

Well Begun is Half Done: Training-Free and Model-Agnostic Semantically Guaranteed User Representation Initialization for Multimodal Recommendation

Jinfeng Xu

jinfeng@connect.hku.hk
The University of Hong Kong
HongKong SAR, China

Zheyu Chen

zheyu.chen@bit.edu.cn
Beijing Institute of Technology
Beijing, China

Shuo Yang

shuoyang.ee@gmail.com
The University of Hong Kong
HongKong SAR, China

Jinze Li

lijinze-hku@connect.hku.hk
The University of Hong Kong
HongKong SAR, China

Hewei Wang

heweiw@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jianheng Tang

tangentheng@gmail.com
Peking University
Beijing, China

Wei Wang

weiwang@mpu.edu.mo
Macao Polytechnic University
Macao SAR, China

Xiping Hu

huxp@bit.edu.cn
Beijing Institute of Technology
Beijing, China

Edith C. H. Ngai*

chngai@eee.hku.hk
The University of Hong Kong
HongKong SAR, China

Abstract

Recent advancements in multimodal recommendations, which leverage diverse modality information to mitigate data sparsity and improve recommendation accuracy, have gained significant attention. However, existing multimodal recommendations overlook the critical role of user representation initialization. Unlike items, which are naturally associated with rich modality information, users lack such inherent information. Consequently, item representations initialized based on meaningful modality information and user representations initialized randomly exhibit a significant semantic gap.

To this end, we propose a Semantically Guaranteed User Representation Initialization (SG-URInit). SG-URInit constructs the initial representation for each user by integrating both the modality features of the items they have interacted with and the global features of their corresponding clusters. SG-URInit enables the initialization of semantically enriched user representations that effectively capture both local (item-level) and global (cluster-level) semantics.

Our SG-URInit is training-free and model-agnostic, meaning it can be seamlessly integrated into existing multimodal recommendation models without incurring any additional computational overhead during training. Extensive experiments on multiple real-world datasets demonstrate that incorporating SG-URInit into advanced multimodal recommendation models significantly enhances recommendation performance. Furthermore, the results show that SG-URInit can further alleviate the item cold-start problem and also accelerate model convergence, making it an efficient and practical solution for multimodal recommendations.

* Corresponding authors



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
<https://doi.org/10.1145/3805712.3809623>

CCS Concepts

• Information systems → Recommender systems;

Keywords

Recommender System, Multimodal, Initialization

ACM Reference Format:

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, Jianheng Tang, Wei Wang, Xiping Hu, and Edith C. H. Ngai. 2026. Well Begun is Half Done: Training-Free and Model-Agnostic Semantically Guaranteed User Representation Initialization for Multimodal Recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, July 20–24, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3805712.3809623>

1 Introduction

The rapid growth of the internet has resulted in an overwhelming abundance of information, making recommender systems crucial for filtering and personalizing content. Traditional systems rely on historical user-item interactions [2, 5, 23, 24], but their performance is often limited by data sparsity, as they struggle to effectively model user preferences and item properties from limited interaction data. To address this, recent works have incorporated multimodal information [22, 26], such as images and textual descriptions, which explicitly enrich user and item representations and alleviate data sparsity. Early work, such as VBPR [4], incorporated visual content with matrix factorization to improve recommendations, while later studies [1, 10, 29, 32] combined visual and textual modalities to enhance item representations. Graph-based methods, such as MMGCN [21], introduced GCNs to model user-item interactions, and GRCN [20] further refined edge pruning to reduce noise. Additional advancements, such as DualGNN [17] and LATTICE [33], leveraged item-item and user-user homogeneous graphs, while FREEDOM [37] stabilized item representations. Recent studies,

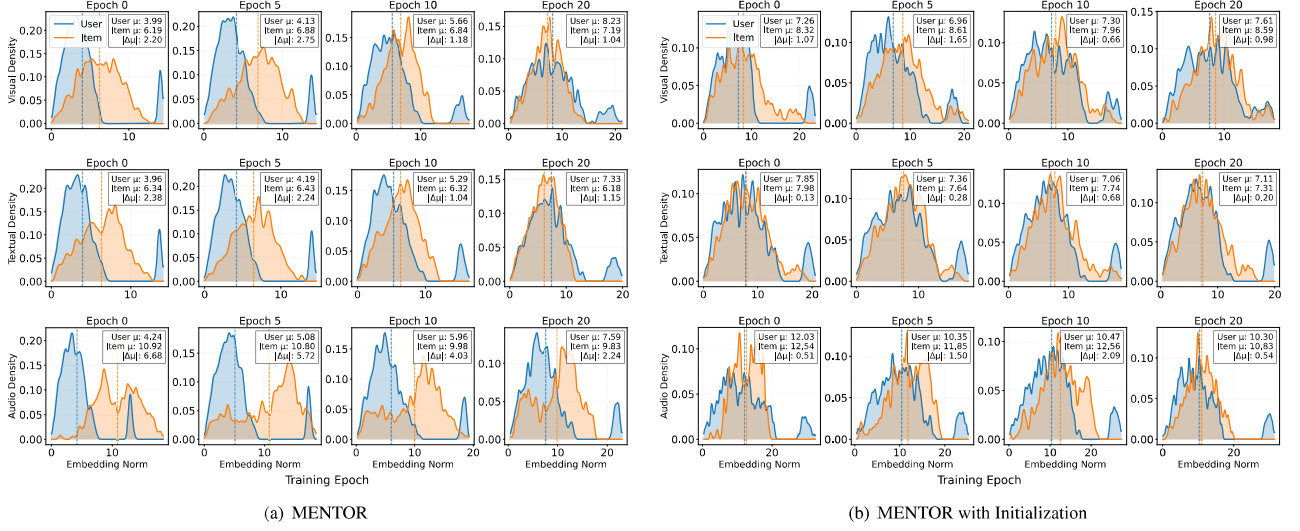


Figure 1: Evolution of discrepancies between user and item representation during model optimization.

including MMSSL [18] and MICRO [34], adopted contrastive learning to enhance modality-specific representation, while BM3 [38] and MENTOR [29] further align representations among different modalities. Furthermore, LGMRec [3] and DiffMM [7] utilized hypergraphs and diffusion models, and COHESION [26] introduced dual-stage modality fusion for enhanced representations.

Despite these advancements, a critical limitation remains largely unaddressed in the field of multimodal recommendation: the initialization of user representations. Unlike items, which are enriched with modality information, user representations are typically initialized randomly, creating a significant semantic gap between user and item embeddings. This gap introduces noise during graph convolution, hindering the learning of effective representations. Since GCNs update node representations by aggregating their neighbors' representations, the process of aggregating pre-trained, semantically guaranteed item embeddings with randomly initialized user embeddings becomes problematic. In this case, the random user embeddings act as sources of noise, propagating into the semantically guaranteed item representations during the initial training phases. In Section 2, we provide a detailed investigation to validate these observations.

