

VAE-Integrated Multiscale Generative Diffusion Modeling for Incomplete Multimodal Emotion Recognition

ZiXuan Wang, ZhuYi Yao, JiaYue Shen, Pan Wang, *Member, IEEE*,
XueJiao Chen, Xu Zhang, XiaoKang Zhou, *Member, IEEE*,

Abstract—Multimodal emotion recognition (MER) is widely applied in scenarios such as intelligent human-computer interaction, affective computing, and smart customer service. However, in real-world settings, multimodal data often suffer from incompleteness or missing modalities. Existing methods, which typically perform conditional generation directly in the high-dimensional raw feature space, are prone to semantic distortion due to noise accumulation and exhibit sharp performance degradation under high missing rates. To address these challenges, we propose HVDER (Hierarchical VAE-Diffusion for Emotion Reconstruction), an incomplete multimodal emotion recognition approach based on a hierarchical latent-space generative diffusion model. The framework operates in three key stages: First, independent variational autoencoders (VAEs) project language, visual, and audio modalities into a unified low-dimensional latent space. Then, within this latent space, a two-stage hierarchical diffusion module (coarse-to-fine) reconstructs missing modalities by recovering global semantics followed by supplementing local details. Finally, the completed latent variables—along with available modalities—are fused via self-attention and cross-modal attention mechanisms. The entire model is optimized end-to-end using a multi-objective loss function that integrates reconstruction loss, diffusion matching loss, cross-modal consistency constraints, and classification cross-entropy to output the final emotion prediction. Experiments on the CMU-MOSEI dataset demonstrate that: Under fixed single-modality missing scenarios, HVDER achieves an average classification accuracy (ACC2) of 76.8% and an F1-score of 75.4%, outperforming the best baseline by approximately 1.5% and 1.7%, respectively. In random missing scenarios with a 50% missing rate, HVDER maintains an ACC2 of 81.2%, surpassing comparable methods by about 3.4%. The proposed method effectively enhances the stability and reliability of emotion recognition in high-missing-rate and weak-modality environments, providing solid technical support for real-world incomplete multimodal emotion recognition applications.

Index Terms—Emotion AI, Multimodal emotion recognition, Fusion reconstruction.

I. INTRODUCTION

IN modern social systems such as education, healthcare, and digital communication, Emotion AI enables intelligent agents to sense and respond to human emotions. Com-

pared with single-modality approaches, leveraging the complementary information among multiple modalities enables a more comprehensive and robust understanding of human emotions. [1] This technique has wide applications in fields such as mental health monitoring, personalized education, intelligent customer service and virtual assistants, as well as service robots. [2]

However, in practical applications, multimodal data often suffer from incompleteness or even missing information. For example, in mental health monitoring, the audio modality may be missing due to equipment limitations; in personalized education, only visual modality data might be collected; and in intelligent customer service, certain modalities may be intentionally omitted for privacy protection. In addition, issues such as sensor malfunction or data acquisition costs can also result in the absence of certain modalities. [3] Traditional models, when trained on complete modality data, become highly sensitive to missing modalities during inference, leading to a significant performance drop. Thus, maintaining accuracy and robustness in emotion recognition despite missing modalities is a key challenge to address.

In recent years, two major types of methods have emerged for the problem of "incomplete modality" emotion recognition: (1) Imputation-Free Methods [4], which do not attempt to reconstruct missing modalities but focus on utilizing existing modality information to train robust models. These methods extract modality-independent features through latent representation learning, graph neural networks, contrastive learning, knowledge distillation, etc., to mitigate the impact of missing modalities on performance. While imputation-free methods can fully exploit available modality information, they fail to address the information gap and imbalance caused by missing modalities. Even if the model learns some modality-invariant features, the unique information of the missing modality is still absent, limiting the accuracy of emotion recognition. (2) Imputation-Based Methods, [5] which aim to generate or infer the missing modality data and then perform emotion recognition. Simple imputation techniques replace missing information with zero or mean values, which are easy to implement but yield low accuracy. More advanced methods train specialized models to generate or infer the features of the missing modality. Some studies use graph neural networks to infer missing modality information from existing modalities, or employ autoencoder frameworks to reconstruct missing modality features. The latest work even introduces diffusion generative models to this field, mapping noise distributions to the missing modality's data distribution through conditional diffusion models to recover missing modality features. These

ZiXuan Wang, ZhuYi Yao, JiaYue Shen, Pan Wang are with the School of Modern Posts, Nanjing University of Post&Telecommunications, Nanjing, China, (e-mail: 2020070135@njupt.edu.cn; 1024072003@njupt.edu.cn; wang-pan@njupt.edu.cn).

XueJiao Chen School of Communications, Nanjing Vocational College of Information Technology, Nanjing, China, (e-mail: chenxj@njcit.cn).

Xu Zhang is with the School of Computing Sciences, University of East Anglia, (e-mail: x.zhang27@uea.ac.uk)

Xiaokang Zhou is with the Faculty of Business Data Science, Kansai University, Osaka 565-0823, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: zhou@kansai-u.ac.jp)

Manuscript received April 19, 2021; revised August 16, 2021.

methods can partially compensate for missing modality information and improve the model’s ability to assess emotions. However, generating missing modalities from observed modalities also presents challenges. The data distributions of different modalities may differ significantly, and if the imputation model does not account for these distribution discrepancies, the generated modality data will be inconsistent with the real data, affecting the subsequent emotion recognition. Additionally, fluctuations in the quality of the imputation model itself can severely impact final recognition accuracy. [6]

Overall, current methods for incomplete multimodal emotion recognition still fall short in terms of balancing cross-modal information and maintaining emotional semantic consistency. Traditional approaches struggle to fully utilize all available modality information to fill in the gaps of missing modalities, limiting the model’s understanding and inference accuracy in emotional semantics. [7]

To address the challenges of semantic loss and reduced recognition accuracy caused by incomplete modalities, this paper proposes a multiscale incomplete multimodal emotion recognition method based on the integration of Variational Autoencoders (VAE) and Diffusion Models, termed HVDER (Hierarchical-VAE Diffusion for Emotion Reconstruction). The proposed method organically combines structured latent space modeling with a multiscale diffusion-based reconstruction mechanism, which not only improves the reconstruction quality of missing modalities, but also effectively preserves the cross-modal emotional consistency. These properties endow HVDER with strong practical applicability in real-world scenarios—such as audio-modality absence in mental health monitoring, vision-only data in personalized education, or partial modality masking due to privacy constraints in intelligent customer service—where it can still maintain stable emotion recognition performance and provide reliable technical support for scenario-based applications. The main contributions of this work are summarized as follows:

This paper proposes a hybrid latent space modeling method that integrates VAE and diffusion processes. Specifically, the available modalities are first encoded into a low-dimensional latent space via a Variational Autoencoder (VAE), which preserves the core semantic features of each modality and forms a well-structured conditional latent distribution. This latent space serves as the generative domain for the diffusion model, enabling a more efficient and stable sampling process and significantly reducing the computational cost associated with generation in high-dimensional modality spaces. Compared to conventional diffusion-based approaches directly applied in data space, the proposed method leverages the KL divergence regularization of the VAE to suppress noise propagation in the latent space, thereby improving the realism and diversity of the generated modalities.

This paper designs a hierarchical coarse-to-fine diffusion mechanism to enhance generation controllability and emotional expressiveness. Within the constructed hybrid latent space, a two-stage diffusion sampling pipeline is introduced: a coarse-grained stage for capturing global semantic structure, followed by a fine-grained stage for refining local and high-frequency details. This hierarchical generation strategy enables

progressive refinement from semantic abstraction to precise detail modeling, improving the emotional detail representation—especially for high-dimensional modalities such as visual and acoustic signals. [8]

This paper constructs a multi-objective composite loss function tailored for incomplete multimodal emotion recognition. The training process incorporates a comprehensive loss framework consisting of reconstruction loss, cross-modal semantic consistency constraints, emotional intensity supervision, and inter-modal contrastive loss. This design jointly optimizes both the semantic fidelity and the emotional discriminability of the generated modalities. In particular, the cross-modal contrastive loss enhances the structural alignment between generated and observed modalities in the latent space, while the emotional intensity supervision further guides the model to accurately distinguish varying levels of emotional strength, ensuring that the reconstructed modalities are not only structurally coherent but also emotionally expressive.

The organization of this paper is as follows: Chapter 1 introduces the research background and problem; Chapter 2 reviews related research progress; Chapter 3 details the proposed HVDER model; Chapter 4 describes the experimental design and results analysis; Chapter 5 discusses the advantages and limitations of the method; and Chapter 6 concludes the paper and outlines future research directions.

II. RELATED WORK

A. Fully Multimodal Emotion Recognition Methods

Fully multimodal methods assume that all modalities are available during both training and testing phases. The focus of research is on designing efficient multimodal feature fusion mechanisms to explore the complementarity and synergy between modalities. Representative methods can be divided into the following two categories:

(1) **Tensor-based Modeling and Low-Rank Decomposition Methods:** Tensor fusion mechanisms have gained widespread attention due to their powerful capability for modeling inter-modal interactions. Zadeh et al. [9] proposed the Tensor Fusion Network (TFN), which explicitly models high-order inter-modal interactions through the tensor outer product of three-modal features. TFN has achieved excellent performance on benchmark datasets such as CMU-MOSI. However, the parameter scale of TFN grows exponentially with the number of modalities and feature dimensions, resulting in enormous computational costs and high sensitivity to missing modalities.

