Generating-Filtering-Ranking: A Three-Stage MultiModal Data Augmentation Framework Under Partial Modality Missing

Zhirui Kuai*, Huan Zhang*, Yang Yang, Yiping Ma, Mingjing Huang, Ning Gui, Li Kuang†

School of Computer Science and Engineering, Central South University, Changsha, China
{kuaizhirui, z_huan, yang978, 8209240115, 8209220101, ninggui, kuangli}@csu.edu.cn

Abstract

Multimodal data significantly improves the performance of pretrained models, but its practical application is often limited by missing or incomplete data across modalities. There are two key challenges that existing methods of synthesizing missing data face: (1) semantic inaccuracies due to model hallucinations and (2) discrepancies in distribution preferences between generated and original data. To address these challenges, we propose a novel three-stage multimodal data augmentation framework (**GFR**), which **G**enerate, **F**ilter, and **R**ank missing modality data. Our framework leverages multimodal large models for diverse data generation, designs a scene graph matching-based filtering algorithm to ensure semantic consistency, and constructs a preference-aware ranking model to align the generated data with both the original distribution and task relevance. Our framework not only enhances semantic diversity and consistency in data generation but also effectively captures the implicit characteristics of the original dataset and the target model. We demonstrate the effectiveness of GFR across multiple datasets by testing different missing types and missing ratios.

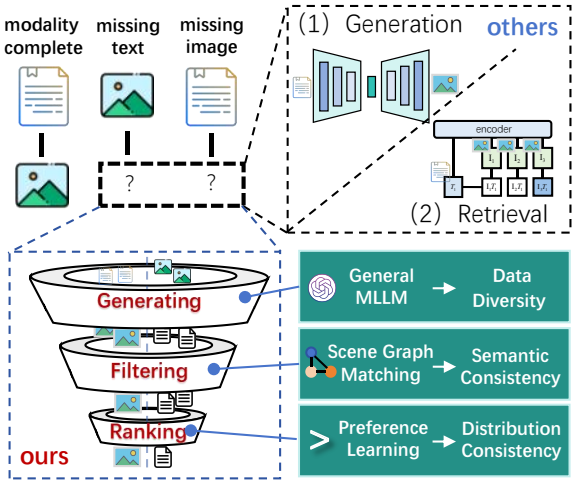

Figure 1: A comparison between GFR and other methods operating at the data-level.

Introduction

In recent years, multimodal data, such as text and images, has become increasingly integral to pretrained models, offering aligned and complementary information that significantly enhances performance across various downstream tasks (Pham et al. 2019; Li et al. 2021). However, real-world applications often encounter challenges due to missing or incomplete data across modalities, stemming from hardware limitations, privacy concerns, environmental interference, or data transmission issues. Studies have demonstrated that such partial modality missing can severely degrade model performance (Kuai et al. 2024; Lee et al. 2023b). For instance, Ma et al. observed that when only 30% of text modality data is available, the performance of multimodal models on movie classification tasks can drop by up to 43.6%, even underperforming models trained solely on image data (Ma et al. 2022).

Some works (Ma et al. 2022; Lee et al. 2023b) address this issue at the model-level, which typically requires architectural modifications and model retraining. More commonly, data-level methods (Kuai et al. 2024; Zhang et al. 2024) focus on reconstructing missing data using existing modality information before multimodal fusion, without modifying the subsequent model pipeline. As shown in Figure 1 (others), data-level methods can be mainly divided into two types: modality retrieval and modality generation. Modality retrieval methods infer missing data from existing modalities, but often fail due to their reliance on matching similar data in the multimodal space. More commonly, modality generation methods synthesize missing data using generative models, either through single-modality or unified multimodal generation approaches. While these methods hold greater potential for recovering modality-specific information, they face two key challenges:

Challenge 1. Semantic inaccuracies due to model hallucinations. (Gunjal, Yin, and Bas 2024; Lee et al. 2023a). Since generative models may learn incorrect features or patterns during training, the generated data may be semantically inconsistent with the original data. Such inaccurately generated data can mislead the training and inference processes of the target model, ultimately degrading its final performance.

*These authors contributed equally.
†Corresponding author.

Challenge 2. Discrepancies in distribution preferences between generated and original data. Generative models, trained on diverse datasets, may generate data that favors specific styles or structures that do not align with the distribution of the original data. For example, if the original dataset consists of realistic images, the generative model may produce anime-style images, and this style discrepancy can also affect the model's performance.

To address these challenges, we propose a novel three-stage multimodal data augmentation framework, **GFR** (**G**enerating-**F**iltering-**R**anking), inspired by recommender systems. As shown in Figure 1 (ours), our framework leverages multimodal large language models (MLLMs) for diverse data generation, designs a scene graph matching-based filtering algorithm to ensure semantic consistency, and constructs a preference-aware ranking model that ranks data based on distribution, style, and task relevance through preference learning. Specifically:

1) **Generating Stage:** We utilize general-purpose MLLMs, pre-trained on large-scale multimodal datasets, to generate missing modality data. These models, with their extensive world knowledge and strong generalization capabilities, can produce diverse and high-quality data candidates.

