Modern Management Science & Engineering
ISSN ISSN 2052-2576
Vol. 8, No. 3, 2026
www.scholink.org/ojs/index.php/mmse

Original Paper

Generative AI Data Augmentation and Missing Value Completion Method for Multi-Source Heterogeneous Big Data

Linfeng Yang[1]

[1] Shanghai Hetang Environmental Protection Technology Co., Ltd, Shanghai, China

Received: May 09, 2026 Accepted: June 12, 2026 Online Published: July 3, 2026

doi:10.22158/mmse.v8n3p66 URL: http://dx.doi.org/10.22158/mmse.v8n3p66

Abstract

The proliferation of multi-source heterogeneous big data across healthcare, industrial Internet of Things (IIoT), and smart city domains has introduced two pervasive bottlenecks in data analytics: widespread missing values and insufficient high-quality training samples. Traditional missing value imputation methods fail to capture complex nonlinear correlations across heterogeneous data modalities, while existing data augmentation techniques are mostly designed for single-modal data and neglect intrinsic cross-source associations. To address these gaps, this paper proposes ***Hetero-GenAI****, a unified generative artificial intelligence framework for joint missing value completion and data augmentation in multi-source heterogeneous big data. First, a heterogeneous data embedding module with cross-modal attention maps numerical, categorical, temporal, and textual features into a shared latent space, explicitly modeling inter-source and intra-source feature dependencies. Second, a missing-aware conditional diffusion model performs adaptive imputation by integrating missing masks as soft constraints, eliminating the need for pre-assumptions about missing mechanisms (MCAR, MAR, MNAR). Third, a distribution-aligned augmentation strategy generates diverse, realistic samples via latent space interpolation and semantic perturbation while preserving cross-modal semantic coherence and distribution consistency. Extensive experiments on two real-world datasets—the MIMIC-IV clinical dataset and NASA C-MAPSS industrial sensor dataset—demonstrate that the proposed method reduces numerical imputation RMSE by 12.3%–21.7% and improves categorical imputation F1-score by 8.5%–15.2% under 30% missing rates compared to state-of-the-art baselines. Furthermore, the augmentation pipeline improves downstream classification and prediction task performance by 9.2%–14.6%, verifying its dual effectiveness in data completion and quality enhancement.*

Keywords

Generative AI, Multi-source heterogeneous data, Missing value imputation, Data augmentation, Diffusion model, Cross-modal learning

1. Introduction

The digital transformation of industries has led to an explosive growth of multi-source heterogeneous big data. In clinical scenarios, patient data spans numerical vital signs, categorical diagnostic codes, temporal monitoring waveforms, and textual clinical notes. In IIoT systems, equipment data includes numerical sensor readings, categorical fault labels, temporal degradation trajectories, and textual maintenance logs. These multi-modal, multi-source datasets contain rich complementary information, which can significantly improve the accuracy of downstream analytical tasks such as disease prediction and equipment remaining useful life (RUL) estimation [1][2].

However, real-world multi-source data universally suffers from two critical quality issues that severely degrade model performance. First, missing values are pervasive due to sensor failures, transmission interruptions, privacy desensitization, and irregular collection cycles. Statistics show that missing rates in electronic health record systems often exceed 30%, and up to 70% of clinical features are partially missing [3]. Worse, missing values follow complex and often mixed mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Traditional imputation methods either rely on strong distribution assumptions or only handle single data types, leading to large biases in heterogeneous scenarios. Second, high-quality labeled samples are scarce, especially in domains where annotation requires expert knowledge. Data augmentation is a standard solution to sample shortage, but existing augmentation techniques are mostly designed for single-modal data (e.g., image flipping, text back-translation, numerical noise injection). When applied directly to multi-source heterogeneous data, these methods often break cross-modal semantic associations and generate unrealistic samples that harm rather than improve downstream performance [4].

Current research addresses missing imputation and data augmentation as two separate pipeline stages. A typical workflow first imputes missing values using statistical or deep learning methods, then applies generative models to the completed data for augmentation. This two-stage approach suffers from error accumulation: imputation errors propagate to the augmentation phase and are amplified during sample generation. Furthermore, few existing methods explicitly model the correlation structure across heterogeneous data sources during either imputation or augmentation, leaving valuable cross-modal information underutilized.

