CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2026.051

MIRAGE-SR: Missing-modality Invariant Representation And Generative Enrichment for multimodal Session-based Recommendation

Mai Thi Cam Nhung, and Nguyen Thai Nghe*

College of Information and Communication Technology, Can Tho University, Viet Nam

**Corresponding author (ntnghe@cit.ctu.edu.vn)*

Article info.

Received 17 Jun 2026
Revised 6 Sep 2026
Accepted 12 Oct 2026

Keywords

Graph neural network, missing modality, multimodal recommendation, session-based recommendation

ABSTRACT

Multimodal session-based recommendation commonly assumes that ID, visual, and textual features are available for every item, but this assumption is difficult to satisfy in real-world e-commerce data. This paper proposes MIRAGE-SR, a session-graph neural network architecture designed to maintain recommendation quality when one or more modalities are missing. The method combines two-stage modality imputation over an item-item co-occurrence graph, modality-invariant and modality-specific representation learning, and a missing-modality-aware late-fusion gate trained with curriculum modality dropout. MIRAGE-SR is evaluated on two subsets of Amazon Reviews 2023 under two settings: simulated modality cold-start on Grocery and naturally missing modalities on Cellphones. In the cold-start scenario, the configurations combining modality imputation and curriculum modality dropout achieve higher P@20 and MRR@20 than the zero-imputation configuration without curriculum modality dropout. On Cellphones, the two-stage GMI configuration achieves a P@20 of 26.24% and an MRR@20 of 14.61%, improving upon zero imputation by 3.91 and 1.53 points, respectively. The results indicate that the configurations combining imputation and training-time modality masking improve performance over the zero-imputation configuration; the current evaluation does not isolate the contribution of each component. These findings provide initial evidence of the potential of MIRAGE-SR for incomplete multimodal sequential recommendation.

1. INTRODUCTION

Session-based recommendation aims to predict the next item with which a user is likely to interact from a short sequence of actions in the current session, without relying on user identities or long-term interaction histories. In recent years, graph neural network models have demonstrated effectiveness in representing item-transition relationships within sessions. In particular, SR-GNN (Cam-Nhung et al., 2024) represents each session as a directed graph and employs graph-based message passing to learn session representations. This session-graph modeling approach is extended by multimodal models such as MMLF-GNNT and its session-level gated variant MMLF-GNNT*, both introduced in the same work (Cam-Nhung & Thai-Nghe, 2027). In these models, each item is represented by combining its ID embedding with content features extracted in advance from images, text, or other data sources.

Despite their promising performance, multimodal methods designed for complete inputs face practical data limitations. Malitesta et al. (2024) document missing visual and textual features in Amazon

product datasets, showing that incomplete product information is relevant to e-commerce recommendation. Removing affected items and imputing their missing features are two possible responses. Removal also discards associated interactions, whereas imputation retains the items but requires a choice of replacement features. Their study compares zero-vector and graph-aware imputation strategies and shows that retaining incomplete items through imputation can be preferable to dropping them.

In general, multimodal recommendation over user-item graphs, the missing-modality problem has received considerable attention through approaches such as single-branch architectures with shared weights, graph-based feature imputation, and early-fusion mechanisms designed specifically for incomplete multimodal data. However, these methods are primarily developed for user-item interaction graphs and do not fully exploit the transition structures inherent in session data. A session-based GNN architecture designed specifically to handle missing modalities in a principled manner therefore remains insufficiently explored. This gap calls for a method that can recover missing information, maintain stable representations under different missing-modality patterns, and adapt the fusion process to the availability of each modality.

To address this problem, we propose MIRAGE-SR (*Missing-modality Invariant Representation And Generative Enrichment for Multimodal Session-based Recommendation*), a multimodal graph neural network architecture for session-based recommendation under missing modalities. The main contributions are summarized as follows:

– We introduce MIRAGE-SR to address missing modalities in a principled manner by combining generative feature imputation with modality-invariant representation learning.

– We construct a global item-item co-occurrence graph from the training sessions and use it as an information-propagation channel for recovering missing features, thereby directly exploiting the relational structure specific to session data.

– We extend the session-level late-fusion gate into a missing-modality-aware fusion mechanism whose channel weights are adjusted for each session according to an availability mask.