2) **Filtering Stage:** We filter the generated data to retain only those instances that maintain semantic consistency, effectively reducing hallucinations. To achieve this, we convert the generated data into scene graphs and design a scene graph matching algorithm to ensure precise semantic alignment with the existing modality data.

3) **Ranking Stage:** To further enhance the quality of the generated data, we rank the filtered data based on their preference consistency. To achieve this, we train a scoring model using partial-order pairs constructed from existing and generated data, ensuring that high-scoring data closely aligns with the original data in terms of distribution characteristics, style consistency, and task relevance.

The main contributions of this paper are as follows:

- To generate diverse and consistent data, based on the diverse candidate data generated by MLLMs, we design a scene graph matching algorithm to filter out data that is semantically inconsistent with the existing modality.
- To further retain data that aligns with the original data preferences, we propose a preference-aware ranking model to capture the implicit characteristics of both the original dataset and target model via preference learning.
- We conduct extensive experiments and ablation studies on multiple multimodal datasets, demonstrating the robust performance recovery of our proposed method.

Related Work

We classify data-level modality enhancement methods into two categories. The first category, **modality retrieval methods**, involves filling missing modality samples using zero/random values, directly copying data from similar instances, or retrieving matching samples to form complete modality samples (Liu et al. 2023; Malitesta et al. 2024; Sun et al. 2024; Zhi et al. 2024). The second category, **modality generation methods**, employs generative models—such as autoencoders, GANs, or diffusion models—to synthesize raw data for the missing modality (Mario Christoudias et al. 2010; Hinton and Zemel 1993; Sohn, Shang, and Lee 2014).

Modality Retrieval Methods Retrieval-based representation combination methods replace missing modality data by copying or averaging data from retrieved samples with the same classification (Yang et al. 2024). For instance, Modalmixup completes the training dataset by randomly supplementing samples with missing modalities of the same category (Yang et al. 2024). Frame-Repeat addresses frame loss in video streams by utilizing previous frames, and KNN or its variants are employed to retrieve the best matching samples for combination (Campos et al. 2015; Yang et al. 2024). Despite KNN-based methods performing relatively well, they are often hampered by high computational complexity, sensitivity to imbalanced data (Campos et al. 2015; Yang et al. 2024). In general, these methods are simple and easy to implement, but their performance is often unsatisfactory due to the limited ability to capture unique modality-specific information (Chen, Wang, and Qian 2020; Liu et al. 2023; Malitesta et al. 2024; Sun et al. 2024).

Modality Generation Methods With the development of deep learning, modality generation methods are now more effective, utilizing powerful representation learning and generative models to synthesize missing modalities. These methods can be divided into individual and unified generation types. Individual modality generation methods train a separate generative model for each potentially missing modality, such as Gaussian processes, Boltzmann machines, and autoencoders (Mario Christoudias et al. 2010; Hinton and Zemel 1993; Sohn, Shang, and Lee 2014). GANs (Arya and Saha 2021; Bischke et al. 2018; Gunasekar, Qiu, and Yang 2020) are widely used for modality generation due to their substantial improvement in image quality. Recently, diffusion models, such as the IMDer method, have further improved image quality by using available modalities as conditions to generate missing ones (Wang, Li, and Cui 2024). Unified data generation methods train a single model to generate all modalities simultaneously, such as the Cascade AE model (Tran et al. 2017). These methods can alleviate performance decline to some extent, but training a generator to produce high-quality missing modalities in scenarios with severely missing data remains a challenge (Tran et al. 2017; Zhang et al. 2024).

Different from the methods discussed above, our proposed GFR framework introduces specific core designs to enhance the quality of generated data. Firstly, in the filtering stage, we design a scene graph matching algorithm to enable data filtering, ensuring semantic consistency and diversity. Secondly, in the ranking stage, we build a preference-aware ranking model via preference learning on partial-order pairs constructed from original and generated modalities to capture the implicit characteristics. This distinguishes our objective from methods like Knowledge Bridger (KB) (Ke et al. 2025): while KB aims to match the semantic and knowledge structure between available and generated modalities, GFR further requires the generated modalities to align with the original dataset's distribution preferences.