While a naive solution might be to simply average the embeddings of interacted items, this solution suffers from local overfitting, restricting the user's representation to only their historical observations. To this end, we propose Semantically Guaranteed User Representation Initialization (SG-URInit). SG-URInit initializes semantically guaranteed user representations by integrating modality information of interacted items and cluster-level global information, effectively capturing both local (item-level) and global (cluster-level) semantics. This combination allows us to initialize a user representation that not only captures specific, localized preferences but also broadens the scope to include generalized preferences toward similar items within clusters. Our SG-URInit is training-free and model-agnostic, enabling seamless integration into existing multimodal recommendation models with no additional computational overhead. Extensive

experiments on real-world datasets across various advanced multimodal recommendations demonstrate that SG-URInit significantly enhances recommendation performance, alleviates the item cold-start problem, and accelerates model convergence, establishing it as an efficient and practical solution for multimodal recommendations. In summary, our contributions can be outlined as follows:

- This work first reveals that the critical limitation remains largely unaddressed in multimodal recommendations: the initialization of user representations
- This work proposes SG-URInit, a semantically guaranteed user representation initialization, which mitigates the semantic gap between item and user initial representations by initializing a user representation that not only captures specific, localized preferences but also broadens the scope to include generalized preferences toward similar items within clusters.
- SG-URInit is training-free and model-agnostic, allowing seamless integration into existing multimodal recommendation models without additional computational overhead.
- We conduct comprehensive experiments on real-world datasets across various advanced multimodal recommendations, demonstrating that SG-URInit significantly enhances recommendation performance, alleviates the item cold-start problem, and accelerates model convergence.

2 Investigation

In this section, we first provide a brief overview of the standard multimodal recommendation process. Then, we conducted an investigation into the discrepancies between user and item representations across different modalities during the model optimization process under two initialization settings: (1) random initialization of user representations and (2) initialization of user representations based on the representations of interacted items and clusters.

2.1 Preliminary

Given a user set $\mathcal{U} = \{u_1, \dots, u_{|\mathcal{U}|}\}$ and an item set $\mathcal{I} = \{i_1, \dots, i_{|\mathcal{I}|}\}$. In multimodal scenarios, each item contains multiple features, we introduce modality-specific item embedding i^m for each item i belonging to the set of modalities $\mathcal{M}$. The user-item interaction matrix is denoted as $\mathcal{R} \in \{0, 1\}^{|\mathcal{U}| \times |\mathcal{I}|}$. Specifically, each entry $\mathcal{R}_{u,i}$ indicates whether the user u is connected to item i , with a value of 1 representing a connection and 0 otherwise. These user-item interactions $\mathcal{R}$ naturally constructs the bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{I}, \mathcal{E})$, where $\mathcal{U}, \mathcal{I}$ serve as the graph vertices, and $\mathcal{E}$ denotes the edge set. For each user-item pair (u, i) that satisfies $\mathcal{R}_{u,i} = 1$, there exists bidirectional edges $(u, i) \in \mathcal{E}$ and $(i, u) \in \mathcal{E}$. Previous works (refer to survey [30]) commonly random initialize $\mathbf{E}_u^m = \{\mathbf{e}_{u_1}^m, \dots, \mathbf{e}_{u_{|\mathcal{U}|}}^m\} \in \mathbb{R}^{d^m \times |\mathcal{U}|}$ to represent user representation with modality m . $\mathbf{E}_i^m = \{\mathbf{e}_{i_1}^m, \dots, \mathbf{e}_{i_{|\mathcal{I}|}}^m\} \in \mathbb{R}^{d^m \times |\mathcal{I}|}$ represents item initialized representation with modality m , which extracted by pre-trained encoders. Here d^m represents the hidden dimensionality.

Subsequently, the initialized user and item representations are enhanced through various structures, such as homogeneous graphs [5, 21, 24] naturally formed by user-item interactions, heterogeneous graphs [26, 29, 33, 37] constructed based on user preferences and item properties, or other advanced structures [3] and techniques [7]. These representations are then optimized using a combination of supervised and self-supervised objective functions. However, the semantic gap between randomly initialized user representations and modality-informed item representations adversely impacts both the optimization process and the final representation quality. We provide empirical support in subsequent subsection.

2.2 Validation

We selected the TikTok dataset, which contains visual, textual, and audio modalities (refer to Table 2), and used MENTOR, a state-of-the-art multimodal recommendation model, for our experiments. The experiments involved two variants: MENTOR and MENTOR with Initialization. The former adopts random initialization for user representations, as in existing multimodal recommendation models, while the latter initializes user representations based on the representations of interacted items (refer to Eq. 1).

We apply t-SNE [16] to map the user and item representations into two-dimensional space for each modality and plotted the distribution of their representation magnitudes (L2 norms). Figure 1 illustrates the distributions of user and item representations across the three modalities at 0, 5, 10, and 20 epochs. By comparing the two curves (blue for users and orange for items), the similarities and discrepancies in their magnitude distributions can be observed. Additionally, each subfigure provides: (1) User μ : the mean magnitude of user representations. (2) Item μ : the mean magnitude of item representations. (3) $|\Delta\mu|$: the absolute difference between the two means. Dashed lines indicate the positions of the mean values (blue dashed line for users and orange dashed line for items). From the results, we observe that user representations initialized based on item representations are more similar to item representations than randomly initialized user representations, regardless of the training stage. It is worth noting that these representations are enhanced through techniques such as heterogeneous and homogeneous graph

convolutions. This indicates that the semantic gap caused by initialization persists even after multiple layers of convolution, ultimately impacting the overall performance of the model.

To this end, we propose a tailored Semantically Guaranteed User Representation Initialization (SG-URInit) to bridge the semantic gap between user and item initial representations during initialization, ensuring a better optimization process and enhanced representations.

3 Methodology

In Section 2, we discussed the drawbacks of the random user representation initialization commonly adopted in prior multimodal recommendation studies. While items are associated with specific modality information, users inherently lack such modality information. Consequently, item initialization leverages modality features, whereas user initialization typically relies on random values, resulting in a significant semantic gap. In this section, we detail our proposed Semantically Guaranteed User Representation Initialization (SG-URInit). SG-URInit constructs the initial representation for each user by integrating the modality information of the items the user has interacted with and the overall features of their corresponding clusters. SG-URInit enables the initialization of a semantically guaranteed user representation that effectively captures both local (item-level) and global (cluster-level) semantics. The overview of SG-URInit¹ can be found in Figure 2.

3.1 SG-URInit

The modality representation of a user should comprehensively reflect their modality-specific preference tendencies. A natural and intuitive approach is to aggregate the modality representations of items they have interacted with, which captures their local preferences at a item-level. However, relying solely on such aggregation risks overfitting to observed preferences, thereby neglecting the user's potential interest in similar but unobserved items. To address this limitation, SG-URInit further adopts a cluster-level aggregation. Firstly, we group items into clusters based on their modality features. Beyond aggregating the local modality representations of interacted items, we further incorporate the global cluster centroids of the corresponding items. This combination allows us to construct a user representation that not only captures specific, localized preferences but also broadens the scope to include generalized preferences toward similar items within clusters. By integrating both local and global semantic information, our SG-URInit ensures a more holistic and semantically guaranteed user representation.