2. RELATED WORK

Early sequential recommendation methods primarily learned representations from interaction sequences and item identities. GRU4Rec (Hidasi & Karatzoglou, 2018) and NARM (Cam-Nhung & Thai-Nghe, 2026) employ recurrent neural networks to encode within-session behavior, whereas SASRec (Ho & Nguyen, 2024) uses unidirectional self-attention to model sequential dependencies. Among graph-based approaches, session-based recommendation with graph neural networks (SR-GNN) (Cam-Nhung et al., 2024) established the foundation for representing each session as a directed graph. It employs a Gated Graph Neural Network (GGNN) to propagate information between items and subsequently aggregates item representations through a soft-attention mechanism. SRGNN-MF and NARM-MF combine session-based encoders with matrix factorization to incorporate latent item factors (Cam-Nhung & Thai-Nghe, 2026) to improve prediction accuracy. GNNT (Cam-Nhung & Thai-Nghe, 2025) extends session-based GNNs by integrating a Transformer, thereby improving the modeling of both local transitions and long-range item dependencies. Despite their improved recommendation performance, these methods rely primarily on interaction signals and ID embeddings and therefore do not fully exploit item-content information.

To overcome this limitation, multimodal sequential recommendation models augment ID representations with visual and textual features. Features from each modality can be projected into a shared latent space, processed by session encoders, and combined through weighted aggregation or late fusion. MMLF-GNNT integrates visual and textual features into a GNN-Transformer architecture (Cam-Nhung & Thai-Nghe, 2027). The same work also introduces MMLF-GNNT*, its session-level gated late-fusion variant, which dynamically learns modality weights for each session (Cam-Nhung & Thai-Nghe, 2027). The asterisk thus denotes a variant from that work, not a separate GNNT* model introduced here.

A complementary direction uses multimodal pre-training and transfer learning. MISSRec (Wang et al., 2023) employs a Transformer-based encoder-decoder to learn interest-aware sequence representations and dynamic fusion to construct user-adaptive item representations. Its relevance here is the use of multimodal information to improve sequence representation and

transferability, rather than its use as evidence for a missing-feature imputation mechanism. MIRAGE-SR instead focuses on integrating feature imputation and missing-modality-aware fusion into a session-graph architecture.

When inputs are incomplete, the treatment of missing features becomes an additional design choice beyond the encoding and fusion architecture. Malitesta et al. (2024) show that dropping incomplete items can weaken the benefits of multimodal recommendation and evaluate alternatives based on feature imputation. These findings motivate explicit handling of missing modalities; they do not establish a universal representation-bias mechanism or demonstrate degradation specifically for MMLF-GNNT and MMLF-GNNT*. Here, the question is how to incorporate imputation and availability-aware fusion into session-graph recommendation.

In general, multimodal recommendation over user-item graphs, the missing-modality problem has been addressed through several directions, including single-branch architectures with weights shared across modalities (Ganhör et al., 2024), generative imputation using diffusion models combined with counterfactual reasoning (Li et al., 2025), graph-based feature imputation, and early-fusion mechanisms specifically designed for incomplete multimodal data (Chen et al., 2026). These studies provide important insights into feature propagation, information recovery, and robust representation learning. However, they primarily operate on static user-item graphs, whereas session data exhibit distinctive item-transition and item-co-occurrence structures. In contrast, our approach constructs an item-item co-occurrence graph from sessions to propagate and recover missing features while integrating generative imputation, missing-modality awareness, and missingness-conditioned training within a unified session-based GNN architecture.

3. PROPOSED APPROACH

3.1. Problem formulation and notation

Given a session $s = (v_1, v_2, \ldots, v_n)$, the goal of the model is to predict the item $\hat{v}_{n+1}$ with which the user will interact next. Here, s denotes an ordered sequence of item interactions, not an encoded vector. Each item v has three modalities (ID, image, and text) with corresponding item-level representations x_v^{id}, x_v^{img} and x_v^{txt}. The ID representation is always available, whereas the availability of the visual and textual modalities is indicated by the mask $m_v = \left(m_v^{img}, m_v^{txt}\right) \in \{0,1\}^2$, where 1 denotes an available modality and 0 denotes a missing modality. The task is to accurately predict the next item when some items in the input session or candidate set satisfy $m_v \neq (1,1)$, while effectively exploiting the available modalities and minimizing the adverse effects of missing information.