To alleviate the computational burden of tensor fusion, Liu et al. [10] proposed the Low-rank Multimodal Fusion (LMF) method, which decomposes the fusion tensor into several low-rank factors, significantly reducing the model parameters while maintaining performance comparable to TFN. LMF offers better scalability and efficiency, but the linear decomposition still limits its ability to model complex modality relationships, resulting in certain performance bottlenecks.

(2) **Self-attention and Modality Feature Modeling Methods:** With the success of Transformer structures in natural language processing, Tsai et al. [11] introduced it into multimodal learning by proposing the Multimodal Transformer (MuT).

This model employs a cross-modal attention mechanism to align and fuse modality sequences, avoiding the need for explicit alignment in traditional methods. MulT effectively captures long-sequence dependencies and implicit relationships between modalities, but its large model structure is highly dependent on the quality and quantity of training data.

Additionally, Hazarika et al. [12] proposed the MISA model, which uses adversarial training to separately model modality-invariant and modality-specific representations, enhancing the discriminability and generalization ability of representations. MISA helps capture shared semantics across modalities while preserving modality uniqueness, but its implementation is complex and still relies on the availability of all modalities. In summary, fully multimodal methods achieve excellent performance under ideal input conditions through complex fusion strategies. However, they are generally highly dependent on modality completeness, and their performance deteriorates significantly when faced with missing modalities, limiting their applicability in practical scenarios.

B. Imputation-Free Incomplete Multimodal Emotion Recognition Methods

Imputation-free methods do not attempt to reconstruct missing modalities. Instead, they improve tolerance and stability in the face of missing data through robust model structures, alignment strategies, or training mechanisms. These methods are mainly divided into the following categories:

Deep Correlation Modeling Methods: Deep Canonical Correlation Analysis (DCCA) [13] and its improved version, Deep Canonical Correlation Autoencoder (DCCAE) [14], perform nonlinear projections on multiple modalities, ensuring maximal correlation in a shared representation space. The former emphasizes alignment relationships between modalities, while the latter introduces reconstruction loss to enhance the preservation of original information. Both methods inherently rely on correlation modeling, showing strong robustness to missing modalities. However, they still face the risk of expression degradation under high proportions of missing data.

Meta-learning and Missing Pattern Simulation Methods: Ma et al. [15] proposed the SMIL model, which adopts a Bayesian meta-learning framework to introduce perturbations in the latent feature space, allowing single-modality embeddings to gradually approach those of the multimodal fusion during training. This enables handling various missing modality combinations. Similarly, Chi et al. [16] proposed the M3S method, which adds missing modality sampling strategies in the MAML framework to simulate various missing combinations and enhance the model's adaptability to mixed missing situations. These methods show good generalization ability in real-world scenarios with frequently changing missing modalities, but the inherent complexity of meta-learning results in high training costs and difficult hyperparameter tuning.

Robust Representation Modeling with Alignment and Contrastive Learning: Lin et al. [17] proposed a classification-based alignment framework that maintains complete modalities during training and builds structurally consistent fusion branches for each missing-modality combination. It aligns incomplete

and complete representations through geometric contrastive, distributional, and affective semantic constraints, achieving robust performance without generative models. While Miss-Modal offers low parameter overhead and easy integration with general fusion architectures, it heavily depends on alignment target quality and label supervision, and struggles to recover modality-specific semantics under distribution shifts. Qian et al. [18] introduced a contrastive masked-attention model that replaces missing-modality features with learnable mask vectors and applies cross-modal contrastive learning to reduce inter-modal gaps. By restricting attention to observed modalities, it prevents noise propagation and provides a unified architecture adaptable to various missing patterns. However, the semantic content of missing modalities is only implicitly represented, and GAN-based augmentation may introduce training instability and parameter sensitivity.

C. Imputation-Based Incomplete Multimodal Emotion Recognition Methods

Imputation-based methods explicitly generate missing modality data or features, enabling the model to perform emotion recognition on nearly complete multimodal inputs. These methods typically leverage generative models such as autoencoders (AE), recurrent neural networks, graph neural networks, generative adversarial networks (GAN), variational autoencoders (VAE), and more recently, diffusion models. [19] Representative methods include:

Residual Reconstruction and Cycle Consistency-based Imagination Models: Tran et al. [20] proposed the Cascade Residual Autoencoder (CRA) model, which uses a cascading structure to iteratively reconstruct missing modality feature residuals, simulating the transformation relationships between modalities. Building on this, Zhao et al. [21] proposed the MMIN (Missing Modality Imagination Network), which introduces cross-modal consistency constraints (such as cycle reconstruction) and adversarial loss to further enhance reconstruction quality and the discriminability of generated features. Both models achieve missing modality imputation through "modality imagination," but challenges remain in stabilizing the generative path and controlling error propagation.

Modality Translation and Sequence Modeling-based Methods: Pham et al. [22] proposed the MCTN (Modality Cyclic Translation Network), which uses a sequence-to-sequence structure to achieve modality-to-modality translation and obtains more robust latent representations through cycle consistency loss. Additionally, Liu [23] introduced a reconstruction branch in the Transformer framework, utilizing an information feedback mechanism to implicitly recover missing modality features, achieving a good balance of performance.

Graph Modeling and Diffusion Generative New Imputation Mechanisms: Recent research has attempted to enhance imputation quality using structured generative models. Lian et al. [24] proposed the GCNet model, which models modality information as a graph structure and completes node features through graph convolutional networks. Wang et al. [25] introduced the IMDer model, which employs a diffusion probabilistic model to gradually generate missing modality features

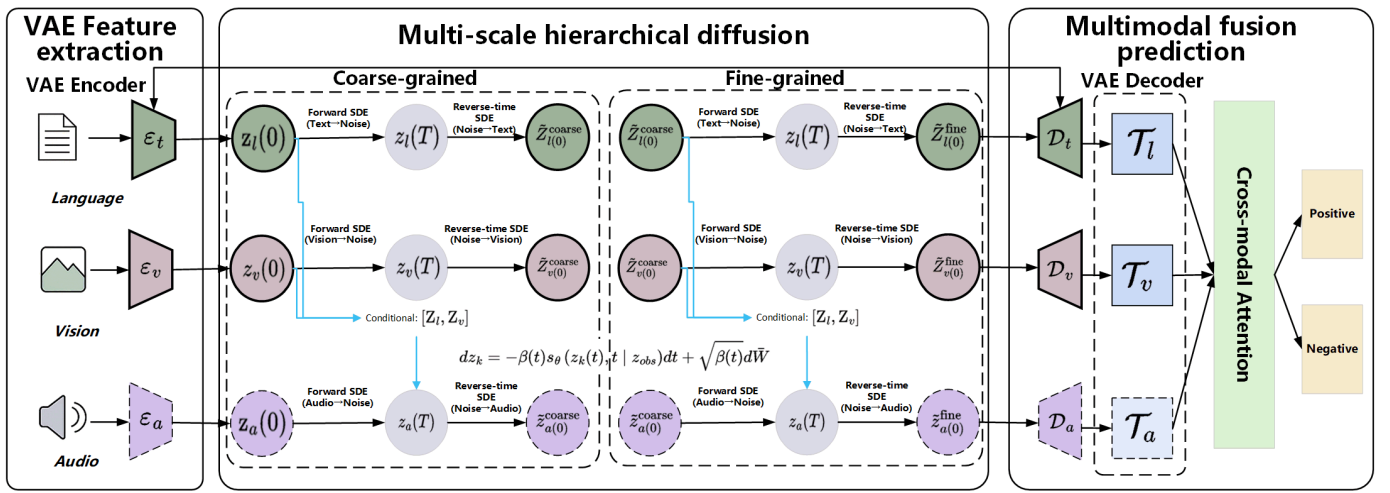


Fig. 1. Architecture of HVDER

through a reverse diffusion process and incorporates cross-modal attention to guide the generation process, demonstrating superior performance in high-missing-rate scenarios. Diffusion models, due to their stable generative mechanisms and ability to model high-dimensional complex distributions, are becoming a hot direction in the next generation of multimodal imputation research. [26]

Dynamic Detection and Multi-Model Fusion-based Ensemble Strategies: Zeng et al. proposed the EMMR (Ensemble Missing Modality Reconstruction) method, which first determines whether the missing modality is critical for emotion judgment. If it is, multiple generative models (such as Transformer, AE) are employed to reconstruct it, and the reconstruction results are fused for classification decision-making. This method offers stronger interpretability and reliability in imputation but involves complex overall system design, with high training and maintenance costs.

In summary, imputation-based methods more directly fill in the information gap through generative mechanisms, providing superior performance in scenarios with high modality missing or critical missing modalities. However, these methods are highly sensitive to generative quality, have large training overheads, and present implementation complexities, with the need for further improvement in generalization to unseen missing patterns.

III. MULTISCALE INCOMPLETE MULTIMODAL EMOTION RECOGNITION METHOD BASED ON VAE-DIFFUSION HYBRID MODEL

A. Overall Architecture

As shown in Fig. 1, the overall architecture of the proposed model is designed for incomplete multimodal emotion recognition involving text, visual, and audio modalities, consisting of three main components: a multimodal VAE encoder module, a multi-scale diffusion generation module, and a multimodal fusion prediction module. First, an independent Variational Autoencoder (VAE) is used for latent space modeling for each modality. For the data of the k -th modality x_k ($k \in l, v, a$),

where l, v , and a denote language, visual, and audio modalities, respectively), its encoder E_k maps the original features to the latent vector space, resulting in the corresponding latent variable z_k . The output dimensions of encoders for each modality are unified to the same scale to facilitate cross-modal fusion. Next, in the shared latent space, a multi-scale diffusion generation module is introduced to perform progressive generation of latent variables from coarse-grained to fine-grained.

