

# Cover Letter for "On LLM-Enhanced Mixed-Type Data Imputation with High-Order Message Passing"

We would like to thank the meta-reviewer and all reviewers for their insightful comments. We have revised the paper thoroughly and addressed all the issues according to the comments. The revised parts are highlighted in blue in the manuscript. Please find a summary of the revisions below.

## RESPONSE TO META-REVIEWER

**Comment M1:** Clarifications (figures, LLMs specially trained for tables, etc).

**Response M1:** Thank you for your insightful comments. We make the following clarifications in the revised paper.

1. *Figures improvements.* In [Figure 1](#) and [Figure 2](#), we enhance the figures with color-coded components, annotated workflows, and example-driven visualizations. Moreover, we add the illustration of the GNN-only methods. More details can be found in [Response 1.1](#) and [Response 3.1](#).
2. *Prompt construction for LLM.* In [Section 3.3](#), we use formal formulations on the prompt construction. Also, in [Exp-8](#), we evaluate different serialization strategies. More details can be found in [Response 2.2](#) and [Response 2.3](#).
3. *Component analysis.* In [Section 3.3](#), [Section 3.5](#) and [Section 3.6](#), we add a deeper discussion of cell-oriented hypergraph modeling, bidirectional high-order message passing and XFusion block with reference to individual datasets. More details can be found in [Response 1.5](#). Moreover, we add more details about the hypergraph properties in [Section 3.3](#). More details can be found in [Response 3.3](#).
4. *Detailed discussion of the strength and weakness of UnIMP.* In [Section 5.1](#), we further clarify the approximation ratio and add more details about the derivation of the approximation ratio. More details can be found in [Response 1.2](#).
5. *Bound analysis.* In [Section 4.3](#), we improve our bound analysis with hierarchical complexity analysis and more intuitive parameter sets. More details can be found in [Response 3.4](#).
6. *Complete proof.* In [Section 4.4](#), we improve our proof by explicitly linking the model components and introducing formal equations. More details can be found in [Response 3.6](#).

**Comment M2:** Experiments (LLM parameters, chunking size, baseline methods, etc).

**Response M2:** Thank you for your valuable comments. We add the following details and experiments in the revised paper.

1. *Experimental setup.* In [Section 5.1](#), we further clarify the parameter settings of the LLM, the training datasets used, and the computational resources involved. More details can be found in [Response 1.2](#), [Response 1.4](#) and [Response 2.1](#).
2. *Hyperparameter generalization.* In [Figure R1-D3-1](#) and [Figure R1-D3-2](#), we further investigate the generalizability of  $\delta$  and  $\kappa$  across various noise distributions as well as different missingness patterns. More details can be found in [Response 1.3](#).
3. *Scalability evaluation.* In [Exp-6 of Section 5](#), we further include an investigation of scalability w.r.t extremely large datasets and high dimensional datasets. More details are in [Response 1.3](#).

4. *Computational resources.* In [Exp-5 of Section 5](#), we add experiments to evaluate the CPU and GPU memory consumption. More details can be found in [Response 2.1](#).
5. *Feature encoding.* In [exp-18](#), we further investigate different feature encoding methods, including one-hot for categorical data and HYREL [8] and BERT [11] for text. More details can be found in [Response 2.2](#).
6. *Serialization strategy.* In [Exp-8](#), we further evaluate different serialization strategies, including attribute-value pairs (adopted by UnIMP), sentences, markdown, and dataframes. More details can be found in [Response 2.3](#).
7. *Imputation on different text linguistics.* In [Exp 8](#), we further evaluate four types of text linguistics: long text, short text, structured text, and colloquial text. More details can be found in [Response 2.3](#).
8. *Bias evaluation.* In [Figure 9](#) and [Table 10](#), we further assess whether the imputation process introduces any bias. More details can be found in [Response 3.11](#).

## RESPONSE TO REVIEWER #1

**Comment 1.1 (W1, D1):** Figures 1 and 2 lack clarity in illustrating key technical innovations (e.g., hypergraph message passing and LLM integration); enhancing them with color-coded components, annotated workflows, and example-driven visualizations would improve readability and highlight the framework's novelty.

**Response 1.1:** Thanks for the insightful comments. In the revision, we substantially refine [Figure 1](#) and [Figure 2](#) to better illustrate the key technical innovations. Details are as follows.

1. *Color-coded components.* We align the colors of components in [Figure 1](#) and [Figure 2](#) to improve visual consistency. Moreover, we add a legend to both figures to aid interpretation.
2. *Annotated workflows.* We added annotations to each component in [Figure 1](#) and [Figure 2](#) to improve readability. In addition, we included detailed descriptions of the workflows to enhance clarity and understanding.
3. *Example-driven visualizations.* In [Figure 1](#), we use a unified abstract dataset to illustrate how different models process the same input. In [Figure 2](#), we present a real-world example to demonstrate how the various modules in UnIMP collaborate to predict the missing name of the president.

**Comment 1.2 (D2):** Not clear if the LLMs were configured with parameters or used out-of-the-box. It is recommended to clarify this in the experimental settings.

**Response 1.2:** Thanks for the valuable comments. In the [experimental setup in Section 5.1](#) of the revised paper, we make the following revision to improve clarity: "*The default LLM for all baselines is Llama2-7B (additional LLMs are evaluated in Exp-10), used out-of-the-box without parameter modifications.*"

**Comment 1.3 (W2,D3):** Key hyperparameters (e.g.,  $\delta=1$  in Huber loss,  $k=35\%$  in progressive masking) are selected via limited experiments without theoretical grounding. This raises concerns

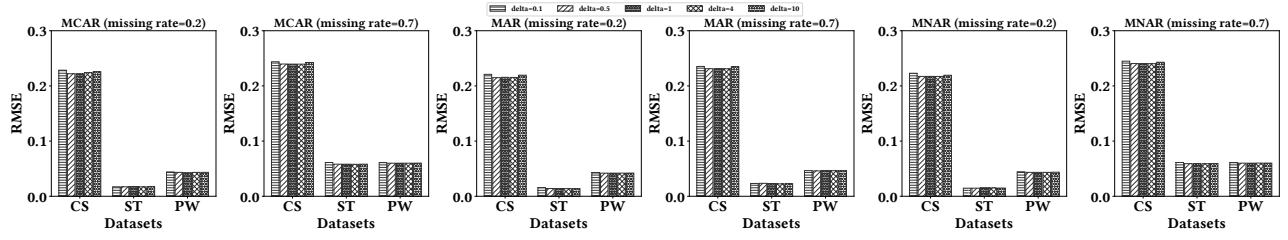


Figure R1-D3-1: Experimental results of hyperparameter generalization on delta.

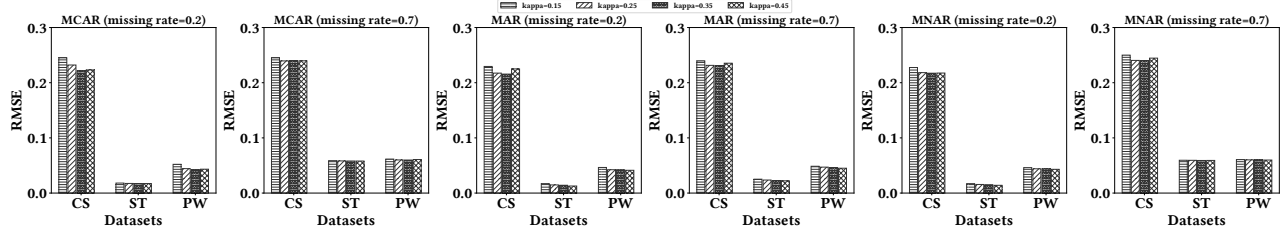


Figure R1-D3-2: Experimental results of hyperparameter generalization on kappa.

about generalizability to datasets with different noise distributions or missingness patterns. While chunking is used for large tables, the scalability of hypergraph construction and message passing for high-dimensional or extremely large datasets (e.g., millions of rows) is not analyzed.

**Response 1.3:** Thanks for the valuable comments. We make the following revisions to strengthen the evaluation w.r.t. generalizability and scalability.

1. *Generalizability across noise distributions and missingness patterns.* In Figure R1-D3-1 and Figure R1-D3-2, we further investigate the generalizability of  $\delta$  and  $\kappa$  across various noise distributions, including MCAR, MAR, and MNAR, as well as different missingness patterns, such as a moderate missing rate (0.2) and a high missing rate (0.7). The tested values of  $\delta$  are 0.1, 0.5, 1, 4, and 10. We also evaluate the imputation RMSE for different values of  $\kappa$ , specifically 0.15, 0.25, 0.35, and 0.45. The results reveal a consistent trend in which RMSE first decreases slightly and then increases as the values of  $\delta$  or  $\kappa$  increase. Overall, our method demonstrates stable performance across a wide range of hyperparameter settings. These observations confirm the strong generalizability and robustness of our approach in handling diverse missing data scenarios. (Due to space limitations, detailed results are provided in the appendix of our online full version.)<sup>1</sup>
2. *Scalability for extremely large datasets and high-dimensional.* In Exp-6 of Section 5.3 of the revised paper, we include an investigation of the scalability of UnIMP which contains the hypergraph construction process and message passing process. To evaluate scalability w.r.t. large datasets, we randomly select 100, 500, 2,000, 20,000, 200,000, and 2,000,000 samples respectively from the Power dataset. To assess scalability w.r.t. high-dimensional datasets, we generate synthetic datasets with 200 samples and feature dimensions of 5, 10, 50, 250, and 500, respectively. The results presented in Figure 4(c) and Figure 4(d) highlight the competitive scalability of UnIMP. For scalability w.r.t. the number of samples, ReMasker has a bad scalability

since it requires remasking and training each sample, while NOMI has a good scalability as it focuses only on local context. Graph-based methods (GRAPE, IGRM, UnIMP) fall in between. Although graph-based methods leverage both global and local information, they benefit from the parallel processing capabilities of GNNs and thus have competitive scalability. For scalability w.r.t. dimension, NOMI has a bad scalability due to column-wise enumeration for imputation, whereas ReMasker has a good scalability as it imputes all missing columns simultaneously. Graph-based methods fall in between and exhibit consistent behavior, as they all formulate imputation as link prediction.

**Comment 1.4 (W3):** The pre-training process uses "all available datasets," but the paper does not clarify whether these include test datasets or overlapping domains. Unintentional data leakage during pre-training could artificially inflate fine-tuning performance.

**Response 1.4:** Thanks for the insightful comments. In Table 3 of the revised paper, we further add the detailed domains of these datasets. As shown in the table, these datasets are from different domains, such as biology, physics and healthcare. Moreover, in the experimental setup in Section 5.1 of the revised paper, we make the following revision to improve clarity: *"The pre-training phase uses only raw incomplete datasets. Moreover, these datasets are distinct and from different domains as shown in Table 3, ensuring no overlap for a fair evaluation."*

**Comment 1.5 (D4):** The component analysis can be improved. A deeper discussion with reference to individual datasets would help to better understand the benefits of individual components.

**Response 1.5:** Thanks for the helpful comments. We make the following revisions to improve clarity and better highlight the benefits of key individual components in UnIMP:

1. *Add deeper discussion between individual components and datasets in the ablation study.* In Table 6, we provide additional details on the improvement each component brings to each dataset. We also add more detailed discussion with overall results and individual dataset as follows: *"The results in Table 6 show that hypergraph information improves ROUGE-1F<sub>1</sub> by 5.96 percentage points (pp) and Cos-Sim by 2.03 pp. BiHMP contributes a 2.55 pp increase in*

<sup>1</sup>The full version is in [https://github.com/guaiyoui/UnIMP/blob/master/full\\_version/UnIMP\\_Full\\_Version.pdf](https://github.com/guaiyoui/UnIMP/blob/master/full_version/UnIMP_Full_Version.pdf)

ROUGE-1 $F_1$  and 1.04 pp in Cos-Sim. Xfusion improves ROUGE-1 $F_1$  by 4.57 pp and Cos-Sim by 1.94 pp. Progressive masking results in a 1.43 pp gain in ROUGE-1 $F_1$  and 0.74 pp in Cos-Sim. Overall, these components are more important on the Walmart dataset. This is because Walmart is larger and more reliant on these components to aggregate information for accurate imputation."

2. *Cell-oriented hypergraph modeling.* In Section 3.3, We provide a more in-depth introduction to hypergraph modeling, with references to the dataset in Figure 2, as follows: "This hypergraph-based modeling effectively captures tabular characteristics. For instance, in Figure 2, a hyperedge captures the high-order relationship among a president's name, nation, and term in office. Moreover, each cell is represented as a distinct node, capturing intra-column heterogeneity, while nodes within the same column are linked by a hyperedge, capturing inter-column homogeneity."
3. *Bidirectional high-order message passing.* In Section 3.5, We provide a deeper discussion of BiHMP, with references to the dataset in Figure 2, as follows: "When predicting the name of the president, as shown in Figure 2, the Node-to-Hyperedge layer aggregates contextual information within hyperedge, such as name formatting. The Hyperedge-to-Node layer then can utilize this information to aid prediction."
4. *XFusion block.* In Section 3.6, We provide a deeper discussion of the XFusion block, with references to the dataset in Figure 2, as follows: "When predicting the name of the president, as shown in Figure 2, the LLM uses its knowledge to infer the name, while the hypergraph provides contextual details, such as name patterns. XFusion combines both to enhance imputation accuracy."

## RESPONSE TO REVIEWER #2

**Comment 2.1 (W1, D1):** The authors acknowledge that some baseline imputation methods could be very demanding in terms of computational resources. They should include more explicit details on the hardware and software requirements, including the number of GPUs, CPU specifications, memory usage, and approximate runtimes for training datasets of different sizes and complexities. In general, more focus should be placed on the practical aspects of implementing the method, and how it can be scaled up.

**Response 2.1:** Thank you for the helpful comments. In the revised paper, we make the following revisions to improve clarity regarding computational resource usage, practical implementation, and scalability across different contexts.

1. *Add more details on implementation and computational resource usage.* In implementation details of Section 5.1, we provide more information on the experimental platform: *We implement our method and baselines with Pytorch and conduct the experiment on a server equipped with an Intel Xeon(R) Silver 4314 CPU, 256GB system memory, and 2 Nvidia A5000 GPUs, each with 24GB GPU memory.* We also add experiments on CPU and GPU memory consumption in Exp-5. Results illustrated in Figure 4(b) show that graph-based methods require slightly more memory due to the graph construction process, while our method shows competitive memory usage compared to other graph-based or LLM-based methods.
2. *Add more experiments on scalability across different contexts.* In Exp-6, we evaluate runtime scalability with varying dataset sizes and dimensions. Specifically, we assess performance using 100, 500, 2,000, 20,000, 200,000, and 2,000,000 samples randomly selected from the Power dataset, and synthetic datasets with 200 samples and 5, 10, 50, 250, and 500 dimensions, respectively.

The results summarized in Figures 4(c) and 4(d) highlight the competitive scalability of UnIMP. For scalability w.r.t. the number of samples, ReMasker has a bad scalability since it requires remasking and training each sample, while NOMI scales the slowest as it focuses only on local context. Graph-based methods (GRAPE, IGRM, UnIMP) fall in between. Although graph-based methods leverage both global and local information, they benefit from the parallel processing capabilities of GNNs and thus have competitive scalability. For scalability w.r.t. dimension, NOMI has a bad scalability due to column-wise enumeration for imputation, whereas ReMasker has a good scalability as it imputes all missing columns simultaneously. Graph-based methods fall in between and exhibit consistent behavior, as they all formulate imputation as link prediction.

3. *Add more details on practical aspects of implementing and scaling the method.* In Exp-11 and Exp-12, we evaluate the chunking and batching techniques, which are critical for scaling up UnIMP. The chunk sizes tested are 128, 256, 512, 1024 and 2048. The batch sizes tested are 16, 32, 64, 128 and 256. The results in Figures 5(b) and 5(c) show that imputation time initially decreases and then increases as chunk size grows, with a chunk size of 512 achieving a good balance between RMSE and runtime. Additionally, imputation time consistently decreases with larger batch sizes, and a batch size of 64 generally offers a favorable trade-off between RMSE and efficiency. These results provide practical guidance for selecting hyperparameters to implement and scale our UnIMP.

**Comment 2.2 (W2, D2):** The paper needs a more detailed explanation of how the prompts are constructed. It is also not completely clear how numerical data are encoded and the impact of the encoding on the effectiveness of the proposed method.

**Response 2.2:** Thanks for the valuable comments. In the revision, we make the following detailed explanations on prompt construction and feature encoding.

1. *Add more details on the prompt construction..* In Section 3.4, we provide a formal formulation of the prompts for nodes and hyperedges as follows:  

$$\text{Serial}(x_{ij}) = \text{Row } i, \text{col\_name}[d \neq i]: x_{dj}, \text{col\_name}[i] \Rightarrow \{x_{ij}\} \text{ EOS}$$

$$\text{Serial}(i\text{-th row}) = \text{This is row: } i \text{ EOS;}$$

$$\text{Serial}(i\text{-th col}) = \text{This is col: col\_name}[i] \text{ EOS;}$$

Moreover, we provide an example in Example 3.1 to better illustrate the prompts construction process as follows: "For the node 'Obama' shown in Figure 2, we serialize as: This is row 0, Nation of the President: USA, Term of the President: 44, Name => Obama EOS. For the hyperedge corresponding to the first column, we serialize as: This is col: Name EOS."

2. *Add more details on numerical data and experiments on the feature encoding.* In Section 3.4, we add more details on the encoding mechanism for numerical data as follows: "For numerical data, following previous works[13,37], we directly encode the input as its value to avoid unnecessary computational overhead." Moreover, in Exp-18, we evaluate different feature encoding methods, including one-hot for categorical data and HYTREL [8] and BERT [11] for text. The results are summarized in Figure 6. The results show that one-hot encoding generally performs better. Our default setting follows previous works [13, 55] and uses label encoding for categorical data. Additionally, the results highlight that HYTREL and BERT perform less effectively compared to naive LLM. The likely reason is that these methods are designed



for representation learning, and the encoded features are not well aligned with the projection head for imputation.