3.2. Architecture overview

As illustrated in Figure 1, MIRAGE-SR consists of four main modules: (i) two-stage Generative Modality Imputation; (ii) a multimodal session-graph encoder that shares GGNN parameters across all channels; (iii) modality-invariant and modality-specific representation learning; and (iv) a missing-modality-aware late-fusion gate trained with curriculum modality dropout.

3.3. Two-stage generative modality imputation

GMI is designed to recover missing modality features through two consecutive stages. In the first stage, a global item-item co-occurrence graph is constructed from all training sessions. The edge weight between two items is determined by the frequency with which they co-occur or appear consecutively within the same session. For each missing modality, an initial representation is estimated by propagating features from neighboring items for which the corresponding modality is available. By directly exploiting session-level co-occurrence relationships, this graph can better capture substitutive and complementary relationships between items.

Stage 2 refines first-stage estimates using available modalities and session context, with masked features as reconstruction targets. Leakage prevention requires excluding these targets from all inputs. Residual MLP and reduced diffusion designs remain unverified; architecture, parameters, and training details require confirmation from implementation records. This functional description does not yet support exact reproduction.

The overall two-stage modality-imputation procedure is illustrated in Figure 2.

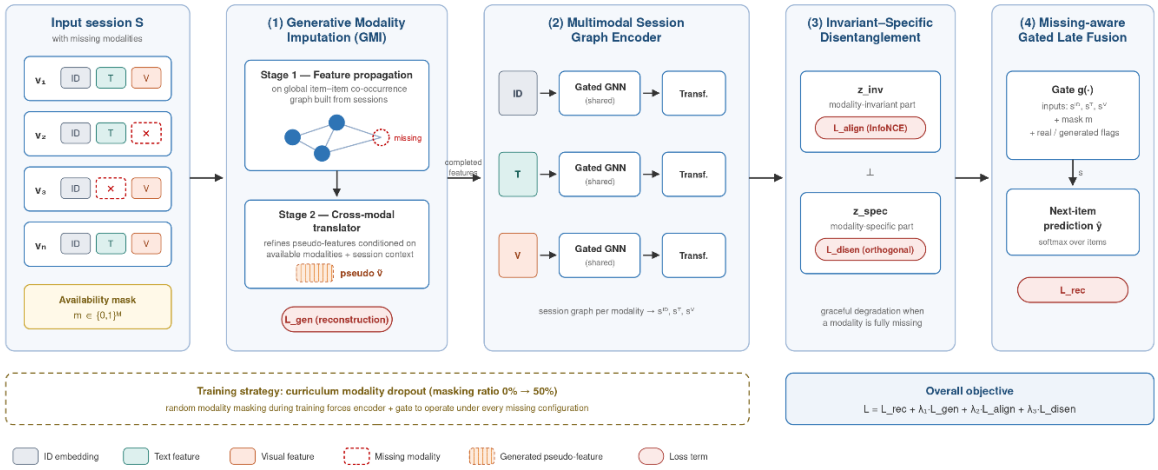


Figure 1. Overall architecture of the proposed MIRAGE-SR model

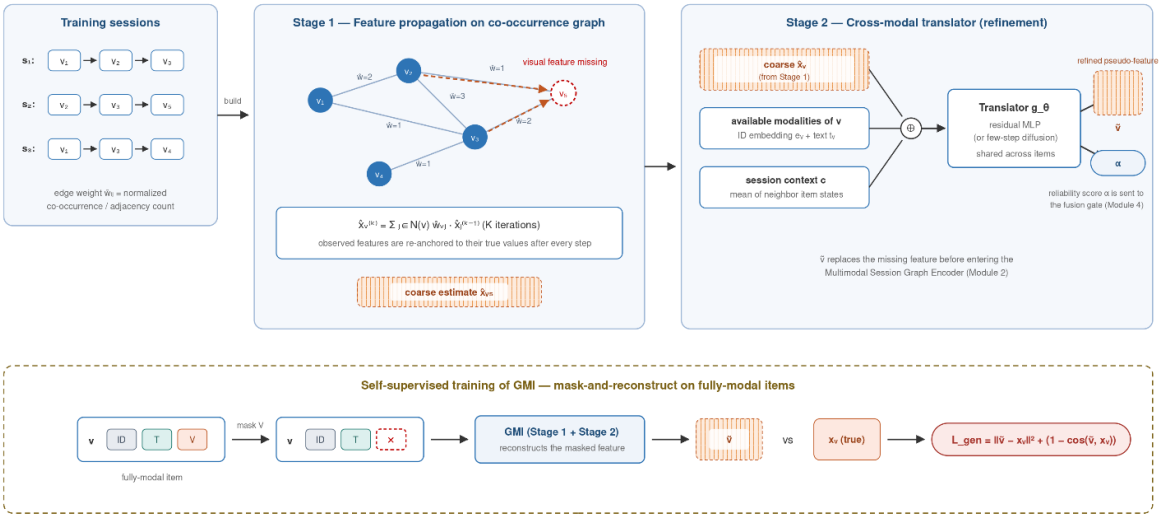


Figure 2. The proposed two-stage modality-imputation procedure

3.4. Multimodal session-graph encoder inherited from MMLF-GNNT

Following modality imputation, the ID, textual, and visual channels are processed by the multimodal session-graph encoder inherited from MMLF-GNNT. Each channel is passed through a Gated Graph Neural Network (GGNN) with parameters shared across modalities to aggregate local transition information over the session graph. The resulting node representations are subsequently fed into a Transformer block to capture long-range dependencies within the interaction sequence. This process produces three channel-specific session vectors, denoted by s_{id}, s_{img} *and* s_{txt} . They encode the same interaction sequence s defined in Section 3.1, using the ID, visual, and textual channels, respectively. They are learned session-level representations, distinct from both the input sequence s and the item-level features x_v^m. Shared GGNN parameters do not imply identical session vectors, because the channel inputs differ.

3.5. Missing-modality-aware late-fusion gate