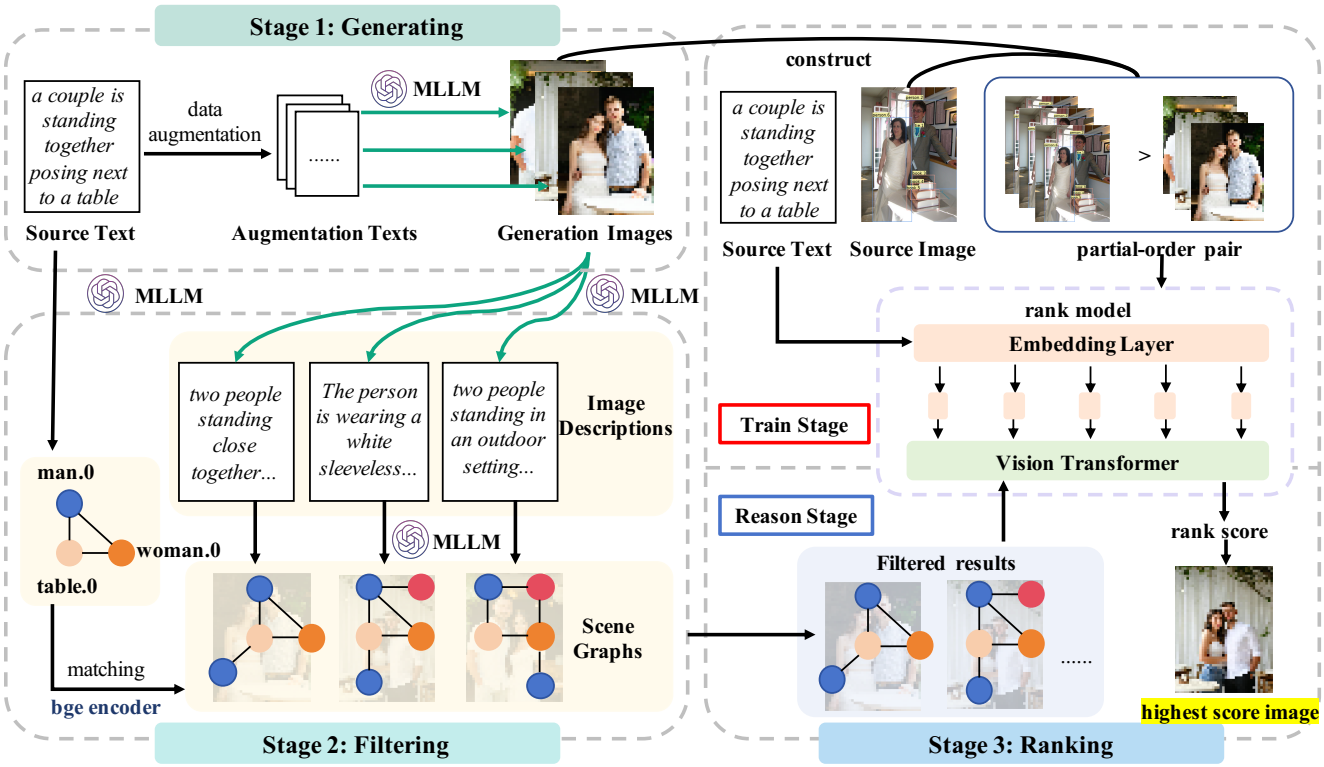

Figure 2: The overall framework of GFR, using missing image modality as an example.

GFR Framework

To alleviate the issue of missing modalities, we propose a novel three-stage data augmentation framework (GFR), which generates, filters, and ranks the missing modality data by leveraging large multimodal models, scene graphs, and preference learning, as shown in Figure 2. It consists of three main stages: 1) Generating stage, 2) Filtering stage, and 3) Ranking stage.

For missing image data, we use an existing multimodal model to generate images conditioned on text in the generation stage. In this stage, the model uses its learned world knowledge to generate a diverse range of images, which can include information not present in the original text, increasing the chance of recovering modality-specific information. In the filtering stage, we use a scene graph generation model to convert the generated images into scene graphs. Then, we design a scene graph matching algorithm to compare generated scene graphs with the graph structure extracted from the original text through entity and relationship extraction, thus filtering out images that are semantically inconsistent with the original text. This step enhances the semantic consistency and diversity of the generated image candidate set. Finally, we construct a ranking model to learn the implicit features of the original dataset and target model through partial order data and score the candidate-generated images. The higher the score, the more the image matches the implicit characteristics of the original dataset and target model, with the highest-scoring image selected as the final generated image. GFR can also recover missing text data using the same process as for missing images. The details of each section are provided below.

MLLM based Candidate Generating

In the generating stage, our goal is to leverage the model's world knowledge to generate diverse images, enhancing the recovery of modality-specific information. We selected two open-source models, Stable Diffusion XL (Podell et al. 2023) for text-to-image generation and Qwen2.5-VL-7B (Wang et al. 2024) for image-to-text generation. The input data includes existing modality data and prompt words. We designed a common prompt for generating missing image modalities from text and missing text modalities from images, combined with data augmentation methods to generate diverse data while maintaining semantic consistency. The prompt includes descriptive information about the original dataset, the classification task, instructions for generating data, and an example (only for image-to-text), helping the model better understand the target task and improve the generated data quality. The detailed generation process is shown in Figure 3. Ultimately, we obtain k candidate samples.