**Comment 2.3 (W3, D3):** The contribution of LLMs could be very sensitive to prompts and changes in the type of text being imputed. In addition, the reliance on a particular pre-trained LLM might also limit the generalisability of the method. These aspects should be better discussed.

**Response 2.3:** Thanks for the useful comments. In the revised paper, we make the following revisions to study prompt formulations, types of text being imputed, and to provide a more detailed discussion on the reliance on a particular pre-trained LLM.

1. *Add experiments on different prompt formulations.* In [Exp-8](#), we evaluate different prompt formulations (i.e., serialization strategies). We tested four commonly used strategies: attribute-value pairs (adopted by UnIMP), sentences, markdown, and dataframes. The results in [Table-8](#) show that, despite its simplicity, the attribute-value pair strategy achieves the best performance. Additionally, UnIMP, benefiting from fine-tuning, demonstrates competitive performance across different serialization strategies.
2. *Add experiments on type of text being imputed.* In [Exp-8](#), we evaluate four types of text linguistics: Long (text with more than 10 characters), Short (text with fewer than 10 characters), Structured (text with a fixed format, e.g., telephone numbers), and Colloquial (unstructured text, e.g., restaurant names). The results in [Table-7](#) show that UnIMP perform better on short text due to lower cumulative errors compared to long text. Additionally, colloquial text imputation outperforms structured text since language models are pre-trained on natural language, which is inherently flexible rather than fixed in format.
3. *Add detailed discussion on the strengths and limitations of UnIMP w.r.t. the reliance on a particular pre-trained LLM.* In [Section 7](#), we add more details on the limitations and comparison with LLM-based imputation as follows: *"UnIMP models granular tabular structures and leverages hypergraphs and HGNN for reasoning, enabling more accurate information aggregation for imputation. However, it has several limitations: 1) Cross-dataset information. Table-GPT, Jellyfish, and DFMs do not explicitly model granular tabular structures, making them more flexible in integrating multiple tabular datasets for imputation. In contrast, UnIMP struggles to capture cross-table information. 2) Complementary information from related tasks. Table-GPT, Jellyfish, and DFMs support various data quality management tasks and can integrate results from related tasks. For example, entity-matching outcomes may provide contextual information that enhances imputation. 3) Flexibility to update LLM. DFMs, based on prompt engineering, may be sensitive to prompt variations but can flexibly utilize different LLMs. In contrast, UnIMP, Table-GPT, and Jellyfish, which rely on fine-tuning, are less sensitive to prompt structures after fine-tuning but depend on a specific pre-trained LLM, potentially impacting their generalizability."*

## RESPONSE TO REVIEWER #3

**Comment 3.1 (D1) D1)** Figure 1 should include a comparison with GNN-only approaches.

**Response 3.1:** Thanks for the constructive comments. In the revision, we include a comparison with GNN-only approaches in [Figure 1\(a\)](#). Moreover, we substantially improve [Figure 1](#) and [Figure](#)

[2](#) with color-coded components, annotated workflows, and example-driven visualizations to enhance presentation and readability.

**Comment 3.2 (D2)** Are there envisioned cases where leveraging the reasoning capabilities of the LLM over a larger portion of the table, as in Table-GPT, could be more effective? What are the pros and cons of limiting LLMs to accessing only one or a few records at a time while relying on the hypergraph + GNN for the reasoning?

**Response 3.2:** Thank you for the valuable comments. There are potential cases, particularly when leveraging cross-dataset information and complementary insights from related tasks, where other LLM-based methods, such as Table-GPT, may be more effective. In [Section 7](#) of the revised paper, we add more details about the pros and cons of UnIMP as follows: *"UnIMP models granular tabular structures and leverages hypergraphs and HGNN for reasoning, enabling more accurate information aggregation for imputation. However, it has several limitations: 1) Cross-dataset information. Table-GPT, Jellyfish, and DFMs do not explicitly model granular tabular structures, making them more flexible in integrating multiple tabular datasets for imputation. In contrast, UnIMP struggles to capture cross-table information. 2) Complementary information from related tasks. Table-GPT, Jellyfish, and DFMs support various data quality management tasks and can integrate results from related tasks. For example, entity-matching outcomes may provide contextual information that enhances imputation."*

**Comment 3.3 (D3)** What are the degree and edge cardinality distributions in the constructed hypergraphs? Is the proposed method robust to skewed distributions?

**Response 3.3:** Thank you for your insightful comments. In the revised paper, we make the following revisions to clarify the degree and edge cardinality distributions, as well as to evaluate the robustness to skewed distributions.

1. *Add details about degree and edge cardinality distributions.* In [Section 3.2](#) of the revised paper, we add a detailed introduction to the degree and edge distributions of the constructed hypergraph, as follows: *"Each node has a degree of 2, connecting to both a row-level and a column-level hyperedge. The edge cardinality of a row-level hyperedge is the number of columns ( $d$ ), while that of a column-level hyperedge is the number of samples ( $n$ )."*
2. *Connection between edge cardinality distributions and chunking techniques.* After chunking, the cardinality of column-level hyperedges is reduced and the edge cardinality distribution becomes more balanced, which provides additional explanation of the effectiveness of our chunking technique. In [Section 4.2](#) of the revised paper, we add a detailed discussion on how edge cardinality distributions can motivate the use of chunking techniques, as follows: *"Moreover, in this way, the cardinality of column-level hyperedges becomes closer to that of row-level hyperedges, ensuring a more balanced edge distribution across chunks compared to previous works. This mitigates skewness and ultimately enhances performance."*
3. *Robustness to skewed distributions.* In [Exp-11](#), we evaluate the robustness of UnIMP to skewed distributions. We use different chunk sizes, which lead to varying edge cardinality distributions with larger chunk sizes resulting in more skewed distributions. The results in [Figure 5\(b\)](#) show that imputation accuracy generally decreases with skewed distributions. Overall, the performance remains stable with modest fluctuations, demonstrating the robustness of our method.

**Comment 3.4 (D4)** The bound in Section 4.3 are quite verbose. Consider making them more readable by using a more intuitive set of parameters.

**Response 3.4:** Thanks for the insightful comments. In Section 4.3 of the revised paper, we enhance the bound analysis with the following improvements: 1) *Hierarchical complexity analysis*: We decompose the overall complex bound by combining the complexities of smaller modules, which increases the hierarchical structure of the bound and improves readability. 2) *More intuitive sets*: We rename the different modules using more readable and interpretable names. The revised part is as follows:

The LLM inference needs  $C_{\text{prompt}} = O(l_{\text{trans}} \cdot t_{\text{max}}^2 \cdot d_{\text{LLM}})$  for inference a prompt, where  $l_{\text{trans}}$  is the number of transformer layers,  $t_{\text{max}}$  is maximum number of tokens, and  $d_{\text{LLM}}$  is the dimension of embeddings in LLM. The feature encoding process requires encoding  $|\mathcal{V}| + |\mathcal{E}|$  prompts. Therefore, the total time complexity of feature encoding is  $O((|\mathcal{V}| + |\mathcal{E}|) \cdot C_{\text{prompt}})$ .

The message passing of  $l_{\text{max}}$  layers of Node-to-Hyperedge and Hyperedge-to-Node operations is with time complexity of  $C_{\text{msg}} = O(l_{\text{max}} \times (d_{\text{LLM}} \times |\mathcal{V}| + 2 \times d_{\text{LLM}} \times |\mathcal{E}|))$ . The attention module has a time complexity of  $C_{\text{fusion}} = (t_{\text{max}}^2 \times d_{\text{LLM}})$ , and the projection has  $C_{\text{proj}} = (d_{\text{LLM}})$ . Therefore, the total time complexity of training Epochs epochs is  $O(\text{Epochs} \times t_{\text{max}} \times (C_{\text{msg}} + C_{\text{fusion}} + C_{\text{proj}}))$ .

**Comment 3.5 (D5)** The claims in Section 3.1 are too general and could apply to any method combining global and local information. They do not seem specific to the techniques proposed in this paper. Moreover, the proofs do not explicitly use the definitions of global and local information, meaning the argument could apply to any two types of information.

**Response 3.5:** Thank you for the constructive comments. Through these general claims, we aim to highlight the motivation behind the design of specific techniques in UnIMP. We present the underlying logic in the following steps. 1) We observe that both global and local information are crucial for effective imputation; 2) We abstract this observation and formally establish a theoretical claim to support it; 3) The claim demonstrates that simultaneously considering both global and local information leads to more effective imputation; 4) The claims directly motivate the design of techniques in UnIMP. For example, the BiHMP module in UnIMP is specifically designed to aggregate global and local information for accurate imputation; and 5) Our experimental results further confirm the effectiveness of our techniques motivated by the claims. To better align these claims with the design of each technique and better present the proofs, we make the following revisions.

1. *Highlight the critical role of global and local information with specific datasets.* In Section 3.1, we further highlight the importance of global and local information with specific datasets as follows: "The global-local information suggests that the value of a cell is influenced by both neighboring cells (local context) and cells distributed across the entire table (global context) [54,55]. For example, as illustrated in Figure 2, the name of the president is determined not only by the associated nation but also by the sequential relationship of terms. Hence, we can infer that the missing value is the president succeeding President Trump."
2. *Add more comprehensive claims.* In Theorem 3.1, we revise the statement to demonstrate that simultaneously leveraging both global and local information yields better performance than using either global or local information alone. The revision

is as follows: "Consider two imputation models,  $\theta^{g+l}$  and  $\theta^l$  (resp.  $\theta^g$ ), where  $\theta^{g+l}$  simultaneously captures both global and local information, and  $\theta^l$  (resp.  $\theta^g$ ) captures only the local (resp. global) information. Assuming that interactions of global and local information are independent, then:

$$\Psi(\theta^{g+l}, X_{\text{miss}}) \leq \Psi(\theta^l, X_{\text{miss}}) \text{ or } \Psi(\theta^g, X_{\text{miss}})$$

indicating that a model capable of capturing both global and local information achieves a lower imputation error."

3. *Highlight our contribution in handling these features.* In Section 3.1, we further highlight our contribution in handling the global and local information in an integrated manner as follows: "Our work systematically combines both types of information and theoretically demonstrates the value of this integration."
4. *Add more details about the connection between claims and specific techniques in UnIMP.* In Section 3.1, we add more details about the connection between claims and specific technique designs in UnIMP as follows: "These claims motivate us to design a graph-based method with multi-hop message passing for effective information aggregation. The BiHMP in UnIMP integrates global-local information to enhance imputation, while hypergraph modeling and BiHMP capture high-order dependencies. Additionally, the cell-oriented, row-level, and column-level designs in UnIMP enable the model to capture inter-column heterogeneity and intra-column homogeneity."
5. *Add more details about the ablation study.* In Exp-7, we add more details about the ablation study to highlight the effectiveness of our techniques and claims. The revision is as follows: "In this experiment, we conduct an ablation study to assess the effectiveness of key components in UnIMP, including hypergraph-aggregated global-local information, the BiHMP module, the Xfusion model, and progressive masking. To evaluate hypergraph-aggregated information, we replace it with a zero tensor while keeping other components unchanged. For BiHMP, we substitute it with the Set Transformer from [8]."
6. *Add more details about the relationship of studied information in the proofs.* In the Proof of Theorem 3.1 in Section 4.4, we further highlight the specific relationship of the studied information, which is essential for the argument to hold. The revision is as follows: " $B^g$  is the global information,  $B^l$  is the local information,  $B^{g+l}$  is the fused global and local information."

**Comment 3.6 (D6)** In Section 4.4: The rationale behind Theorem 3.4 is presented at a high level, but a complete proof is not provided.

**Response 3.6:** Thank you for the valuable comments. In Proof of Theorem 3.4 in Section 4.4 of the revision, we enhance the proof with 1) explicitly linking the proof with the model components, and 2) introducing formal equations, thereby improving readability and clarity. The revised content is as follows:

"1): Global-local information. We prove that any two cells  $x_{ij}$  and  $x_{rs}$  can interact via a two-hop path in the hypergraph. Since  $v_{ij}$  and  $v_{is}$  both lie in the same row hyperedge, there is a direct connection:  $d(v_{ij}, v_{is}) = 1$ . Next, since  $v_{is}$  and  $v_{rs}$  share the same column hyperedge  $E_s^c$ , they are directly connected:  $d(v_{is}, v_{rs}) = 1$ . Thus, the overall distance is  $d(v_{ij}, v_{rs}) \leq d(v_{ij}, v_{is}) + d(v_{is}, v_{rs}) = 2$ . Moreover, the BiHMP has 2 layers, allowing it to aggregate information from the 2-hop neighborhood."

"2): High-order information. The message passing of Node-to-Hyperedge follows  $z_{ej}^{\text{temp}} = \frac{1}{|e_j|} \sum_{v_i \in e_j} \sigma(f_1^l(z_{vi}^l))$ . It aggregates

information over a set of cells at each layer, thereby capturing high-order relationships among cells."

"3): Inter-column heterogeneity and intra-column homogeneity. Each cell  $v_{ij}$  is represented as a distinct node and is connected to a specific column hyperedge  $e_j$ . According to the Node-to-Hyperedge in Equation 2, the network to learn distinct transformations for different columns, thus capturing heterogeneity between columns. Moreover, according to the Hyperedge-to-Node in Equation 3, cells in the same column are learned with the same column representation, thus capturing intra-column homogeneity."

**Comment 3.7 (D7)** Why is the method not compared with [8]? Also Table-GPT is not specifically designed for imputation.

**Response 3.7:** Thank you for your helpful comments. We would like to clarify that the method in [8] is designed for tabular representation learning rather than imputation. Moreover, Table-GPT is fine-tuned on tabular tasks, including imputation. In the revision, we add Exp-18, where the representations learned by HYTREL [8] are used to replace those generated by our UnIMP. As shown in Figure 6, HYTREL-based encoding performs less effectively, likely because its encoded features are not well aligned with the projection head. Moreover, in Exp-7, we replace UnIMP with the set transformer, a critical component in HYTREL[8]. The results in Table-6 show that UnIMP brings a 2.55% performance increase as measured by ROUGE-1 $F_1$ .

**Comment 3.8 (D8)** In the experiments, can we quantify whether the LLM or the GNN contributes more significantly to the imputation process?

**Response 3.8:** Thank you for your constructive feedback. Directly comparing the contributions of the LLM and GNN is challenging because, without the LLM, the missing data can not be inferred. However, as demonstrated in the Exp-7 of ablation study, the GNN's information aggregation contributes 5.96% to the ROUGE-1 $F_1$  score. In contrast, Exp-10 and Figure 5(a) show that replacing Llama 2 with a smaller language model (e.g., LiteLlama) leads to a performance drop of over 10%. These indicate that the LLM plays a more significant role in the imputation process.

**Comment 3.9 (D9)** In the ablation study, could you replace the LLM with more traditional BERT embeddings?

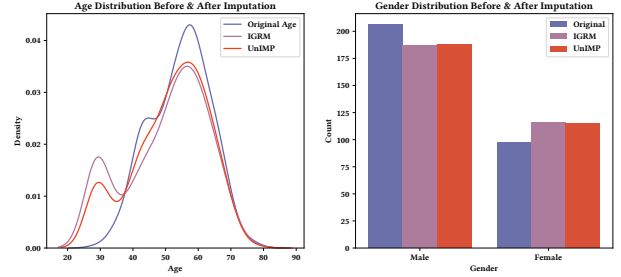
**Response 3.9:** Thank you for the useful comments. In the revision, we add Exp-18, where we replace the LLM with the traditional BERT embeddings. As shown in Figure 6, BERT-based encoding performs less effectively, likely because its encoded features excel at textual understanding but are not well-suited for generation tasks.

**Comment 3.10 (M1)** what is 81 in the query of Figure 2? shouldn't be 46?

**Response 3.10:** Thank you for pointing these out. We revise the typo in the revised paper and perform thorough proofreading.

**Comment 3.11 (M2)** As a suggestion, it would be interesting to evaluate the method on datasets containing sensitive attributes (e.g., gender, age) to assess whether the imputation process introduces any bias.

**Response 3.11:** Thank you for the constructive comments. In the revision, we add experiments using the Heart Disease dataset from the UCI [14] to assess whether the imputation process introduces any bias. We focus our analysis on the age and gender columns



**Figure R3-M2: Bias evaluation.**

**Table R3-M2: Bias evaluation**

Model	Age-KL	Age-WL	GenderBias
IGRM	0.1896	4.0495	0.1254
UnIMP	0.1166	2.9043	0.1188

to examine whether the imputation process introduces bias into sensitive attributes. To simulate missingness, we randomly mask 20% of all the cell values under the MCAR setting. The missing values are then imputed by two methods: UnIMP and IGRM, where IGRM is the previous state-of-the-art on small dataset.

To assess the bias in the age column, which contains numerical data, we employ both Kullback-Leibler (KL) divergence and Wasserstein distance to measure the difference between the original and imputed distributions. Specifically, we discretize the age values into  $k = 10$  bins to obtain the normalized histograms of the original distribution  $P = \{p_1, p_2, \dots, p_k\}$  and the imputed distribution  $Q = \{q_1, q_2, \dots, q_k\}$ . The KL divergence is computed as:  $KL(P \parallel Q) = \sum_{i=1}^k p_i \log \frac{p_i}{q_i}$ , where a lower value indicates closer alignment between the two distributions. Additionally, we compute the Wasserstein distance directly from the continuous age values:

$$\text{Wasserstein}(X, Y) = \inf_{\gamma \in \Gamma(X, Y)} \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|],$$

where  $X$  and  $Y$  denote the original and imputed age samples, and  $\Gamma(X, Y)$  is the set of all joint distributions with marginals  $X$  and  $Y$ . Together, these two metrics provide complementary insights into the distributional bias introduced during imputation.

To quantify the distributional shift in the imputed gender values, we compute the gender bias as the total variation distance between the original and imputed gender distributions. Given the original distribution  $P = \{p_1, p_2, \dots, p_k\}$  and the imputed distribution  $Q = \{q_1, q_2, \dots, q_k\}$  over  $k$  categories, the gender bias is defined as:

$$\text{GenderBias}(P, Q) = \sum_{i=1}^k |p_i - q_i|.$$

A lower value of GenderBias indicates that the imputed gender values better preserve the original distribution, suggesting reduced demographic distortion.