The fusion mechanism extends the session-level late-fusion gate of MMLF-GNNT* to account for missing modalities. In the current implementation, its inputs are the three encoded session vectors produced by the encoder in Section 3.4, s_{id} , s_{img} and s_{txt} , together with the availability mask m_{v_n} of the last item in the session. The fusion-weight vector is computed as

$$g = softmax\,(W_{gate}\,[s_{id};\, s_{img};\, s_{txt};\, m_{v_n}] \quad (1)$$

where $g = (g_{id}, g_{img}, g_{txt})$ represents the contribution of each channel. The final logits are obtained through the linear combination

$$l = (g_{id} l_{id} + g_{img} l_{img} + g_{txt} l_{txt}) \quad (2)$$

where l_{id}, l_{img} and l_{txt} are the prediction logits of the ID, visual, and textual channels, respectively. When the last item in the session lacks an image, the gate can reduce g_{img} because that channel relies on a less reliable imputed feature and can increase the weights of the remaining channels. When the availability mask is not used, the fusion weights become

$$g = softmax(\alpha) \quad , \quad \alpha \in R^3 \quad (3)$$

where α contains three learned scalar parameters, preserving compatibility with the previous late-fusion architecture. A current limitation is that the gate only uses the mask of the last item, does not model the missing-modality ratio over the entire session, and lacks an explicit provenance indicator that distinguishes observed from imputed features. Training-time masking follows the curriculum described below; the maximum masking probability is a cap, not a fixed rate throughout training.

3.6. Training with modality dropout

Training only on complete multimodal data may prevent the encoder and fusion gate from adapting to missing-modality configurations encountered at deployment, especially in modality cold-start scenarios. To address this issue, selected item modalities are randomly masked during training with a probability $p(e)$ that increases linearly from 0 to p_{max} during the warm-up period, where the default value is $p_{max} = 0.5$. For masked modalities, the original features are replaced with features estimated by GMI to simulate inference-time missingness and prevent information leakage from the original representations. At test time, we retain the observed availability masks and no longer apply the curriculum strategy.

The cap $p_{max} = 0.5$ is treated as a heuristic setting rather than an empirically optimized value. At the level of a masking decision, it gives equal probabilities of masking and retaining an eligible modality at the end of warm-up, allowing training to include both observed and imputed features. This is a design rationale, not evidence that the cap is optimal. The masking probability during training is distinct from the fraction of test items masked in the Grocery scenario. No sensitivity analysis of this cap is reported; the warm-up duration and precise masking granularity also require documentation for reproduction.