Specifically, a diffusion generative model tailored to the latent variables of each modality is trained. When a modality is missing, the model takes the latent variables of the remaining available modalities as conditions, and gradually generates and completes the latent variable of the missing modality in its latent space through the diffusion model; for non-missing modalities, their latent variables can be directly obtained by the encoder. The multi-scale diffusion module adopts a coarse-to-fine two-stage diffusion process: first, generating coarse-grained latent variables reflecting global semantics, and then restoring local details on this basis to generate fine-grained latent variable representations.

The conditional guidance mechanism runs through all stages of the diffusion process. When generating the latent variable of the missing modality, the information of latent variables from available modalities is integrated into the diffusion model to ensure that the completed latent representation is semantically consistent with the known modalities. Finally, after processing by the diffusion module, we obtain the latent representation $\tilde{z}_k$ of each modality. These latent representations are respectively mapped back to the reconstructed modal features $\tilde{x}_k$ through the VAE decoder D_k of the corresponding modality, thus obtaining the complete set of multimodal features after completion $\tilde{x}_l, \tilde{x}_v, \tilde{x}_a$. Eventually, the multimodal fusion prediction module applies self-attention and cross-modal attention mechanisms to the completed features to extract the fused emotional semantic representation F_{fusion} , which is then input into the classifier to predict the final emotion category $\hat{y}$.

The entire architecture is designed to fully utilize the latent representations of each modality and the inter-modal

complementary information: VAE provides a robust latent representation distribution and regularization constraints; the multi-scale diffusion model simulates the evolution of data distribution in the latent space to achieve modal completion; and the attention fusion strategy is used to improve the accuracy and robustness of emotion recognition.

B. Hybrid Latent Space Modeling with VAE-Diffusion

For each type of modal data, the model uses an independent Variational Autoencoder (VAE) to compress it into the latent variable space. The dimensionality reduction process serves as the first layer of robustness assurance in high missing-rate scenarios: high-dimensional raw modal data are prone to information sparsification due to modality absence, where noise can rapidly accumulate and amplify during direct modeling. In contrast, the VAE maps each modality into a low-dimensional latent space of uniform dimensionality through its encoder, which filters redundant noise from the raw data while preserving the core emotional semantics of the modalities, thereby significantly reducing the complexity and noise interference in subsequent generative tasks. Given the input feature x_k of the k -th modality, the encoder E_k extracts its high-level representation and outputs a mean vector μ_k and a variance vector σ_k , thereby constructing the posterior distribution:

$$q_\phi(z_k | x_k) = \mathcal{N}(z_k | \mu_k, \sigma_k^2) \quad (1)$$

where z_k is the latent vector corresponding to this modality, and ϕ are the encoder parameters. The z_k obtained through sampling serves as the initial input to the diffusion model. During training, the VAE also introduces a KL divergence regularization term to encourage $q_\phi(z_k | x_k)$ to be close to the prior distribution $p(z_k)$, ensuring the stability and consistency of the latent variable distribution across modalities. This constraint ensures that even when a modality is missing, the latent variables of the remaining modalities still lie within a region consistent with the global distribution, preventing latent-space distribution shift caused by single-modality absence and providing a stable structural foundation for subsequent missing-modality generation.

After obtaining the modal latent variables, this paper combines the diffusion probabilistic model to model the data distribution in the latent space. The diffusion model gradually injects noise into the latent variables through a forward diffusion process and learns the reverse diffusion process to recover data from the noise. The diffusion model addresses the issue of semantic disconnection between generated content and available modalities under high missing rates by coupling a forward diffusion process (progressively injecting noise into latent variables) with a reverse diffusion process (restoring data from noise). In the continuous-time framework, the forward diffusion can be represented by a Stochastic Differential Equation (SDE):

$$dz_k(t) = \sigma dW \quad (2)$$

where $t \in [0, 1]$, σ controls the noise intensity, and W is standard Brownian motion. This process gradually perturbs

the latent variables of the original distribution into diffusion distributions at each time t . When $t \rightarrow 1$, the latent variables approximate an isotropic Gaussian noise distribution. Through a unified noise perturbation mechanism, latent variables under different missing patterns are mapped into a shared noise distribution space, preventing the model from overfitting to specific missing configurations and enhancing its generalization ability in diverse high-missing scenarios. To regenerate data in the diffused latent space, we need to solve the corresponding reverse diffusion SDE. By introducing a learnable score function s_θ , the general form of the reverse SDE can be written as:

$$dz_k = -\beta(t)s_\theta(z_k(t), t | z_{\text{obs}})dt + \sqrt{\beta(t)}\bar{W} \quad (3)$$

where $\beta(t)$ is a noise time scheduling function related to σ , $\bar{W}$ is Brownian motion in reverse time, and z_{obs} represents all available modal latent conditional information. The above reverse SDE describes the evolution process of gradually denoising and reconstructing the latent variable of the missing modality from the noise distribution: the first drift term $-\beta s_\theta$ uses the gradient estimation provided by the score network to guide z_k toward the high-probability direction, and the second diffusion term re-injects an appropriate amount of randomness to satisfy distribution constraints. By training the score network to approximate $\nabla_{z_k} \log p_t(z_k | z_{\text{obs}})$, the model can effectively sample and generate the latent variable of the missing modality given the conditional information of the available modalities.

The integration of the diffusion generation process with the VAE latent space establishes a dual robustness mechanism: the VAE constructs a consistent and low-noise latent foundation through dimensionality reduction and KL regularization, mitigating the interference of high-dimensional data and distribution shifts; meanwhile, the diffusion model leverages conditional score functions and progressive denoising to fully exploit the available modality information and guide the generation of missing modalities. During training, the model jointly optimizes the VAE reconstruction and KL objectives together with the score-matching objective of the diffusion process, thereby ensuring both the consistency of latent representations and the accuracy of generation under high missing rates.

C. Multi-Scale Diffusion Generation Module

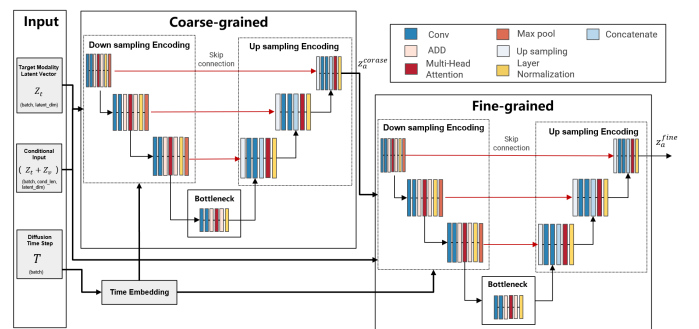


Fig. 2. Multiscale Hierarchical Diffusion Generation Architecture

Inspired by the concept of multi-scale generation in fields such as computer vision [27] and speech synthesis [28], the generation process is decomposed into two stages. As shown in Fig. 2, the first stage focuses on restoring a stable and coherent global structure (coarse-grained semantics), providing reliable guidance and constraints for subsequent generation. The second stage builds upon this semantic framework to enhance the realism and discriminability of fine-grained details. This cooperative division of labor has been shown to improve generation quality, stabilize the training process, and strengthen the model’s ability to capture critical information. [29]

To enhance the semantic consistency between generated missing modalities and existing modalities, this paper proposes a multi-scale diffusion generation strategy, which realizes the step-by-step refinement of latent variables from coarse-grained to fine-grained. The module is composed of two cascaded diffusion sub-modules: a coarse-scale diffusion generator and a fine-scale diffusion generator, corresponding to the global semantic modeling and local detail supplementation of modal features, respectively. [30] “Coarse-grained information” refers to global semantic features that determine the overall emotional category or polarity, such as the general sentiment tendency of a text, the overall happy or sad category of a facial expression, or the emotional tone of a speech segment (e.g., excited or calm). In contrast, “fine-grained information” denotes local discriminative cues that help distinguish between similar emotions and enhance classification precision—for example, the intensity and weighting of specific emotional words in text, subtle movements around the mouth or eyebrows in facial expressions, or key pitch fluctuations and prosodic patterns in speech. This distinction closely aligns the generation process with the emotion recognition task: the coarse stage ensures that the generated content captures the correct emotional direction, establishing a reliable semantic backbone, while the fine stage refines this foundation by adding detailed cues that enhance classification confidence, making the reconstructed modality features more expressive and discriminative.

For the missing modality m , the coarse-scale diffusion generator first takes the latent variables of all available modalities $\{z_k \mid k \neq m\}$ as conditions to generate the coarse-grained latent representation z_m^c . The generation process follows the reverse SDE framework mentioned in Section 3.2, but the score network s_θ^c in this stage is designed as a cross-modal attention structure. It first fuses the conditional information of available modalities through a self-attention mechanism, and then guides the generation of z_m^c through cross-attention. The objective function of the coarse-scale diffusion process is:

$$\mathcal{L}_c = \mathbb{E}_{t_c, z_m^0, z_{\text{obs}}} \left[\left\| s_\theta^c(z_m(t_c), t_c \mid z_{\text{obs}}) - \nabla_{z_m(t_c)} \log p_{t_c}(z_m(t_c) \mid z_{\text{obs}}) \right\|_2^2 \right] \quad (4)$$

where $t_c \in [0, 1]$ is the time step of the coarse-scale diffusion, z_m^0 is the initial latent variable of the missing modality, and p_{t_c} is the conditional distribution of the coarse-grained latent variable at time t_c .