The results of IGRM and UnIMP are summarized in Figure R3-M2 and Table R3-M2. Overall, UnIMP consistently demonstrates a closer alignment with the original data distribution, indicating better performance in preserving demographic fairness during the imputation process. (Due to space limitations, detailed results are provided in the appendix of our online full version.)<sup>2</sup>

<sup>2</sup>The full version is in [https://github.com/guaiyoui/UnIMP/blob/master/full\\_version/UnIMP\\_Full\\_Version.pdf](https://github.com/guaiyoui/UnIMP/blob/master/full_version/UnIMP_Full_Version.pdf)



# On LLM-Enhanced Mixed-Type Data Imputation with High-Order Message Passing

Jianwei Wang<sup>1</sup>, Kai Wang<sup>2</sup>, Ying Zhang<sup>3</sup>, Wenjie Zhang<sup>1</sup>, Xiwei Xu<sup>4</sup>, Xuemin Lin<sup>2</sup>

<sup>1</sup>The University of New South Wales, Sydney, Australia

<sup>2</sup>Antai College of Economics & Management, Shanghai Jiao Tong University, Shanghai, China

<sup>3</sup>Zhejiang Gongshang University, Hangzhou, China

<sup>4</sup>Data61, CSIRO, Australia

{jianwei.wang1, wenjie.zhang}@unsw.edu.au, {w.kai, xuemin.lin}@sjtu.edu.cn, ying.zhang@zjgsu.edu.cn, xiwei.xu@data61.csiro.au

## ABSTRACT

Missing data imputation, which aims to impute the missing values in the raw datasets to achieve the completeness of datasets, is crucial for modern data-driven models like large language models (LLMs) and has attracted increasing interest over the past decades. Despite its importance, existing solutions for missing data imputation either 1) only support numerical and categorical data or 2) show an unsatisfactory performance due to their design prioritizing text data and overlooking intrinsic characteristics of tabular data.

In this paper, we propose UniIMP, a **Unified IMP**utation framework that leverages LLM and high-order message passing to enhance the imputation of mixed-type data including numerical, categorical, and text data. Specifically, we first introduce a cell-oriented hypergraph to model the table. We then propose BiHMP, an efficient **Bidirectional High-order Message-Passing** network to aggregate global-local and high-order information on the constructed hypergraph while capturing the inter-column heterogeneity and intra-column homogeneity. To effectively and efficiently align the capacity of the LLM with the information aggregated by BiHMP, we introduce Xfusion, which, together with BiHMP, acts as adapters for the LLM. We follow a pre-training and fine-tuning pipeline to train UniIMP, integrating two optimizations: chunking technique, which divides tables into smaller chunks to enhance efficiency; and progressive masking technique, which gradually adapts the model to learn more complex data patterns. Both theoretical proofs and empirical experiments on 10 real-world datasets highlight the superiority of UniIMP over existing techniques.

## PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/guaiyoui/UniIMP>.

## 1 INTRODUCTION

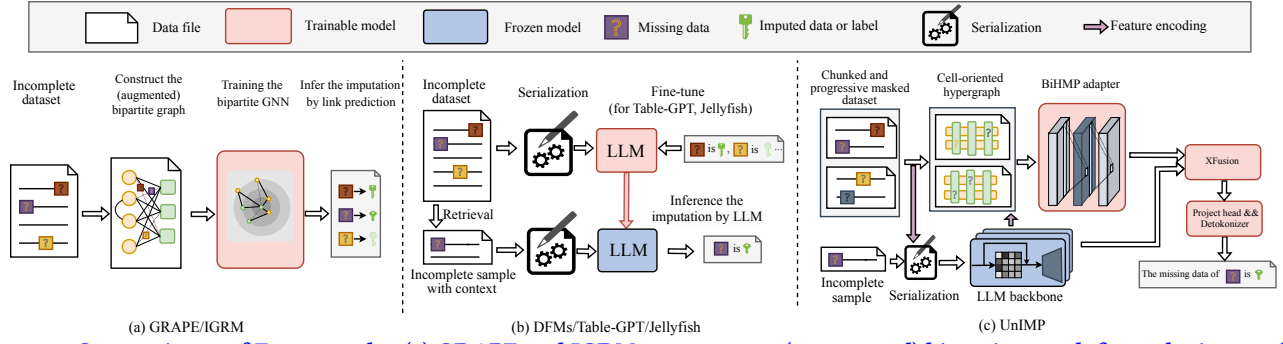
Data quality has gained increasing attention due to its pivotal role in developing and training data-driven models, particularly those built on artificial intelligence (AI) technology. For instance, the performance of large language models (LLMs) can be significantly enhanced by high-quality training datasets, without requiring substantial changes to the model architecture [15]. The missing data problem, as one of the most critical issues of data quality, is ubiquitous in real-world raw datasets due to various factors such as poor data collection practices and device malfunctions [37]. To handle these issues, missing data imputation (a.k.a. missing value imputation) [54, 66] has been introduced, which aims to fill in missing values using information from observed data samples and features.

**Table 1: Comparison of representative imputation methods. “—” denotes not applicable. Abbr: GAN (generative adversarial network), GNN (graph neural network), OT (optimal transport), AE (autoencoder), LLMs (large language models).**

Methods	Handle Data Types			Generalization	Backbone
	Num.	Cate.	Text		
Mean/Mode [24]	★★	★★	—	★★★★	Statistics
KNNI [70]	★★★★	★★★★	—	★★★★	Similarity
MICE [57]	★★★★	★★★★	—	★	Regression
GAIN [65]/VGAIN [37]	★★★★	★★★★	—	★	GAN
GRAPE [66]/IGRM [72]	★★★★	★★★★	—	★	GNNs
TDM [71]	★★★★	★★★★	—	★	OT
ReMasker [13]	★★★★	★★★★	—	★	AE
NOMI [54]	★★★★	★★★★	—	★	Similarity
DFMs [42]	★	★★	★★★★	★★★★	LLMs
Table-GPT [30]/Jellyfish [68]	★	★★	★★★★	★★★★	LLMs
UniIMP (ours)	★★★★	★★★★	★★★★	★★★★	GNNs+LLMs

When applying imputation techniques to real-world datasets, it is desirable that these techniques can achieve high accuracy and support a variety of data types. High imputation accuracy is crucial, as existing studies have demonstrated that a higher accuracy would generally lead to a higher downstream task performance [37, 54, 65, 71]. Supporting multiple data types is also important, as real-world datasets in fields like e-commerce [62] and healthcare [25] often contain a mix of numerical, categorical and text data [46, 63]. Moreover, with advances in the Internet of Things (IoT [33]) and big data, the proportion of mixed-type data continues to grow [1, 56]. **Existing works.** Given the importance and practical value of missing data imputation, a set of techniques has been developed [13, 54, 66]. Some representative and recent methods are summarized in Table 1, and more details can be found in the surveys [32, 37]. The early-stage imputation methods were primarily rule-based, including statistics-based approaches such as Mean and Mode [24], and similarity-based methods such as  $K$  Nearest Neighbors Imputation (KNNI) [70]. These methods have good generalization capacity, as the rules can be applied to various datasets. However, their imputation accuracy is often limited since these fixed rules often struggle to capture the underlying patterns in diverse datasets effectively.

To better model relationships among samples and features within datasets for imputation, various learning-based methods are introduced, such as regression algorithm in MICE [57], generative adversarial net (GAN) [60] in GAIN [65] and VGAIN [37], optimal transport (OT) [59] in TDM [71], auto-encode (AE) in ReMasker [13], and graph neural networks (GNNs) in GRAPE [66] and IGRM [72]. **The GNN-based method, which stands out for its feature aggregation capacity, is illustrated in Figure 1(a).** While these methods achieve outstanding performance in handling numerical and categorical



**Figure 1: Comparisons of Frameworks:** (a) GRAPE and IGRM construct an (augmented) bipartite graph from the incomplete dataset and train a bipartite GNN, modeling imputation as link prediction. (b) DFMs, Table-GPT and Jellyfish select incomplete samples with context and infer imputations using LLMs. Table-GPT and Jellyfish fine-tune LLMs for tabular tasks, while DFMs use them directly. (c) Our model builds a cell-oriented hypergraph and encodes features with an LLM. The incomplete sample is serialized, and information is aggregated via both the LLM and BiHMP before imputation.

data, they struggle to process text data effectively. Furthermore, these approaches typically train a separate model for each dataset or even for an individual feature, hindering their generalizability.

Recently, LLM-based methods are emerging in the field, leveraging knowledge encoded in LLM to predict missing data. LLM-based imputation can be broadly categorized into two primary categories: in-context-learning-based methods (e.g., [5, 42, 43, 67]) and fine-tuning-based methods (e.g., [30, 68]). The in-context-learning-based methods, with Direct Foundation Models (DFMs) [42] as a prominent representative, directly infer missing data by pre-trained LLM with carefully designed prompts. The fine-tuning-based methods, with Table-GPT and Jellyfish as prominent representatives, fine-tune LLM to adapt the model for tabular tasks including imputation. The LLM-based method is illustrated in Figure 1(b).

**Motivations.** By serializing tabular data into text and generating the next token auto-regressively from prompts, existing LLM-based methods can handle mixed-type data with good generalization capacity. However, they typically exhibit limited performance, especially with numerical data. They directly recast the problem of imputation as text generation while the intrinsic characteristics of tabular data remain insufficiently explored: 1) *Global-local information*. The value of a cell is influenced by both global patterns across the entire table and localized details of the cell. These LLM-based methods focus on individual samples and several examples as prompt, overlooking global information; 2) *High-order dependencies*. The relationships in the table are not necessarily pairwise but may involve three or more entities simultaneously. However, LLM-based methods, built on the attention mechanism [51], mainly focus on pairwise relationships. 3) *Inter-column heterogeneity and intra-column homogeneity*. Features across columns can be highly diverse, and cells within the same column generally exhibit consistent semantics. LLM-based methods, designed for sequential inputs, struggle to align the specific contextual relationships between columns and within columns.

As summarized in Table 1, existing imputation methods generally fall into two categories: 1) LLM-based methods exhibit an unsatisfactory performance, particularly for numerical data, and 2) other learning-based methods and rule-based methods only support numerical and categorical data. To better support real-world applications, an imputation method capable of handling mixed-type data

while maintaining high accuracy is in demand. A straightforward approach would be to impute each data type using the outperforming method for that type, such as applying MICE [45] or NOMI [54] for numerical and categorical data, and Jellyfish [68] for text data. However, imputing each data type separately may ignore the intricate dependencies between them, leading to inconsistencies and reduced overall accuracy.

**Challenges.** To design an accurate imputation method for mixed-type data, two challenges exist below:

*Challenge I: How to aggregate global-local and high-order information for imputation while capturing inter-column and intra-column patterns?* The existence of missing data and large-scale tables complicate the aggregation process. A direct solution is to input the entire table into the LLM. However, computing complexity grows quadratically as the size of input increases. Another promising alternative is to apply the global to local framework [55] to learn the multi-level features for imputation. However, it relies on similarity search on samples with missing features, leading to inaccurate results. Methods such as retrieval-augmented generation [29] show promise for aggregating information. However, they primarily focus on local information while overlooking global information. Therefore, it is challenging to aggregate these key information.

*Challenge II: How to effectively integrate the aggregation module and train the model to support mixed-type imputation?* The presence of mixed-type data complicates the integration process. To handle mixed-type data, a direct approach would be to sequentially apply LLM, aggregation module and then another LLM. This pipeline first uses the LLM to embed the input into a shared latent space, followed by information aggregation and finally generates the output by another LLM. However, treating LLM and the aggregation module as separate components may limit the alignment of the capabilities encoded in LLM with the aggregated information. Moreover, training this sequential pipeline on tables of varying sizes may fail to adequately utilize computational resources, resulting in inefficiencies. Hence, it is challenging to integrate the aggregation module and train the model for mixed-type imputation.

**Our solutions.** Guided by the above challenges, we propose UnIMP, a Unified IMPutation framework as depicted in Figure 1(c) that employs LLM and the high-order message passing for accurate mixed-type data imputation.



To address *Challenge I*, we design a cell-oriented hypergraph in UnIMP to model tabular datasets and introduce BiHMP, an efficient and effective Bidirectional High-order Message-Passing network, to aggregate information on the constructed hypergraph. By leveraging high-order message passing on the hypergraph, BiHMP learns to aggregate global-local and high-order information while capturing intra- and inter-column relationships. Most existing hypergraphs for tables are value-oriented, modeling distinct values as nodes [12]. However, this approach struggles with modeling unseen values in the context of missing data and handling numerical and textual data where co-occurrence patterns are not easily defined, limiting its applicability in mixed-type imputation. Therefore, we adopt the cell-oriented hypergraph for imputation following [8], where each cell in the dataset is modeled as a node and nodes within the same row or column are connected by hyperedges. This structure enables the flexible integration of missing and mixed-type data. Furthermore, BiHMP is designed to efficiently yet effectively aggregate information. It comprises iteratively applied linear hyperedge-to-node and node-to-hyperedge layers to refine and propagate features.

To address *Challenge II*, UnIMP is designed with the LLM as the backbone, while BiHMP and Xfusion (a fusion module based on attention mechanism [51]) serve as adapters [22, 69] for the LLM. Specifically, prompts are together propagated by both the LLM and the BiHMP on the embedded hypergraph. The information aggregated by BiHMP, along with the LLM-generated data, is sent to the Xfusion module to enhance the alignment and integration of both sets of information. The fused data is subsequently sent to the projection head of LLM to generate output. Moreover, the adapter strategy trains only the BiHMP, Xfusion and projection head, freezing most LLM parameters to enhance efficiency and better preserve pre-trained knowledge in LLM while incorporating BiHMP-aggregated information.

To more effectively and efficiently train UnIMP for mixed-type imputation, we follow the pre-train and fine-tune strategy [2, 31, 52] and integrate two optimizations, i.e., the chunking and the progressive masking motivated by [13, 64, 66]. Specifically, the chunking technique splits all tables into smaller, uniform chunks to handle large tables and facilitate efficient batch processing. This approach allows for more effective use of computational resources, such as GPU memory and parallel computation capabilities. The progressive masking technique, on the other hand, gradually increases the complexity of the imputation task by progressively masking more data during training. This technique enables the model to learn complex patterns progressively, improving its understanding of data relationships and generalization capacity.

**Theoretical and empirical studies.** We theoretically and empirically demonstrate the superiority of UnIMP. Theoretically, we prove that a model capable of capturing global-local information, high-order dependencies, and inter- and intra-column patterns leads to improved imputation accuracy. Furthermore, we show that UnIMP effectively incorporates these critical properties. Empirical evaluations on 10 real-world datasets with varying table sizes and data types also validate the excellent performance of UnIMP in terms of both accuracy and efficiency. For numerical and categorical data, UnIMP reduces imputation RMSE by an average of 12.11% to 45.22% compared to previous state-of-the-art methods and other

**Table 2: Symbols and Descriptions**

Notation	Description
$X, \tilde{X}, \bar{X}$	raw dataset, imputed dataset and ground-truth
$n, d$	the row number and column number
$x_i, x_{ij}$	$i$ -th sample in $X$ and its $j$ -th column data
$M \in \{0, 1\}^{n \times d}$	the mask matrix indicating incompleteness
$m_i, m_{ij}$	$i$ -th sample in $M$ and its $j$ -th column data
$HG(\mathcal{V}, \mathcal{E})$	hypergraph with nodes $\mathcal{V}$ and hyperedges $\mathcal{E}$
$\theta$	model parameters.
$z_{v_i}, z_{e_i}$	the hidden embedding of node $v_i$ & hyperedge $e_i$
$\sigma$	non-linear function (ReLU)
$f, g$	functions or layers in the neural networks

baselines. In the case of imputing text data, UnIMP improves imputation accuracy by 30.53, 28.04 and 23.47 percentage points (pp) over DFM [42], Table-GPT [30] and Jellyfish [68] respectively, as measured by ROUGE-1<sub>F1</sub>. Furthermore, UnIMP can handle large datasets and exhibits robust generalization ability.

**Contributions.** The main contributions are as follows.

- We propose a unified framework UnIMP to accurately impute mixed-type data, including numerical, categorical and text data.
- We propose a cell-oriented hypergraph to model tabular data and design BiHMP, an efficient and effective bidirectional high-order message-passing network, to aggregate information while capturing the key intrinsic characteristics of tabular data.
- We introduce Xfusion, along with BiHMP, as adapters for LLM to process diverse data and enable accurate mix-type imputation. Additionally, we integrate chunking and progressive masking techniques to train UnIMP, enhancing performance.
- We theoretically prove the critical role of global-local information, high-order relationships, and inter- and intra-column patterns to the accuracy of imputation.
- Extensive empirical experiments on 10 real-world datasets validate the outstanding performance of UnIMP.

## 2 PRELIMINARIES

### 2.1 Problem Statement

Following previous works [37, 54], the task of missing data imputation is defined over tabular datasets  $X$  [30, 37, 54, 71], where each  $X \in \mathcal{X}$  consists of  $n$  data samples (rows) and  $d$  features (columns). We denote the  $j$ -th feature of the  $i$ -th sample as  $x_{ij}$ , which can be of numerical, categorical or text type. A mask matrix  $M \in R^{n \times d}$  is used to indicate missing components in  $X$ .  $m_{ij}$  equals 1 if  $x_{ij}$  is observed and 0 if  $x_{ij}$  is missing, i.e.,  $X = X_{obs} \cup X_{miss}$  where  $X_{obs} = \{x_{ij} | x_{ij} \in X, m_{ij} = 1\}$  and  $X_{miss} = \{x_{ij} | x_{ij} \in X, m_{ij} = 0\}$ . The imputed data and the ground-truth data are denoted as  $\tilde{X}$  and  $\bar{X} = X_{obs} \cup \bar{X}_{miss}$ , respectively. Note that the indexing in this paper starts from zero, which is a common convention in many programming languages.

The hypergraph is a generalization of the graph where each hyperedge may contain an arbitrary number of nodes. We denote a hypergraph by  $HG(\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  is a set of nodes, and  $\mathcal{E}$  is the set of hyperedges.  $\mathcal{E} \subseteq P^*(\mathcal{V}) \setminus \{\emptyset\}$  where  $P^*(\mathcal{V})$  is the power set on the nodes [8]. We use  $x_v$  to denote the cell w.r.t. the node  $v$ . With the above notations, we define the task of mixed-type missing data imputation as follows.