3.7. Modality-invariant and modality-specific representation learning

To improve representation robustness under different missing-modality configurations, each modality representation is decomposed into a shared component and a modality-specific component. The shared component is expected to preserve semantic information that is consistent across modalities, whereas the modality-specific component captures information unique to each channel. An InfoNCE contrastive loss aligns the shared components associated with the same item or session, while an orthogonality constraint reduces redundancy between the shared and modality-specific components. Consequently, when a modality is entirely missing or its imputed feature is unreliable, shared components learned from the remaining modalities can still preserve the core preference signal of the session.

3.8. Training objective

The overall training objective is defined as

$$L = (L_{rcc} + \gamma_1 L_{gen} + \gamma_2 L_{align} + \gamma_3 L_{disen}) \quad (4)$$

where L_{rcc} is the cross-entropy loss for next-item prediction, L_{gen} is the modality-reconstruction loss, L_{align} is the cross-modal contrastive loss, and L_{disen} is the orthogonality loss between the shared and modality-specific components. The coefficients γ_1, γ_2 and γ_3 are selected through grid search on the validation set.

4. EXPERIMENTAL RESULTS

In the MIRAGE-SR experiments, the model is trained with a hidden size of 100, a learning rate of 0.001, a batch size of 100, and one GGNN propagation step. Early stopping is applied with a patience of 10 epochs; training terminates when neither P@20 nor MRR@20 improves for 10 consecutive epochs. These two metrics are used as the primary evaluation criteria. Using the same metrics as MMLF-GNNT and MMLF-GNNT* does not, by itself, establish comparable experimental conditions.

4.1. Data and feature preparation

The experimental workflow builds on our earlier MMLF-GNNT and MMLF-GNNT* study (Cam-Nhung & Thai-Nghe, 2027), extending the

multimodal session-graph approach to missing-modality conditions. This methodological inheritance does not imply identical run configurations or hardware; the settings reported for MIRAGE-SR are stated explicitly here.

We use the Cell Phones and Accessories (Cellphones) and Grocery and Gourmet Food (Grocery) subsets of Amazon Reviews 2023, a dataset of user reviews and item metadata released by McAuley Lab (Hou et al., 2024). The 5-core rule filters the user--item interaction data so that each remaining user and item has at least five interactions; users and items below this threshold are removed iteratively until the condition holds (McAuley Lab, 2026). Interactions are then divided into sessions with a maximum inactivity interval of 24 hours and split chronologically, with the most recent 10% of sessions used for testing. This session-based split should not be confused with the ready-made benchmark splits distributed with the dataset.

Textual features are obtained using *all-MiniLM-L6-v2*, a Sentence Transformers model that maps sentences and paragraphs to 384-dimensional embeddings (Sentence Transformers, 2026). Visual features are extracted using ResNet50, the 50-layer residual convolutional architecture of He et al. (2016). Principal component analysis (PCA) projects features onto orthogonal directions of greatest variance to reduce dimensionality (Jolliffe & Cadima, 2016); each modality is reduced to 100 dimensions in the reported setup. Items lacking modalities are retained and identified through availability masks. Exact feature-extractor checkpoints, text/image preprocessing options, and the data used to fit PCA still require documentation from the original implementation records.

4.2. System configuration and code availability

The recommender settings reported above specify the hidden size, learning rate, batch size, GGNN propagation steps, and early-stopping patience. The CPU/GPU models and counts, memory, operating system, software versions, and measured training times are not documented in the available records for this report. No runtime estimate is inferred from the earlier MMLF-GNNT study, because its execution environment and workload do not match those of MIRAGE-SR. A verified MIRAGE-SR code repository link is not available for inclusion in this revision; the cited dataset and model resources are third-party resources, not the implementation of MIRAGE-SR. Documenting the system configuration, per-epoch and total training times with preprocessing costs distinguished, and the code version remains necessary for full reproducibility.

Cellphones contain naturally missing visual and textual features for 14.7% and 14.6% of its items, respectively, and most items without images also lack text. Cellphones is therefore used to evaluate the model under naturally missing modalities. In contrast, Grocery has a very low natural missingness rate and is used as a controlled environment for simulated missing-modality scenarios with an adjustable rate ρ. Dataset statistics are reported in Table 1.