Scene Graph Matching based Candidate Filtering

In the filtering stage, our goal is to filter out generated data that contradicts the original modality data using the scene graph generation model and graph comparison, thereby enhancing the semantic consistency and diversity of the generated data candidate set. Scene graphs represent text and

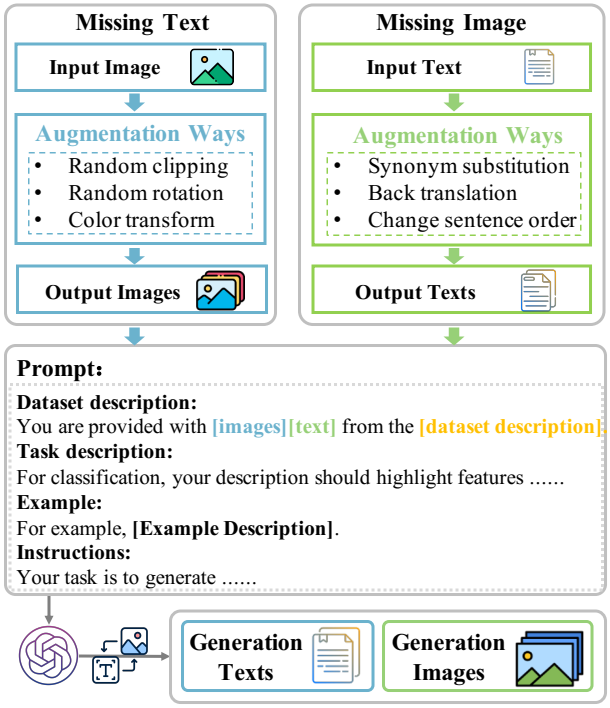

Figure 3: MLLM based Candidate generation.

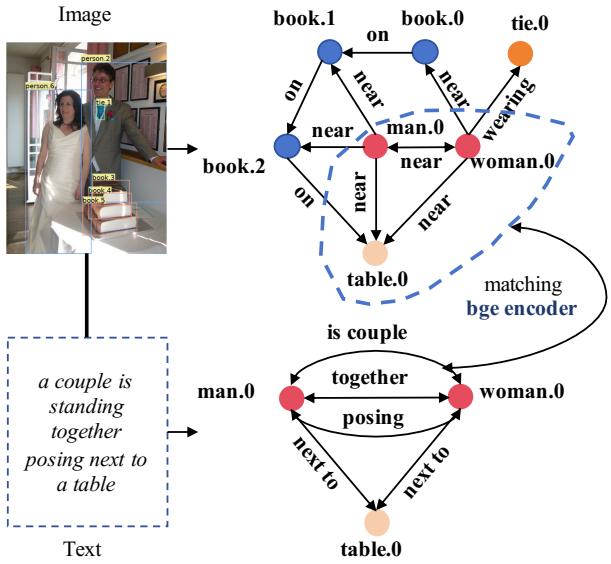

Figure 4: An example of scene graphs for an image-text pair.

image data as graph structures, allowing the filtering process for both types of generated data to be handled in the same way, and increasing the interpretability of the process. We use a unified method to generate scene graphs for both existing modality data and all generated candidates. Figure 4 shows an example of generating scene graphs for a text-image pair.

To enhance the semantic consistency and diversity of the generated data candidate set, we use the aforementioned scene graphs for graph matching-based filtering. For missing image data, we check whether each generated image scene graph G_i contains all entity nodes of the original text scene graph G_{target} and ensure that there is at least one similar relationship between the same nodes. If G_i meets the requirements, it is added to G_{filtered}. After ensuring node matching, we further enhance diversity. For each scene graph in G_{filtered}, we extract nodes and edges that exceed those in G_{target} (i.e., $extra$). If this $extra$ has not appeared in previous scene graphs, the scene graph is considered diverse and added to G_{diverse}. Finally, the list of scene graphs G_{diverse}, processed by subgraph matching and diversity processing, is returned as the final list of retained scene graphs.

One of the main challenges in the filtering process is establishing correspondence between nodes in different scene graphs and identifying conflicting relationships. For example, the text scene graph may contain a triple like (”jack”, ”hit”, ”roce”), while the image scene graph may use ”man” instead of ”jack”, complicating the matching process. To address these challenges, we use Qwen2.5-VL to standardize the triples for consistency. The model replaces entities with corresponding nouns, converts plurals to singulars, numbers entities, and removes non-noun entities (e.g., adjectives) and relationships that are neither verbs nor prepositions. This standardization ensures consistency in the matching process. Once standardized, we match entities and relationships using the following strategies: **String Matching**: We first attempt to match triples using string matching. **Semantic Similarity Matching**: For triples that cannot be directly matched (e.g., synonyms like ”next to” and ”near”), we combine them into phrases and calculate cosine similarity using the BGE (Chen et al. 2024) vector encoder. A similarity score above threshold is considered a successful match.

We also tested methods that directly use text similarity matching and large model judgment for the filtering process, the results of which will be discussed in the experimental section. Using these methods, we filter out data that does not match the semantics of the original text, further ensuring the semantic consistency and diversity of the generated data candidate set.