**Definition 2.1:**(Mixed-type Missing Data Imputation [54, 66]). *Mixed-type missing data imputation* aims to impute the unobserved elements in the raw data, i.e.,  $X_{miss}$ , and make the imputed matrix  $\tilde{X}$  as close to the real complete dataset  $\bar{X}$  as possible. The raw data matrix  $X$  may contain numerical, categorical and text data.

## 2.2 Large Language Models

LLMs, such as GPT [6] and Llama [15] which contain billions of parameters, have demonstrated remarkable emergent behaviors and impressive zero-shot generalization capabilities across diverse tasks. The execution process of LLMs typically comprises four key components: 1) a tokenizer that segments input text into discrete tokens, 2) an LLM backbone that is used for feature propagation, 3) a projection head that predicts the next token, and 4) a decoding process that translates token IDs back into human-readable text.

These LLMs are often auto-regressive, trained on large text corpora with the objective of maximizing the log-likelihood of the next word given the previous words.

$$\theta_{LLM} = \arg \max_{\theta} \sum_i \log P(t_i | t_{i-k}, \dots, t_{i-1}; \theta) \quad (1)$$

where  $t_i$  stands the token and  $k$  is the context window size. There are two typical ways to adapt the model to new tasks and boost performance: supervised-fine-tuning (SFT) and in-context-learning. For SFT, given a training query  $T = \{t_0, t_1, \dots, t_s\}$  and training target  $y$ , the objective is to maximize the following log-likelihood:  $\sum_{(T, y)} \log P(y | t_0, t_1, \dots, t_s)$ . The in-context-learning-based methods handle new tasks during inference by constructing a prompt that includes examples (i.e., context). Given a set of example queries  $\{T^0, T^1, \dots, T^w\}$ , targets  $\{y^0, y^1, \dots, y^w\}$ , and a new query text  $T^{w+1}$ , a prompt  $(T^0, y^0, \dots, T^w, y^w, T^{w+1})$  is constructed and sent to the LLM to infer the results.

## 3 THE OVERALL FRAMEWORK

In this section, we first introduce three key intrinsic characteristics of tabular data and prove its critical role for imputation. Then, we introduce the details of UnIMP.

### 3.1 Intrinsic Characteristics of Tabular Data

**Global-local information.** The global-local information suggests that the value of a cell is influenced by both neighboring cells (local context) and cells distributed across the entire table (global context) [54, 55]. For example, as illustrated in Figure 2, the name of the president is determined not only by the associated nation but also by the sequential relationship of terms. Hence, we can infer that the missing value is the president succeeding President Trump.

**High-order relationship.** The high-order relationship suggests that the value of the cell is influenced simultaneously by multiple columns as a whole. For example, as shown in Figure 2, the name of a president can be determined simultaneously by the nation and the term as a whole. However, neither the nation nor the term alone is sufficient to fully determine the value.

**Intra-column heterogeneity and intra-column homogeneity.** Features across columns can be highly diverse, and cells within the same column typically exhibit consistent semantics. For example, as shown in Figure 2, the name is text data and the nation is categorical data. Furthermore, the name remains consistent across rows.

Next, we formally define the imputation error.

**Definition 3.1:**(Imputation Error). Given a model  $\theta_{X_{obs}}$  trained on the observed data  $X_{obs}$ , the imputation error  $\Psi(\theta_{X_{obs}}, X_{miss})$  captures the expected performance of  $\theta_{X_{obs}}$  on the test unobserved data  $X_{miss}$ , i.e.,  $\Psi(\theta_{X_{obs}}, X_{miss}) = \mathbb{E}[\|\theta_{X_{obs}}(X_{miss}) - \bar{X}_{miss}\|]$ . When the context is clear, we denote  $\theta_{X_{obs}}$  as  $\theta$  to improve readability.

We then have the following three theorems highlighting the importance of global-local information, high-order information and column patterns, respectively. The proofs are in Section 4.

**Theorem 3.1:** Consider two imputation models,  $\theta^{g+l}$  and  $\theta^l$  (resp.  $\theta^g$ ), where  $\theta^{g+l}$  simultaneously captures both global and local information, and  $\theta^l$  (resp.  $\theta^g$ ) captures only the local (resp. global) information. Assuming that interactions of global and local information are independent, then:

$$\Psi(\theta^{g+l}, X_{miss}) \leq \Psi(\theta^l, X_{miss}) \text{ or } \Psi(\theta^g, X_{miss})$$

indicating that a model capable of capturing both global and local information achieves a lower imputation error.

Our work systematically combines both types of information and theoretically demonstrates the value of this integration.

**Theorem 3.2:** Consider two imputation models,  $\theta^{[0:r]}$  and  $\theta^{[0:s]}$ , where  $\theta^{[0:r]}$  captures interactions up to order  $r$  in the latent space, and  $\theta^{[0:s]}$  captures interactions up to order  $s$ , with  $r > s$ . We have

$$\Psi(\theta^{[0:r]}, X_{miss}) \leq \Psi(\theta^{[0:s]}, X_{miss}),$$

indicating that the model capable of capturing higher-order interactions exhibits a lower imputation error.

**Theorem 3.3:** Consider two imputation models,  $\theta^{cP}$  and  $\theta$ , where  $\theta^{cP}$  captures the column patterns including intra-column heterogeneity and intra-column homogeneity, while  $\theta$  does not. Then, we have:

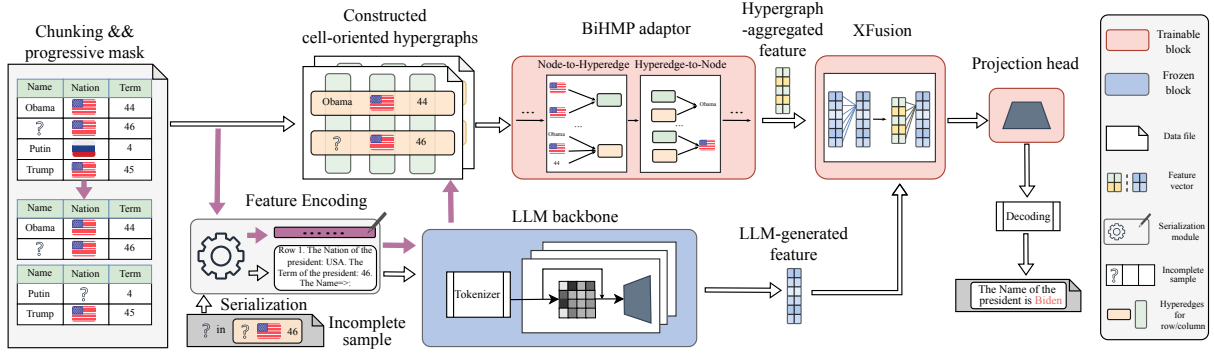
$$\Psi(\theta^{cP}, X_{miss}) \leq \Psi(\theta, X_{miss}),$$

indicating that model  $\theta^{cP}$  capturing the column patterns achieves a lower imputation error.

These claims motivate us to design a graph-based method with multi-hop message passing for effective information aggregation. The BiIMP in UnIMP integrates global-local information to enhance imputation, while hypergraph modeling and BiHMP capture high-order dependencies. Additionally, the cell-oriented, row-level, and column-level designs in UnIMP enable the model to capture inter-column heterogeneity and intra-column homogeneity.

### 3.2 Overview of UnIMP

Guided by the above characteristics of tabular data, we propose UnIMP for accurate mixed-type data imputation. The overall architecture is illustrated in Figure 2. Given the incomplete raw dataset, UnIMP first utilizes the chunking technique and progressive masking to process data. Then, cell-oriented hypergraphs are constructed. UnIMP utilizes the tokenizer and LLM backbone for feature initialization and propagation, embedding both raw data and query prompts. These obtained features are processed through the BiHMP module, which iteratively runs the node-to-hyperedge layer and the hyperedge-to-node layer to aggregate both local and global information. The aggregated feature is then fused with the embedded prompts using an XFusion module. Finally, a projection head maps



**Figure 2: Framework overview of UnIMP.** Given an incomplete dataset, UnIMP chunks the data and applies progressive masking. A hypergraph is constructed for each chunk, with node and hyperedge features extracted by an LLM from serialized data. The incomplete sample is serialized, processed by the LLM, and aggregates information on the hypergraph via the BiHMP adaptor, which iteratively applies Node-to-Hyperedge and Hyperedge-to-Node layers. Features from the LLM and hypergraph are fused through the Xfusion module, followed by a projection head for imputation.

the predicted data back to token space, and a decoding process is employed to translate the tokens into human-readable results.

### 3.3 Cell-Oriented Hypergraph Modeling

**Motivation.** To capture the structural properties of tabular data, existing methods often employ graph-based approaches, such as bipartite graphs [66, 72] and similarity graphs [7]. While effective, these methods fall short in comprehensively representing the complex interactions and higher-order relationships inherent in tabular data. To better capture the intrinsic characteristics of tabular data, we adopt a hypergraph-based approach for modeling. While some hypergraph-based methods, such as the cell-oriented hypergraph in [8] and the value-oriented method in [12], have been introduced for general tabular representation learning, they have not yet been employed for imputation. The value-oriented approach models each distinct value as a node, with nodes in the same row sharing a hyperedge. However, this method struggles to deal with mixed-type data and effectively handles missingness. Moreover, missing data imputation focuses on the granularity of individual cells, as missingness occurs at the cell level. Therefore, we adopt the cell-oriented hypergraph following [8]. It is important to note that the work in [8] focuses on tabular representation learning, our work is specifically designed for missing data imputation. Additionally, their method treats hypergraph neural network and language models as separate components, whereas we introduce BiHMP and Xfusion, which serve as the adapters for LLM, to better aggregate information and combine both advantages.

Given a tabular dataset  $X$  with  $n$  samples, each containing  $d$  features, we construct a hypergraph  $HG(\mathcal{V}, \mathcal{E})$  as follows:

- For each cell  $x_{ij} \in X$ , we create a corresponding node  $v_{idx} \in \mathcal{V}$ , where  $idx = i * d + j$ .
- For nodes in the same column (i.e., nodes corresponding to  $\{x_{0j}, x_{1j}, \dots\}$ ), we construct a hyperedge  $e_j \in \mathcal{E}$ .
- Similarly, nodes in the same row (i.e., nodes corresponding to  $\{x_{i0}, x_{i1}, \dots\}$ ) form a hyperedge  $e_{i+d} \in \mathcal{E}$ .

Each node has a degree of 2, connecting to both a row-level and a column-level hyperedge. The edge cardinality of a row-level hyperedge is the number of columns ( $d$ ), while that of a column-level hyperedge is the number of samples ( $n$ ). This hypergraph-based

modeling effectively captures tabular characteristics. For instance, in Figure 2, a hyperedge captures the high-order relationship among a president’s name, nation, and term in office. Moreover, each cell is represented as a distinct node, capturing intra-column heterogeneity, while nodes within the same column are linked by a hyperedge, capturing inter-column homogeneity.

### 3.4 Feature Encoding

As the neural network models need vectorized inputs, we introduce a vectorization technique to project data from the text space into the vector space. We use the LLM backbone to encode the text data. **Serialization.** In order to obtain the embedding, we need to feed the proper prompt to the LLM backbone. Recent research has explored various approaches to serializing tabular data, like the sentences-transformation and the markdown-transformation [16]. Here, we follow the widely-used serialization method of attribute-value pair [16], which is shown as follows:

$$\text{Serial}(x_{ij}) = \text{Row } i, \text{col\_name}[d \neq i]: x_{dj}, \text{col\_name}[i] \Rightarrow \{x_{ij}\} \text{ EOS}$$

Here, EOS (i.e., end of the sentence) is a special token to denote the end of the prompt. We use the following serialization prompt for hyperedge data.

$$\text{Serial}(i\text{-th row}) = \text{This is row: } i \text{ EOS;}$$

$$\text{Serial}(i\text{-th col}) = \text{This is col: col\_name}[i] \text{ EOS;}$$

**Example 3.1:** For the node ‘Obama’ shown in Figure 2, we serialize as: *This is row 0, Nation of the President: USA, Term of the President: 44, Name => Obama EOS*. For the hyperedge corresponding to the first column, we serialize as: *This is col: Name EOS*.

**Tokenization.** Then, the text is sent into the tokenizer to split the prompt into tokens. There are many commonly used tokenizers like BPE [4] and SentencePiece [28]. In our experiments, we use the SentencePiece tokenizer associated with Llama2.

$$\{t_0, t_1, \dots, t_s\} = \text{tokenizer}(\text{prompt-text})$$

**Propagation of LLM backbone.** The tokens are processed through the LLM backbone for feature initialization and propagation. Initially, the LLM backbone performs a lookup in the embedding



matrix, a predefined matrix with dimensions  $|\text{Vocab}| \times d$ , where  $|\text{Vocab}|$  denotes the vocabulary size and  $d$  represents the hidden dimension. Each token ID corresponds to a specific row in this embedding matrix, serving as an index to retrieve the corresponding embedding vector. These embedding vectors are then fed into the transformer layer [51] for feature propagation.

$$z_p : \{z_{t_0}, z_{t_1}, \dots, z_{t_s}\} = \text{LLM-backbone}(t_0, t_1, \dots, t_s)$$

Each token  $t_i$  is encoded as a feature representation  $z_{t_i}$ . For the prompt text corresponding to node  $v_i$  (or hyperedge  $e_i$ ), we use the representation of the last token as the initial feature  $z_{v_i}^0$  (or  $z_{e_i}^0$ , respectively). **For numerical data, following previous works [13, 37], we directly encode the input as its value to avoid unnecessary computational overhead.** For categorical features, we employ label encoding, which assigns unique integers to each category, aligning with recent works [21, 37, 54, 55, 71]. Encoded numerical and categorical data are padded into the same latent dimension.

### 3.5 Bidirectional High-order Message Passing

**Motivation.** In tabular data, interactions often involve multiple entities simultaneously. For instance, multiple cells within the same row and column exhibit high-order interactions. To capture these complexities, Hypergraph Neural Networks (HGNNs) (e.g., [18]) can be employed for message passing, but they typically come with high computational costs. While a hypergraph-structure-aware transformer was proposed in [8], it not only faces efficiency challenges but also tends to lose fine-grained local information due to its maximal invariant properties. A simplified and fast HGNN [49] can accelerate training, but it only handles node information, overlooking hyperedges. Therefore, we propose BiHMP, a novel Bidirectional High-order Message Passing network that operates bidirectionally between nodes and hyperedges in the constructed hypergraph. BiHMP comprises two key components: 1) Node-to-Hyperedge layer, which propagates information from each node to its associated hyperedges through a linear layer. 2) Hyperedge-to-Node layer, which propagates information from hyperedges back to their constituent nodes through a linear layer. Through the iterative application of these two layers, BiHMP captures high-order and multi-hop relationships within the data, enhancing its ability to model complex interactions.

**Node-to-Hyperedge.** For each hyperedge  $e_j \in \mathcal{E}$  and node  $v_i \in e_j$ , with their respective embeddings  $z_{e_j}^l$  and  $z_{v_i}^l$  in  $l$ -th Node-to-Hyperedge layer, we use the following equation to learn to update the representation of hyperedge:

$$z_{e_j}^{temp} = \frac{1}{|e_j|} \sum_{v_i \in e_j} \sigma \left( f_1^l(z_{v_i}^l) \right) \quad (2)$$

$$z_{e_j}^{l+1} = \sigma \left( f_2^l \left( \text{CONCAT} \left( z_{e_j}^l, z_{e_j}^{temp} \right) \right) \right)$$

where  $f_1^l(\cdot)$  and  $f_2^l(\cdot)$  represents learnable linear transformation functions.  $\sigma(\cdot)$  is a non-linear activation function (ReLU in UnIMP).  $\text{CONCAT}(\cdot)$  is the concatenation function that stacks multiple inputs into a single, longer vector. For each node in the hyperedge, we apply a learnable transformation followed by a non-linear activation  $\phi(\cdot)$ . We then compute the mean of these transformed node

features to obtain  $z_{e_j}^{temp}$ , which captures the average information from all nodes in the hyperedge. This temporary representation is concatenated with the representation of the hyperedge from the previous layer,  $z_{e_j}^l$ . Finally, we apply another transformation layer and non-linear activation layer to this concatenated vector to obtain the updated hyperedge representation  $z_{e_j}^{l+1}$ .

**Hyperedge-to-Node.** The Hyperedge-to-Node layer updates node representations by aggregating information from their incident hyperedges. For each node  $v_i \in \mathcal{V}$  with its two incident hyperedges i.e., the row-level hyperedge  $e_{v_i}^r$  and column-level hyperedge  $e_{v_i}^c$ , we update the node representation by the following equation in the  $l$ -th Hyperedge-to-Node layer:

$$z_{v_i}^{l+1} = \sigma \left( f_3^l \left( \text{CONCAT} \left[ z_{v_i}^l, z_{e_{v_i}^c}^l, z_{e_{v_i}^r}^l \right] \right) \right) \quad (3)$$

where  $f_3^l(\cdot)$  is a learnable linear function. The Hyperedge-to-Node layer first concatenates the embedding of the node  $z_{v_i}^l$  with the embeddings of all hyperedges that contain the node, i.e.,  $z_{e_{v_i}^c}^l$  and  $z_{e_{v_i}^r}^l$ . The concatenated representation is then processed through a learnable linear transformation function  $f_3^l(\cdot)$ . Finally, a non-linear activation function  $\sigma(\cdot)$  is applied to obtain the updated node representation in  $(l+1)$ -th layer  $z_{v_i}^{l+1}$ . Note that, the main component in our BiHMP is the linear transformation which is efficient yet effective for information propagation.