Table 1. Statistics of the two datasets used in the experiments

Dataset	Cellphones	Grocery
Number of items	101,419	126,689
Missing image modality	14,877 (14.7%)	52 (0.04%)
Missing text modality	14,860 (14.6%)	2 (0.002%)



Figure 3. Experimental scenarios

As illustrated in Figure 3, MIRAGE-SR is evaluated under two complementary experimental scenarios. The first scenario uses Grocery to simulate modality cold-start: the source training data are treated as complete before configuration-specific modality dropout is applied, while visual or textual features of ρ% of test items are masked under a mixed configuration. This setting evaluates the model's ability to adapt to the discrepancy between training and deployment conditions and examines the performance of configurations combining imputation, modality dropout, and missing-modality-aware fusion. The second scenario retains the naturally incomplete Cellphones data without removing items that lack images or text, thereby evaluating MIRAGE-SR under real missing-modality conditions.

The result tables compare three MIRAGE-SR configurations rather than isolated component ablations: **MIRAGE-SR-0** is the control variant that replaces missing features with zero vectors and does not use GMI or curriculum modality dropout. **MIRAGE-SR-GMI1** uses the first stage of Generative Modality Imputation, namely feature propagation over the item-item co-occurrence graph, together with curriculum modality dropout. **MIRAGE-SR-GMI2** extends MIRAGE-SR-GMI1 with the second-stage cross-modal translator and modality-invariant and modality-specific representation learning. These names are used consistently in Tables 2 and 3.

4.3. Comparison protocol and scope

MMLF-GNNT and MMLF-GNNT* scores are reference results, not controlled comparisons, because matched data, missing-modality masks, and training conditions have not been verified. They therefore do not establish direct superiority or relative robustness. On Grocery, complete training data refers to source data before modality masking. MIRAGE-SR-0 excludes curriculum modality dropout, whereas MIRAGE-SR-GMI1 and MIRAGE-SR-GMI2 include it; MIRAGE-SR-GMI2 also adds cross-modal refinement and representation learning. The results thus compare combined configurations rather than isolate component contributions. Matched evaluation protocols and component-wise ablations remain necessary.

4.4. Scenario 1: Modality Cold-Start and Recovery on Grocery

This scenario evaluates the model's adaptability to modality cold-start. The source training data are treated as complete before configuration-specific training-time modality masking is applied, whereas 50% of the items in the test set have modalities masked under a mixed configuration. Because Grocery has a very low natural missingness rate, it serves as a controlled environment for examining the performance of MIRAGE-SR configurations with incomplete test inputs. Results are reported in Table 2.

MIRAGE-SR-0 obtains a P@20 of 31.47% and an MRR@20 of 26.15%, the lowest values among the MIRAGE-SR variants. This observation concerns the zero-imputation configuration without curriculum modality dropout; it does not isolate the effect of the imputation strategy alone. When the first GMI stage and curriculum modality dropout are applied, MIRAGE-SR-GMI1 increases P@20 to 43.05% and MRR@20 to 32.42%, absolute improvements of 11.58 and 6.27 points over MIRAGE-SR-0. These results show that the configuration combining feature propagation and curriculum modality dropout improves performance over MIRAGE-SR-0; however, the contributions of these two components cannot be isolated from this comparison.

Table 2. MIRAGE-SR results on Grocery (complete source training data before masking; mixed modality masking for 50% of test items)

Methods	P@20	MRR@20
MMLF-GNNT	44.56	35.51
MMLF-GNNT*	45.71	39.21
Our proposed result		
MIRAGE-SR-0	31.47	26.15
MIRAGE-SR-GMI1	43.05	32.42
MIRAGE-SR-GMI2	42.99	35.60

With cross-modal refinement and modality-invariant and modality-specific representation learning introduced together, MIRAGE-SR-GMI2 achieves a P@20 of 42.99% and an MRR@20 of 35.60%. Compared with MIRAGE-SR-GMI1, P@20 remains nearly unchanged, whereas MRR@20 increases by 3.18 points. These values describe higher MRR@20 for the extended configuration, without establishing statistical significance. Because this configuration adds both cross-modal refinement and representation learning, the improvement cannot be attributed solely to the second imputation stage. The MMLF-GNNT and MMLF-GNNT* values are retained only as reference results and are not used to infer relative robustness to missing modalities.