On the basis of obtaining z_m^c , the fine-scale diffusion generator further optimizes the latent representation to generate

the fine-grained latent variable z_m^f . Different from the coarse-scale stage, the fine-scale diffusion takes both the available modal latent variables and the coarsely generated z_m^c as joint conditions, and uses a more complex score network s_θ^f with residual connections to capture local detail features. The objective function is:

$$\mathcal{L}_f = \mathbb{E}_{t_f, z_m^c, z_{\text{obs}}} \left[\left\| s_\theta^f(z_m(t_f), t_f \mid z_m^c, z_{\text{obs}}) - \nabla_{z_m(t_f)} \log p_{t_f}(z_m(t_f) \mid z_m^c, z_{\text{obs}}) \right\|_2^2 \right] \quad (5)$$

Where $t_f \in [0, 1]$ is the time step of the fine-scale diffusion, and p_{t_f} is the conditional distribution of the fine-grained latent variable given the coarse-grained result and available modalities.

The multi-scale design enables the diffusion process to first establish the global semantic framework of the missing modality under the guidance of available modalities, and then gradually supplement local details on this basis, which effectively reduces the semantic deviation caused by direct generation. During the inference phase, for the missing modality, the model first generates z_m^c through the coarse-scale diffusion generator, then inputs z_m^c and z_{obs} into the fine-scale diffusion generator to obtain z_m^f , and finally takes z_m^f as the completed latent variable $\tilde{z}_m$ of the missing modality. For the non-missing modality, its latent variable z_k is directly used as $\tilde{z}_k$ after being processed by a simple linear mapping to align with the scale of the generated latent variable. Through this two-stage generation mechanism, the model can generate high-quality complementary latent representations even in the case of multiple modalities missing, laying a foundation for subsequent multimodal fusion.

D. Multimodal Fusion and Prediction

After the above steps, the model has obtained the completed features of each modality $\tilde{x}_k (k \in 1, v, a)$. The multimodal fusion and prediction module fuses these features to output the final result of emotion recognition. In this module, we introduce self-attention and cross-modal attention mechanisms to fully explore the intra-modal and inter-modal correlation information. First, self-attention mechanism is applied to the feature sequence of each modality to model the long-range dependencies within the sequence, thereby improving the representation ability of unimodal features. Through self-attention, elements within the same modal sequence calculate attention weights with each other, capturing relationships between different words in a sentence, temporal relationships between video frames, and correlations between sound features at different moments in audio signals, thus obtaining more discriminative modal representations.

After obtaining the enhanced unimodal representations, cross-modal attention mechanism is adopted to realize inter-active fusion between features of different modalities. Cross-modal attention enables the features of one modality to act as query vectors (Query) to focus on the keys and values (Key/Value) generated by the features of another modality (or multi-modal combined features), thereby learning the corresponding relationships between modalities. The calculation of cross-modal attention can be defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (6)$$

where Q , K , and V represent the query matrix, key matrix, and value matrix respectively, and d is the dimension of the key vector. For the fusion of any pair of modalities, the features of one modality are mapped through a learnable projection as Q , and the features of another modality (or multi-modal combined features) are mapped as K and V . The attention weight matrix $\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$ obtained by softmax calculation depicts the correlation strength between each element in the query modality and each element in the key modality, and accordingly weights and summarizes the value matrix, thus realizing the capture and fusion of relevant information from the value modality into the query modality representation. After multiple rounds or multi-head attention interactions, the representations of different modalities already contain each other's information, achieving consistency and complementarity in emotional semantics. Then, these fusion-rich features are further transformed and integrated through the fusion and reconstruction sub-module, which uses a series of feed-forward networks or convolutional layers to compress the multi-modal information into a fixed-length fusion representation F_{fusion} . This fusion representation aggregates the emotional clues of each modality of the current sample, and finally, the emotion classifier discriminates F_{fusion} to output the emotion prediction value $\hat{y}$. Through the above fusion and prediction process, the model can comprehensively consider the information of text, visual, and audio modalities. Even if some of these modalities were originally missing, they have been completed, thus enabling more robust emotion recognition.

E. Multi-Objective Loss Function Design

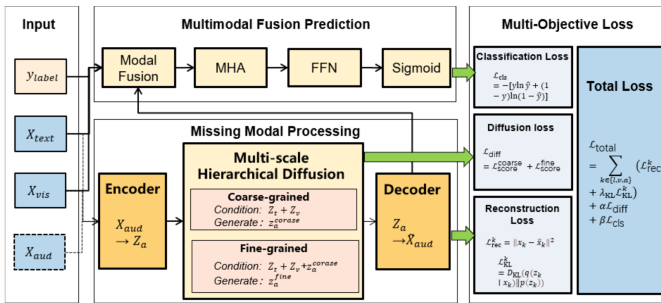


Fig. 3. Multi-Objective Loss Function

To achieve the joint optimization of each module in the model, this paper designs a loss function containing multiple objective terms, which is used to constrain the performance of both generative reconstruction and discriminative classification. As shown in Fig. 3, the loss function consists of the following parts: the reconstruction error and KL divergence regularization of the VAE for each modality, the score matching loss of the coarse and fine-stage diffusion models, and the cross-entropy loss of the final emotion classification. The

specific forms and symbol explanations of each part of the loss are as follows.

For the VAE encoder-decoder modules of text, visual, and audio modalities, the reconstruction loss $\mathcal{L}_{\text{rec}}^k$ and KL divergence regularization term $\mathcal{L}_{\text{KL}}^k$ are defined for each modality $k (k \in \{l, v, a\})$, representing language, visual, and audio respectively). Let the original input be x_k , the latent variable $z_k \sim q(z_k | x_k)$ output by the encoder, and the reconstructed result by the decoder be $\hat{x}_k$. The reconstruction loss uses Mean Squared Error (MSE) to measure the difference between the reconstructed sample and the original input:

$$\mathcal{L}_{\text{rec}}^k = \|x_k - \hat{x}_k\|^2 \quad (7)$$

It ensures that the latent representation z_k retains the ability to reconstruct the original modal information. On the other hand, the Kullback-Leibler divergence (KL divergence) regularization term is introduced:

$$\mathcal{L}_{\text{KL}}^k = D_{\text{KL}}(q(z_k | x_k) \| p(z_k)) \quad (8)$$

It is used to measure the difference between the posterior distribution $q(z_k | x_k)$ and the prior distribution $p(z_k)$. The prior $p(z_k)$ is usually set as a standard Gaussian distribution $\mathcal{N}(0, I)$ with independent dimensions. By minimizing $\mathcal{L}_{\text{KL}}^k$, the latent variable distributions of each modality are encouraged to be close to the prior, so as to ensure the stability and consistency of the latent variable distributions across modalities. [31]

After obtaining the modal latent variables, this paper combines the diffusion probability model to model the data distribution in the latent space. The diffusion model gradually injects noise into the latent variables through the forward diffusion process and restores the data from the noise by learning the reverse diffusion process. In the continuous-time framework, the forward diffusion can be represented by a stochastic differential equation (SDE).

For the multi-scale diffusion generation module of each modality, the score matching loss (score matching loss) is set for the coarse and fine stages respectively to train the corresponding score networks. In the coarse-grained stage, the score network is denoted as $s_{\theta}^{\text{coarse}}$, and the optimization goal is to make its output match the true score (noise gradient) after adding noise to the latent variables. Specifically, let $z_k^{(0)} = z_k^{\text{coarse}}$ represent the target latent vector generated in the coarse stage, $z_k^{(t)}$ represent the latent variable at time step t after adding noise disturbance to $z_k^{(0)}$, $\epsilon \sim \mathcal{N}(0, I)$ be the standard Gaussian noise, and $\sigma(t)$ be the noise standard deviation coefficient of the diffusion process at time t . The coarse-stage score matching loss is defined as:

$$\mathcal{L}_{\text{score}}^{\text{coarse}} = \mathbb{E}_{z_k^{\text{coarse}}, z_k^{\text{obs}}, \epsilon, t} \left[\frac{1}{2\sigma(t)^2} \left\| s_{\theta}^{\text{coarse}} \left(z_k^{(t)}, t, z_k^{\text{obs}} \right) - \frac{z_k^{(t)} - z_k^{\text{coarse}}}{\sigma(t)} \right\|_2^2 \right] \quad (9)$$

where z_k^{obs} represents the set of latent variables of the available modalities (non-missing modalities) in the current sample, which are injected into the score network as conditional information. The above loss function describes the gap between the predicted residual of the coarse-grained latent

variable after adding noise (equivalent to the true score function $\nabla_{z_k^{(t)}} \log p(z_k^{(t)} | z_k^{(0)})$) and the model output s_θ^{coarse} , thereby guiding s_θ^{coarse} to learn accurate score estimation.