To overcome these limitations, this paper proposes Hetero-GenAI, an end-to-end generative AI framework that unifies missing value completion and data augmentation for multi-source heterogeneous data. The core innovations are threefold:

1. A cross-modal attention embedding mechanism that unifies numerical, categorical, temporal, and textual features into a shared latent space, capturing fine-grained inter-source dependencies without manual feature alignment.
2. A missing-aware conditional diffusion imputation module that treats missing masks and observed values as conditional constraints, enabling adaptive imputation across all three missing mechanisms without prior assumptions.

3. A distribution-aligned latent space augmentation strategy that jointly optimizes sample diversity and cross-modal consistency, avoiding error accumulation in two-stage pipelines.

Extensive experiments on two real-world heterogeneous datasets validate the superiority of the proposed method over eight baseline algorithms in both imputation accuracy and downstream task improvement. Ablation studies further verify the contribution of each core module.

The remainder of this paper is organized as follows. Section 2 reviews related work on missing imputation, generative data augmentation, and heterogeneous data fusion. Section 3 details the proposed Hetero-GenAI framework, including problem formulation, module architecture, and optimization objectives. Section 4 presents experimental setup, results, and ablation analysis. Section 5 discusses the implications, limitations, and practical applicability of the method. Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1 Missing Value Imputation Methods

Missing value imputation has been studied for decades, evolving from classical statistical methods to modern deep learning approaches.

Classical statistical methods include mean/mode imputation, hot-deck imputation, and multiple imputation by chained equations (MICE) [5]. MICE, the most widely used classical method, builds separate regression models for each feature and iteratively updates imputed values. While it handles mixed data types, it assumes linear or generalized linear relationships between features and performs poorly on high-dimensional, nonlinearly correlated data. Machine learning-based methods such as K-nearest neighbor (KNN) imputation and random forest imputation capture nonlinear patterns but suffer from exponential computational cost with increasing dimensionality and fail to model long-range feature dependencies.

Deep learning has advanced imputation performance significantly. Generative Adversarial Imputation Nets (GAIN) [6] use adversarial training to learn data distributions and generate realistic imputations. Variational autoencoder (VAE)-based methods model feature distributions in latent space and perform imputation via posterior inference [7]. For temporal data, self-attention-based methods such as SAITS [8] achieve state-of-the-art performance by capturing long-range temporal dependencies. However, most deep imputation models are designed for single data types—pure numerical, pure categorical, or pure temporal. They lack a unified encoding mechanism for heterogeneous data and cannot leverage cross-modal complementary information to guide imputation. Moreover, most methods implicitly assume MAR and degrade sharply under MNAR, which is common in real-world data.

2.2 Generative AI for Data Augmentation

Data augmentation improves model generalization by expanding training sample diversity. Traditional augmentation applies deterministic transformations: for images, flipping, cropping, and color jittering; for text, synonym replacement and back-translation; for numerical data, Gaussian noise injection and

synthetic minority oversampling technique (SMOTE). These transformations are simple but limited in diversity and cannot generate novel, realistic samples.

Generative AI enables data-driven augmentation by learning the underlying data distribution. Generative adversarial networks (GANs) were the first widely adopted generative augmentation method, used to address class imbalance and sample shortage in computer vision and tabular data [9]. However, GANs suffer from training instability and mode collapse. Diffusion probabilistic models [10], which generate data via iterative denoising, have recently surpassed GANs in generation quality and stability. They have achieved remarkable success in image, audio, and text generation and are gradually being applied to tabular and temporal data augmentation [11].

Despite these advances, generative augmentation research for multi-source heterogeneous data remains limited. Most existing methods process each modality independently and ignore cross-modal semantic constraints, resulting in generated samples with mismatched modalities (e.g., a generated patient record where vital signs contradict the diagnosis text). Additionally, current augmentation pipelines assume complete input data and cannot directly handle missing values, requiring separate pre-imputation that introduces cumulative errors.

2.3 Multi-Source Heterogeneous Data Fusion

Multi-source heterogeneous data fusion aims to integrate information from different sources to obtain a unified, comprehensive data representation. Early methods used simple feature concatenation or weighted fusion, which ignore distribution differences across modalities and easily introduce noise. Modern cross-modal representation learning maps different modalities to a shared latent space via metric learning or contrastive learning. For example, CLIP [12] aligns image and text representations using contrastive loss on large-scale paired data. Multi-view learning methods also explore consensus and complementary information across different data views [13].

However, most heterogeneous fusion methods assume fully observed data. When missing values are widespread, the quality of cross-modal alignment degrades drastically. Few studies have jointly modeled missing imputation and heterogeneous fusion. To fill this gap, this paper integrates cross-modal fusion into a generative imputation-augmentation framework, using cross-modal dependencies to guide both missing value completion and synthetic sample generation.