Preference Learning based Candidate Ranking

In the ranking stage, our goal is to construct a model to score and sort the candidate data from the generation stage, selecting the data that best aligns with the implicit characteristics of the original dataset and target model. Although the filtering stage enhances the semantic consistency of the generated data in terms of entities and content, the data distribution can significantly impact the performance of the target multimodal model. For example, if the original dataset consists of realistic images, the generative model might produce cartoon-style images. While these images may satisfy the filtering criteria for entities and content, they do not align with the implicit characteristics of the original dataset in style and features. Our objective is to enable the target model to achieve performance comparable to the base-

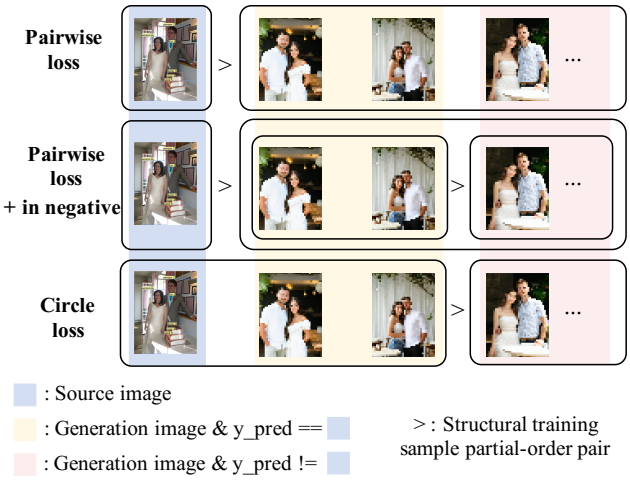

Figure 5: Constructing positive and negative samples.

line model trained with complete modality data, even with missing modalities. The ground truth for training the ranking model should align with the classification results of the target multimodal model for the corresponding text-image pairs, not with the actual ground truth labels. We refer to this as aligning with the implicit characteristics of the target multimodal model.

To achieve this, we focus on learning the implicit characteristics of both the dataset and target model using preference learning. We pair each generated sample with the original data from the other modality and feed the pairs into the target multimodal classification model for classification results. We experimented with three methods for constructing positive and negative samples, as shown in Figure 5. The ranking model aims to maximize the gap between the scores of positive and negative samples. With multiple positive and negative samples for each instance, we chose the Circle loss function to optimize the ranking model. Circle loss is designed to effectively handle the relationships among multiple positive and negative samples. The final loss function is defined as Equation 1:

$$\mathcal{L} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}\Bigg[(1-\delta_{ij})\cdot\Big(1-r_\theta(x_i,y_i)\cdot r_\theta(x_j,y_j)\Big)^2 + \delta_{ij}\cdot\Big(\big(1-r_\theta(x_i,y_i)\big)^2+\big(1-r_\theta(x_j,y_j)\big)^2\Big)\Bigg]. \quad (1)$$

K denotes the total number of samples. x_i and y_i represent the i-th original and generated data pairs, respectively. $r_\theta(x_i, y_i)$ is the score assigned by the model to the pair (x_i, y_i). The indicator function δ_{ij} equals 1 if $i = j$ and 0 otherwise. This loss function optimizes the model parameters to maximize the score gap between positive and negative samples, ensuring that positive samples receive much higher scores than negative ones. This approach captures the implicit characteristics of the original dataset and target model, leading to high-quality and relevant generated data.

Experiments

Experimental Setup

Datasets Our experiments were conducted on several widely used multimodal classification datasets. The primary comparisons were performed on two multi-label datasets: MM-IMDb (Arevalo et al. 2017) and IU-Xray (Demner-Fushman et al. 2015). We also tested our method on three additional multi-class/single-label datasets: Food101 (Wang et al. 2015), MVSA (Niu et al. 2016), and Pascal VOC (Hoiem, Divvala, and Hays 2009).

Evaluation Metrics For multi-label classification on MM-IMDb and IU-Xray, we use F1-Score as the primary metric. For classification tasks on Food101, MVSA, and Pascal VOC, we use Accuracy.

Baselines We compare GFR with several state-of-the-art (SOTA) methods for handling missing modalities, including MMIN(Zhao, Li, and Jin 2021), DiCMoR (Wang, Cui, and Li 2023), MPLMM (Guo, Jin, and Zhao 2024), and KB (Ke et al. 2025). Our GFR framework, along with MACP (Kuai et al. 2024) and Missing-aware-prompts (Lee et al. 2023b), was implemented and evaluated on the ViLT model. The results for SOTA methods are cited from KB, which may use different backbones. This does not affect the validity of our comparative analysis. Backbone is an intermediate step for feature extraction, but we focus on classification task under missing modalities. To eliminate backbone bias, we use Performance Recovery Capability ($\Delta = (M-N)/N$, where N denotes full-modality performance and M represents post-completion performance) as the primary metric.

Experimental Implementation

Input Processing Our datasets mainly consist of images and text. For missing text, an empty string is used to represent the missing input. For missing images, a uniform image with pixel values set to 1 is used as a placeholder. The parameter ϵ controls the missing rate. When one modality is missing, the input is reconstructed with ϵ incomplete data and $(1-\epsilon)$ complete data. Text modality preprocessing, including tokenizers and image data processing, follows the protocols established in the ViLT study.

Model Configuration During the generating phase, we generate 10 candidate samples for each missing data. We used the Stable Diffusion XL (Podell et al. 2023) model as the general image generator and Cheff (Weber et al. 2023) for Xray image generator. Meanwhile, we used the Qwen2.5-VL-7B (Wang et al. 2024) as the text generator. We use Qwen2.5-VL to extract scene graph triples. In the scene graph filtering process, we use bge-base-ev-v1.5 (Chen et al. 2024) to generate text vectors.