**When predicting the name of the president, as shown in Figure 2, the Node-to-Hyperedge layer aggregates contextual information within hyperedge, such as name formatting. The Hyperedge-to-Node layer then can utilize this information to aid prediction.**

### 3.6 XFusion Block and Projection Head

**XFusion block.** After encoding the prompt and aggregating local and global information by BiHMP, it is crucial to effectively integrate these features for accurate imputation. We designed XFusion to synthesize the rich information derived from the prompt feature (denoted as  $z_p = \{z_{t_0}, z_{t_1}, \dots, z_{t_s}\}$ ) with the features extracted from the hypergraph representation (denoted as  $z_g$  where  $z_g \in \mathcal{Z}_g$  is a concatenation of  $z_{e_{v_i}}^{lmax}$  and  $z_{v_i}^{lmax}$  for imputing node  $v_i$ ). XFusion is based on the attention mechanism [51], a critical component of LLMs. The attention is computed as follows:

$$\text{Attn}(Q, K, V; w) = w \cdot \text{Softmax} \left( \frac{QK^T}{\sqrt{d_K}} \right) V \quad (4)$$

where  $Q, K$  and  $V$  are the input feature matrices,  $w$  is a learnable weight matrix, and  $d_K$  is the dimensionality of the vectors in  $K$ .

Building on this attention mechanism, our XFusion module leverages both token embeddings and hypergraph-derived embeddings to construct a rich, context-aware representation. The forward process of XFusion is as follows:

$$z_p^{temp} = \text{Attn}(z_p W_Q, z_p W_K, z_p W_V; w_1) \quad (5)$$

$$z_{output} = \text{Attn}(z_p^{temp} W_Q, z_g W_K, z_g W_V; w_2)$$

Here,  $W_Q, W_K$  and  $W_V$  are learnable weight matrices. We first calculate the self-attention of the prompt embedding. We then use the obtained prompt feature  $h_p^{temp}$  as the query and the graph feature

---

**Algorithm 1:** Pre-Training Pipeline

---

**Input:** The tabular datasets  $\mathcal{T}$  with its masks  $\mathcal{M}$ , chunk size  $c_{size}$ , batchsize  $b_{size}$ , model parameters  $\theta_{\text{UnIMP}}$ , learning rate  $\varphi$ , Epoch num  $Epochs$ , mask rate  $\kappa$ .

- 1  $\{X^c, M^c\} \leftarrow \text{Split } X, \mathcal{M} \text{ into batched chunks by } c_{size} \text{ \& } b_{size}$
- 2 Initialize optimizer  $opt_\theta$  with learning rate  $\varphi$
- 3 **for**  $Epoch \in \{0, 1, \dots, Epochs\}$  **do**
- 4      $M^c \leftarrow \text{mask } (\kappa + \frac{Epoch}{Epochs} \times 0.30) \text{ of } M^c$
- 5     **for each**  $X^c \in \{X^c\}$  **do**
- 6          $\tilde{X}^c \leftarrow \text{Algorithm } 2(X^c, M^c, \theta_{\text{UnIMP}})$
- 7         **for each**  $x_{ij} \in T^c \text{ and } M_{ij}^c = 0$  **do**
- 8             **if**  $x_{ij}$  being numerical or categorical **then**
- 9                  $\mathcal{L} \leftarrow \text{Compute loss by Eqn 7}$
- 10             **else**
- 11                  $\mathcal{L} \leftarrow \text{Compute loss by Eqn 6}$
- 12             Update  $\theta_{\text{UnIMP}}$  by  $opt_\theta$  with loss  $\frac{\mathcal{L}}{|X^c|}$ .
- 13 **Return**  $\theta_{\text{UnIMP}}$

---

$h_g$  as both the key and value matrices. The attention mechanism between prompt embeddings and graph embeddings allows the model to effectively align and integrate the prompt information with the local and global information from the graph embeddings.

When predicting the name of the president, as shown in Figure 2, the LLM uses its knowledge to infer the name, while the hyper-graph provides contextual details, such as name patterns. Xfusion combines both to enhance imputation accuracy.

**Projection head.** The obtained  $z_{\text{output}} \in \mathbb{R}^{(s+1) \times d_h}$  is in a high-dimensional space, and the projection head is a key component that maps this high-dimensional embedding to a target space. It is modeled as a linear layer  $f^{\text{head}} : z_{\text{output}} \mapsto o \in \mathbb{R}^{(s+1) \times d_o}$ . For numerical and categorical data,  $d_o = 1$ , as it requires only a single value as the imputation. In contrast, for text data, the task involves predicting a token ID, making it a classification task. Hence,  $d_o = |\text{Vocab}|$ , where  $|\text{Vocab}|$  denotes the size of the vocabulary.

With the above introduction of key components in UnIMP, we now show that the UnIMP captures the key intrinsic characteristics analyzed in Section 3.1. The proof is in Section 4.

**Theorem 3.4:** *The UnIMP framework captures the global-local information, high-order relationship, inter-column heterogeneity and intra-column homogeneity for imputation.*

## 4 TRAINING AND ANALYSIS

### 4.1 Training Objectives

**Motivation.** The training objective guides the model towards desired behaviors and outputs during the learning process. For the text data, it may contain an arbitrary length of tokens. Therefore, it is often modeled as the classification of the token ID of the next token and is processed in an auto-regressive manner. For the numerical data and the label-encoded categorical data, the imputation is often modeled as the task of regression.

The learning process for text data is handled using an auto-regressive approach. For a sequence of tokens  $(t_0, t_1, \dots, t_s)$ , an

autoregressive model tries to predict the next element in the sequence, given all the previous tokens. This objective is based on Cross Entropy (CE) [10, 53] as follows:

$$\begin{aligned} \mathcal{L}(T; \theta) &= P(t_0) \cdot P(t_1|t_0) \cdots P(t_s|t_0, t_1, \dots, t_{s-1}) \\ &= \prod_{i=0}^s \text{CE}(t_i, \theta(t_0 \cdots t_{i-1})) = \prod_{i=0}^s \sum_{j=0}^{d_o} -t_i[j] \cdot \log(\theta(t_0 \cdots t_{i-1})[j]) \end{aligned} \quad (6)$$

Where  $t_i[j]$  and  $\theta(\cdot)[j]$  are the probabilities of corresponding values belonging to  $j$ -th class.

The learning objective for numerical and categorical data is formulated using the Huber Loss [23]. For a true value  $x_{ij}$  and predicted value  $\tilde{x}_{ij}$ , Huber Loss is defined:

$$\mathcal{L}(x_{ij}, \tilde{x}_{ij}) = \begin{cases} \frac{1}{2}(x_{ij} - \tilde{x}_{ij})^2 & \text{for } |x_{ij} - \tilde{x}_{ij}| \leq \delta \\ \delta|x_{ij} - \tilde{x}_{ij}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \quad (7)$$

where  $\delta$  is a positive hyper-parameter. Compared to other loss functions, Huber Loss offers the following advantages: 1) For errors smaller than  $\delta$ , Huber Loss behaves like Mean-Squared-Error (MSE), being sensitive to small errors; 2) For errors larger than  $\delta$ , it behaves like Mean-Absolute-Error (MAE), being less sensitive to potential outliers. Moreover, Huber Loss is differentiable at all points, which is beneficial for optimization. By adjusting the value of  $\delta$ , it is flexible to balance between MSE and MAE, adapting to different contexts. We set  $\delta = 1$  suggested by the experiment in Section 5.

### 4.2 Training Pipeline

**Motivation.** Previous works in imputation [54, 65, 66, 71] typically train a separate model for each dataset. While effective, this practice significantly increases computational costs, as training a model can be time-intensive. Additionally, models trained on individual datasets often struggle to generalize to other datasets. A common alternative in the field of LLM is the "pre-train and finetune" strategy. However, the variability in table sizes and the propensity of neural models to overfit present challenges. In this work, we follow the "pre-train and finetune" paradigm for imputation, introducing two plug-and-play optimizations: the chunking technique and the progressive masking technique.

**Optimization 1: Chunking technique.** We split the entire dataset into smaller, uniformly sized chunks (512 rows in each chunk in our experiments). This approach ensures that all chunks are processed efficiently and in parallel, enabling batch training and reducing the heavy IO time associated with handling large and irregularly sized datasets. By transforming the dataset into manageable chunks, we also facilitate more effective memory utilization, preventing bottlenecks during model training. Moreover, in this way, the cardinality of column-level hyperedges becomes closer to that of row-level hyperedges, ensuring a more balanced edge distribution across chunks compared to previous works. This mitigates skewness and ultimately enhances performance.

**Optimization 2: Progressive masking technique.** Starting with the raw dataset, we gradually increase the proportion of masked data. The newly masked ratio is set from  $\kappa$  to  $\kappa + 30\%$  where  $\kappa$  is hyperparameter and is set to 35% suggested by the experiments in Section 5. As training progresses, we mask more data, gradually exposing the model to more challenging imputation scenarios. This

progressive masking strategy allows the model to incrementally learn patterns from simpler to more complex missing data scenarios and thus improve the generalization ability.

We summarize the overall pre-training pipeline in Algorithm 1. Following the SFT, the fine-tuning process has a similar procedure to the pre-training. The pre-training stage utilizes all available datasets, while fine-tuning focuses on specific datasets. The pipeline inputs training datasets, chunk size, batch size, initial model parameters, learning rate and the number of epochs. It first splits the training datasets into batched chunks (Line 1) and initializes the optimizer with the given learning rate (Line 2). The model is trained for  $Epochs$  epochs (Lines 3–12). During each epoch, we mask more data, and for each batch, the algorithm performs forward propagation (as described in Algorithm 2) to compute the imputation results (Line 6). For numerical or categorical data, the loss is calculated using Eqn 7, while for text data, the loss is computed using Eqn 6. After processing all cells in a chunk, the model parameters are updated using the optimizer and the loss (Line 12). Finally, the learned parameters are returned (Line 13).

### 4.3 Complexity Analysis

The LLM inference needs  $C_{prompt} = O(l_{trans} \cdot t_{max}^2 \cdot d_{LLM})$  for inference a prompt, where  $l_{trans}$  is the number of transformer layers,  $t_{max}$  is maximum number of tokens, and  $d_{LLM}$  is the dimension of embeddings in LLM. The feature encoding process requires encoding  $|\mathcal{V}| + |\mathcal{E}|$  prompts. Therefore, the total time complexity of feature encoding is  $O((|\mathcal{V}| + |\mathcal{E}|) \cdot C_{prompt})$ .

The message passing of  $l_{max}$  layers of Node-to-Hyperedge and Hyperedge-to-Node operations is with time complexity of  $C_{msg} = O(l_{max} \times (d_{LLM} \times |\mathcal{V}| + 2 \times d_{LLM} \times |\mathcal{E}|))$ . The attention module has a time complexity of  $C_{fusion} = (t_{max}^2 \times d_{LLM})$ , and the projection has  $C_{proj} = (d_{LLM})$ . Therefore, the total time complexity of training  $Epochs$  epochs is  $O(Epochs \times t_{max} \times (C_{msg} + C_{fusion} + C_{proj}))$ .

### 4.4 Theoretical Proofs

In this part, we present the proofs for Theorem 3.1, Theorem 3.2, Theorem 3.3 and Theorem 3.4. We first introduce the concepts of entropy and mutual information which are the base of the proof.

**Definition 4.1:**  $H(X) = -\sum_{x \in X} p(x) \log p(x)$  represents the entropy that quantifies the average level of information associated with  $X$ .  $I(X; Y) = H(X) - H(X|Y)$  is the mutual information between the  $X$  and  $Y$ .

**Lemma 4.1:**  $H(X|Y) \geq H(X|Y, Z)$  and the equation get if the information of  $Z$  is encoded in  $Y$ .

**PROOF.**  $H(X, Z | Y) = H(Z | Y) + H(X | Y, Z)$ . Therefore  $H(X | Y, Z) = H(X, Z | Y) - H(Z | Y)$ . As  $H(Z | Y) \geq 0$ , we can obtain the inequality.  $\square$

**Proof of Theorem 3.1.** The imputation error reflects how well the model trained on the training dataset generalizes to the unobserved data. According to the information bottleneck theory [27, 58],  $\Psi(\theta_{X_{obs}}, X_{miss})$  is proven to scale as  $\tilde{O}(\sqrt{\frac{I(X_{miss}; Z_{miss}) + 1}{n_{obs}}})$  where  $n_{obs}$  is the number of training observed data. For model  $\theta^{g+l}$ , its mutual information is  $I^{g+l}(X_{miss}; Z_{miss}) = I(X_{miss}; Z_{miss}|B^{g+l}) = I(X_{miss}; Z_{miss}|B^g, B^l)$  where  $B^g$  is the global information,  $B^l$  is

the local information,  $B^{g+l}$  is the fused global and local information. For model  $\theta^l$ , its mutual information is  $I^l(X_{miss}; Z_{miss}) = I(X_{miss}; Z_{miss}|B^l)$ . Then we have

$$\begin{aligned} I(X_{miss}; Z_{miss}|B^l) &= I(X_{miss}; Z_{miss}, B^l) - I(X_{miss}; B^l) \\ &= H(X_{miss}) - H(X_{miss}|Z_{miss}, B^l) - H(X_{miss}) + H(X_{miss}|B^l) \\ &= H(X_{miss}|B^l) - H(X_{miss}|Z_{miss}, B^l) \\ &\approx H(X_{miss}|B^l) - H(X_{miss}|Z_{miss}, B^l, B^g) \\ &\geq H(X_{miss}|B^l, B^g) - H(X_{miss}|Z_{miss}, B^l, B^g) \\ &= I(X_{miss}; Z_{miss}|B^l, B^g) \end{aligned}$$

The first line is obtained by the definition of conditional mutual information. The second line is obtained by the definition of mutual information. The fourth line and fifth line are based on Lemma 4.1. The fourth line is obtained as  $Z_{miss}$  contains the information in  $B^g$ , adding this term will not increase the entropy. The sixth line can be obtained by the reversal of the first three lines. Therefore, we have  $\Psi(\theta^{g+l}, X_{miss}) \leq \Psi(\theta^l, X_{miss})$ . Similarly, we can have  $\Psi(\theta^{g+l}, X_{miss}) \leq \Psi(\theta^g, X_{miss})$ .

**Proofs of Theorems 3.2 and 3.3.** The proofs of Theorem 3.2 and Theorem 3.3 are similar to that of Theorem 3.1. Given  $\theta^{[0:r]}$ , its mutual information can be represented as  $I^{[0:r]}(X_{miss}; Z_{miss}) = I(X_{miss}; Z_{miss}|B^{[0:r]}) = I(X_{miss}; Z_{miss}|B^{[0:s]}, B^{[s+1:r]})$ . Here, we use  $B^{[0:r]}$  to represent the captured interactions from order 0 to order  $r$ . For model  $\theta^{[0:s]}$ , its mutual information is  $I^{[0:s]}(X_{miss}; Z_{miss}) = I(X_{miss}; Z_{miss}|B^{[0:s]})$  where  $B^{[0:s]}$  is the captured interactions from order 0 to order  $s$  and  $r \geq s$ . Then, we can get  $\Psi(\theta^{[0:r]}, X_{miss}) \leq \Psi(\theta^{[0:s]}, X_{miss})$  by replacing  $B^{g+l}$  and  $B^l$  in the proof of Theorem 3.1 with  $B^{[0:r]}$  and  $B^{[0:s]}$ , respectively. Similarly, given a model  $\theta^{cp}$ , its mutual information can be calculated as  $I^{cp}(X_{miss}; Z_{miss}) = I(X_{miss}; Z_{miss}|B^{cp}, B^{base})$  where  $B^{cp}$  is the captured column patterns beside the base pattern  $B^{base}$ . Then, we can get the inequality  $\Psi(\theta^{cp}, X_{miss}) \leq \Psi(\theta, X_{miss})$  by replacing  $B^{g+l}$  and  $B^l$  in the proof of Theorem 3.1 with  $B^{cp}$  and  $B^{base}$ , respectively.

**Proof of Theorem 3.4. 1): Global-local information.** We prove that any two cells  $x_{ij}$  and  $x_{rs}$  can interact via a two-hop path in the hypergraph. Since  $v_{ij}$  and  $v_{is}$  both lie in the same row hyperedge, there is a direct connection:  $d(v_{ij}, v_{is}) = 1$ . Next, since  $v_{is}$  and  $v_{rs}$  share the same column hyperedge  $E_c^s$ , they are directly connected:  $d(v_{is}, v_{rs}) = 1$ . Thus, the overall distance is  $d(v_{ij}, v_{rs}) \leq d(v_{ij}, v_{is}) + d(v_{is}, v_{rs}) = 2$ . Moreover, the BiHMP has 2 layers, allowing it to aggregate information from the 2-hop neighborhood.

**2): High-order information.** The message passing of Node-to-Hyperedge follows  $z_{e_j}^{temp} = \frac{1}{|e_j|} \sum_{v_i \in e_j} \sigma(f_1^l(z_{v_i}^l))$ . It aggregates information over a set of cells at each layer, thereby capturing high-order relationships among cells.

**3): Inter-column heterogeneity and intra-column homogeneity.** Each cell  $v_{ij}$  is represented as a distinct node and is connected to a specific column hyperedge  $e_j$ . According to the Node-to-Hyperedge in Equation 2, the network to learn distinct transformations for different columns, thus capturing heterogeneity between columns. Moreover, according to the Hyperedge-to-Node in Equation 3, cells in the same column are learned with the same column representation, thus capturing intra-column homogeneity.



**Table 3: The profiles of datasets**

Dataset	Alias	Domain	$n$	$d_{\text{num}}$	$d_{\text{cate}}$	$d_{\text{text}}$
Blogger	BG	Social	100	0	6	0
ZOO	ZO	Biology	101	0	17	0
Parkinsons	PK	Healthcare	195	22	1	0
Bike	BK	Trans.	8,760	12	1	0
Chess	CS	Game	28,055	3	4	0
Shuttle	ST	Physics	43,500	0	10	0
Power	PW	Energy	2,049,280	6	0	0
Buy	BY	Retail	651	1	1	2
Restaurant	RR	Service	864	0	2	3
Walmart	WM	Location	4,654	0	3	2

## 5 EXPERIMENTAL EVALUATION

### 5.1 Experiment Setup

**Datasets description.** The evaluation uses 10 real-world datasets from UCI [14] and Kaggle [26], following the previous works [13, 36, 37, 54]. The profiles of these datasets are presented in Table 3.