2.4 Research Gaps

Three key gaps remain in existing research:

1. **Task separation**: Missing imputation and data augmentation are treated as independent stages, causing error propagation and suboptimal overall performance.
2. **Heterogeneity gap**: Most imputation and augmentation methods target single data types and cannot effectively leverage cross-modal correlations in multi-source heterogeneous data.
3. **Missing mechanism assumption**: Most imputation methods require pre-specification of the missing mechanism and perform poorly under mixed or MNAR scenarios common in practice.

This work addresses all three gaps with a unified generative framework.

3. Proposed Method

3.1 Problem Formulation

Let a multi-source heterogeneous dataset be denoted as $D=D^{(1)},D^{(2)},\ldots,D^{(M)}$, where is the number of data sources and each source contains samples. Each sample consists of four types of features:

Numerical features: $\mathbf{x}_i^{num} \in \mathbb{R}^{d_{num}}$, e.g., vital signs, sensor readings

Categorical features: $\mathbf{x}_i^{cat} \in 1,2,\ldots,K^{d_{cat}}$, e.g., gender, fault type

Temporal features: $\mathbf{X}_i^{ts} \in \mathbb{R}^{T\times d_{ts}}$, e.g., time-series monitoring data

Textual features: $\mathbf{x}_i^{text}$, e.g., clinical notes, maintenance logs

A binary missing mask matrix (where is the total feature dimension) indicates missing positions: if feature of sample is missing, and otherwise. Missing values can follow MCAR, MAR, or MNAR mechanisms.

The dual objectives of this work are:

1. **Missing value completion**: Learn an imputation function that maps incomplete data and mask to a complete dataset that approximates the ground-truth complete data.

2. **Data augmentation**: Learn an augmentation function that generates synthetic samples such that: (a) follows the true data distribution; (b) cross-modal semantic consistency is preserved within each synthetic sample; (c) downstream model performance improves when trained on $\hat{D} \cup D_{aug}$.

3.2 Overall Architecture of Hetero-GenAI

The Hetero-GenAI framework consists of three cascaded and jointly optimized modules, as shown in Figure 1:

1. **Heterogeneous Data Embedding Module**: Encodes four types of heterogeneous features into a shared latent space and fuses cross-modal information via multi-head cross-attention.

2. **Missing-Aware Conditional Diffusion Module**: Takes the incomplete latent representation and missing mask as conditions, and reconstructs the complete latent representation via an iterative denoising process.

3. **Distribution-Aligned Augmentation Module**: Performs latent space interpolation and semantic perturbation on the complete latent representations to generate diverse synthetic samples, with distribution alignment and cross-modal consistency constraints.

The entire framework is trained end-to-end with a multi-component loss function, jointly optimizing imputation accuracy and augmentation quality.

3.3 Heterogeneous Data Embedding Module

This module maps heterogeneous features of different types and dimensions to a unified latent space while preserving modality-specific semantics and modeling cross-modal dependencies. Each feature type uses a dedicated encoding sub-network, followed by a cross-modal attention fusion layer.

3.3.1 Modality-Specific Encoding

Numerical encoding: Numerical features are first standardized to zero mean and unit variance. A

two-layer multilayer perceptron (MLP) with LayerNorm maps each numerical vector to a d-dimensional latent embedding:

$$\mathbf{z}_i^{num}=\text{MLP}_{num}(\text{LayerNorm}(\mathbf{x}_i^{num})) \quad (1)$$

Categorical encoding: Each categorical feature is passed through a learnable embedding layer to obtain dense vectors, which are then concatenated and projected to the *d*-dimensional space via a linear layer:

$$\mathbf{z}_i^{cat}=\mathbf{W}_{cat}\cdot\ \text{Concat}(\text{Emb}(x_{i,1}^{cat}),\ldots,\text{Emb}(x_{i,d_{cat}}^{cat})) \quad (2)$$

Temporal encoding: Temporal sequences are encoded using a 2-layer Transformer encoder with sinusoidal positional encoding to capture temporal dependencies. Global average pooling is applied to the output sequence to produce a fixed-dimensional latent vector:

$$\mathbf{z}_i^{ts} = \text{GlobalAvgPool}(\text{TransformerEnc}(\mathbf{X}_i^{ts} + \mathbf{P})) \quad (3)$$

where is the positional encoding matrix.