Training Details When training the baseline model, all settings were consistent with Missing-aware-prompts. The initial learning rate was 0.01, weight decay was 0.02, the learning rate was warmed up to 10% of the total training steps, and then linearly decayed to zero. It is also necessary to train the ranking model. We implemented a ViLT-structured ranking model that scores text-image pairs via linear projection of the final hidden state. During the experiment, each dataset was divided into a training set and a test

Dataset		MM-IMDb			IU-Xray	
Missing Rate		30% (Δ ↓)	50% (Δ ↓)	70% (Δ ↓)	30% (Δ ↓)	70% (Δ ↓)
CLIP-based	Baseline		56.2		57.0	
	Baseline+missing	51.8 (-7.8%)	50.3 (-10.5%)	47.2 (-16.0%)	49.1 (-12.6%)	31.5 (-44.7%)
	MPLMM	53.9 (-4.1%)	52.8 (-6.0%)	49.1 (-12.6%)	49.3 (-12.3%)	35.2 (-38.2%)
	MMIN	50.1 (-10.9%)	49.5 (-11.9%)	44.6 (-20.6%)	37.3 (-33.6%)	26.7 (-53.2%)
	DiCMoR	49.2 (-12.5%)	43.7 (-22.2%)	30.5 (-45.7%)	40.5 (-27.9%)	29.8 (-47.7%)
	KB	52.4 (-6.8%)	51.9 (-7.7%)	50.3 (-10.5%)	51.4 (-8.5%)	41.1 (-27.9%)
ViLT-based	Baseline		56.4		58.8	
	Baseline+missing	50.29 (-10.8%)	45.36 (-19.6%)	40.78 (-27.7%)	55.5 (-1.6%)	50.5 (-14.1%)
	GFR	**54.5 (-3.4%)**	**53.8 (-4.6%)**	**52.7 (-6.6%)**	**59.1 (+4.8%)**	**58.7 (-0.2%)**

Table 1: Performance comparison of models at varying missing image rates on MM-IMDb and IU-Xray datasets.

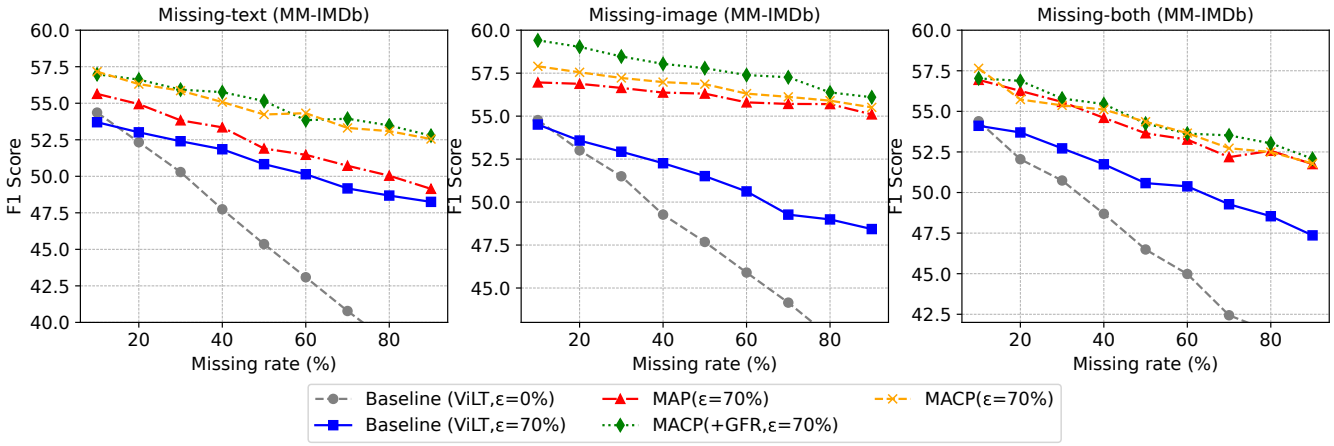

Figure 6: Performance comparison of our method and baselines across three missing data scenarios at varied deletion rates. ϵ=70% means that all models were trained with a 70% missing rate.

set. The ranking model was trained only on the training set and did not affect the testing of the missing modality task.

Main Results and Analysis

As shown in Table 1, all models exhibit performance degradation as the missing modality rate increases across both datasets. However, the GFR framework consistently recovers this lost performance. For example, on the IU-Xray dataset, under a 70% missing rate, GFR enables the ViLT model to recover a significant portion of its performance, achieving an F1-score notably higher than both the baseline with missing data removed and most other competing methods. This result underscores that GFR's effectiveness is not limited to a specific backbone architecture; it serves as a robust data augmentation pipeline that enhances model resilience in missing modality scenarios.