3.1.1 Item-level Aggregation. We first perform item-level aggregation of the modality features of the items each user has interacted with. Formally:

$$\bar{\mathbf{e}}_u^m = \sum_{\tilde{i} \in N(u)} \frac{1}{|N(\tilde{i})|} \mathbf{e}_{\tilde{i}}^m, \quad (1)$$

where $N(u)$ and $N(\tilde{i})$ refers to the set of items or users that interact with user u and item $\tilde{i}$, respectively. The weight of each interacted item is influenced by its degree, inspired by the design of LightGCN [5]. We empirically validate in Section 4.11 that this weight allocation strategy outperforms uniform weighting.

¹Code is available at <https://github.com/Jinfeng-Xu/SG-URInit>.

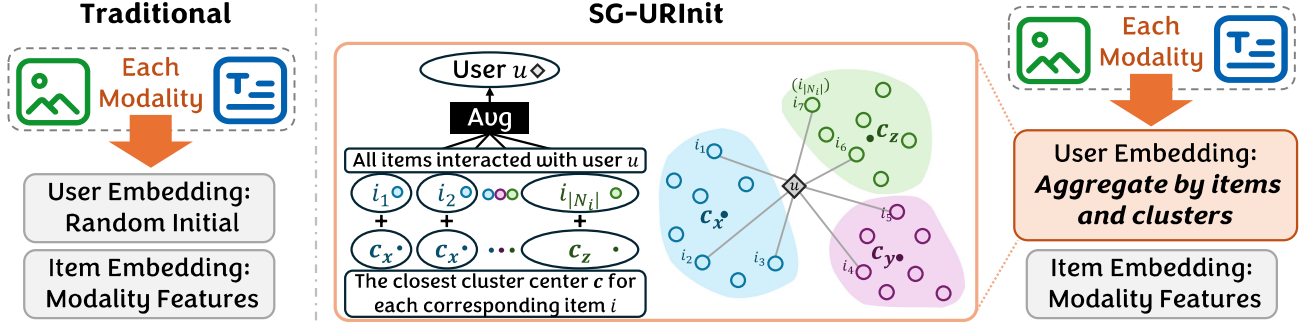


Figure 2: Overview of user representation initialization. Left: traditional user representation initialization; Right: our SG-URInit.

3.1.2 Cluster-level Aggregation. Focusing solely on item-level aggregation risks overfitting to the user’s observed preferences, which may overlook their potential interest in similar but unobserved items. To this end, SG-URInit introduces a cluster-level aggregation mechanism. We first partition all items into K clusters based on their modality features through K-Means clustering:

$$C^m = \text{K-Means}(\{e_i^m \mid \forall i \in I\}, K), \quad (2)$$

where $C^m = \{c_1^m, c_2^m, \dots, c_K^m\}$ denotes the cluster centroids for modality m . K is the number of cluster centroids, which we empirically discussed in Section 4.8.

For each item i , we compute its cluster-aligned representation by assigning it to the nearest cluster centroid:

$$\tilde{e}_i^m = \arg \min_{c_k^m \in C^m} \|e_i^m - c_k^m\|_2. \quad (3)$$

The cluster-level user representation is then obtained by aggregating these cluster-aligned item representations:

$$\tilde{e}_u^m = \sum_{\tilde{i} \in N(u)} \frac{1}{|N(\tilde{i})|} \tilde{e}_{\tilde{i}}^m, \quad (4)$$

where $N(u)$ and $N(\tilde{i})$ denote neighbor sets for user u and item $\tilde{i}$, respectively. To ensure consistency, we adopt the same weighting strategy as that used at the item-level.

3.2 Fusion

We combine the user representations generated by item-level aggregation and cluster-level aggregation to form the final user representation, which simultaneously incorporates both local and global semantics. Formally:

$$e_u^m = (1 - \lambda) \tilde{e}_u^m + \lambda \tilde{e}_u^m, \quad (5)$$

where $\lambda \in [0, 1]$ is a hyper-parameter that adjusts the scaling and balances the contributions of the item-level ($\tilde{e}_u^m$) and cluster-level ($\tilde{e}_u^m$) user representations. Since both $\tilde{e}_u^m$ and $\tilde{e}_u^m$ share the same representation scalar level as the item representations, directly summing the two would lead to an increase in the scalar magnitude of the user representation. To this end, we introduce a scaling factor λ to balance and normalize the aggregation, ensuring the final user representation maintains the same scalar level as the original representations.

3.3 Discussion

This form of initialization is conceptually analogous to the behavior of graph convolution, as the essence of representation enhancement in graph convolution lies in utilizing the similar representations of neighboring nodes (interacted nodes) to enrich a node’s own representation. We point out that the effectiveness of graph convolution can be significantly improved by our semantically guaranteed user representation initialization.

Specifically, prior works often initialize item representations using meaningful modality information, while user representations are randomly initialized [30]. As a result, in a direct single-layer graph convolution operation, the aggregation of randomly initialized user representations for items can essentially be viewed as injecting random noise (randomly initialized user representation) into the item representations. On the other hand, the aggregation of modality-informed item representations for users can be interpreted as a process akin to our semantically guaranteed user representation initialization, but with the additional introduction of noise through the randomly initialized user representations.

To address these shortcomings, our semantically guaranteed user representation initialization can effectively mitigate the propagation of noise into item representations during the convolution process, while also avoiding the influence of users’ initial random noise. This ensures a more robust and semantically aligned representation learning process, enhancing both the quality of the representations and the overall performance of graph-based models.

4 Evaluation

We conduct extensive experiments² aiming to answer the following research questions: **RQ1:** Can SG-URInit enhance the performance of multimodal recommender systems? **RQ2:** How do the different components in SG-URInit affect performance enhancement? **RQ3:** Can SG-URInit have a positive impact on the convergence speed? **RQ4:** Can SG-URInit be compatible with robust training and data augmentation strategies? **RQ5:** Can SG-URInit enhance multimodal recommendations in different sparse data scenarios? **RQ6:** Does SG-URInit perform satisfactorily in the item cold-start scenario? **RQ7:** What is the impact of key hyper-parameters in SG-URInit? **RQ8:** What is the cost of user initialization in SG-URInit?

²To avoid the instability of K-Means, all results reported in this paper are averaged over 10 K-Means runs.

Table 1: Performance comparison of baselines with or without SG-URInit on all datasets in terms of Recall@K (R@K) and NDCG@K (N@K). * indicates the improvement is statistically significant, where the p-value is less than 0.01.