Textual encoding: A lightweight pre-trained BERT-base model extracts the token embedding as the sentence representation, which is then projected to the d-dimensional shared space via a linear projection layer:

$$\mathbf{z}_i^{text}=\mathbf{W}_{text}\cdot\ \text{BERT}() \quad (4)$$

3.3.2 Cross-Modal Attention Fusion

The four modality embeddings are stacked into a sequence $\mathbf{Z}=[\mathbf{z}^{num},\mathbf{z}^{cat},\mathbf{z}^{ts},\mathbf{z}^{text}]\in\mathbb{R}^{4\times d}$. A multi-head cross-attention layer computes pairwise attention weights across modalities, capturing fine-grained feature correlations:

$$\mathbf{Z}_{fused}=\text{MultiHeadAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) \quad (5)$$

where are linear projections of $\mathbf{Z}$. The output is a fused latent representation that integrates information from all data sources.

This fusion mechanism allows missing information in one modality to be compensated by information from other modalities. For example, missing blood pressure values can be inferred from textual descriptions of hypertension in clinical notes.

3.4 Missing-Aware Conditional Diffusion Imputation

We base our imputation module on the denoising diffusion probabilistic model (DDPM) framework, modified to accept missing-related conditions for targeted imputation.

3.4.1 Forward Diffusion Process

The forward process gradually adds Gaussian noise to the complete latent representation over steps:

$$\mathbf{z}_t=\sqrt{\alpha_t}\mathbf{z}_{t-1}+\sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t,\boldsymbol{\epsilon}_t\sim\ N(0,\mathbf{I}) \quad (6)$$

where is a pre-defined noise schedule parameter that decreases monotonically from near 1 to near 0. After steps, approximates standard Gaussian noise.

3.4.2 Conditional Reverse Denoising Process

The reverse process predicts the noise added at each step and gradually recovers from $\mathbf{z}_T$. To guide imputation, we construct a condition vector from two components:

1.Observed feature embedding $\mathbf{z}_{obs}$: the latent embedding of known (non-missing) features.

2.Missing mask embedding $\mathbf{e}_M$: a learnable embedding of the binary mask vector, encoding which positions are missing.

The condition is injected into every layer of the denoising U-Net via cross-attention. At each denoising step t, the network predicts the noise conditioned on $\mathbf{z}_t$, t, and $\mathbf{c}$:

$$\hat{\boldsymbol{\epsilon}}_t = \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$$

A hard constraint is applied to observed positions: after each denoising step, the values at non-missing positions are replaced with the original observed values (in latent space). This ensures that imputed values are consistent with known information and prevents the model from altering observed data.

3.4.3 Adaptive Mechanism for Unknown Missing Mechanisms

To handle unknown and mixed missing mechanisms, we design an adaptive condition weighting strategy. For each missing position, the model computes a relevance score based on the number and similarity of neighboring observed features across all modalities. The constraint strength for that position is adjusted dynamically: positions with more supporting information receive stronger constraints, while highly uncertain positions allow more generation flexibility.

For MNAR scenarios (where missingness depends on the unobserved value itself), we add an auxiliary missing prediction head that predicts the mask from the latent representation. This head is trained jointly with the diffusion model, implicitly capturing the dependency between missingness and true values and reducing MNAR bias.

3.5 Distribution-Aligned Data Augmentation Strategy

After obtaining complete latent representations, we perform augmentation directly in the latent space to ensure cross-modal consistency—any change in the latent space affects all modalities synchronously, avoiding mismatched generated samples.

3.5.1 Latent Space Augmentation Operations

Two complementary augmentation operations are combined to balance diversity and realism:

1. **Spherical Linear Interpolation (Slerp)**: For pairs of real sample latent vectors and $\mathbf{z}_b$, we interpolate along the hypersphere arc:

$$\mathbf{z}_{int} = \frac{\sin((1-\lambda)\theta)}{\sin\theta}\mathbf{z}_a + \frac{\sin(\lambda\theta)}{\sin\theta}\mathbf{z}_b \tag{7}$$

where and $\lambda \in [0,1]$. Slerp preserves the magnitude structure of latent space and generates more realistic intermediate samples than linear interpolation.

2. **Principal Component-Guided Perturbation**: We perform principal component analysis (PCA) on the latent representations of real samples. Perturbations are applied along the top-k principal component directions, which correspond to the main axes of data variation. This ensures that perturbations produce semantically meaningful variations rather than random noise:

$$\mathbf{z}_{pert} = \mathbf{z} + \epsilon \cdot \sum_{i=1}^{k} w_i \mathbf{v}_i \tag{8}$$

where are principal components, are random weights, and controls perturbation magnitude.

