentiment analysis with image-text interaction network. *IEEE transactions on multimedia*, 25:3375–3385, 2022. 2