The fine-grained stage uses an independent score network s_ϕ^{fine} , and its training method is similar to that of the coarse stage, but additionally uses the coarse-grained generation result as conditional input to focus on local detail reconstruction. Let z_k^{fine} be the target latent variable of the fine stage (the latent representation with details added on the basis of the coarse stage), $z_k^{(t')}$ be the latent variable at time t' in the fine-stage diffusion process, $\epsilon' \sim \mathcal{N}(0, I)$, and $\sigma(t')$ be the fine-stage noise standard deviation coefficient. Then the fine-grained stage score matching loss is:

$$\mathcal{L}_{\text{score}}^{\text{fine}} = \mathbb{E}_{z_k^{\text{fine}}, z_k^{\text{obs}}, z_k^{\text{coarse}}, \epsilon', t'} \left[\frac{1}{2\sigma(t')^2} \left\| s_\phi^{\text{fine}} \left(z_k^{(t')}, t', z_k^{\text{obs}}, z_k^{\text{coarse}} \right) - \frac{z_k^{(t')} - z_k^{\text{fine}}}{\sigma(t')} \right\|_2^2 \right] \quad (10)$$

In the fine-stage training, in addition to using the available modal latent variables z_k^{obs} , the coarse-grained generation result z_k^{coarse} is also used as conditional input to ensure that the fine-grained generation result maintains consistency with the coarse-grained semantics while enhancing local details. The sum of the coarse and fine-stage score matching losses constitutes the total generation loss of the diffusion model:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{score}}^{\text{coarse}} + \mathcal{L}_{\text{score}}^{\text{fine}} \quad (11)$$

In the multi-modal fusion prediction module, we use cross-entropy loss to constrain the model's final output of emotion category prediction. Let the fused emotion prediction result be $\hat{y}$ and the true emotion label be y (this task is a binary classification problem, where $y \in \{0, 1\}$ can represent negative/positive emotions). Then the classification loss is defined as:

$$\mathcal{L}_{\text{cls}} = -[y \ln \hat{y} + (1 - y) \ln(1 - \hat{y})] \quad (12)$$

where when $y=1$, $\hat{y}$ represents the probability that the model outputs a positive emotion; when $y=0$, $\hat{y}$ represents the probability of outputting a negative emotion. By minimizing $\mathcal{L}_{\text{cls}}$, the model's ability to discriminate emotion categories can be improved, and the fused multi-modal features can be aligned with the true labels in the emotion classification space.

Finally, the total loss function of the entire model is obtained by weighted summation of the loss terms of the above modules:

$$\mathcal{L}_{\text{total}} = \sum_{k \in \{l, v, a\}} (\mathcal{L}_{\text{rec}}^k + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}^k) + \alpha \mathcal{L}_{\text{diff}} + \beta \mathcal{L}_{\text{cls}} \quad (13)$$

where k traverses the three modalities of text (l), visual (v), and audio (a). The coefficient λ_{KL} controls the weight of the KL regularization term relative to the reconstruction error in the VAE, and α and β are the weight hyperparameters of the generative diffusion loss and classification loss respectively, which are used to balance the contribution of the generative module and the classification module to the overall goal. During model training, appropriate selection of α and β can avoid a single loss dominating the optimization, thereby taking

into account both the quality of reconstruction completion and the accuracy of emotion recognition.

In this work, the specific values of these hyperparameters are determined through grid search on the validation set. The optimization objective is to achieve the best balance between emotion recognition accuracy (ACC2) and the fidelity of the generated modalities (indirectly measured by reconstruction error). Specifically, λ_{KL} is searched within the range of (0.1, 0.5, 1.0), while α and β are searched within (0.01, 0.1, 1.0). The final optimal configuration is $\lambda_{\text{KL}} = 0.5$, $\alpha = 0.1$, and $\beta = 0.1$.

To sum up, the design of this multi-objective loss function enables the model to focus on minimizing reconstruction errors, constraining latent distributions, generating quality of missing modalities, and emotion classification performance during training, realizing the collaborative optimization of each module to improve the overall effect of incomplete multi-modal emotion recognition.

IV. EXPERIMENTAL SETUP

This experiment aims to verify the performance of the proposed multi-scale incomplete multimodal emotion recognition method (HVDER) based on the VAE-diffusion hybrid model in scenarios with missing modalities. By comparing with current mainstream imputation and non-imputation methods, the robustness and recognition accuracy of the proposed model under different missing strategies are evaluated.

A. Dataset

This study employs the widely used benchmark dataset CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) for experimental validation. The dataset, constructed by Zadeh et al., contains 22,856 English video segments collected from online video platforms. Each segment provides manually annotated text transcripts, visual frame sequences, and audio waveforms, forming three modalities: Text (T), Vision (V), and Audio (A).

The sentiment annotations are continuous-valued emotion scores ranging from -3 to $+3$, where negative values indicate negative emotions and positive values indicate positive emotions. To adapt the dataset for a binary emotion recognition task, we follow common practice in the field by discretizing the sentiment scores using 0 as the threshold: samples with scores greater than 0 are labeled as Positive, and those with scores less than 0 are labeled as Negative. Samples with a neutral score of exactly 0 are excluded to ensure task clarity. The dataset split follows the official partitioning, with 16,326 samples used for training, 1,871 samples for validation, and 4,659 samples for testing, ensuring comparability with prior research.

For modality-specific preprocessing, we adopt standard procedures consistent with existing studies to obtain unified feature representations:

Text modality (T): Each sample's transcript is represented as a tensor of shape (50, 300), where each sentence is padded or truncated to 50 tokens, and each token is represented by a 300-dimensional word embedding, forming a complete sentence-level temporal embedding matrix.

Visual modality (V): Each sample’s visual information is represented as a 35-dimensional global feature vector, typically obtained by analyzing facial features from video frame sequences and applying aggregation or dimensionality reduction to capture emotion-related facial expression cues.

Audio modality (A): Each sample’s audio information is represented as a 74-dimensional acoustic feature vector, typically extracted using COVAREP or similar acoustic analysis toolkits, containing low-level descriptors such as fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCCs), and spectral features.

All modality features are standardized before being fed into the model to ensure training stability.

B. Comparative Experimental Setup

To evaluate the model’s adaptability and robustness in different modal missing scenarios, two types of comparative experiments are designed: fixed modal missing and random modal missing. In terms of model comparison, the proposed method is compared with seven existing representative methods, including two non-imputation methods (DCCA [13] and DCCAE [14]) that directly construct robust representations without imputing missing modalities, and five imputation methods (MCTN, [22] MMIN, [21] GCNet, [24] RecFormer, [23] and IMDer [25]) that recover features of missing modalities through various mechanisms.

Comparison under Fixed Modal Missing: In this setup, specific combinations of missing modalities are predefined, including single-modal missing (T/V/A), dual-modal missing (T+V, T+A, V+A), and full-modal availability (T+V+A), totaling seven combinations, which are uniformly applied to test samples.

Comparison under Random Modal Missing: To simulate the uncertainty of modal missing in real scenarios, the missing combination of each sample in the random modal missing experiment is controlled by the Missing Rate (MR). The definition of MR is as follows:

$$MR = \frac{1}{N \times M} \sum_{i=1}^N (M - m_i) \quad (14)$$

where N is the total number of samples, $M=3$ represents the total number of modalities, and m_i is the number of actually available modalities in the i -th sample. MR increases from 0.0 to 0.7 with a step size of 0.1, ensuring that each sample retains at least one modality. The same MR setting is adopted in the training, validation, and testing phases, and the experiment is repeated 5 times to report the average results.

C. Experimental Metrics

Since this task is a binary emotion recognition task, the main evaluation metrics include binary classification accuracy (ACC2) and F1 score. These two metrics can comprehensively reflect the classification performance of the model, especially in cases of unbalanced sample distribution or modal missing, where the F1 score is more robust.

D. Experimental Environment

In the experimental implementation, to ensure fair comparison and result reproducibility, all experiments are conducted using the TensorFlow framework on a server equipped with an NVIDIA RTX 5070 Ti GPU. The experimental code is developed in Python 3.8, and data analysis is assisted by the Scikit-learn library. All models are trained using the Adam optimizer with an initial learning rate of 1×10^{-4} , a batch size of 64, and a maximum of 80 epochs, with early stopping applied when the validation performance ceases to improve. The latent dimension of the VAE encoder is set to 64, and the diffusion process adopts a coarse-to-fine two-stage structure, each stage consisting of 100 time steps. To balance generation quality and task performance, the generation and classification terms in the loss function are assigned weight coefficients of $\alpha = 0.1$ and $\beta = 0.1$, respectively, which are determined through validation set tuning.

V. EXPERIMENTAL RESULTS

This chapter will evaluate the performance of the proposed HVDER method under two incomplete multimodal scenarios (fixed modal missing and random modal missing) and conduct a comparative analysis with existing methods.