4.5. Scenario 2: Naturally Missing Modalities on Cellphones

This scenario evaluates MIRAGE-SR directly on naturally missing modalities in Cellphones without removing items that lack images or text. Unlike the simulated setting on Grocery, missingness in Cellphones is described as highly correlated, with most items without images also lacking textual information. This setting reflects naturally incomplete product data, but within-item co-missingness alone does not establish the availability of useful neighbors for graph propagation. Results are presented in Table 3.

Among the proposed variants, MIRAGE-SR-0 obtains a P@20 of 22.33% and an MRR@20 of

13.08%, providing a reference for the configuration without GMI or curriculum modality dropout. The configuration combining feature propagation over the item-item co-occurrence graph with curriculum modality dropout achieves a P@20 of 23.39% and an MRR@20 of 13.15%, absolute gains of 1.06 and 0.07 points. This modest improvement provides a positive signal for the configuration combining the first GMI stage and curriculum modality dropout, although the increase in MRR@20 is small. The effect of feature propagation alone cannot be isolated.

Table 3. MIRAGE-SR results under naturally missing modalities on Cellphones

Methods	P@20	MRR@20
MMLF-GNNT	27.81	14.52
MMLF-GNNT*	28.38	16.34
Our proposed result		
MIRAGE-SR-0	22.33	13.08
MIRAGE-SR-GMI1	23.39	13.15
MIRAGE-SR-GMI2	26.24	14.61

MIRAGE-SR-GMI2 achieves the best results among the MIRAGE-SR variants, with a P@20 of 26.24% and an MRR@20 of 14.61%. Compared with MIRAGE-SR-GMI1, it improves P@20 by 2.85 points and MRR@20 by 1.46 points; compared with MIRAGE-SR-0, the improvements are 3.91 and 1.53 points, respectively. The reported P@20 and MRR@20 values for MIRAGE-SR-GMI2 are higher than those for the other two MIRAGE-SR configurations on Cellphones. However, cross-modal refinement and representation learning are introduced together, so their individual contributions cannot be isolated. The baseline results serve only as references because comparable missing-modality and training conditions have not been established.

4.6. Interpreting the cross-dataset difference

The P@20 gain of MIRAGE-SR-GMI1 over MIRAGE-SR-0 is 11.58 percentage points on Grocery and 1.06 points on Cellphones. Correlated missingness is a possible explanation, not a quantitatively verified cause of this difference. The marginal missingness rates in Table 1 do not determine the proportion of modality-complete neighbors or the availability of the target modality among neighbors of an incomplete item. These neighborhood statistics have not been measured in the present study. A follow-up analysis should report both proportions under the actual propagation masks, separately for items missing images and text, together with the fraction having no observed target-modality neighbors. This distinction matters because first-stage propagation requires the target modality, not necessarily all modalities, to be observed in a neighbor. Differences in graph structure and evaluation settings may also contribute, and the comparison changes both imputation and curriculum modality dropout. The current results therefore do not isolate a neighborhood-availability effect.

4.7. Limitations of configuration-level comparisons

The reported comparisons are not a complete component-wise ablation study. MIRAGE-SR-GMI2 introduces the cross-modal translator together with modality-invariant and modality-specific representation learning, including the alignment and orthogonality objectives. Consequently, we cannot assign its observed gains to the translator, representation learning, either auxiliary objective, or their interactions individually. The results also do not establish that each added component is necessary or that the components provide synergistic benefits. Similarly, the comparison between MIRAGE-SR-0 and MIRAGE-SR-GMI1 changes both imputation and curriculum modality dropout. The current evidence therefore supports only comparisons of the combined configurations.

A controlled follow-up should compare configurations with and without the cross-modal translator and with and without modality-invariant and modality-specific representation learning, while holding first-stage propagation, curriculum modality dropout, the encoder, and the fusion gate fixed. Separate removal of the alignment and orthogonality losses would then examine their contributions. These comparisons should use matched data splits and missing-modality masks, consistent training budgets and checkpoint-selection rules, and multiple random seeds. These ablations have not been conducted in the present study and remain a limitation and a priority for future work.

4.8. Statistical uncertainty

The reported results are point estimates; no across-run means, standard deviations, confidence intervals, or significance tests are provided. Their variability across random seeds and statistical significance therefore cannot be assessed from the current report. All performance differences are interpreted descriptively, and small gaps should not

be taken as evidence of reliable superiority. Evaluation across multiple seeds under matched conditions, with uncertainty estimates and appropriate paired significance tests, remains future work; no additional runs or statistical tests were performed for this revision.