3.5.2 Distribution and Consistency Constraints

To ensure that augmented samples follow the true data distribution and maintain cross-modal consistency, two regularization terms are added:

Distribution alignment loss: The maximum mean discrepancy (MMD) [14] between the latent distributions of real and augmented samples is minimized to align the overall distributions:

$$\mathcal{L}_{mmd}=\mathrm{MMD}(\mathbf{z}_{real},\mathbf{z}_{aug}) \tag{9}$$

Cross-modal consistency loss: For each generated sample, we decode the latent vector back to each modality and compute the mutual information between modality pairs. A contrastive loss pulls paired modalities from the same sample closer and pushes modalities from different samples apart, preserving cross-modal semantic association:

$$\mathcal{L}_{cmc}=-\log f_0\frac{\exp f_0(\mathrm{sim}(\mathbf{z}^{num},\mathbf{z}^{text})/\tau)}{\sum_j \exp f_0(\mathrm{sim}(\mathbf{z}^{num},\mathbf{z}_j^{text})/\tau)} \tag{10}$$

where is cosine similarity and is the temperature parameter.

3.6 Optimization Objective

The total loss function combines four components for joint optimization:

$$\mathcal{L}_{total}=\mathcal{L}_{diff}+\alpha\mathcal{L}_{imp}+\beta\mathcal{L}_{cmc}+\gamma\mathcal{L}_{mmd} \tag{11}$$

1. **Diffusion loss ($\mathcal{L}_{diff}$)**: The MSE between predicted noise and true noise, the standard DDPM training objective:

$$\mathcal{L}_{diff}=\mathrm{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}_t}[|\boldsymbol{\epsilon}_t-\hat{\boldsymbol{\epsilon}}_t|^2] \tag{12}$$

2. **Imputation reconstruction loss ($\mathcal{L}_{imp}$)**: Measures the accuracy of decoded imputed values relative to ground truth, evaluated only on missing positions. MSE is used for numerical and temporal features, and cross-entropy for categorical features.

3. **Cross-modal consistency loss ($\mathcal{L}_{cmc}$)**: As defined in Section 3.5.2.

4. **MMD distribution loss ($\mathcal{L}_{mmd}$)**: As defined in Section 3.5.2.

α,β,γ are hyperparameters weighting each loss component, tuned via validation set. The model is trained using the AdamW optimizer with a cosine learning rate scheduler.

4. Experiments and Results

4.1 Datasets

Two real-world multi-source heterogeneous datasets from different domains are used to verify the generalizability of the proposed method.

4.1.1 MIMIC-IV Clinical Dataset

The MIMIC-IV database [15] is a large-scale electronic health record dataset from the Beth Israel Deaconess Medical Center. We extracted a subset of 10,000 adult inpatient stays with the following multi-modal features:

Numerical (12 features): Heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, body temperature, oxygen saturation, etc.

Categorical (8 features): Gender, admission type, ICU type, primary diagnosis category, etc.

Textual: Discharge summary notes (average length 180 tokens).

Missing values are artificially generated following three mechanisms:

MCAR: Features are missing with uniform random probability.

MAR: Missing probability of a feature depends on the values of other observed features (e.g., missing blood pressure is more likely for older patients).

MNAR: Missing probability depends on the feature's own true value (e.g., extremely high blood pressure values are more likely to be missing due to sensor saturation).

Missing rates are set to 10%, 30%, and 50%. The downstream task is 30-day hospital readmission prediction (binary classification).

4.1.2 NASA C-MAPSS Turbofan Engine Dataset

The C-MAPSS dataset [16] is a benchmark for engine degradation prediction. We use the FD001 subset with 100 engines and augment it with synthetic maintenance log text generated based on fault conditions to form heterogeneous data:

Temporal (21 sensor features): Time-series sensor readings tracking engine degradation.

Numerical (3 features): Three operational setting parameters.

Categorical: Fault type label.

Textual: Simulated maintenance log text describing observed engine conditions.

Missing mechanisms and rates are identical to the MIMIC-IV experiment. The downstream task is remaining useful life (RUL) prediction (regression).

4.2 Baseline Methods

We compare the proposed method with eight baselines divided into three categories:

Traditional imputation methods

1.**Mean/Mode**: Numerical features imputed with mean, categorical with mode.