Datasets	Baby				Sports				Clothing				TikTok			
Metrics	R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10
MMGCN	0.0240	0.0378	0.0160	0.0200	0.0216	0.0370	0.0143	0.0193	0.0130	0.0218	0.0088	0.0110	0.0250	0.0485	0.0125	0.0200
+SG-URInit	0.0263*	0.0447*	0.0171*	0.0231*	0.0268*	0.0439*	0.0175*	0.0232*	0.0147*	0.0249*	0.0100*	0.0126*	0.0341*	0.0589*	0.0162*	0.0238*
Improv.	9.58%	18.25%	6.88%	15.50%	24.07%	18.65%	22.38%	20.21%	13.08%	14.22%	13.64%	14.55%	36.40%	21.44%	29.60%	19.00%
SLMRec	0.0343	0.0529	0.0226	0.0290	0.0429	0.0663	0.0288	0.0365	0.0292	0.0452	0.0196	0.0247	0.0309	0.0529	0.0165	0.0234
+SG-URInit	0.0360*	0.0556*	0.0235*	0.0307*	0.0454*	0.0708*	0.0305*	0.0388*	0.0312*	0.0488*	0.0210*	0.0265*	0.0344*	0.0577*	0.0182*	0.0256*
Improv.	4.96%	5.10%	3.98%	5.86%	5.83%	6.79%	5.90%	6.30%	6.85%	7.96%	7.14%	7.29%	11.33%	9.07%	10.30%	9.40%
FREEDOM	0.0374	0.0627	0.0243	0.0330	0.0446	0.0717	0.0291	0.0385	0.0388	0.0629	0.0257	0.0341	0.0367	0.0575	0.0250	0.0316
+SG-URInit	0.0391*	0.0657*	0.0256*	0.0351*	0.0471*	0.0759*	0.0310*	0.0411*	0.0406*	0.0662*	0.0270*	0.0360*	0.0404*	0.0616*	0.0269*	0.0338*
Improv.	4.55%	4.78%	5.35%	6.36%	5.61%	5.86%	6.53%	6.75%	4.64%	5.25%	5.06%	5.57%	10.08%	7.13%	7.60%	6.96%
DRAGON	0.0380	0.0662	0.0249	0.0345	0.0449	0.0752	0.0296	0.0413	0.0401	0.0671	0.0270	0.0365	0.0335	0.0568	0.0228	0.0308
+SG-URInit	0.0396*	0.0697*	0.0262*	0.0365*	0.0473*	0.0791*	0.0311*	0.0438*	0.0426*	0.0714*	0.0286*	0.0388*	0.0379*	0.0632*	0.0250*	0.0336*
Improv.	4.21%	5.29%	5.22%	5.80%	5.35%	5.19%	5.07%	6.05%	6.23%	6.41%	5.93%	6.30%	13.13%	11.27%	9.65%	9.09%
LGMRec	0.0374	0.0626	0.0249	0.0333	0.0446	0.0719	0.0288	0.0387	0.0371	0.0555	0.0246	0.0302	0.0341	0.0583	0.0235	0.0311
+SG-URInit	0.0394*	0.0660*	0.0262*	0.0350*	0.0472*	0.0763*	0.0308*	0.0418*	0.0399*	0.0607*	0.0267*	0.0328*	0.0376*	0.0636*	0.0259*	0.0343*
Improv.	5.35%	5.43%	5.22%	5.11%	5.83%	6.12%	6.94%	8.01%	7.55%	9.37%	8.54%	8.61%	10.26%	9.09%	10.21%	10.29%
MENTOR	0.0422	0.0678	0.0281	0.0362	0.0485	0.0763	0.0323	0.0409	0.0411	0.0668	0.0274	0.0360	0.0364	0.0634	0.0247	0.0334
+SG-URInit	0.0444*	0.0713*	0.0296*	0.0382*	0.0511*	0.0802*	0.0340*	0.0431*	0.0432*	0.0707*	0.0290*	0.0381*	0.0463*	0.0719*	0.0292*	0.0374*
Improv.	5.21%	5.16%	5.34%	5.52%	5.36%	5.11%	5.26%	5.38%	5.11%	5.84%	5.84%	5.83%	27.20%	13.41%	18.22%	11.98%

Table 2: Statistics of all evaluation datasets.

Datasets	#Users	#Items	#Interactions	Sparsity	Modality
Baby	19,445	7,050	160,792	99.88%	V,T
Sports	35,598	18,357	296,337	99.95%	V,T
Clothing	39,387	23,033	278,677	99.97%	V,T
TikTok	9,319	6,710	59,541	99.90%	V,T,A

4.1 Settings

4.1.1 Datasets. The experiments are conducted on three real-world datasets: Baby, Sports, and Clothing from Amazon [12]. All the datasets comprise textual and visual features in the form of item descriptions and images. To further evaluate the performance of SG-URInit in scenarios involving multiple modalities, we also conduct experiments on the TikTok dataset [7]. Our data preprocessing methodology follows the approach outlined in MMRec [36]. Table 2 shows the statistics of these datasets.

4.1.2 Baselines. We extensively examine the performance of our SG-URInit across a variety of multimodal recommendation models:

- **MMGCN** [21] applies GCN for each data modality to learn modality-specific features and then integrates all user-predicted ratings across modalities to produce the final rating.
- **SLMRec** [15] leverages a self-supervised learning framework for multimodal recommendations by establishing a tailored node self-discrimination task, which reveals hidden multimodal patterns.
- **FREEDOM** [37] refines LATTICE by freezing the item-item graph to stabilize item relationships and reducing noise in the user-item graph to enhance recommendation accuracy.
- **DRAGON** [35] leverages heterogeneous and homogeneous graphs to learn high-quality user/item representations.

- **LGMRec** [3] integrates local embeddings, which capture fine-grained topological embeddings, with global embeddings considering hypergraph dependencies among items.
- **MENTOR** [29] proposes multi-level cross-modal alignment tasks to effectively improve final representation and achieve state-of-the-art recommendation accuracy.

Moreover, we test the compatibility of our SG-URInit with the adversarial training strategy (AMR [14]) and LLM-based data augmentation strategy (GPT-4o [31]).

4.1.3 Metrics. For a fair comparison, we follow the settings of previous works [30, 37, 38] to adopt two widely-used evaluation metrics for top- K recommendation: Recall@ K and NDCG@ K . We report the average scores for all users in the test dataset under both $K = 5$ and $K = 10$, respectively.

4.1.4 Implementation Details. We retain the standard settings for all baselines and fix batch size B to 2048. We perform a complete grid search for all baselines to find the optimal setting based on their published papers. For SG-URInit, we apply a grid search on hyper-parameter K in Eq 2 in $\{2, 4, 8, 16\}$, and balancing hyper-parameter λ in Eq. 5 in $\{0, 0.0001, 0.001, 0.01, 0.1, 1\}$. The common optimizer is Adam [8] and all training and evaluation of all models are conducted on an RTX4090 GPU. For the GPT-4o data augmentation strategy, we utilize GPT-4o [31] to augment the raw description via the items' image for all datasets to improve the correlation between textual and visual modalities. We designed the prompt as: "[V] Here is the description of an item and the corresponding picture, please combine the picture to improve the description quality in one paragraph. The description is as follows: [T].", where $[V]$ and $[T]$ are the raw image and description for each item, respectively.