When models are retrained on data completed by GFR, their advantages are amplified, demonstrating that the high-quality data it generates significantly benefits model training. Figure 6 shows a more challenging setup, where all models (GFR, Missing-aware-prompts, and MACP) are based on a unified ViLT baseline and retrained on the training set processed by their respective methods. The model trained on GFR-completed data consistently outperforms all other methods across all missing rates. This provides strong evidence that GFR not only generates semantically accurate samples but also produces a high-quality dataset whose distribution closely aligns with the original, offering a superior foundation for downstream tasks. Overall,*GFR framework effectively mitigates performance degradation caused by missing modalities, showing superior recovery compared to other SOTA methods.*

Ablation Studies

Table 2 shows a detailed component-wise analysis of the GFR framework. The data show that removing either the Filtering or Ranking stage results in a significant performance drop, proving the necessity of our three-stage design. Further analysis shows that in the filtering stage, Scene Graph Matching, which utilizes structured semantic information, is more effective than relying solely on Text Similarity. In the ranking stage, the model trained with Circle Loss, which handles complex partial-order relationships, outperforms both traditional Pairwise Loss and the absence of a

ranking strategy. These results validate that each component of GFR is both essential and highly effective. Overall, *the Filtering and Ranking stages are essential components of the GFR framework, with Scene Graph Matching and Circle Loss-based ranking being key strategies for optimal performance.*

Stage	Method	Missing Rate		
		30%	50%	70%
Filtering	Only rank	43.09	42.19	40.68
	Text sim	42.99	42.32	39.78
	LLM filter	43.20	42.19	41.07
	Scene graph	43.22	42.29	41.14
Ranking	Only filter	42.50	40.96	40.13
	Pairwise	43.81	42.40	40.56
	Pairwise*	42.79	41.36	40.68
	Circle loss	43.22	42.29	41.14

Table 2: Contributions of GFR's filtering and ranking stages.

Generalizability and Visualization

To validate the broad applicability of the GFR framework, we conducted experiments on additional benchmark datasets, where GFR demonstrated strong performance in recovering from missing modalities across various data domains and tasks. Table 3 presents the results of generalization tests on the Food101, MVSA, and Pascal VOC datasets, each posing unique challenges. The results, evaluated using Accuracy, affirm our central claim: whether the missing modality is image or text, GFR effectively bridges the performance gap caused by data incompleteness, bringing the model's performance closer to that of complete data. This shows that *GFR is not domain-specific, but a generalizable framework for multimodal data augmentation.*

Qualitative analysis through visualizations provides strong support for our quantitative results, clearly demonstrating the high fidelity and semantic coherence of the data generated by GFR. In Figure 7, we compare the ground-truth data, GFR-generated samples, and samples from a basic generative model. The images produced by GFR maintain high semantic consistency with the source modality and accurately reconstruct details, styles, and entity relationships. In contrast, simpler generative methods tend to produce hallucinations, such as missing key details, fabricating content, or generating stylistically inconsistent outputs. These visual examples provide an intuitive explanation for GFR's superior performance in our quantitative evaluations.

Limitations and Discussion

GFR is applicable to visual-language tasks that require cross-modal semantic alignment, but its design principles can be extended to other multimodal learning scenarios. Of course, GFR assumes semantic consistency, which may not fully apply to certain tasks with modality conflicts, such as multimodal satire and hate meme analysis (Kiela et al. 2020). Despite its limitations, our work remains of academic value and practical potential.

Dataset	Missing Type	Missing Rate	
		30 %(Δ ↓)	70% (Δ ↓)
Food101	No missing	82.43	
	Image	81.53 (-1.1%)	80.23 (-2.7%)
	Image+GFR	81.88 (-0.6%)	81.28 (-1.4%)
	Text	60.64 (-26.4%)	31.72 (-61.5%)
	Text+GFR	68.48 (-16.9%)	49.75 (-39.6%)
MVSA	No missing	70.25	
	Image	64.25 (-8.5%)	60.75 (-13.5%)
	Image+GFR	68.00 (-3.2%)	64.75 (-7.8%)
	Text	62.50 (-11.0%)	52.25 (-25.6%)
	Text+GFR	65.50 (-6.7%)	60.25 (-14.2%)
Pascal	No missing	84.00	
	Image	75.00 (-10.7%)	63.00 (-25%)
	Image+GFR	85.00 (+1.2%)	83.00 (-1.2%)
	Text	69.00 (-17.9%)	34.00 (-59.5%)
	Text+GFR	82.00 (-2.4%)	77.00 (-8.3%)

Table 3: GFR performance on additional datasets.

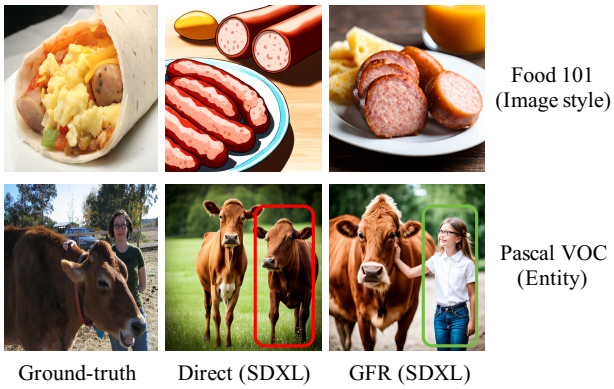

Figure 7: Visualization results of GFR performance.