**Missing mechanisms.** Original data are fully observed, we thus follow previous works [37, 71] to generate the mask matrix. Three mechanisms are utilized, i.e., missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

**Baseline methods.** We include 13 baselines, including state-of-the-art methods, for a comprehensive evaluation. Specifically, we incorporate 10 baselines for numerical and categorical data, including MEAN [24], KNNI [70], MICE [57], VGAIN [37], GAIN [65], GINN [47], GRAPE [66], IRGM [72], TDM [71], NOMI [54] and ReMasker [13]. Additionally, we compare our method to three LLM-based approaches that handle multi-type data imputation, including DFM [42], Table-GPT [30] and Jellyfish [68].

For our model, we include two variants, i.e., UnIMP which is a generalized model trained on all datasets using one unified parameter set and UnIMP-ft which is fine-tuned on specific datasets. **The pre-training phase uses only raw incomplete datasets. Moreover, these datasets are distinct and from different domains as shown in Table 3, ensuring no overlap for a fair evaluation.**

**Metrics.** For numerical and categorical data, we adopt Root-Mean-Square-Error (RMSE) and Mean-Absolute-Error (MAE) as metrics, following previous works [37, 55, 65], to assess accuracy. For text data, we employ two metrics to assess accuracy: Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1) [41] and Cosine Similarity (Cos-Sim) [44], covering both lexical-wise and semantic-wise approaches. The detailed definitions are in the appendix.

For the RMSE and MAE, a lower value indicates a better imputation. For the ROUGE-1 $F_1$  and Cos-Sim, a higher value indicates a better imputation. Note that for the first seven datasets (BG, ZO, PK, BK, CS, ST, PW) with only numerical and categorical features, imputation is assessed using RMSE and MAE. For the last three datasets (BY, RR, WM) that include text, evaluation is based on ROUGE-1 $F_1$  and Cos-Sim.

**Implementation details.** The missing rate is set as 20% and the missing mechanism is MCAR by default. **The default LLM for all baselines is Llama2-7B [50] (additional LLMs are evaluated in Exp-10), used out-of-the-box without parameter modifications.** We use Jellyfish-7B to keep the same size of LLM in [68]. As the model

in [30] is not released, we fine-tune Llama2-7B using the provided datasets. The number of layers is 3 in BiHMP. For the numerical and categorical data (*resp.* text data), we set the chunk size as 512 (*resp.* 32) with a batch size of 64 (*resp.* 2) and set the latent dimension as 64 (*resp.* 4096). We trained for 4000 (*resp.* 100) epochs and fine-tuned with 1000 (*resp.* 20) epochs. **We implement our method and baselines with Pytorch and conduct the experiment on a server equipped with an Intel Xeon(R) Silver 4314 CPU, 256GB system memory, and 2 Nvidia A5000 GPUs, each with 24GB GPU memory.**

### 5.2 Accuracy Comparison

**Exp-1: Accuracy over numerical and categorical data.** In this experiment, we evaluate the imputation accuracy for numerical and categorical data. The results are summarized in Table 4, with a missing rate of 20% under the MCAR mechanism. Some results are omitted due to imputation processes exceeding 10 hours (OOT) or running out of memory (OOM). As shown in the table, LLM-based methods such as DFMs and Table-GPT demonstrate poor performance, while our method achieves good performance across both data types. Graph-based methods, including GINN, GRAPE, IGRM and UnIMP show an outstanding performance. Furthermore, our fine-tuned version, UnIMP-ft, further enhances performance of UnIMP. Specifically, in terms of RMSE, UnIMP-ft shows average performance gains of 12.11%, 28.27%, 28.32% and 29.61% compared to IGRM, MICE, NOMI and ReMasker, respectively. Regarding MAE, UnIMP-ft demonstrates average performance gains of 8.74%, 36.49%, 30.86% and 35.36% over IGRM, MICE, NOMI, and ReMasker, respectively. These results highlight the excellence of UnIMP and UnIMP-ft in imputing numerical and categorical data.

**Exp-2: Accuracy over text data.** In this experiment, we evaluate the performance of imputing text data, comparing UnIMP with LLM-based methods, as other methods struggle with text data. The results at a 20% missing rate and MCAR are shown in Table 5. Our methods, UnIMP and UnIMP-ft, outperform previous LLM-based methods. Specifically, UnIMP shows a 30.53, 28.04, and 23.47 percentage point (pp) gain in ROUGE-1 $F_1$  over DFMs, Table-GPT, and Jellyfish, respectively. For Cos-Sim, it achieves an 8.68 pp, 5.58 pp, and 6.88 pp improvement. UnIMP-ft further boosts UnIMP. These results demonstrate the high performance of UnIMP and UnIMP-ft.

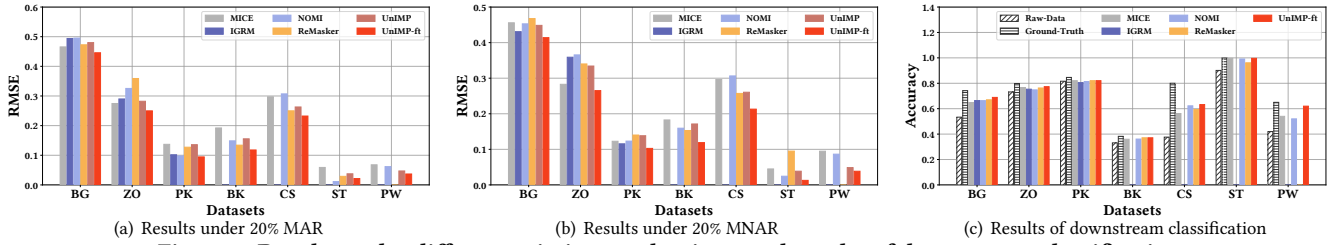
**Exp-3: Performance under MAR and MNAR.** We evaluate performance under MAR and MNAR, following the missing patterns defined in [37, 54]. Following [37, 54, 65], we model missingness using a logistic model based on randomly selected features, with a 20% missing rate as in Exp-1 and Exp-2. Figure 3 shows the results. UnIMP-ft achieves superior performance, reducing RMSE under MAR by 27.58%, 10.22%, 4.39%, and 17.86% compared to MICE, IGRM, NOMI, and ReMasker, respectively. Under MNAR, it further reduces RMSE by 30.23%, 10.72%, 27.32%, and 29.57%, highlighting the effectiveness of UnIMP and UnIMP-ft.

**Exp-4: Downstream classification accuracy.** In this experiment, we evaluate the downstream classification performance on datasets imputed by MICE, IGRM, NOMI, ReMasker and UnIMP-ft. We also include results from the ground-truth dataset (with no missing data) and the raw dataset (where missing values are replaced with 0, following [54]). We employ Random Forest as our classifier for all imputed datasets. The datasets are split into 80% training and 20% testing sets. Figure 3(c) presents the classification results. The

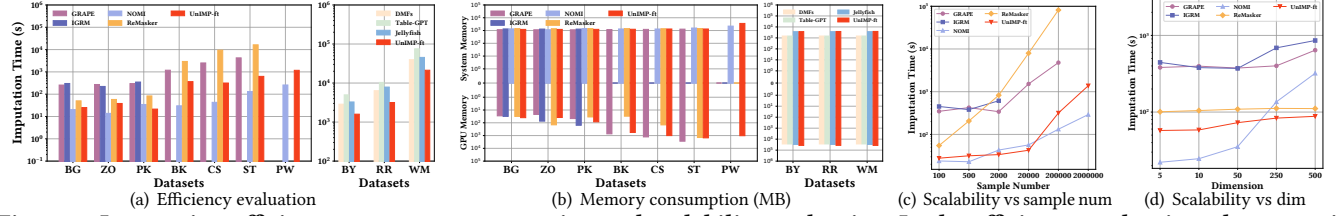
**Table 4: Results of missing data imputation (20% MCAR)**

Model	RMSE								MAE							
	Blogger	Zoo	Parkinsons	Bike	Chess	Shuttle	Power	Improve%	Blogger	Zoo	Parkinsons	Bike	Chess	Shuttle	Power	Improve%
MEAN	0.4315	0.4260	0.2062	0.2239	0.3079	0.0914	0.0869	45.22%	0.3631	0.3663	0.1473	0.1561	0.2584	0.0467	0.0559	57.24%
KNNI	0.4417	0.2749	0.1891	0.2023	0.3246	0.0456	OOT	35.75%	0.3337	0.1473	0.1207	0.1277	0.2606	0.0192	OOT	41.33%
MICE	0.4134	0.2645	0.1263	0.1796	0.2966	0.0426	0.0650	28.27%	0.3605	0.1731	0.0667	0.1097	0.2453	0.0131	0.0370	36.49%
VGAIN	0.4316	0.4114	0.1913	0.2219	0.2797	0.0786	0.0811	42.48%	0.3643	0.1891	0.1156	0.1363	0.2537	0.0352	0.0509	48.88%
TDM	0.4384	0.2949	0.1862	0.2444	0.3027	0.0769	0.0926	41.48%	0.3229	0.1490	0.0830	0.1499	0.2345	0.0350	0.0600	42.96%
GINN	0.4657	0.2761	0.1466	0.1641	0.2911	0.0734	OOM	32.41%	0.3444	0.1445	0.0884	0.0921	0.2339	0.0434	OOM	36.97%
GRAPE	0.4304	0.3211	0.1064	0.1481	0.2749	0.0242	OOM	20.89%	0.3151	0.1605	0.0535	0.0796	0.2200	0.0073	OOM	21.39%
IGRM	0.4551	0.3063	<b>0.1035</b>	OOM	OOM	OOM	OOM	12.11%	0.3423	0.1621	<b>0.0497</b>	OOM	OOM	OOM	OOM	8.74%
DFMs	0.4413	0.4445	0.2412	0.2483	OOT	OOT	OOT	36.76%	0.3676	0.2934	0.1647	0.1529	OOT	OOT	OOT	49.81%
Table-GPT	0.4237	0.4315	0.2246	0.2547	OOT	OOT	OOT	39.71%	0.3572	0.2713	0.1761	0.1442	OOT	OOT	OOT	48.99%
Jellyfish	0.4133	0.4177	0.2127	0.1935	OOT	OOT	OOT	36.43%	0.3557	0.2719	0.1548	0.1478	OOT	OOT	OOT	47.38%
NOMI	0.4112	<b>0.2576</b>	0.1322	0.1582	0.3042	<b>0.0237</b>	0.0731	28.32%	<b>0.3102</b>	<b>0.1442</b>	0.0710	0.0740	0.2298	<b>0.0071</b>	0.0463	30.86%
ReMasker	<b>0.4068</b>	0.3309	0.1508	<b>0.1277</b>	0.2662	0.1111	OOT	29.61%	0.3293	0.1807	0.0995	<b>0.0655</b>	0.2113	0.0545	OOT	35.36%
UnIMP	0.4171	0.2979	0.1407	0.1730	<b>0.2628</b>	0.0398	<b>0.0485</b>	25.84%	0.3384	0.1822	0.0966	0.1121	<b>0.2050</b>	0.0238	<b>0.0256</b>	36.08%
UnIMP-ft	<b>0.3972</b>	<b>0.2474</b>	<b>0.0990</b>	<b>0.1172</b>	<b>0.2142</b>	<b>0.0134</b>	<b>0.0425</b>	—	<b>0.3082</b>	<b>0.1428</b>	<b>0.0475</b>	<b>0.0602</b>	<b>0.1438</b>	<b>0.0040</b>	<b>0.0225</b>	—

\* Red text indicates the best result. Blue text indicates the second best result. 'OOT' indicates out of time (with a limit of 10 hours). 'OOM' indicates out of memory.



**Figure 3: Results under different missing mechanisms and results of downstream classification**



**Figure 4: Imputation efficiency, memory consumption and scalability evaluation. In the efficiency evaluation, the numerical/categorical data and text data need 20824 seconds and 73221 seconds for pre-training UnIMP, respectively. 45381 seconds are needed for pre-training Table-GPT.**

**Table 5: Results of imputation over text data**

Model	ROUGE-1 <sub>F1</sub>			Cos-Sim		
	Buy	Restaurant	Walmart	Buy	Restaurant	Walmart
DFMs	0.1535	0.0822	0.1420	0.8251	0.7609	0.7943
Table-GPT	0.1784	0.1398	0.1344	0.8345	0.8137	0.8254
Jellyfish	0.2153	0.1675	0.2067	0.8418	0.8145	0.778
UnIMP	0.3327	0.4017	0.5594	0.8610	0.8774	0.9025
UnIMP-ft	0.4273	0.4326	0.5931	0.8892	0.8923	0.9177

results indicate a higher imputation accuracy generally leads to a higher downstream classification, and datasets imputed by our proposed method achieve superior classification results compared to datasets imputed by other baselines.

### 5.3 Efficiency Evaluation

**Exp-5: Efficiency and memory consumption.** In this experiment, we evaluate imputation efficiency and memory consumption, as shown in Figure 4(a) and (b). We compare UnIMP-ft with graph-based methods (GRAPE, IGRM), LLM-based methods (DFMs, Table-GPT, Jellyfish), and high-accuracy baselines (NOMI, ReMasker). The results demonstrate that UnIMP-ft achieves competitive efficiency. It outperforms graph-based methods with a 3.07× and 5.38× speedup over GRAPE and IGRM, respectively. Additionally, the use of BiHMP as an adapter and fewer prompt tokens for inference improve imputation efficiency compared to LLM-based methods.

These results highlight the superior efficiency of UnIMP. Moreover, the results of memory consumption show that graph-based methods require slightly more memory due to graph construction, while our method demonstrates competitive memory consumption compared to other graph-based or LLM-based approaches.

**Exp-6: Scalability.** This experiment evaluates scalability w.r.t. the number of samples and data dimensions. We randomly select 100, 500, 2,000, 20,000, 200,000, and 2,000,000 samples from the Power dataset and generate synthetic datasets with 200 samples and dimensions of 5, 10, 50, 250, and 500. The results in Figure 4(c) and (d) highlight the competitive scalability of UnIMP. For sample scalability, ReMasker has a bad scalability, while NOMI scales the slowest. Graph-based methods (i.e., GRAPE, IGRM, UnIMP) show similar growth. For dimensional scalability, NOMI grows the fastest due to iterative column-wise enumeration, while ReMasker scales the slowest as it imputes all missing columns simultaneously. Graph-based methods exhibit similar trends.

### 5.4 Ablation Study

**Exp-7: Ablation study.** In this experiment, we conduct an ablation study to assess the effectiveness of key components in UnIMP, including hypergraph-aggregated global-local information, the BiHMP

**Table 6: Results from the ablation study. Color highlights the percentage point (pp) decline compared to the full model.**

Model	ROUGE-1 <sub>F<sub>1</sub></sub>			Cos-Sim		
	Buy	Restaurant	Walmart	Buy	Restaurant	Walmart
UnIMP	0.3327	0.4017	0.5594	0.8610	0.8774	0.9025
w/o Hypergraph	0.2580 (↓7.47pp)	0.3757 (↓2.60pp)	0.4812 (↓7.82pp)	0.8508 (↓1.02pp)	0.8611 (↓1.63pp)	0.8681 (↓3.44pp)
w/o BiHMP	0.3014 (↓3.13pp)	0.3877 (↓1.40pp)	0.5282 (↓3.12pp)	0.8571 (↓0.39pp)	0.8702 (↓0.72pp)	0.8822 (↓2.03pp)
w/o Xfusion	0.2867 (↓4.60pp)	0.3648 (↓3.69pp)	0.5052 (↓5.42pp)	0.8517 (↓0.93pp)	0.8581 (↓1.93pp)	0.8729 (↓2.96pp)
w/o P.g. mask	0.3263 (↓0.64pp)	0.3908 (↓1.09pp)	0.5337 (↓2.57pp)	0.8602 (↓0.08pp)	0.8654 (↓1.20pp)	0.8932 (↓0.93pp)

**Table 7: Imputation results of text linguistic**

Metrics	Overall	Long	Short	Structured	Colloquial
ROUGE-1 <sub>F<sub>1</sub></sub>	0.4551	0.2636	0.6339	0.2995	0.4693
Cos-Sim	0.8801	0.8391	0.9184	0.8404	0.8837

**Table 8: Results of serialization study**

Method	ROUGE-1 <sub>F<sub>1</sub></sub>			Cos-Sim		
	Buy	Restaurant	Walmart	Buy	Restaurant	Walmart
Attr-Value Pairs	0.3327	0.4017	0.5594	0.8610	0.8774	0.9025
Sentence	0.3016	0.3760	0.4338	0.8221	0.8167	0.8735
Markdown	0.3312	0.4230	0.5232	0.8599	0.8502	0.8894
Dataframe	0.2123	0.3163	0.4586	0.8416	0.8353	0.8833

**Table 9: RMSE results of inductive generalization**

Model	Ionosphere	Breast	Spam	News	Connect	Metro
MICE	0.2487	0.2183	0.0712	0.0237	0.2539	0.1225
IGRM	0.2502	0.2107	0.0571	OOM	OOM	OOM
NOMI	0.2320	0.2115	0.0611	0.0252	0.2473	0.1485
ReMasker	0.2360	0.2058	0.0595	0.0282	OOT	OOT
UnIMP	0.2444	0.2426	0.0626	0.0306	0.2569	0.1308

module, the Xfusion model, and progressive masking. To evaluate hypergraph-aggregated information, we replace it with a zero tensor while keeping other components unchanged. For BiHMP, we substitute it with the Set Transformer from [8]. For Xfusion, we replace it with a weighted sum fusion mechanism. To analyze progressive masking, we train the model without masking. The results in Table 6 show that hypergraph information improves ROUGE-1<sub>F<sub>1</sub></sub> by 5.96 pp and Cos-Sim by 2.03 pp. BiHMP contributes a 2.55 pp increase in ROUGE-1<sub>F<sub>1</sub></sub> and 1.04 pp in Cos-Sim. Xfusion improves ROUGE-1<sub>F<sub>1</sub></sub> by 4.57 pp and Cos-Sim by 1.94 pp. Progressive masking results in a 1.43 pp gain in ROUGE-1<sub>F<sub>1</sub></sub> and 0.74 pp in Cos-Sim. Overall, these components are more important on the Walmart dataset. This is because Walmart is larger and more reliant on these components to aggregate information for accurate imputation.