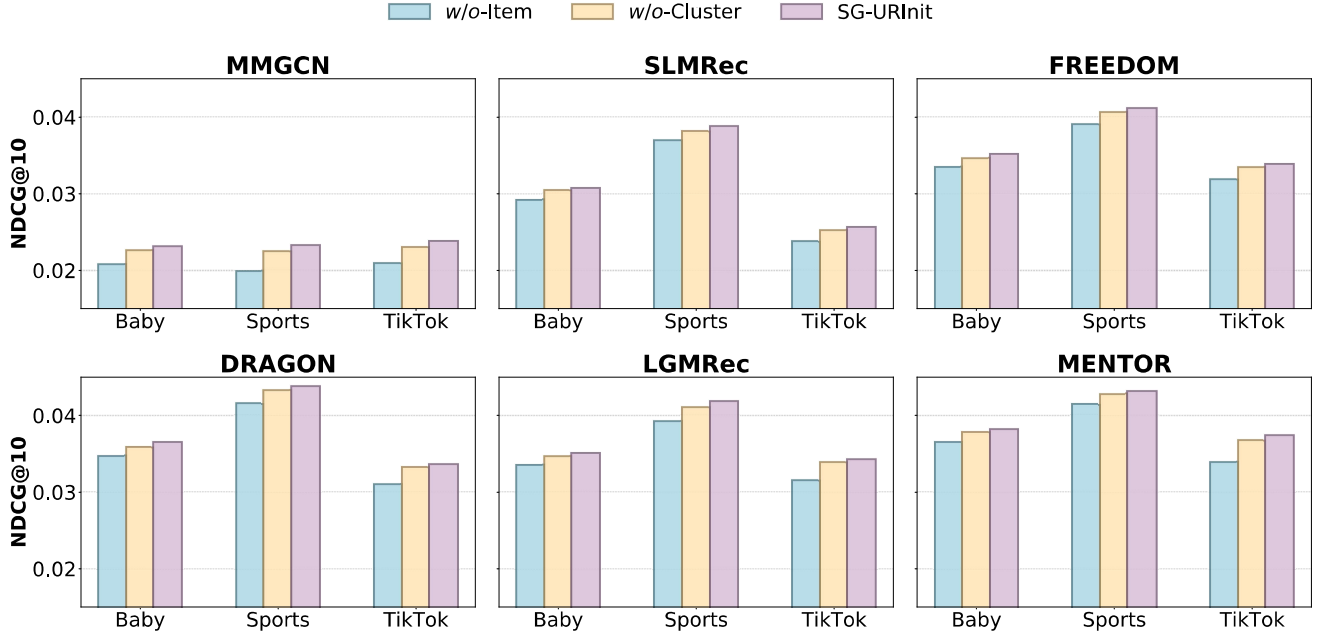


Figure 3: Ablation study on key components of SG-URInit in terms of NDCG@10.

4.2 Overall Performance (RQ1)

We evaluate the effectiveness of our SG-URInit on various models across multiple real-world datasets in multimodal recommendation scenarios. From Table 1, we find the following observations:

Observation 1: SG-URInit can enhance the performance of various multimodal recommendation models across all datasets.

As shown in Table 1, we conducted extensive experiments using SG-URInit on six multimodal recommendation models across four distinct public datasets. The experimental results demonstrate that SG-URInit achieves significant improvements over all baselines across all evaluation metrics. In summary, these results validate that leveraging item multimodal information to initialize user representations can effectively enhance recommender performance by mitigating the semantic gap between user and item representations.

Observation 2: SG-URInit is more effective on datasets with a greater variety of modality information. The performance improvement brought by SG-URInit is particularly significant on the TikTok dataset, which we attribute to that TikTok contains a richer variety of modality information. In contrast, the random initialization of user representations faces more pronounced challenges when handling datasets with more diverse modality information. SG-URInit addresses this concern effectively, providing a promising solution for real-world scenarios that incorporate richer multimodal information.

4.3 Ablation Study (RQ2)

To discern the impact of our SG-URInit’s core components, we conducted an ablation study with various configurations:

- *w/o-Cluster*: It removes the cluster-level aggregation, resulting in a user representation initialization with only local semantic. This configuration is equivalent to a recent study (AlphaRec [13]).

- *w/o-Item*: It removes the item-level aggregation, resulting in a user representation initialization with only global semantic.

We conduct extensive experiments for our SG-URInit across all six multimodal recommendation models on three datasets for various configurations. The findings presented in Figure 3 clearly demonstrate that our SG-URInit surpasses all its modified configurations, thereby confirming the essential role each component plays in initialing high-quality user representations. It is worth noting that *w/o-Item* consistently underperforms compared to *w/o-Cluster*, indicating that users’ local multimodal information plays a dominant role. Meanwhile, the fact that *w/o-Cluster* consistently underperforms compared to SG-URInit demonstrates that global multimodal information also provides valuable semantic information for the initialization process.

4.4 Convergence Speed (RQ3)

In addition, SG-URInit helps accelerate model training convergence. We visualized the training loss of three advanced multimodal recommendation models (MMGCN, FREEDOM, and DRAGON) on the TikTok dataset. Following previous training settings [37, 38], we adopted an early stopping strategy with a patience of 20 epochs and set the maximum number of epochs to 1,000. As shown in Figure 4, SG-URInit effectively improves the convergence speed of all models. We attribute this improvement to the reduction in the gap between user and item representations achieved by our user representation initialization, which accelerates model training convergence.

4.5 Compatibility Analysis (RQ4)

Existing studies improve the robustness of multimodal recommendation systems through adversarial training strategies [9, 14] and

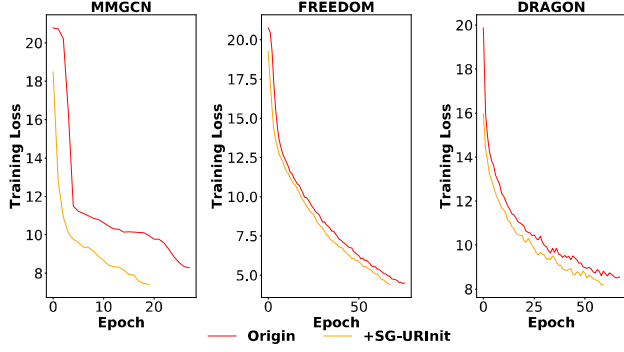


Figure 4: Convergence study on the TikTok dataset.

Table 3: Performance comparison for strategies on three datasets under Recall@5 (R@5) and NDCG@5 (N@5). +S, +A, and +G denote +SG-URInit, +AMR, and +GPT-4o, respectively.

Models	Datasets	Baby		Sports		Clothing	
	Metrics	R@5	N@5	R@5	N@5	R@5	N@5
MMGCN	origin	0.0240	0.0160	0.0216	0.0143	0.0130	0.0110
	+S	0.0263	0.0171	0.0268	0.0175	0.0147	0.0126
	+S+A	0.0267	0.0174	0.0273	0.0177	0.0151	0.0128
	+S+G	<u>0.0271</u>	<u>0.0176</u>	<u>0.0276</u>	<u>0.0179</u>	<u>0.0152</u>	<u>0.0131</u>
	+S+A+G	0.0272	0.0176	0.0279	0.0180	0.0154	0.0135
LGMRec	origin	0.0374	0.0249	0.0446	0.0288	0.0371	0.0246
	+S	0.0394	0.0262	0.0472	0.0308	0.0399	0.0267
	+S+A	0.0400	0.0267	0.0476	0.0311	0.0406	0.0272
	+S+G	0.0404	0.0269	0.0481	0.0314	0.0410	0.0274
	+S+A+G	0.0408	0.0272	0.0484	0.0317	0.0413	0.0277
MENTOR	origin	0.0422	0.0281	0.0485	0.0323	0.0411	0.0274
	+S	0.0444	0.0296	0.0511	0.0340	0.0432	0.0290
	+S+A	0.0447	0.0298	0.0514	0.0343	0.0433	0.0290
	+S+G	<u>0.0450</u>	<u>0.0300</u>	<u>0.0518</u>	<u>0.0346</u>	<u>0.0437</u>	<u>0.0293</u>
	+S+A+G	0.0452	0.0303	0.0520	0.0347	0.0440	0.0296

data augmentation methods [6, 11, 19]. Building on this, we further evaluate the compatibility of SG-URInit with the adversarial training strategy (AMR [14]) and the LLM-based data augmentation strategy (GPT-4o [31]). We conduct extensive experiments on two multimodal recommendation models across three public datasets. As shown in Table 3, combining SG-URInit with both AMR and GPT-4o further enhances model performance. Furthermore, GPT-4o consistently outperforms AMR across all datasets, which we attribute to its ability to better bridge the inherent gap among modalities information of items. Notably, using both AMR and GPT-4o simultaneously achieves the best performance, surpassing the results of employing either method individually.