CONCLUSION

This paper proposes MIRAGE-SR, a session-based GNN architecture that combines two-stage modality imputation over an item--item co-occurrence graph, a GGNN-Transformer encoder, modality-invariant and modality-specific representation learning, and a missing-modality-aware fusion gate. Experimental results under simulated modality cold-start and naturally missing modalities provide initial evidence that the combined MIRAGE-SR configurations improve ranking performance over the zero-imputation configuration. These descriptive comparisons establish neither statistical significance nor the individual contributions, necessity, or interactions of the components introduced together in MIRAGE-SR-GMI2. Future work should prioritize controlled component-wise ablations of cross-modal refinement and representation learning, followed by separate ablations of the alignment and orthogonality losses, under matched data and masking conditions and consistent training and checkpoint-selection protocols. Multiple random seeds and statistical significance tests, together with simulated correlated missingness in which multiple modalities are absent from the same item, are needed to strengthen the evaluation. The fusion gate should also be enhanced to capture missingness across the entire session and explicitly distinguish observed features from imputed ones. In addition, tuning the loss coefficients, reducing the computational cost of the cross-modal translator, and evaluating the model on more datasets and modality types will provide stronger evidence of the stability, effectiveness, and generalizability of MIRAGE-SR.

ACKNOWLEDGMENT

This study was funded by Can Tho University, Code: T2026-85.

CONFLICT OF INTEREST

There are no competing interests.

REFERENCES

Cam-Nhung, M. T., & Thai-Nghe, N. (2025). Graph neural networks with transformer for sequential recommendation. *SN Computer Science, 6*(7), 792.

Cam-Nhung, M. T., & Thai-Nghe, N. (2026). A fusion model of neural attention mechanism and matrix factorization for sequential recommendation systems. *In The 14th Conference on Information Technology and its Applications* (pp. 203–215). Springer Nature Switzerland.

Cam-Nhung, M. T., & Thai-Nghe, N. (2027). Mmlf-gnnt: A multimodal late-fusion graph neural network with transformer for sequential recommendation. *In Advances and Trends in Artificial Intelligence. Theory and Applications* (pp. 238–249). Springer Nature Singapore.

Chen, L., Yi, P., Hou, X., Yang, C., & Cai, X. (2026). When modalities go missing: Early fusion for multimodal recommendation. *In Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining* (pp. 1094–1098).

Ganhör, C., Moscati, M., Hausberger, A., Nawaz, S., & Schedl, M. (2024). A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios. *In Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 380–390).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hidasi, B., & Karatzoglou, A. (2018). Recurrent neural networks with top-k gains for session-based recommendations. *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 843–852).

Ho, T. D. H., & Nguyen, S. T. T. (2024). Self-attentive sequential recommendation models enriched with more features. *In Proceedings of the 2024 8th International Conference on Deep Learning Technologies* (pp. 49–55).

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). Bridging language and items for retrieval and recommendation. arXiv preprint arXiv:2403.03952.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, *374*(2065), 20150202.

Li, J., Wang, S., Zhang, Q., Yu, S., & Chen, F. (2025). Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. *In Proceedings of the ACM on Web Conference 2025* (pp. 2787–2798).

Malitesta, D., Rossi, E., Pomo, C., Di Noia, T., & Malliaros, F. D. (2024). Do we really need to drop

items with missing modalities in multimodal recommendation*? In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 3943–3948).

McAuley Lab. (2026). Amazon Reviews 2023: Pure IDs (5-core). https://amazon-reviews-2023.github.io/data_processing/5core.html Sentence Transformers. (2026). all-MiniLM-L6-v2: Model card. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Cam-Nhung, M., Thuy Anh, N., & Thai-Nghe, N. (2024). Sequential recommendation using graph neuron networks. *In Future Data and Security Engineering* (pp. 66–79). Springer Nature Singapore.

Wang, J., Zeng, Z., Wang, Y., Wang, Y., Lu, X., Li, T., Yuan, J., Zhang, R., Zheng, H. T., & Xia, S. T. (2023). Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation*. In Proceedings of the 31st ACM International Conference on Multimedia* (pp. 6548–6557).