A. Comparison under Fixed Modal Missing

In the fixed modal missing scenario, we simulate situations where specific modalities are missing, that is, a predefined modality is fixed as missing in each experiment. Specifically, we consider 7 modal combination cases: only text modality (denoted as t), only visual modality (v), only audio modality (a), pairwise available modalities (t,v, t,a, v,a), and finally, the result with all three modalities intact (t,v,a) as a control. The model adopts the same modal availability settings in both training and testing phases. For each modal combination, we calculate the ACC2 and F1 of each method on the test set respectively to compare their emotion recognition performance under that specific missing pattern. Table. I presents the results of HVDER and the comparative methods under the above modal combinations. Each column represents the ACC2 and F1 values of different methods under fixed modal missing. The best performer in each row is highlighted in bold. (t = text, v = video, a = audio modality)

As can be seen from Table. I, the HVDER method achieves the best ACC2 and F1 scores under all combinations of fixed modal missing, showing an improvement of approximately 1–2% compared to the second-best method. When all modalities are available, the ACC2 of HVDER reaches 86.0%, which is about 1.0 percentage point higher than the best baseline method; in the case of only text modality available, the ACC2 of HVDER is 85.5%, which is approximately 1.1% higher than the second-best IMDer (85.5% vs 84.4%). Compared with DCCA and DCCAE that do not adopt modal completion, their performance is significantly lower (e.g., the average ACC2 is only about 72%), verifying that modal missing will lead to significant information loss. Each completion method, by generating or reconstructing missing modalities, makes up for the information deficiency to a certain extent, thus greatly

TABLE I
COMPARISON OF EMOTION RECOGNITION PERFORMANCE UNDER FIXED MODAL MISSING ON THE CMU-MOSEI DATASET (ACC2 / F1)

Modal Combination	DCCA	DCCAE	MCTN	MMIN	GCNet	RecFormer	IMDer	HVDER
t	78.6 / 78.5	79.6 / 79.4	82.7 / 82.9	82.5 / 82.2	83.0 / 83.0	82.3 / 82.1	84.4 / 84.5	85.5 / 85.4
v	61.8 / 55.5	61.0 / 57.1	62.5 / 57.2	59.4 / 60.1	61.8 / 61.5	62.3 / 60.9	62.9 / 61.3	63.5 / 62.0
a	61.1 / 50.3	61.5 / 53.7	62.4 / 54.6	59.0 / 59.6	60.0 / 60.2	62.1 / 60.0	62.0 / 60.7	62.8 / 61.2
t,v	80.4 / 79.8	80.5 / 80.4	83.3 / 83.1	83.9 / 83.5	84.2 / 84.3	84.0 / 83.6	84.7 / 84.3	85.5 / 85.2
t,a	79.6 / 79.3	80.1 / 80.0	83.4 / 83.2	83.6 / 83.4	84.4 / 84.3	83.9 / 83.4	84.4 / 84.2	85.3 / 85.5
v,a	62.5 / 57.0	62.8 / 57.3	63.6 / 57.2	63.4 / 57.1	64.2 / 57.3	64.1 / 57.0	64.0 / 57.5	64.9 / 58.2
t,v,a	80.8 / 81.0	81.1 / 81.3	84.1 / 84.2	84.4 / 84.3	85.1 / 85.0	84.6 / 84.4	85.2 / 85.1	86.0 / 85.8
Avg	72.4 / 68.9	72.6 / 70.3	74.8 / 72.6	73.9 / 73.5	74.8 / 73.7	74.6 / 73.0	76.0 / 74.6	76.8 / 75.4

improving the performance. Among them, the latest methods such as IMDer and RecFormer perform better than earlier methods (MCTN, MMIN, etc.), reflecting the effectiveness of more advanced modal completion mechanisms.

It is worth noting that when key modalities are missing, the performance degradation of different methods varies more significantly. Especially in scenarios where only visual or audio is available (i.e., text is missing), the overall recognition accuracy is significantly lower than when text is available; however, our HVDER still maintains relatively leading performance. When only the visual modality is available, the ACC2 of HVDER is about 1.0 percentage point higher than the second-best method, improving the robustness of the model on average. In contrast, non-completion methods can hardly obtain text semantic information in this situation, with an ACC2 of less than 62%, which is significantly lower than methods relying on completion. When the text modality is missing, the advantage of HVDER over other methods is more obvious. This is because text/language modality often carries the richest semantic information, which is crucial for emotion recognition. The hierarchical diffusion completion mechanism proposed by HVDER can more accurately recover text modal features consistent with the real distribution, thereby effectively making up for the performance loss caused by text missing. In combinations lacking text such as t,a and v,a, the F1 of HVDER is about 1.0–1.5% higher than that of the second place, which proves that the proposed method still has strong recognition ability when key modalities are missing.

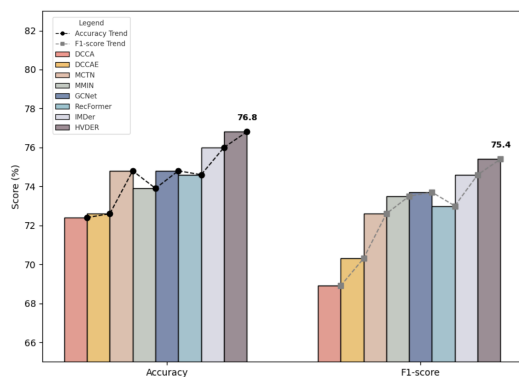


Fig. 4. Model Performance with Separated Accuracy F1 Trends for Incomplete Multimodal Emotion Recognition

Fig. 4 shows the Accuracy and F1-score performance of eight representative models in the incomplete multimodal emotion recognition task on the CMU-MOSEI dataset. The horizontal axis represents model names, and the vertical axis denotes performance percentages. Two trend lines respectively depict the overall changing trends of Accuracy and F1-score. The proposed HVDER model achieves the highest scores in both metrics, with an Accuracy of 76.8% and an F1-score of 75.4%. Compared with non-imputation methods such as DCCA and DCCAE, HVDER has an average improvement of more than 4 percentage points; even compared with advanced imputation methods like MMIN, RecFormer, and IMDer, it still has an advantage of 1 to 2 percentage points. This result verifies the effectiveness of the proposed multi-scale hierarchical diffusion generation mechanism in improving the quality of missing modality reconstruction and maintaining cross-modal semantic consistency.

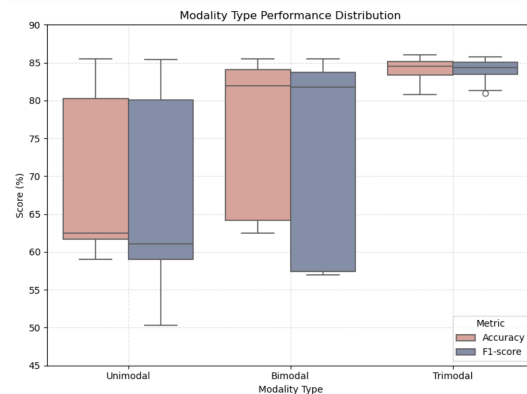


Fig. 5. Modality Type Performance Distribution for Incomplete Multimodal Emotion Recognition

Fig. 5 analyzes the performance distribution under different modal combinations. As the number of modalities increases from single-modal to three-modal, both Accuracy and F1-score show a continuous upward trend, and the performance variance gradually decreases. The median performance under the three-modal setting is close to 85%, significantly higher than that of single-modal and dual-modal configurations, indicating that modal complementarity plays a key role in emotional semantic modeling, and further demonstrating the stability and discriminative ability of HVDER in coping with modal missing and multimodal fusion scenarios.

TABLE II
COMPARISON OF EMOTION RECOGNITION PERFORMANCE UNDER RANDOM MODAL MISSING ON THE CMU-MOSEI DATASET (ACC2 / F1)

Missing rate MR	DCCA	DCCAE	MCTN	MMIN	GCNet	RecFormer	IMDer	HVDER
0	80.7 / 80.9	81.1 / 81.2	84.3 / 84.1	84.4 / 84.3	85.0 / 85.1	84.6 / 84.4	85.3 / 85.0	86.1 / 85.8
0.1	77.5 / 77.2	78.3 / 78.1	81.6 / 81.5	82.5 / 81.8	82.9 / 82.0	82.8 / 82.5	83.4 / 83.1	84.2 / 83.8
0.2	75.0 / 74.1	75.5 / 75.7	79.2 / 78.5	79.9 / 79.0	80.5 / 79.9	80.8 / 80.0	81.6 / 80.8	82.4 / 81.5
0.3	71.2 / 71.5	72.0 / 72.4	76.0 / 75.4	77.3 / 75.6	78.5 / 77.3	79.0 / 78.0	80.2 / 79.5	81.6 / 80.5
0.4	69.4 / 69.6	70.4 / 70.0	74.1 / 74.0	75.2 / 72.8	75.8 / 74.9	77.3 / 75.5	78.0 / 76.8	79.4 / 78.1
0.5	67.3 / 65.0	68.9 / 66.1	73.7 / 72.5	73.6 / 70.8	74.4 / 73.0	74.2 / 72.5	75.9 / 74.6	77.0 / 75.4
0.6	66.0 / 62.8	67.4 / 63.5	73.1 / 71.0	73.3 / 70.2	74.0 / 72.1	73.8 / 71.5	75.0 / 73.3	76.0 / 74.2
0.7	65.1 / 60.5	66.2 / 62.0	70.4 / 68.9	72.0 / 69.4	73.5 / 70.5	73.0 / 70.1	74.0 / 71.2	75.4 / 72.5
Avg	70.3 / 70.2	71.4 / 71.0	76.0 / 75.3	77.2 / 75.5	78.0 / 77.1	78.0 / 76.8	79.2 / 78.1	80.3 / 79.0

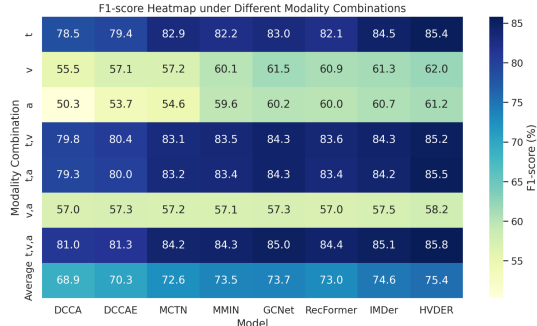


Fig. 6. F1-score Heatmap under Different Modality Combinations for Incomplete Multimodal Emotion Recognition

Fig. 6 presents the F1-score heatmap of different models under seven fixed modal combination configurations. It can be clearly seen that HVDER achieves the optimal or near-optimal results in all modal settings, further proving its robustness and universality in environments with modal missing.