**Exp-8: Text linguistic and serialization strategy.** In this experiment, we evaluate UnIMP under different text linguistic and serialization strategies, as shown in Table 7 and Table 8. We consider four types of text linguistics: Long (text with more than 10 characters), Short (text with fewer than 10 characters), Structured (text with a fixed format, e.g., telephone numbers), and Colloquial (unstructured text, e.g., restaurant names). More details can be found in the full version<sup>1</sup>. The results indicate that models perform better on short text due to lower cumulative errors compared to long text. Additionally, colloquial text imputation outperforms structured text since language models are pre-trained on natural language, which is inherently flexible rather than fixed in format. For serialization strategies, we evaluate attribute-value pairs, sentences, markdown, and dataframes [16]. Definitions of these formats are provided in the full version. The results show that, despite its simplicity, the attribute-value pair strategy achieves the best performance.

**Exp-9: Generalization.** In this experiment, we assess the performance on datasets that are not used during pre-training. We use six unseen UCI datasets: Ionosphere, Breast, Spam, News, Connect,

and Metro [14]. As shown in Table 9, UnIMP achieves competitive performance despite not being trained on these datasets. Notably, for large datasets like Connect and Metro, its RMSE differs by only  $\pm 10\%$  from models trained directly on them.

## 5.5 Hyper-parameters Analysis

**Exp-10: Varying LLMs.** We evaluate the imputation RMSE using several representative LLMs, including LiteLlama [61], TinyLlama, Phi2 [38], Llama2 [50] and Llama3 [15], with the results shown in Figure 5(a). As depicted, more advanced LLM generally leads to better imputation performance. We selected Llama2, as it was one of the most outstanding open-source LLMs with extensive resources and support available when conducting experiments.

**Exp-11: Chunk sizes.** We evaluate imputation RMSE and imputation time with varying chunk sizes, as shown in Figure 5(b). The chunk sizes tested are 128, 256, 512, 1024 and 2048. The evaluation is conducted on the Chess, Shuttle and Power datasets which have the most samples. As shown, RMSE fluctuates with different chunk sizes, while imputation time first decreases and then increases. This may be because very small chunks underutilize the GPU resources, while very large chunks require excessive I/O resources. A chunk size of 512 strikes a good balance between RMSE and time.

**Exp-12: Batch sizes.** We evaluate the imputation RMSE and imputation time with varying batch sizes, as shown in Figure 5(c). The batch sizes tested are 16, 32, 64, 128 and 256. As in Exp-9, the Chess, Shuttle, and Power datasets are used. The results show that RMSE fluctuates with different batch sizes, while imputation time decreases as batch size increases. A batch size of 64 generally provides a good imputation RMSE and time.

**Exp-13: Varying  $\delta$ .** We evaluate the imputation RMSE with different values of  $\delta$ , which is used in the loss function as defined in Eqn 7. The values of  $\delta$  tested are 0.1, 0.5, 1, 4 and 10, and the results are shown in Figure 5(d). As illustrated, RMSE fluctuates with varying  $\delta$ , with  $\delta = 1$  generally providing an excellent RMSE.

**Exp-14: Varying ratios of progressive masked data,  $\kappa$ .** We evaluate imputation RMSE for different values of  $\kappa$  (0.15, 0.25, 0.35, 0.45) which is used to control the newly masked ratio. The results in Figure 5(e) show that RMSE decreases and then increases, with  $\kappa = 0.35$  providing the best performance.

**Exp-15: Missing rates.** In Figure 5(f), we experiment with missing rates ranging from 0.1 to 0.8 on the Power dataset which has the most samples. We compare UnIMP with MICE and NOMI, as they show superior performance on Power as in Exp-1. As illustrated, the RMSE increases with higher missing rates, and UnIMP-ft consistently outperforms the baselines across varying missing rates.

**Exp-16: Pre-training epochs.** We assess the performance across pre-training epochs, varying from 1000 to 4000 in increments of 1000 in Figure 5(g). As shown, RMSE generally decreases with more pre-training epochs, and 4000 epochs is generally sufficient to achieve superior performance.

**Exp-17: Fine-tuning epochs.** We evaluate the performance under varying fine-tuning epochs, ranging from 500 to 2000 at an interval of 500 in Figure 5(h). As shown, RMSE generally decreases with more fine-tuning epochs, and 1000 epochs is generally adequate to achieve excellent performance.

**Exp-18: Encoding methods.** We evaluate different feature encoding methods, including one-hot for categorical data and HYTREL [8]

<sup>1</sup>The full version is in [https://github.com/guaiyoui/UnIMP/blob/master/full\\_version/UnIMP\\_Full\\_Version.pdf](https://github.com/guaiyoui/UnIMP/blob/master/full_version/UnIMP_Full_Version.pdf)



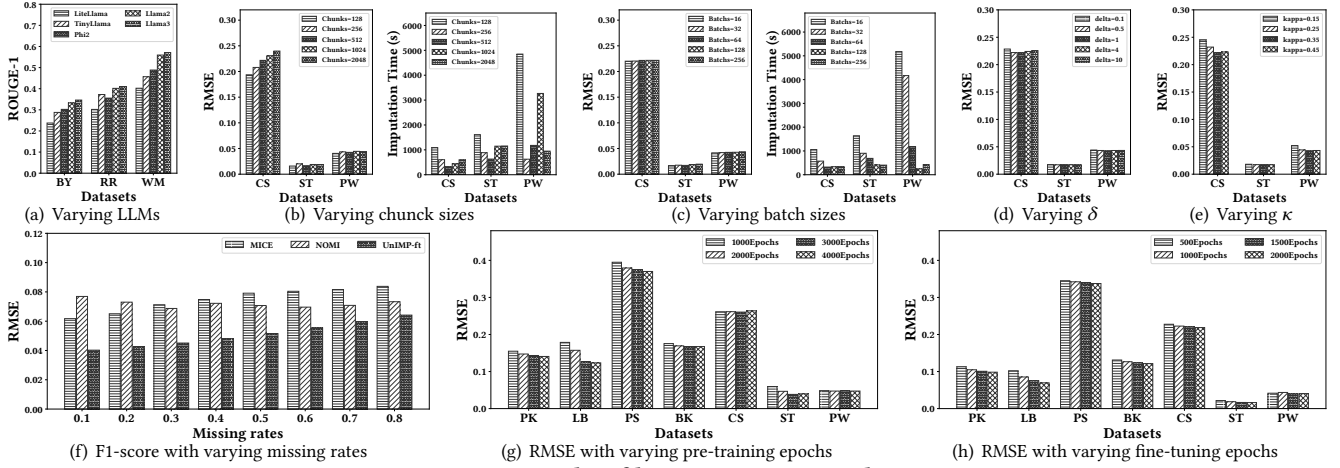


Figure 5: Results of hyper-parameter analysis

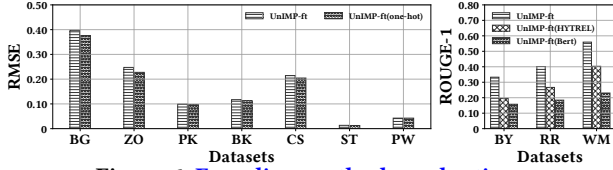


Figure 6: Encoding methods evaluations.

and BERT [11] for text, in Figure 6. The results indicate that one-hot generally performs better, while HYTREL and BERT perform less effectively, likely because their encoded features are not well aligned with the projection head. Our default setting follows previous works [13, 55] and uses label encoding for categorical data.

## 6 RELATED WORK

**Statistics-based and similarity-based imputation.** Statistics-based algorithms impute missing data using statistical values (e.g., mean or mode) for each feature [17, 39]. Similarity-based algorithms impute missing data using closely related data. KNNI [3] selects K nearest neighbors of the sample with missing features for imputation. An iterative method proposed in [70] locates a more reliable local context. NOMI [54] uses similarity to iteratively augment the input and employs an uncertainty-driven network for imputation. **Graph-based models for imputation.** Graph structures have been employed to capture complex relationships for data imputation. One popular approach involves using similarity graphs, as demonstrated in GINN [47] and GEDI [7], where edges are constructed between samples to facilitate imputation. Another popular approach leverages bipartite graphs, as seen in GRAPE [66] and IGRM [72], where each row and column is represented as a node, and the cell values are modeled as edges in the graph.

**Learning-based imputation.** Various AI techniques are used to better capture the underlying relationships in raw datasets. Traditional methods include models like XGBoost [9], random forests in MissForest [48], multi-layer perceptrons [19], and linear regression in MICE [45]. Generative models are also widely employed, with autoencoders being particularly popular. Examples include deep denoising autoencoders in MDAE [20] and ReMasker [13], variational autoencoders [35], and importance-weighted autoencoders in MIWAE[34]. GANs are another commonly used approach for missing data imputation, as seen in GAIN [65] and VGAIN [37].

Recently, distribution-matching-based imputation methods have been introduced for missing data imputation. These approaches align missing values with observed data distributions in the original space (e.g., OTImputer [40]) or in the latent space (e.g., TDM [71]). **LLMs-based imputation.** Recently, LLMs have shown a remarkable ability to understand and generate human-like text across a diverse array of tasks. Two main approaches have emerged: in-context learning and fine-tuning. In the in-context-learning-based approach, LLMs are prompted to predict missing data directly with task-related examples, as demonstrated in [42, 43]. Fine-tune-based methods, such as Table-GPT [30] and Jellyfish [68], fine-tune the LLMs to adapt the model to tasks in the field of tabular data.

## 7 CONCLUSION AND DISCUSSION

We study mixed-type data imputation and propose UnIMP, an LLM-enhanced framework with high-order message passing. We introduce a cell-oriented hypergraph and design BiHMP to aggregate global-local information while capturing column patterns. Xfusion aligns BiHMP with LLMs, serving as an adapter. To train UnIMP, we employ chunking and progressive masking in a pre-training and fine-tuning strategy. Both theoretical and empirical results demonstrate its superiority.

**Limitations and Comparison with other LLM-based Imputation.** UnIMP models granular tabular structures and leverages hypergraphs and HGNN for reasoning, enabling more accurate information aggregation for imputation. However, it has several limitations: 1) Cross-dataset information. Table-GPT, Jellyfish, and DFMs do not explicitly model granular tabular structures, making them more flexible in integrating multiple tabular datasets for imputation. In contrast, UnIMP struggles to capture cross-table information. 2) Complementary information from related tasks. Table-GPT, Jellyfish, and DFMs support various data quality management tasks and can integrate results from related tasks. For example, entity-matching outcomes may provide contextual information that enhances imputation. 3) Flexibility to update LLM. DFMs, based on prompt engineering, may be sensitive to prompt variations but can flexibly utilize different LLMs. In contrast, UnIMP, Table-GPT, and Jellyfish, which rely on fine-tuning, are less sensitive to prompt structures after fine-tuning but depend on a specific pre-trained LLM, potentially impacting their generalizability.

## REFERENCES

- [1] 2024. Mastering Mixed Methods Research in 2024: Combining Qualitative and Quantitative Approaches. <https://editverse.com/mastering-mixed-methods-research-in-2024/>
- [2] Nikolich Alexandr, Oslakova Irina, Kudinova Tatyana, Kappusheva Inessa, and Puchkova Arina. 2021. Fine-tuning GPT-3 for Russian text summarization. In *Data Science and Intelligent Systems: Proceedings of 5th Computational Methods in Systems and Software 2021*, Vol. 2. Springer, 748–757.
- [3] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor non-parametric regression. *The American Statistician* 46, 3 (1992), 175–185.
- [4] Martin Berglund and Brink van der Merwe. 2023. Formalizing BPE Tokenization. *arXiv preprint arXiv:2309.08715* (2023).
- [5] Fabian Biester, Mohamed Abdelaal, and Daniel Del Gaudio. 2024. Llmclean: Context-aware tabular data cleaning via llm-generated ofds. In *European Conference on Advances in Databases and Information Systems*. Springer, 68–78.
- [6] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [7] Katrina Chen, Xiuqin Liang, Zheng Ma, and Zhibin Zhang. 2023. GEDI: A graph-based end-to-end data imputation framework. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 723–730.
- [8] Pei Chen, Soumajyoti Sarkar, Leonard Lausen, Balasubramaniam Srinivasan, Sheng Zha, Ruihong Huang, and George Karypis. 2024. HYTEL: Hypergraph-enhanced tabular data representation learning. *Advances in Neural Information Processing Systems* 36 (2024).
- [9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [10] Pieter-Tjerk de Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134 (2005), 19–67.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [12] Kounianhua Du, Weinan Zhang, Ruiwen Zhou, Yangkun Wang, Xilong Zhao, Jiarui Jin, Quan Gan, Zheng Zhang, and David P Wipf. 2022. Learning enhanced representation for tabular data via neighborhood propagation. *Advances in Neural Information Processing Systems* 35 (2022), 16373–16384.
- [13] Tianyu Du, Luca Melis, and Ting Wang. 2024. ReMasker: Imputing Tabular Data with Masked Autoencoding. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net. <https://openreview.net/forum?id=K19NqjLVDT>
- [14] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>
- [15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [16] Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H Sengamedu, and Christos Faloutsos. [n.d.]. Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding-A Survey. *Transactions on Machine Learning Research* ([n.d.]).
- [17] Alireza Farhangfar, Lukasz A Kurgan, and Witold Pedrycz. 2007. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37, 5 (2007), 692–709.
- [18] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3558–3565.
- [19] Pedro J Garcia-Laencina, José-Luis Sancho-Gómez, and Anibal R Figueiras-Vidal. 2010. Pattern classification with missing data: a review. *Neural Computing and Applications* 19, 2 (2010), 263–282.
- [20] Lovedeep Gondara and Ke Wang. 2018. Mida: Multiple imputation using denoising autoencoders. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3–6, 2018, Proceedings, Part III*. Springer, 260–272.
- [21] John T Hancock and Taghi M Khoshgoftaar. 2020. Survey on categorical data for neural networks. *Journal of big data* 7, 1 (2020), 28.
- [22] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation. 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)*. 2208–2222.
- [23] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 492–518.
- [24] Mortaza Jamshidian and Matthew Mata. 2007. Advances in analysis of mean and covariance structure when data are incomplete. In *Handbook of latent variable and related models*. Elsevier, 21–44.
- [25] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
- [26] Kaggle. 2021. Flipkart Products. <https://www.kaggle.com/datasets/PromptCloudHQ/flipkart-products/data>
- [27] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023. How does information bottleneck help deep learning? In *International Conference on Machine Learning*. PMLR, 16049–16096.
- [28] T Kudo. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226* (2018).
- [29] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttel, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [30] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-GPT: Table Fine-tuned GPT for Diverse Table Tasks. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–28.
- [31] Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208* (2023).
- [32] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* 53 (2020), 1487–1509.
- [33] Somayya Madakam, Ramya Ramaswamy, and Siddharth Tripathi. 2015. Internet of Things (IoT): A literature review. *Journal of Computer and Communications* 3, 5 (2015), 164–173.
- [34] Pierre-Alexandre Mattei and Jes Frelsen. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*. PMLR, 4413–4423.
- [35] John T McCoy, Steve Kroon, and Lidia Auret. 2018. Variational autoencoders for missing data imputation with application to a simulated milling circuit. *IFAC-PapersOnLine* 51, 21 (2018), 141–146.
- [36] Yinan Mei, Shaoxu Song, Chenguang Fang, Haifeng Yang, Jingyun Fang, and Jiang Long. 2021. Capturing semantics for imputation with pre-trained language models. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 61–72.
- [37] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, and Jianwei Yin. 2022. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering* 35, 7 (2022), 6630–6650.
- [38] Microsoft. 2024. microsoft/phi-2. <https://huggingface.co/microsoft/phi-2>
- [39] Karel GM Moons, Rogier ART Donders, Theo Stijnen, and Frank E Harrell Jr. 2006. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology* 59, 10 (2006), 1092–1101.
- [40] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. 2020. Missing data imputation using optimal transport. In *International Conference on Machine Learning*. PMLR, 7130–7140.
- [41] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11–12, 2016*, Yoav Goldberg and Stefan Riezler (Eds.). ACL, 280–290. <https://doi.org/10.18653/V1/K16-1028>
- [42] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment* 16, 4 (2022), 738–746.
- [43] Yichen Qian, Yongyi He, Rong Zhu, Jintao Huang, Zhijian Ma, Haibin Wang, Yaohua Wang, Xiuyu Sun, Defu Lian, Bolin Ding, et al. 2024. UniDM: A Unified Framework for Data Manipulation with Large Language Models. *Proceedings of Machine Learning and Systems* 6 (2024), 465–482.
- [44] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. <https://doi.org/10.18653/V1/D19-1410>
- [45] Patrick Royston and Ian R White. 2011. Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of statistical software* 45 (2011), 1–20.
- [46] Rajat Singh and Srikanta Bedathur. 2023. Embeddings for Tabular Data: A Survey. *arXiv preprint arXiv:2302.11777* (2023).
- [47] Indro Spinelli, Simone Scardapane, and Aurelio Uncini. 2020. Missing data imputation with adversarially-trained graph convolutional networks. *Neural Networks* 129 (2020), 249–260.