4.6 Sparsity Analysis (RQ5)

To assess the impact of SG-URInit on advanced multimodal recommendation models under varying levels of data sparsity, we performed experiments on subsets of four datasets with different sparsity conditions. We compared the performance of three state-of-the-art multimodal recommendation models—SLMRec, FREEDOM, and LGMRec—both with and without the integration of SG-URInit. To better understand the effect of data sparsity, users were grouped

Table 4: Cold start analysis across Baby and TikTok datasets.

Datasets	Baby		TikTok	
	Recall@10	NDCG@10	Recall@10	NDCG@10
MMGCN	0.0103	0.0062	0.0214	0.0112
+SG-URInit	0.0124	0.0073	0.0250	0.0132
SLMRec	0.0172	0.0101	0.0230	0.0117
+SG-URInit	0.0207	0.0123	0.0259	0.0133
FREEDOM	0.0348	0.0195	0.0301	0.0149
+SG-URInit	0.0390	0.0219	0.0336	0.0166
DRAGON	0.0378	0.0212	0.0402	0.0205
+SG-URInit	0.0419	0.0238	0.0445	0.0224
LGMRec	0.0371	0.0208	0.0333	0.0169
+SG-URInit	0.0409	0.0230	0.0370	0.0187
MENTOR	0.0391	0.0219	0.0415	0.0212
+SG-URInit	0.0430	0.0249	0.0459	0.0233

based on the number of interactions in the training data (e.g., the first group includes users with 1–5 interactions). As shown in Figure 5, SG-URInit consistently improves the performance of all models across datasets and user groups, confirming its robustness under sparse conditions. Notably, the performance boost provided by SG-URInit is especially prominent in highly sparse scenarios. This can be attributed to the fact that a well-initialized user representation reduces the dependency on extensive data during training in these challenging settings.

4.7 Cold-Start Analysis (RQ6)

We report the results for the item cold-start scenario on the Baby and TikTok datasets. We adopt the widely used experimental settings described in [34]. Specifically, we remove all user-item interaction pairs associated with a randomly selected 20% item set from the training set. We further divide the half of the items (10%) into the validation set and half (10%) into the testing set. In other words, these items are entirely unseen in the training set. As shown in Table 4, SG-URInit enhances the performance of multimodal recommendation models in item cold-start situations. This improvement can be attributed to the semantic consistency introduced by our user representation initialization, which narrows the semantic gap between user and item representations.

4.8 Hyper-parameter Analysis (RQ7)

We evaluate the impact of the key hyper-parameters (K and λ) on MDVT’s performance across four distinct datasets in terms of Recall@10. Results are presented in Figure 6 and Figure 7.

Hyper-parameter K : In Figure 6, we observe that for hyper-parameter K , all models require the same optimal setting on each dataset, although the optimal setting varies across different datasets. The choice of K depends on the complexity of the scenario: for the more complex Clothing dataset, $K = 8$ is the best choice, while for the other datasets, $K = 4$ yields the best performance.

Hyper-parameter λ : In Figure 7, we observe that for hyper-parameter λ , the optimal value is consistent across different models and less influenced by the dataset. For models incorporating the item-item

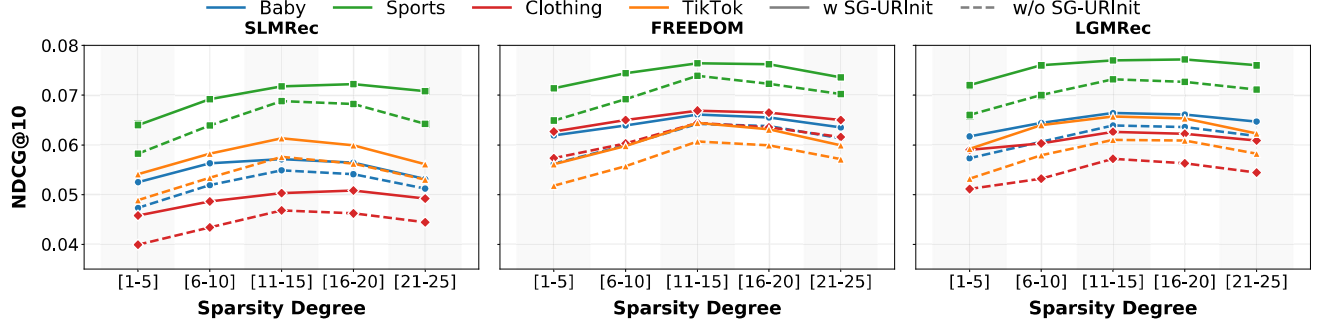
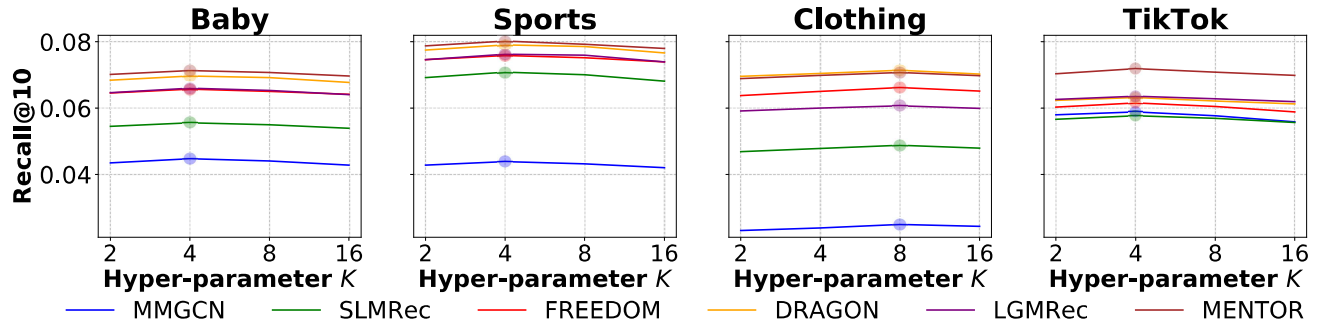
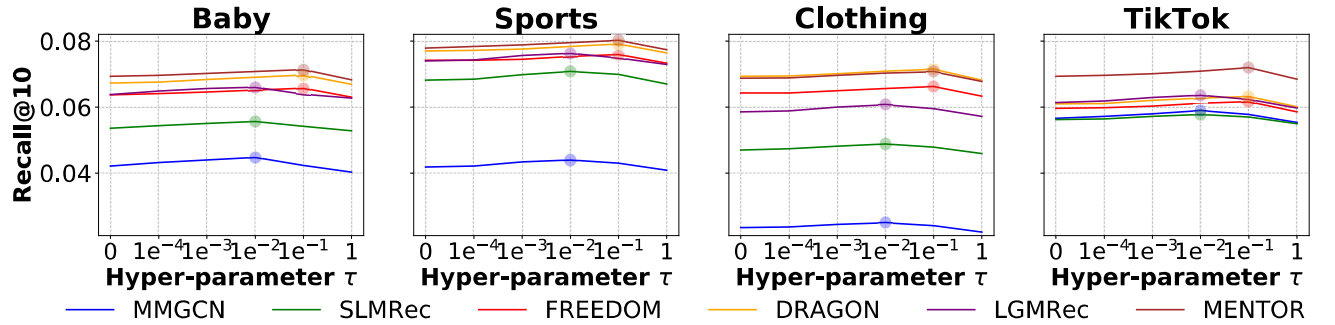


Figure 5: Sparsity study on three advanced multimodal recommendation models across four distinct datasets.