2.**MICE**: Multiple imputation by chained equations [5].

3.**KNN Imputer**: K-nearest neighbor imputation with k=5.

Deep learning imputation methods

4. **GAIN**: Generative adversarial imputation network [6].

5. **VAE-Imp**: VAE-based missing imputation [7].

6. **SAITS**: Self-attention-based temporal imputation, extended to handle heterogeneous features [8].

Two-stage augmentation methods (imputation + generation)

7. **MICE + GAN-Aug**: MICE imputation followed by tabular GAN augmentation.

8. **GAIN + DiffAug**: GAIN imputation followed by diffusion-based data augmentation.

4.3 Evaluation Metrics

4.3.1 Imputation Metrics

Numerical features: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), lower is better.

Categorical features: Macro F1-score and Accuracy, higher is better.

Temporal features: Temporal RMSE (TRMSE), computed across all time steps, lower is better.

4.3.2 Augmentation Metrics

Downstream task performance: For MIMIC-IV readmission classification: AUC-ROC and Accuracy. For C-MAPSS RUL prediction: RMSE and MAE. We compare performance before and after augmentation to measure improvement.

Generation quality: Frechet Inception Distance (FID) computed in the latent space, measuring distribution similarity between generated and real samples (lower is better). Density and Coverage metrics measure sample diversity and fidelity.

4.4 Experimental Setup

All experiments are implemented in PyTorch 2.0 and run on an NVIDIA A100 40GB GPU. The proposed Hetero-GenAI uses the following key hyperparameters: latent dimension d=256, diffusion steps T=1000, 8 attention heads, 2 Transformer encoder layers, learning rate 10^{-4}, batch size 64, and 200 training epochs with early stopping (patience=15). Loss weights are set to α=1.0, β=0.5, γ=0.3, tuned via 5-fold cross-validation.

All baseline methods are tuned with grid search to their optimal hyperparameters. For fair comparison, all deep learning methods use the same training-validation-test split (70%-10%-20%). Each experiment is repeated 5 times with different random seeds, and the mean and standard deviation are reported.

4.5 Experimental Results and Analysis

4.5.1 Missing Value Imputation Performance

Table 1 presents imputation performance under 30% MCAR missing rate on both datasets.

Table 1. Imputation Performance Comparison (30% MCAR)

Method	MIMIC-IV	MIMIC-IV	MIMIC-IV	C-MAPSS
	Num RMSE	Cat F1	Text Sim	TRMSE
Mean/Mode	0.892±0.021	0.583±0.017	0.421±0.022	1.204±0.031
MICE	0.674±0.015	0.672±0.012	0.518±0.019	0.976±0.024
KNN	0.713±0.018	0.645±0.014	0.497±0.020	1.012±0.027
VAE-Imp	0.562±0.013	0.724±0.010	0.603±0.015	0.815±0.021
GAIN	0.518±0.011	0.751±0.009	0.637±0.013	0.762±0.018
SAITS	0.541±0.012	0.718±0.011	0.592±0.016	0.729±0.016
Hetero-GenAI	**0.439±0.009**	**0.847±0.007**	**0.742±0.010**	**0.637±0.013**

The results show that:

1. Traditional statistical methods perform the worst, as they cannot model nonlinear feature correlations. MICE outperforms KNN and mean imputation but still lags far behind deep learning methods.

2. Deep learning methods achieve better imputation accuracy. GAIN performs best among baselines due to its adversarial distribution learning. SAITS excels at temporal features but performs poorly on categorical and textual features, as it is designed primarily for time series.

3. The proposed Hetero-GenAI consistently achieves the best performance across all metrics and all feature types. Compared to the second-best baseline (GAIN), it reduces numerical RMSE by 15.2%, improves categorical F1 by 12.8%, and reduces temporal RMSE by 16.4%. This confirms that cross-modal attention fusion effectively leverages complementary information across modalities to guide imputation.

4.5.2 Performance Under Different Missing Rates

Figure 2 shows numerical RMSE on MIMIC-IV under missing rates from 10% to 50%. All methods degrade as missing rate increases, but Hetero-GenAI degrades the slowest. At 50% missing rate, Hetero-GenAI still outperforms GAIN by 21.7% in RMSE. This robustness stems from the conditional diffusion mechanism, which can still generate reasonable imputations even with few observed features, supported by cross-modal information.

4.5.3 Performance Under Different Missing Mechanisms

Table 2 compares imputation RMSE across MCAR, MAR, and MNAR mechanisms at 30% missing rate on MIMIC-IV.