- [48] Daniel J Stekhoven and Peter Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.
- [49] Bohan Tang, Zexi Liu, Keyue Jiang, Siheng Chen, and Xiaowen Dong. 2024. Simplifying Hypergraph Neural Networks. <https://api.semanticscholar.org/CorpusID:267547514>
- [50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [51] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [52] Jianwei Wang, Kai Wang, Xuemin Lin, Wenjie Zhang, and Ying Zhang. 2024. Efficient Unsupervised Community Search with Pre-trained Graph Transformer. *Proc. VLDB Endow.* 17, 9 (2024), 2227–2240. <https://doi.org/10.14778/3665844.3665853>
- [53] Jianwei Wang, Kai Wang, Xuemin Lin, Wenjie Zhang, and Ying Zhang. 2024. Neural Attributed Community Search at Billion Scale. *Proceedings of the ACM on Management of Data* 1, 4 (2024), 1–25.
- [54] Jianwei Wang, Ying Zhang, Kai Wang, Xuemin Lin, and Wenjie Zhang. 2024. Missing Data Imputation with Uncertainty-Driven Network. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–25.
- [55] Wei Wang, Yimeng Chai, and Yue Li. 2022. A Global to Local Guiding Network for Missing Data Imputation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4058–4062.
- [56] Sharada Prasad Wasti, Padam Simkhada, Edwin R van Teijlingen, Brijesh Sathian, and Indrajit Banerjee. 2022. The growing importance of mixed-methods research in health. *Nepal journal of epidemiology* 12, 1 (2022), 1175.
- [57] Ian R White, Patrick Royston, and Angela M Wood. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine* 30, 4 (2011), 377–399.
- [58] Wikipedia contributors. 2024. Information bottleneck method – Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/w/index.php?title=Information\\_bottleneck\\_method&oldid=1260066467](https://en.wikipedia.org/w/index.php?title=Information_bottleneck_method&oldid=1260066467) [Online; accessed 8-December-2024].
- [59] Yangyang Wu, Xiaoye Miao, Zi-ang Nan, Jinshan Zhang, Jianhu He, and Jianwei Yin. 2024. Effective and Efficient Multi-View Imputation With Optimal Transport. *IEEE Trans. Knowl. Data Eng.* 36, 11 (2024), 6029–6041. <https://doi.org/10.1109/TKDE.2024.3387439>
- [60] Yangyang Wu, Jun Wang, Xiaoye Miao, Wenjia Wang, and Jianwei Yin. 2024. Differentiable and Scalable Generative Adversarial Models for Data Imputation. *IEEE Trans. Knowl. Data Eng.* 36, 2 (2024), 490–503. <https://doi.org/10.1109/TKDE.2023.3293129>
- [61] Xiaotian Han, Xia Ben Hu. 2024. ahxt/LiteLlama-460M-1T. <https://huggingface.co/ahxt/LiteLlama-460M-1T>
- [62] Shuangyin Xie. 2019. *Sentiment Analysis using machine learning algorithms: online women clothing reviews*. Ph.D. Dissertation. Dublin, National College of Ireland.
- [63] Yang Yang, Yizhou Sun, Jie Tang, Bo Ma, and Juanzi Li. 2015. Entity matching across heterogeneous sources. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1395–1404.
- [64] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Leah Li. 2024. Financial Report Chunking for Effective Retrieval Augmented Generation. *arXiv preprint arXiv:2402.05131* (2024).
- [65] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.
- [66] Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. 2020. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 19075–19087.
- [67] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. 2023. Large language models as data preprocessors. *arXiv preprint arXiv:2308.16361* (2023).
- [68] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. 2024. Jellyfish: Instruction-Tuning Local Large Language Models for Data Preprocessing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 8754–8782.
- [69] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).
- [70] Shichao Zhang. 2012. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software* 85, 11 (2012), 2541–2552.
- [71] He Zhao, Ke Sun, Amir Dezfouli, and Edwin V Bonilla. 2023. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*. PMLR, 42159–42186.
- [72] Jiajun Zhong, Ning Gui, and Weiwei Ye. 2023. Data imputation with iterative graph reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11399–11407.



**Algorithm 2: Forward Propagation of UnIMP**


---

**Input:** Raw dataset  $X$ , Mask  $M$ , model parameters  $\theta_{\text{UnIMP}}$ .  
**Output:** The imputed dataset  $\tilde{X}$ .  
 // Hypergraph Construction and Feature Encoding.

- 1  $HG(\mathcal{V}, \mathcal{E}) \leftarrow$  Construct hypergraph from  $X$ ;  $\tilde{X} \leftarrow X$
- 2 **for**  $v_i \in \mathcal{V}, e_j \in \mathcal{E}$  **do**
- 3    $\{t_0, \dots, t_s\} \leftarrow$  Serilization and tokenization.
- 4    $z_p : \{z_{t_0}, z_{t_1}, \dots, z_{t_s}\} \leftarrow \text{LLM-backbone}(t_0, t_1, \dots, t_s)$
- 5    $z_{v_i}^0$  or  $z_{e_j}^0 \leftarrow z_{t_s}$
- // High-order Message Passing.
- 6 **for**  $l = 0, \dots, l_{\max}$  **do**
- 7    $z_{e_j}^{\text{temp}} = \frac{1}{|e_j|} \sum_{v_i \in e_j} \sigma \left( f_1^l(z_{v_i}^l) \right)$
- 8    $z_{e_j}^{l+1} = \sigma \left( f_2^l \left( \text{CONCAT} \left( z_{e_j}^l, z_{e_j}^{\text{temp}} \right) \right) \right)$
- 9    $z_{v_i}^{l+1} = \sigma \left( f_3^l \left( \text{CONCAT} \left[ z_{v_i}^l, z_{e_{v_i}^c}^l, z_{e_{v_i}^r}^l \right] \right) \right)$
- // Xfusion and projection head.
- 10 **for**  $v_i$  with corresponding mask being 0 **do**
- 11    $\tilde{x}_{v_i} = \emptyset$
- 12   **while** **True** **do**
- 13     Update  $z_p$  with  $\tilde{x}_{v_i}$
- 14      $z_p^{\text{temp}} = \text{Attn}(z_p W_Q, z_p W_K, z_p W_V; w_1)$
- 15      $z_{\text{output}} = \text{Attn}(z_p^{\text{temp}} W_Q, z_g W_K, z_g W_V; w_2)$
- 16      $o = f^{\text{head}}(z_{\text{output}})[-1]; \tilde{x}_{v_i}. \text{append}(o)$
- 17     **if**  $\text{Decode}(o) = \text{EOS}$  or  $z_{v_i}$  is num. or cate. **then**
- 18       Break
- 19 **return**  $\tilde{X}$

---

## 8 APPENDIX

### 8.1 Forward Propagation of UnIMP

The overall forward process of UnIMP is summarized in Algorithm 2. UnIMP takes the raw dataset  $X$  and mask  $M$  as inputs and outputs the imputed dataset  $\tilde{X}$ . We first construct a hypergraph according to the process introduced in Section 3.3 (Line 1). It then initializes the feature of each node and hyperedge (Lines 2 to 5). The initialized features are then propagated by the high-order message-passing framework (Lines 6 to 9). It contains  $l_{\max}$  layers and each layer contains both the Node-to-Hyperedge layer (Lines 7 and 8) and Hyperedge-to-Node (Line 9). The aggregated feature and the prompt representation are then fused by the Xfusion for projection (Lines 10 to 18). Prompt embedding is first updated by previous predicted  $\tilde{x}_{v_i}$  (Line 13). The Xfusion is applied to the obtained features (Lines 13 to 14). The projection head is applied to predict the imputation, using the last token as the newly predicted token (Line 16). If the predicted token is EOS or its type is numerical or categorical, it terminates the auto-regressive process (Lines 17 to 18). At last, the algorithm outputs the imputed matrix.

### 8.2 Missing Mechanisms

We utilize three widely used missing mechanisms in our experiments, i.e. missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), following previous

works [13, 37]. The missing data of MCAR is in the following form where the missing probability of the data is only dependent on itself. Here,  $X = X^o \cup X^m$  is the union of the observed data  $X^o$  and miss data  $X^m$ .

$$P_{\text{MCAR}}(M|X) = P(M|X^o, X^m) = P(M) \quad (8)$$

While missing data of MAR is conditioned on the observed data  $X^o$ , which is shown as follows:

$$P_{\text{MAR}}(M|X^o, X^m) = P(M|X^o) \quad (9)$$

By contrast, the missing data of MNAR is dependent on the missing data  $X^m$  in the following form:

$$P_{\text{MNAR}}(M|X^o, X^m) = P(M|X^m) \quad (10)$$

### 8.3 Evaluation Metrics

For numerical and categorical data, we adopt Root-Mean-Square-Error (RMSE) and Mean-Absolute-Error (MAE) as metrics, following previous works [37, 55, 65], to assess accuracy. The definitions of RMSE and MAE are as follows:

$$\text{RMSE}(X, \tilde{X}) = \sqrt{\frac{\sum_{j=1}^d \sum_{i=1}^n (1 - m_{ij})(x_{ij} - \tilde{x}_{ij})^2}{\sum_{j=1}^d \sum_{i=1}^n (1 - m_{ij})}} \quad (11)$$

$$\text{MAE}(X, \tilde{X}) = \frac{\sum_{j=1}^d \sum_{i=1}^n (1 - m_{ij})|x_{ij} - \tilde{x}_{ij}|}{\sum_{j=1}^d \sum_{i=1}^n (1 - m_{ij})} \quad (12)$$

For text data, we employ two metrics to assess accuracy: Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1) [41] and Cosine Similarity (Cos-Sim) [44], covering both lexical-wise and semantic-wise approaches. ROUGE-1 measures unigram overlap, assessing vocabulary similarity between imputed and original text. However, it overlooks semantic equivalence (e.g., "METRO" vs. "SUBWAY"). Hence, we introduce Cosine Similarity, which captures deeper contextual relationships via vector representations. Given the generated text  $T_G$  and the ground-truth reference text  $T_R$ , the metrics are computed as:

$$\begin{aligned} \text{ROUGE-1}_{\text{recall}} &= \frac{|T_R \cap T_G|}{|T_G|}; \text{ROUGE-1}_{\text{precision}} = \frac{|T_R \cap T_G|}{|T_R|} \\ \text{ROUGE-1}_{F_1} &= 2 \cdot \frac{\text{ROUGE-1}_{\text{recall}} \cdot \text{ROUGE-1}_{\text{precision}}}{\text{ROUGE-1}_{\text{recall}} + \text{ROUGE-1}_{\text{precision}}} \\ \text{Cos-Sim} &= \frac{\sum_{i=1}^{d_{\text{Bert}}} \text{BERT}(T_G)_i \cdot \text{BERT}(T_R)_i}{\sqrt{\sum_{i=1}^{d_{\text{Bert}}} \text{BERT}(T_G)_i^2} \sqrt{\sum_{i=1}^{d_{\text{Bert}}} \text{BERT}(T_R)_i^2}} \end{aligned} \quad (13)$$

Where  $\text{BERT}(\cdot)$  represents the embedding vectors given input.  $d_{\text{Bert}}$  is the dimension of the embeddings.

### 8.4 Text Linguistic and Serialization Strategy

In Exp-8, we evaluate UnIMP under different text linguistic characteristics and serialization strategies using the Buy, Restaurant, and Walmart datasets. We evaluate 4 text linguistic characteristics, including short, long, structured, and colloquial, as follows.

- Short. Text with fewer than 10 characters, including manufacturer and price from Buy, city, state, zip\_code, and country from Walmart, and cuisine types from Restaurant.

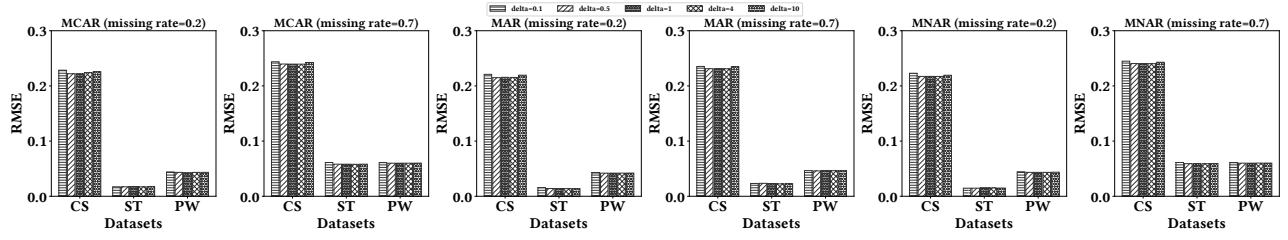


Figure 7: Experimental results of hyperparameter generalization on delta.

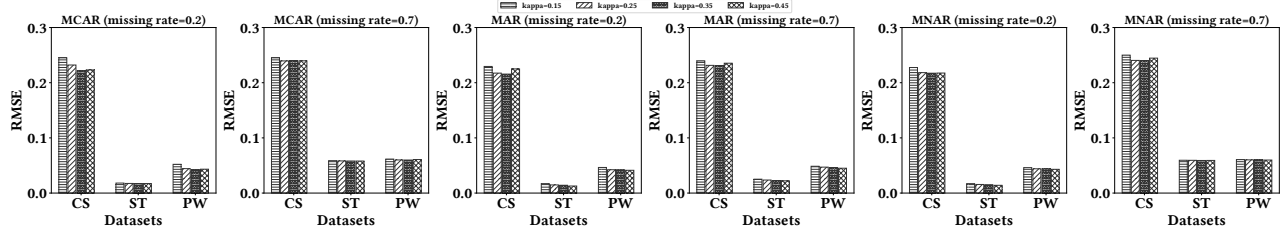


Figure 8: Experimental results of hyperparameter generalization on kappa.

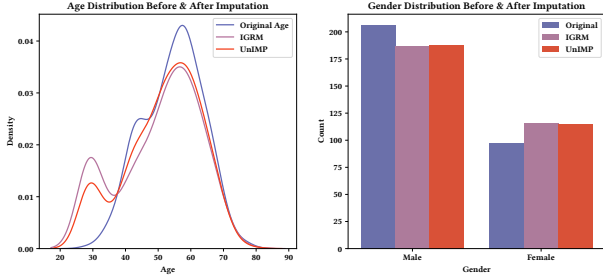


Figure 9: Bias evaluation.

Table 10: Bias evaluation

Model	Age-KL	Age-WL	GenderBias
IGRM	0.1896	4.0495	0.1254
UnIMP	0.1166	2.9043	0.1188

- Long. Text with more than 10 characters, covering attributes not classified as Short.
- Structured. Text following a fixed format, including: price from Buy (formatted as "\$"), and phone numbers from Restaurant (formatted as "xxx-xxx-xxxx").
- Colloquial. Unstructured text, including attributes not falling under Structured, such as restaurant names.

Moreover, we also evaluate 4 serialization strategies for the tabular data, including Attribute-Value Pairs, Sentence, Markdown and Dataframe, as follows:

- Attribute-Value Pairs. Concatenation of paired column names and cell values. For example: Name:"Obama"; Nation:"USA"
- Sentence. Rows are converted into sentences. For example, the Name is Obama, the Nation is USA.
- Markdown. Rows in the Markdown are line-separated, and columns are separated by "|". For example,

```
|   | Name | Nation |
|:--|:-----|-----|
| 0 | Obama | USA   |
```

- Dataframe. using the pandas format of Python to load the table. For example, `pd.DataFrame(Name=['Obama'], Nation=['USA'])`

## 8.5 Hyperparameter Generalization

In Figure 7 and Figure 8, we systematically investigate the generalizability of the hyperparameters  $\delta$  and  $\kappa$  under a variety of missing data conditions. Specifically, we examine performance across three widely studied noise mechanisms: MCAR, MAR, and MNAR. We also consider two representative levels of missingness: a moderate missing rate of 0.2 and a high missing rate of 0.7. For  $\delta$ , we evaluate five candidate values: 0.1, 0.5, 1, 4, and 10. For  $\kappa$ , we assess four values: 0.15, 0.25, 0.35, and 0.45. Imputation performance is measured using RMSE under each setting.

The results reveal a consistent trend in which RMSE first decreases slightly and then increases as the values of  $\delta$  or  $\kappa$  increase. Overall, our method demonstrates stable performance across a wide range of hyperparameter settings. These observations confirm the strong generalizability and robustness of our approach in handling diverse missing data scenarios.

## 8.6 Bias Evaluation

We conduct a bias evaluation using the Heart Disease dataset from the UCI repository [14]. The Heart Disease dataset contains 13 attributes (including both numerical features such as age and categorical features such as gender) along with a binary target variable indicating the presence or absence of heart disease in each patient. To simulate missingness, we randomly mask 20% of all the cell values under the MCAR setting. The missing values are then imputed using two methods: UnIMP and IGRM, where IGRM is the previous state-of-the-art model for imputation on small datasets. We focus our analysis on the age and gender columns to examine whether the imputation process introduces bias into sensitive attributes.

To assess the bias in the age column, which contains numerical data, we employ both Kullback-Leibler (KL) divergence and Wasserstein distance to measure the difference between the original and imputed distributions. Specifically, we discretize the age

values into  $k$  bins to obtain the normalized histograms of the original distribution  $P = \{p_1, p_2, \dots, p_k\}$  and the imputed distribution  $Q = \{q_1, q_2, \dots, q_k\}$ . The KL divergence is computed as:

$$\text{KL}(P \parallel Q) = \sum_{i=1}^k p_i \log \frac{p_i}{q_i},$$

where a lower value indicates closer alignment between the two distributions. Additionally, we compute the Wasserstein distance directly from the continuous age values, defined as:

$$\text{Wasserstein}(X, Y) = \inf_{\gamma \in \Gamma(X, Y)} \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|],$$

where  $X$  and  $Y$  denote the original and imputed age samples, and  $\Gamma(X, Y)$  is the set of all joint distributions with marginals  $X$  and  $Y$ . Together, these two metrics provide complementary insights into the distributional bias introduced during imputation.

To quantify the distributional shift in the imputed gender values, we compute the gender bias as the total variation distance between the original and imputed gender distributions. Given the original distribution  $P = \{p_1, p_2, \dots, p_k\}$  and the imputed distribution  $Q = \{q_1, q_2, \dots, q_k\}$  over  $k$  categories, the gender bias is defined as:

$$\text{GenderBias}(P, Q) = \sum_{i=1}^k |p_i - q_i|.$$

A lower value of GenderBias indicates that the imputed gender values better preserve the original distribution, suggesting reduced demographic distortion.

The results of IGRM and UnIMP are summarized in Figure 9 and Table 10. Overall, the findings indicate that UnIMP introduces less bias compared to IGRM. In terms of age bias, the original data is more concentrated in the 50–60 range, whereas the imputed values from both IGRM and UnIMP are more dispersed across the full range of age bins. Regarding gender bias, the original distribution is skewed toward male patients, while the imputed distributions from both methods appear more balanced between male and female categories. This suggests that both imputation methods tend to smooth out skewed or imbalanced distributions, shifting them toward more uniform representations. Nonetheless, UnIMP consistently demonstrates a closer alignment with the original data distribution, indicating better performance in preserving demographic fairness during imputation.