Figure 6: Performance *w.r.t.* hyper-parameter K .Figure 7: Performance *w.r.t.* hyper-parameter λ .

graph (FREEDOM, DRAGON, and MENTOR), the optimal λ is 0.1, whereas for other models, the optimal value is $\lambda = 0.01$. Notably, when $\lambda = 0$, the SG-URInit degenerates into the variant w/o-Cluster, and when $\lambda = 1$, it degenerates into the variant w/o-Item.

Insights: We further provide guidance on the hyper-parameter settings for SG-URInit. For the hyper-parameter K , setting $K = 4$ consistently yields satisfactory performance improvements across all models and datasets. In more complex scenarios, exploring $K = 8$ may offer additional performance gains. For the hyper-parameter λ , we recommend using $\lambda = 0.1$ for models incorporating item-item graphs, while $\lambda = 0.01$ is more suitable for other models. These insights significantly reduce the cost of hyper-parameter tuning when implementing SG-URInit.

4.9 Discussion for Training-Free and Model-Agnostic

SG-URInit features two key advantages: it is training-free and model-agnostic. Its initialization is pre-constructed based on the interaction data and multimodal information of items in each dataset, requiring no additional overhead during the training phase. The user representations initialized for each modality can be directly applied to any existing multimodal recommendation model. Notably, as demonstrated in Section 4.4, SG-URInit not only imposes no burden on the training process but also accelerates model convergence, thereby reducing the overall training time.

Table 5: Performance comparison for weight allocation strategies on three datasets under Recall@5 (R@5) and NDCG@5 (N@5).

Models	Datasets	Baby		Sports		Clothing	
	Metrics	R@5	N@5	R@5	N@5	R@5	N@5
MMGCN	origin	0.0240	0.0160	0.0216	0.0143	0.0130	0.0110
	SG-URInit	0.0263	0.0171	0.0268	0.0175	0.0147	0.0126
	-bi-degree	0.0261	0.0170	0.0269	0.0177	0.0141	0.0123
	-equal	0.0258	0.0168	0.0261	0.0170	0.0140	0.0122
LGMRec	origin	0.0374	0.0249	0.0446	0.0288	0.0371	0.0246
	SG-URInit	0.0394	0.0262	0.0472	0.0308	0.0399	0.0267
	-bi-degree	0.0391	0.0260	0.0468	0.0305	0.0395	0.0264
	-equal	0.0388	0.0258	0.0466	0.0303	0.0391	0.0262
MENTOR	origin	0.0422	0.0281	0.0485	0.0323	0.0411	0.0274
	SG-URInit	0.0444	0.0296	0.0511	0.0340	0.0432	0.0290
	-bi-degree	0.0442	0.0295	0.0507	0.0337	0.0427	0.0286
	-equal	0.0438	0.0292	0.0503	0.0334	0.0424	0.0284

Table 6: Time cost for user initialization.

Datasets	Baby	Sports	Clothing	TikTok
$K = 4$	15.23s	37.02s	43.01s	4.75s
$K = 8$	17.88s	42.30s	49.32s	5.40s

4.10 Cost of User Initialization (RQ8)

We report the pre-training user initialization times for four datasets, specifically focusing on $K = 4$ and $K = 8$, as these settings consistently delivered the best performance. The results, summarized in Table 6, show the time costs associated with user initialization. Notably, the user initialization process is conducted entirely before the training phase and can be reused across multiple runs, significantly reducing the overall impact of these time costs.

4.11 In-depth Analysis

Ablation studies on weighting strategies during aggregation and the impact of setting different numbers of cluster centers K for different modalities on SG-URInit, are topics worthy of further exploration.

4.11.1 Ablation Study. We conduct an empirical discussion on the weight allocation strategies for aggregating items or clusters during user representation initialization. Specifically, in addition to the degree-based aggregation used by SG-URInit, we also adopt other strategies, such as bi-degree weighting (as in LightGCN) and equal weighting. The detailed formulas are as follows:

- SG-URInit: $\sum_{i \in N(u)} \frac{1}{|N(i)|}$.
- -bi-degree: $\sum_{i \in N(u)} \frac{1}{\sqrt{|N(i)||N(u)|}}$.
- -equal $\sum_{i \in N(u)} \frac{1}{|N(u)|}$.

Table 5 presents the performance comparison of SG-URInit and other weight allocation strategy variants across three datasets and three different advanced models. We observe that SG-URInit generally outperforms all variants, while the bi-degree weighting strategy also significantly outperforms the equal weighting strategy. We attribute this to the importance of popularity-based weighting in the recommendations.

4.11.2 Hyper-parameter Analysis. We further explore whether setting different numbers of cluster centers K for different modalities can enhance the performance of SG-URInit. The value of K for each modality is searched within $\{2, 4, 8, 16\}$. We report the optimal performance under dynamic K for all four datasets across four models, along with the corresponding K selection for each modality.

According to the experimental results in Table 7, we find that dynamic control of K achieves better performance for datasets other than Clothing. We attribute this to the varying levels of information richness across different modalities. For instance, the textual modality generally contains more fine-grained information than the visual modality, and the degree of granularity for each modality varies across different scenarios. Therefore, dynamically adjusting K can further enhance the performance of SG-URInit.

5 Related Work

Many recent studies have incorporated multimodal information to address the data sparsity problem in recommendation systems. Early work such as VBPR [4] leveraged visual content alongside matrix factorization, while subsequent studies [1, 10, 25, 27, 28, 32] integrated both visual and textual modalities for improved item representation. Evolving from traditional architectures, methods like MMGCN [21] and GRCN [20] employed graph convolutional networks (GCNs) to model user-item interactions, with GRCN refining edge pruning to reduce noise. Other approaches such as DualGNN [17], LATTICE [33], and FREEDOM [37] enhanced user and item representations through co-occurrence graphs, semantic graphs, and stabilized item representations, respectively. Further advancements include contrastive learning-based models like MMSSL [18] and MICRO [34], modal fusion techniques in BM3 [38] and MENTOR [29], and innovations using hypergraphs and diffusion models in LGMRec [3] and DiffMM [7]. COHESION [26] further introduced dual-stage modality fusion to enhance user and item representations.