Table 2. Numerical RMSE under Different Missing Mechanisms (30% missing rate)

Method	MCAR	MAR	MNAR
MICE	0.674	0.721	0.895
VAE-Imp	0.562	0.604	0.783
GAIN	0.518	0.567	0.712
Hetero-GenAI	**0.439**	**0.472**	**0.558**

All methods perform best on MCAR and worst on MNAR, as MNAR introduces the strongest bias. However, Hetero-GenAI shows the smallest performance drop from MCAR to MNAR (27.1% relative increase in RMSE) versus GAIN (37.5%). The adaptive missing-aware mechanism and auxiliary missing prediction head effectively mitigate MNAR bias by implicitly modeling the dependency between missingness and true values.

4.5.4 Data Augmentation Performance

Table 3 presents downstream task performance with and without data augmentation. The "Baseline" row shows performance after imputation but before augmentation.

Table 3. Downstream Task Performance with Data Augmentation

Method	MIMIC-IV (AUC)	C-MAPSS (RMSE)
Original (no imputation)	0.712	18.73
MICE (no augmentation)	0.765	15.42
GAIN (no augmentation)	0.782	13.67
MICE + GAN-Aug	0.803	14.01
GAIN + DiffAug	0.831	12.45
Hetero-GenAI (full)	**0.896**	**11.23**

The results demonstrate:

1. Missing value imputation alone improves downstream performance significantly, and GAIN's superior imputation translates to better baseline performance than MICE.

2. Two-stage augmentation methods provide further improvement but with limited gains. MICE + GAN-Aug even slightly hurts C-MAPSS performance, likely because GAN generates low-quality samples that introduce noise.

3. Hetero-GenAI achieves the best downstream performance: 14.6% AUC improvement over the GAIN baseline on MIMIC-IV and 17.8% RMSE reduction on C-MAPSS. Compared to the best two-stage baseline (GAIN + DiffAug), it improves AUC by 7.8% and reduces RMSE by 9.8%. This validates the advantage of the unified joint optimization framework over two-stage pipelines, as it avoids error accumulation and preserves cross-modal consistency in generated samples.

Latent space FID results further confirm generation quality: Hetero-GenAI achieves an FID of 12.37, compared to 28.54 for GAIN + DiffAug and 41.29 for MICE + GAN-Aug, indicating that samples generated by the proposed method are much closer to the real data distribution.

4.6 Ablation Study

To verify the contribution of each core module, we conduct ablation studies on the MIMIC-IV dataset with 30% MNAR missing rate. Four variants are compared:

Full model: The complete Hetero-GenAI framework.

w/o Cross-Modal Attention: Replace cross-modal attention with simple feature concatenation.

w/o Missing-Aware Condition: Remove mask conditioning, use unconditional diffusion for imputation.

w/o Distribution Alignment: Remove MMD and cross-modal consistency losses from augmentation.

Table 4. Ablation Study Results

Variant	Num RMSE	Cat F1	Downstream AUC	FID
Full model	0.558	0.812	0.896	12.37
w/o Cross-Modal Attn	0.647 (-15.9%)	0.748 (-7.9%)	0.834 (-6.9%)	19.82
w/o Missing-Aware	0.683 (-22.4%)	0.723 (-10.9%)	0.817 (-8.8%)	21.45
w/o Dist. Alignment	0.561 (-0.5%)	0.809 (-0.4%)	0.852 (-4.9%)	23.71

Key observations:

1. Removing cross-modal attention causes the largest performance drop in imputation accuracy, confirming that cross-modal complementary information is critical for heterogeneous data imputation.

2. Removing the missing-aware condition severely degrades performance, especially under MNAR, demonstrating that mask conditioning is essential for targeted, adaptive imputation.

3. Distribution alignment has minimal impact on imputation accuracy but significantly affects augmentation quality (FID increases by 91.5%) and downstream task performance. This shows that distribution and consistency constraints are crucial for generating high-quality, useful augmented samples.

All three modules contribute meaningfully to the overall performance, validating the rationality of the framework design.

5. Discussion

5.1 Implications of the Findings

The experimental results lead to three key conclusions with practical implications.

First, cross-modal correlation is a valuable and underutilized resource for missing value imputation in heterogeneous data. Most existing imputation methods treat features independently or only model within-modality correlations. Our results show that explicitly modeling cross-modal dependencies via attention mechanisms can yield substantial imputation improvements, especially when certain modalities are heavily missing. In clinical practice, for example, textual notes can compensate for missing laboratory results, reducing diagnostic bias.