Conclusion

In this study, we proposed a three-stage data augmentation framework (GFR) based on multimodal large models, scene graphs, and preference learning, aiming to recover missing modality data in multimodal datasets and enhance model performance in the case of missing modalities. Through extensive experiments and ablation studies on multiple classification datasets, we have verified the effectiveness of the GFR framework and deeply analyzed the impact of modality missing on model performance and the performance of GFR under different conditions. In addition, GFR is essentially a general framework for data augmentation and can be widely applied in the field of multimodal data augmentation. Overall, our research provides an effective solution to the problem of missing multimodal data and offers new ideas for future research.

Acknowledgments

This work has been supported by the National Key R&D Program of China under grant No.2022YFF0902500, the National Natural Science Foundation of China under grant No.62472447, Hunan Provincial Natural Science Foundation of China under grant No.2024JK2006, the Science and Technology Innovation Program of Hunan Province under grant No.2023RC1023. This work was carried out in part using computing resources at the High Performance Computing Center of Central South University.

References

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; et al. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Arya, N.; and Saha, S. 2021. Generative incomplete multi-view prognosis predictor for breast cancer: GIMPP. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4): 2252–2263.

Bischke, B.; Helber, P.; Koenig, F.; Borth, D.; and Dengel, A. 2018. Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, 1–6. IEEE.

Campos, S.; Pizarro, L.; Valle, C.; Gray, K. R.; Rueckert, D.; and Allende, H. 2015. Evaluating imputation techniques for missing data in ADNI: a patient classification study. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9-12, 2015, Proceedings 20*, 3–10. Springer.

Chen, J.; Xiao, S.; Zhang, P.; et al. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Chen, Z.; Wang, S.; and Qian, Y. 2020. Multi-Modality Matters: A Performance Leap on VoxCeleb. In *Interspeech*, 2252–2256.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Gunasekar, K.; Qiu, Q.; and Yang, Y. 2020. Low to high dimensional modality hallucination using aggregated fields of view. *IEEE Robotics and Automation Letters*, 5(2): 1983–1990.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. *arXiv preprint arXiv:2407.05374*.

Hinton, G. E.; and Zemel, R. 1993. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, 6.

Hoiem, D.; Divvala, S. K.; and Hays, J. H. 2009. Pascal VOC 2008 challenge. *World Literature Today*, 24(1): 1–4.

Ke, G.; He, S.; Wang, X.; Wang, B.; Chao, G.; Zhang, Y.; Xie, Y.; and Su, H. 2025. Knowledge Bridger: Towards Training-Free Missing Modality Completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25864–25873.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

Kuai, Z.; Zhou, Y.; Xie, Q.; and Kuang, L. 2024. Multi-Source Augmentation and Composite Prompts for Visual Recognition with Missing Modality. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 543–551.

Lee, S.; Park, S. H.; Jo, Y.; and Seo, M. 2023a. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Lee, Y.; Tsai, Y.; Chiu, W.; et al. 2023b. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14943–14952.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Liu, H.; Wei, D.; Lu, D.; Sun, J.; Wang, L.; and Zheng, Y. 2023. M3AE: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1657–1665.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18177–18186.

Malitesta, D.; Rossi, E.; Pomo, C.; Malliaros, F. D.; and Di Noia, T. 2024. Dealing with Missing Modalities in Multimodal Recommendation: a Feature Propagation-based Approach. *arXiv preprint arXiv:2403.19841*.

Mario Christoudias, C.; Urtasun, R.; Salzmann, M.; and Darrell, T. 2010. Learning to recognize objects from unseen modalities. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, 677–691. Springer.

Niu, T.; Zhu, S.; Pang, L.; and El-Saddik, A. 2016. Sentiment Analysis on Multi-View Social Data. In *MultiMedia Modeling*, 15–27.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint

representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6892–6899.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Sohn, K.; Shang, W.; and Lee, H. 2014. Improved multimodal deep learning with variation of information. *Advances in neural information processing systems*, 27.

Sun, Y.; Liu, Z.; Sheng, Q. Z.; Chu, D.; Yu, J.; and Sun, H. 2024. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110: 102454.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1405–1414.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Kumar, D.; Thome, N.; et al. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Wang, Y.; Cui, Z.; and Li, Y. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22025–22034.

Wang, Y.; Li, Y.; and Cui, Z. 2024. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

Weber, T.; Ingrisch, M.; Bischl, B.; and Rügamer, D. 2023. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Pacific-Asia conference on knowledge discovery and data mining*, 180–191. Springer.

Yang, Y.; Chen, H.; Chang, Z.; Xiang, Y.; Ye, C.; and Ma, T. 2024. Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In *Medical Imaging With Deep Learning*, 1006–1018. PMLR.

Zhang, Y.; Peng, C.; Wang, Q.; Song, D.; Li, K.; and Zhou, S. K. 2024. Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*.

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

Zhi, Z.; Liu, Z.; Elbadawi, M.; Daneshmend, A.; Orlu, M.; Basit, A.; Demosthenous, A.; and Rodrigues, M. 2024. Borrowing Treasures from Neighbors: In-Context Learning for Multimodal Learning with Missing Modalities and Data Scarcity. *arXiv preprint arXiv:2403.09428*.