Overall, in various scenarios of fixed modal missing, the HVDER method is comprehensively superior to the comparison methods. On the one hand, HVDER has achieved a significant improvement compared with non-imputation methods, indicating that generating and completing missing modalities can provide additional discriminative information; on the other hand, compared with existing multiple imputation strategies, HVDER achieves the highest ACC2/F1 in all combinations, which reflects the effectiveness of the imputation method integrating the diffusion model in this paper in maintaining the consistency of modal distribution and semantic integrity. Especially when text is available, HVDER gains a significant leading advantage in overall performance by virtue of more accurate completion and fusion of non-text modalities. Therefore, it can be concluded that in the scenario of fixed modal missing, the proposed method exhibits stronger robustness and higher emotion recognition accuracy compared with existing methods.

B. Comparison under Random Modal Missing

In the scenario of random modal missing, we use the missing rate (MR for short) to characterize the overall incompleteness of the data. The missing rate MR is defined as the average proportion of the number of missing modalities in each

sample in the dataset. In the three-modal task, MR=0.3 means that each sample has an average of 30% of modalities missing. We set the same missing rate in the training, validation, and testing phases, and simulate situations from MR=0.0 (no missing) to MR=0.7 (highly missing). It is necessary to ensure that each sample has at least one modality available (so the maximum missing rate is chosen to be around 0.7). For each missing rate, we randomly mask the corresponding proportion of modal data to evaluate the performance changes of each model under different missing intensities. Table. II reports the average results of ACC2 and F1 of each method on the CMU-MOSEI dataset when MR ranges from 0.0 to 0.7. Among them, MR=0.0 is equivalent to the normal situation of complete modalities, and can also be used as a reference for the ideal upper limit of each model.

It can be found from Table. II that in the case of random modal missing, the performance of all methods shows a downward trend as the missing rate increases, but the performance of the proposed HVDER method declines the slowest. When the missing rate increases from 0.0 (no missing) to 0.7 (highly missing), the ACC2 of HVDER only decreases by about 10.7 percentage points (from 86.1% to 75.4%), while the decline of other methods is mostly between 12 and 15 percentage points. The ACC2 of GCNet decreases from 85.0% to 73.5%, a drop of 11.5 percentage points; RecFormer decreases from 84.6% to 73.0%, a drop of 11.6 percentage points; IMDer decreases from 85.3% to 74.0%, a drop of 11.3 percentage points. However, HVDER has the smallest performance decline, reflecting its strong adaptability to high missing rates. This trend is roughly the same for the F1 score: the F1 of HVDER decreases from 85.8% to 72.5%, a drop of about 13.3 percentage points, which is significantly smaller than the F1 decline of most comparative methods. Overall, HVDER maintains higher accuracy and F1 under high missing rates, indicating that its modal completion mechanism can still provide effective support when there is severe information missing.

Further observation of the results under specific missing rates shows that HVDER always outperforms all comparative methods at any missing rate (including low and high missing rates). At MR=0.1 (mild missing), the ACC2 of HVDER reaches 84.2%, which is about 0.7 percentage points higher than that of IMDer, the second-highest; at MR=0.4 (moderate missing), the ACC2 of HVDER is 79.4%, which is about 1.4

percentage points higher than that of the second-best method; even in the case of extreme missing (MR=0.7), HVDER still maintains an advantage of at least 1.4% over other methods (75.4% vs 74.0% compared to IMDer). This indicates that regardless of the degree of missing, HVDER can always utilize existing modalities and complete missing information in a better way, thus stably maintaining high emotion recognition performance.

In contrast, non-imputation methods (DCCA, DCCAE) perform significantly worse than imputation methods in random missing scenarios, and the gap becomes more obvious as the missing rate increases (for example, when MR=0.7, the ACC2 of DCCA is only 65.1%, which is more than 10 percentage points lower than that of HVDER). This once again highlights the necessity of explicit modal imputation for alleviating the impact of incomplete modalities. Among imputation methods, early methods such as MCTN and MMIN show a rapid decline in accuracy at medium and high missing rates, indicating their limited ability to generate and complete. Newer methods such as IMDer and RecFormer have significantly improved overall performance compared to the former, benefiting from the introduction of diffusion models and attention mechanisms. Nevertheless, these methods still have an accuracy decline of about 11-13% under high missing rates, while the decline of HVDER is only about 10%, showing better robustness. It can be inferred that HVDER better aligns the distribution of completed modalities with real modalities during training, so that even if most modalities are missing during testing, the model can effectively use conditional information to reconstruct key features, avoiding a sharp degradation in performance.

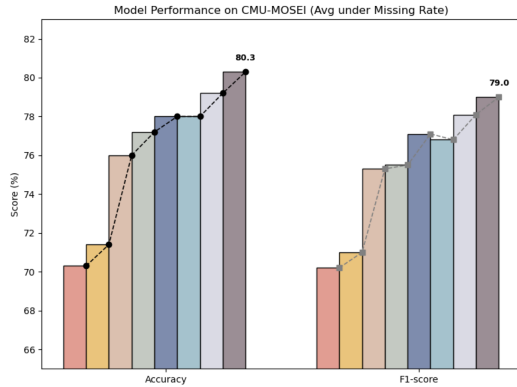


Fig. 7. Model Performance with Separated Accuracy F1 Trends under Increasing Modality Missing Rates

Fig. 7 presents the performance comparison of the eight models in terms of Accuracy and F1-score. The proposed HVDER model achieves the optimal performance in both metrics, with an Accuracy of 80.3 and an F1-score of 79.0, verifying the effectiveness of the proposed multi-scale hierarchical diffusion generation mechanism in improving the quality of missing modality completion and maintaining semantic consistency.

Fig. 8 illustrates the performance distribution under different modal combination configurations, covering single-modal, dual-modal, and tri-modal input settings. As the number of

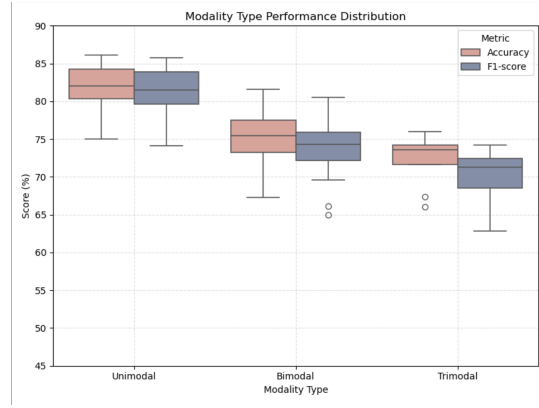


Fig. 8. Modality Type Performance Distribution under Increasing Modality Missing Rates

available modalities increases, the model's performance in both Accuracy and F1-score shows a steady upward trend, while the performance variance gradually decreases, indicating that modal complementarity has a significant positive effect on emotional modeling.

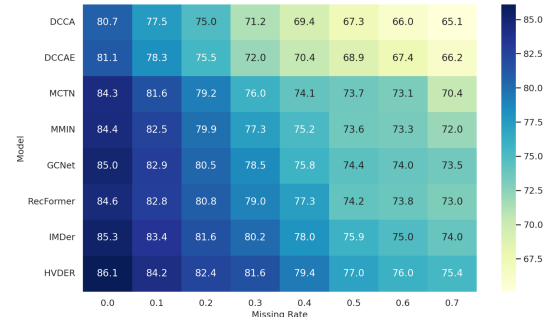


Fig. 9. F1-score Heatmap under Different Modality under Increasing Modality Missing Rates

Fig. 9 is a heatmap of the F1-score of each model under different modal missing rates. As the missing rate increases from 0 to 0.7, all models show a downward trend. HVDER remains leading at all stages, with its F1-score decreasing from 85.8 to 72.5, with an average of 79.0, which is higher than IMDer's 78.1 and MMIN's 75.5, and significantly better than DCCA's 70.2, demonstrating stronger robustness and discriminative ability in the context of modal missing.

Based on the above analysis, in the context of missing on-machine modes, HVDER also demonstrates the best comprehensive performance and the most robust anti-degradation capability. As the missing rate increases, although the recognition accuracy of all models declines, the performance of HVDER remains leading, and the decline is significantly smaller than that of other methods. This proves that the method proposed in this paper can effectively mine and utilize the remaining modal information under different degrees of information deficiency, generate high-quality missing modal features, and thereby significantly improve the robustness and accuracy of multimodal emotion recognition.