Despite these advancements, existing works largely overlook the critical role of user representation initialization. Unlike items, which are enriched with modality information, users lack such inherent features. This disparity results in a semantic gap between meaningful item representations and randomly initialized user representations. To this end, we propose Semantically Guaranteed User Representation Initialization (SG-URInit). SG-URInit integrates modality features of users' interacted items and cluster-level global features to construct semantically guaranteed user representations. SG-URInit effectively captures both local and global semantics, providing a robust initialization for multimodal recommendations.

6 Conclusion

In this paper, we aim to address the critical semantic gap arising from random user initialization in multimodal recommendation, which further introduces noise during graph convolution. We propose SG-URInit, a training-free and model-agnostic semantically guaranteed user representation initialization that integrates interacted items' modality information and cluster-level information, effectively capturing both local (item-level) and global (cluster-level) semantics. Extensive experiments on multiple real-world datasets across various advanced models demonstrate that SG-URInit significantly boosts

Table 7: Analysis for dynamic K for different modalities on two datasets under Recall@5 (R@5) and NDCG@5 (N@5).

Models	Variant	Baby					Sports					Clothing					TikTok				
	Metrics	R@5	N@5	T	V	A	R@5	N@5	T	V	A	R@5	N@5	T	V	A	R@5	N@5	T	V	A
MMGCN	Origin	0.0240	0.0160	-	-	-	0.0216	0.0143	-	-	-	0.0130	0.0088	-	-	-	0.0250	0.0125	-	-	-
	SG-URInit	0.0263	0.0171	4	4	-	0.0268	0.0175	4	4	-	0.0147	0.0100	8	8	-	0.0341	0.0162	4	4	4
	Dynamic K	0.0270	0.0174	8	4	-	0.0271	0.0177	8	4	-	0.0147	0.0100	8	8	-	0.0352	0.0168	8	4	8
FREEDOM	Origin	0.0374	0.0243	-	-	-	0.0446	0.0291	-	-	-	0.0388	0.0257	-	-	-	0.0367	0.0250	-	-	-
	SG-URInit	0.0391	0.0256	4	4	-	0.0471	0.0310	4	4	-	0.0406	0.0270	8	8	-	0.0404	0.0269	4	4	4
	Dynamic K	0.0395	0.0259	8	4	-	0.0477	0.0314	8	4	-	0.0406	0.0270	8	8	-	0.0412	0.0274	8	4	8
DRAGON	Origin	0.0380	0.0249	-	-	-	0.0449	0.0296	-	-	-	0.0401	0.0270	-	-	-	0.0335	0.0228	-	-	-
	SG-URInit	0.0396	0.0262	4	4	-	0.0473	0.0311	4	4	-	0.0426	0.0286	8	8	-	0.0379	0.0250	4	4	4
	Dynamic K	0.0402	0.0266	8	4	-	0.0480	0.0314	8	4	-	0.0426	0.0286	8	8	-	0.0390	0.0256	8	4	8
MENTOR	Origin	0.0422	0.0281	-	-	-	0.0485	0.0323	-	-	-	0.0411	0.0274	-	-	-	0.0364	0.0247	-	-	-
	SG-URInit	0.0444	0.0296	4	4	-	0.0511	0.0340	4	4	-	0.0432	0.0290	8	8	-	0.0463	0.0292	4	4	4
	Dynamic K	0.0452	0.0301	8	4	-	0.0515	0.0344	8	4	-	0.0432	0.0290	8	8	-	0.0471	0.0296	8	4	8

recommendation performance, alleviates the item cold-start problem, and accelerates convergence, offering an efficient and practical solution for multimodal recommendations.

Acknowledgments

This work was supported by the UGC General Research Fund no. 17209822 and the Innovation and Technology Commission Fund no. ITS/383/23FP from Hong Kong.

References

- [1] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.
- [2] Zheyu Chen, Jinfeng Xu, Yutong Wei, and Ziyue Peng. 2025. Squeeze and Excitation: A Weighted Graph Contrastive Learning for Collaborative Filtering. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2769–2773.
- [3] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8454–8462.
- [4] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
- [5] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.
- [6] Feiran Huang, Zhenghang Yang, Junyi Jiang, Yuanchen Bei, Yijie Zhang, and Hao Chen. 2024. Large Language Model Interaction Simulator for Cold-Start Item Recommendation. *arXiv preprint arXiv:2402.09176* (2024).
- [7] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. (2024).
- [8] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [9] Ruiqi Li, Xian Wu, and Wei Wang. 2020. Adversarial learning to compare: Self-attentive prospective customer recommendation in location based social networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 349–357.
- [10] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *The world wide web conference*. 3020–3026.
- [11] Sichun Luo, Yuxuan Yao, Bowei He, Yinya Huang, Aojun Zhou, Xinyi Zhang, Yuanzhang Xiao, Mingjie Zhan, and Linqi Song. 2024. Integrating large language models into recommendation via mutual augmentation and adaptive aggregation. *arXiv preprint arXiv:2401.13870* (2024).
- [12] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.
- [13] Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. 2025. Language Representations Can be What Recommenders Need: Findings and Potentials. In *ICLR*.
- [14] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2019. Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 855–867.
- [15] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).
- [16] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [17] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).
- [18] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.
- [19] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 806–815.
- [20] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.
- [21] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.
- [22] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, Yijie Li, Mengran Li, Puzhen Wu, and Edith CH Ngai. 2025. Mdv: Enhancing multimodal recommendation with model-agnostic multimodal-driven virtual triplets. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 3378–3389.
- [23] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, and Edith CH Ngai. 2024. AlignGroup: Learning and Aligning Group Consensus with Member Preferences for Group Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2682–2691.
- [24] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Wei Wang, Xiping Hu, and Edith Ngai. 2025. Enhancing Graph Collaborative Filtering with FourierKAN Feature Transformation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 5376–5380.
- [25] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Wei Wang, Xiping Hu, Raymond Chi-Wing Wong, and Edith CH Ngai. 2025. Enhancing Robustness and Generalization Capability for Multimodal Recommender Systems via Sharpness-Aware

- Minimization. *IEEE Transactions on Knowledge and Data Engineering* (2025).
- [26] Jinfeng Xu, Zheyu Chen, Wei Wang, Xiping Hu, Sang-Wook Kim, and Edith CH Ngai. 2025. COHESION: Composite Graph Convolutional Network with Dual-Stage Fusion for Multimodal Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1830–1839.
- [27] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, and Edith CH Ngai. 2025. The Best is Yet to Come: Graph Convolution in the Testing Phase for Multimodal Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6325–6334.
- [28] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Zitong Wan, Hewei Wang, Weijie Liu, Yijie Li, and Edith CH Ngai. 2025. VI-MMRec: Similarity-Aware Training Cost-free Virtual User-Item Interactions for Multimodal Recommendation. *arXiv preprint arXiv:2512.08702* (2025).
- [29] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith CH Ngai. 2025. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12908–12917.
- [30] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2026. A survey on multimodal recommender systems: Recent advances and future directions. *IEEE Transactions on Multimedia* (2026).
- [31] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.
- [32] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6576–6585.
- [33] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.
- [34] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9154–9167.
- [35] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*. IOS Press, 3123–3130.
- [36] Xin Zhou. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497* (2023).
- [37] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.
- [38] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.