Second, unifying imputation and augmentation in a single generative framework provides significant advantages over two-stage pipelines. Beyond the intuitive benefit of reduced error accumulation, joint training allows the augmentation objective to regularize the imputation model, leading to better

distribution learning and more realistic imputations. This paradigm can be extended to other data preprocessing tasks such as denoising and outlier removal.

Third, adaptive missing mechanism handling greatly improves model applicability in real-world scenarios. In practice, data analysts rarely know the exact missing mechanism in advance, and mechanisms are often mixed. Our method's strong performance across MCAR, MAR, and MNAR eliminates the need for manual mechanism identification and reduces the risk of model mismatch.

5.2 Limitations

Despite its effectiveness, the proposed method has several limitations.

1. **Computational efficiency**: Diffusion models require many sampling steps, making inference slower than traditional imputation methods. For ultra-large datasets with millions of samples, this could become a bottleneck.

2. **Modal scope**: This work covers numerical, categorical, temporal, and textual data but does not address unstructured modalities such as images or audio. Extending to these modalities would require additional encoding structures.

3. **Interpretability**: Like most deep generative models, Hetero-GenAI is a black-box model. It cannot provide explicit reasoning for imputed values, which may hinder adoption in highly regulated domains such as clinical medicine.

5.3 Practical Applicability

The proposed framework has broad application prospects. In healthcare, it can improve the quality of electronic health record data for predictive analytics and clinical research. In manufacturing, it can enhance IIoT sensor data for predictive maintenance and quality control. In finance, it can impute missing transaction data and augment fraud detection datasets. The modular design allows easy adaptation to different domain-specific modality combinations.

6. Conclusion

This paper presents Hetero-GenAI, a novel generative AI framework for joint missing value completion and data augmentation in multi-source heterogeneous big data. By integrating cross-modal attention embedding, missing-aware conditional diffusion imputation, and distribution-aligned latent augmentation, the framework addresses key limitations of existing methods: modality isolation, two-stage error accumulation, and missing mechanism dependency. Extensive experiments on two real-world datasets from clinical and industrial domains demonstrate that the proposed method achieves state-of-the-art performance in both imputation accuracy and downstream task improvement under various missing rates and mechanisms.

Future work will proceed in four directions:

1. **Efficiency optimization**: Explore fast diffusion sampling algorithms (e.g., DDIM) and latent diffusion architectures to reduce inference time for large-scale data.

2. **Modal expansion**: Extend the framework to support image, audio, and video modalities for richer

heterogeneous scenarios.

3. **Privacy-preserving extension**: Combine the framework with federated learning to enable distributed multi-source data imputation and augmentation without raw data sharing.

4. **Interpretability enhancement**: Integrate attention visualization and feature attribution methods to provide interpretable explanations for imputation decisions.

References

[1] Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314-347.

[2] Zhang, Y., et al. (2022). Multi-source data fusion for industrial IoT: A survey. *IEEE Internet of Things Journal*, 9(20), 19864-19883.

[3] Goldstein, B. A., et al. (2016). Missing data in clinical prediction models: A review of current practice and recommendations. *Journal of the American Medical Informatics Association*, 23(5), 936-943.

[4] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1-48.

[5] Van Buuren, S., & Oudshoorn, K. (2000). Flexible multivariate imputation by MICE. *Statistica Neerlandica*, 54(3), 294-307.

[6] Yoon, J., Jordon, J., & Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. *Proceedings of the 35th International Conference on Machine Learning*, 5689-5698.

[7] Rezende, D. J., et al. (2014). Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Conference on Machine Learning*, 1278-1286.

[8] Du, W., Côté, D., & Liu, Y. (2023). SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219, 119619.

[9] Frid-Adar, M., et al. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, 321-331.

[10] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840-6851.

[11] Huang, X., et al. (2023). TabDDPM: Modelling tabular data with diffusion models. *Proceedings of the 40th International Conference on Machine Learning*, 13731-13745.

[12] Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748-8763.

[13] Xu, C., et al. (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

[14] Gretton, A., et al. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(1), 723-773.

[15] Johnson, A. E., et al. (2021). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 8(1), 1-7.

[16] Saxena, A., & Goebel, K. (2008). *Turbofan engine degradation simulation data set*. NASA Ames Prognostics Data Repository.

[17] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[18] Song, J., et al. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

[19] Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.

[20] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.