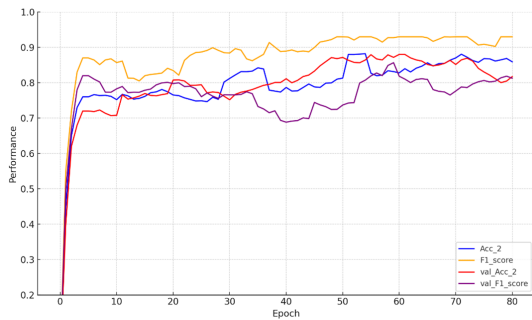


Fig. 10. Model performance variation diagram

The Fig. 10 shows the performance evolution trend of the proposed multi-scale incomplete multimodal emotion recognition model (HVDER) during training, including accuracy, F1 score, and their corresponding indicators on the validation set. It can be observed that all performance indicators rise rapidly in the early stage of training and quickly break through key performance thresholds within the first 5 epochs, indicating that the model has good initial convergence. This rapid convergence benefits from the collaborative mechanism of variational autoencoder (VAE) latent space modeling and multi-scale diffusion generation introduced in the HVDER architecture: VAE provides stable latent variable initialization and regularization, while the diffusion model effectively complements missing modalities through multi-stage fine-grained generation, thereby significantly improving the model's ability to capture emotional signals.

In the subsequent training process, although each indicator shows a certain degree of fluctuation and disturbance, the overall trend is stably upward and eventually tends to converge. This volatility is partly due to the random disturbances caused by the dynamic modeling of modal missing in the diffusion generation process. However, under the guidance of the coarse-fine two-stage diffusion structure adopted by HVDER, the model can still stably reconstruct latent variables and complete semantic alignment when multiple modalities are incomplete, ensuring the discriminability and consistency of the fused representation. Eventually, all indicators basically stabilized around 60 epochs, verifying the advantages of the model in training stability and convergence efficiency when dealing with incomplete multimodal emotion recognition tasks.

VI. CONCLUSION

Aiming at the problem of partial modal data missing in the common "multimodal emotion recognition" scenario in practical applications, this paper proposes a hybrid latent space generation method (HVDER) that integrates variational autoencoder (VAE) and diffusion model. Different from previous conditional diffusion modeling directly in high-dimensional data space, this method first uses VAE to map available modalities to a low-dimensional latent space, and ensures the smoothness and stability of the space distribution through KL divergence regularization. On this basis, a two-stage diffusion process is designed - the coarse-grained stage focuses on restoring the global semantic structure, and the fine-grained

stage supplements local details, so as to realize high-quality reconstruction of latent variables of missing modalities. Finally, the completed latent variables and the latent variables of complete modalities are input into the multimodal fusion prediction network together, and the model is optimized end-to-end with the help of a multi-objective loss function.

Experimental verification on the public CMU MOSEI dataset shows that HVDER has significant advantages in both fixed modal missing and random modal missing test scenarios. Specifically, in scenarios where a single modality (text, visual, or audio) is missing, the classification accuracy and F1 score of HVDER are averagely improved by 1-2 percentage points compared with existing mainstream methods; in random scenarios with a missing rate as high as 70%, the ACC2 indicator only decreases by about 10 percentage points, which is much better than the performance decay rate of comparative methods. In addition, this method also performs well in terms of training convergence speed and generation stability, fully reflecting the effectiveness of the hybrid latent space and multi-scale diffusion strategy in incomplete multimodal emotional modeling.

The theoretical significance of this work lies in its successful integration of structured latent space modeling with hierarchical diffusion generation, offering a novel technical pathway for addressing the challenge of modality incompleteness. The proposed approach demonstrates that performing conditional generation within a well-regularized latent distribution enables more stable and efficient cross-modal semantic transfer and completion. On a practical level, HVDER substantially enhances the robustness of emotion recognition systems under real-world imperfect data conditions, advancing the transition of affective computing from controlled laboratory settings toward practical applications. This provides direct technical support for developing more reliable systems in psychological state monitoring, classroom attention analysis, and natural human-computer interaction, and offers a valuable paradigm for addressing the challenges of data uncertainty in the field of affective AI.

Despite the aforementioned achievements, HVDER still has certain limitations. [32] First, the two-stage diffusion model requires a large number of iterative steps during inference, resulting in considerable computational and memory overhead, which makes it less suitable for scenarios demanding strict real-time performance or operating under limited resources. Second, the current training paradigm relies on a strong assumption—that the model can be pretrained on datasets containing all complete modality pairs—which restricts its applicability in extreme cases where fully paired multimodal data are unavailable. Finally, the conditional information fusion strategy adopted in this work remains relatively simple, relying primarily on concatenation followed by multilayer perceptron (MLP) mapping. Future work should explore more advanced cross-modal attention or graph-based fusion mechanisms to enhance semantic guidance during generation. [33]

Looking ahead, we plan to conduct further research along several directions: Model efficiency: incorporate accelerated sampling algorithms such as DDIM, as well as model pruning and knowledge distillation techniques, to reduce inference

cost; Data dependency and generalization: develop adaptive generation strategies that do not require pretraining on fully complete data and extend the framework to more extreme scenarios where one modality is consistently missing; Handling complex missing patterns: introduce meta-learning frameworks that enable the model to adapt online to dynamically changing missing-modality combinations; [34] and Enhanced fusion mechanisms: explore graph attention networks or bidirectional conditional diffusion to better exploit inter-modal complementarity and further push the model's performance limits in complex real-world environments.

ACKNOWLEDGMENTS

The paper is sponsored by Foundation of State Key Laboratory of Public Big Data under Grant No. PBD2022-10, Future Network Innovation Research and Application Projects under Grant No.2021FNA02006, Development of an Ultra-large-scale Ubiquitous Network Quality Monitoring System Based on Trusted Edge Intelligence under Grant SYG202311, and the Royal Society International Exchanges Award [IESR3243049].

REFERENCES

- [1] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 6, p. e1563, 2024.
- [2] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, "A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges," *IEEE Access*, vol. 12, pp. 103 976–104 019, 2024.
- [3] Z. Hu, W. Zheng, Y. Zong, M. Wei, X. Jiang, and M. Shi, "A novel decoupled prototype completion network for incomplete multimodal emotion recognition," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [4] Q. Zhang, Y. Wei, Z. Han, H. Fu, X. Peng, C. Deng, Q. Hu, C. Xu, J. Wen, D. Hu *et al.*, "Multimodal fusion on low-quality data: A comprehensive survey," *arXiv preprint arXiv:2404.18947*, 2024.
- [5] W. Luo, M. Xu, and H. Lai, "Multimodal reconstruct and align net for missing modality problem in sentiment analysis," in *International conference on multimedia modeling*. Springer, 2023, pp. 411–422.
- [6] Y. Zhan, R. Yang, J. You, M. Huang, W. Liu, and X. Liu, "A systematic literature review on incomplete multimodal learning: techniques and challenges," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2467083, 2025.
- [7] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," *arXiv preprint arXiv:2409.07825*, 2024.
- [8] X. Zhou, W. Liang, A. Kawai, K. Fueda, J. She, and K. I.-K. Wang, "Adaptive segmentation enhanced asynchronous federated learning for sustainable intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 6658–6666, 2024.
- [9] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.
- [10] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.
- [11] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, 2019, p. 6558.
- [12] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.
- [13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.
- [14] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.
- [15] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.
- [16] H. Chi, M. Yang, J. Zhu, G. Wang, and G. Wang, "Missing modality meets meta sampling (m3s): An efficient universal approach for multimodal sentiment analysis with missing modality," *arXiv preprint arXiv:2210.03428*, 2022.
- [17] R. Lin and H. Hu, "Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1686–1702, 2023.
- [18] S. Qian and C. Wang, "Com: Contrastive masked-attention model for incomplete multimodal learning," *Neural Networks*, vol. 162, pp. 443–455, 2023.
- [19] X. Zhou, J. Wu, W. Liang, K. I.-K. Wang, Z. Yan, L. T. Yang, and Q. Jin, "Reconstructed graph neural network with knowledge distillation for lightweight anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 11 817–11 828, 2024.
- [20] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1405–1414.
- [21] J. Zhao, R. Li, and Q. 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)*, 2021, pp. 2608–2618.
- [22] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.
- [23] C. Liu, J. Wen, Z. Wu, X. Luo, C. Huang, and Y. Xu, "Information recovery-driven deep incomplete multiview clustering network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 15 442–15 452, 2023.
- [24] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "Gcnet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 45, no. 7, pp. 8419–8432, 2023.
- [25] Y. Wang, Y. Li, and Z. Cui, "Incomplete multimodality-diffused emotion recognition," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17 117–17 128, 2023.
- [26] X. Zhou, X. Zheng, T. Shu, W. Liang, K. I.-K. Wang, L. Qi, S. Shimizu, and Q. Jin, "Information theoretic learning-enhanced dual-generative adversarial networks with causal representation for robust ood generalization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [28] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Dif-fwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.
- [29] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.
- [30] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.
- [31] X. Zhou, Q. Yang, Q. Liu, W. Liang, K. Wang, Z. Liu, J. Ma, and Q. Jin, "Spatial-temporal federated transfer learning with multi-sensor data fusion for cooperative positioning," *Information Fusion*, vol. 105, p. 102182, 2024.
- [32] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal emotion recognition using visual, vocal and physiological signals: a review," *Applied Sciences*, vol. 14, no. 17, p. 8071, 2024.
- [33] S. R. Sanku and B. Sandhya, "An effective data fusion methodology for multi-modal emotion recognition: A survey," *International Journal*, vol. 12, no. 7, 2024.
- [34] X. Zhou, Q. Yang, X. Zheng, W. Liang, K. I.-K. Wang, J. Ma, Y. Pan, and Q. Jin, "Personalized federated learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 4, pp. 817–831, 2024.