arXiv:2505.16665v1 [cs.IR] 22 May 2025

MDVT: Enhancing Multimodal Recommendation with Model-Agnostic Multimodal-Driven Virtual Triplets

Jinfeng Xu
jinfeng@connect.hku.hk
Department of Electrical and Electronic Engineering,
The University of Hong Kong
HongKong SAR, China

Zheyu Chen
zheyu.chen@connect.polyu.hk
Department of Electrical and Electronic Engineering,
The Hong Kong Polytechnic University
HongKong SAR, China

Jinze Li
lijinze-hku@connect.hku.hk
Department of Electrical and Electronic Engineering,
The University of Hong Kong
HongKong SAR, China

Shuo Yang
shuoyang.ee@gmail.com
Department of Electrical and Electronic Engineering,
The University of Hong Kong
HongKong SAR, China

Hewei Wang
heweiw@andrew.cmu.edu
Robotics Institute,
Carnegie Mellon University
Pittsburgh, PA, United States

Yijie Li
yijieli@andrew.cmu.edu
Robotics Institute,
Carnegie Mellon University
Pittsburgh, PA, United States

Mengran Li
limr39@mail2.sysu.edu.cn
School of Intelligent Systems Engineering,
Sun Yat-sen University
Shenzhen, China

Puzhen Wu
puw4002@med.cornell.edu
Population Health Sciences,
Weill Cornell Medicine
New York, NY, United States

Edith C. H. Ngai*
chngai@eee.hku.hk
Department of Electrical and Electronic Engineering,
The University of Hong Kong
HongKong SAR, China

Abstract

The data sparsity problem significantly hinders the performance of recommender systems, as traditional models rely on limited historical interactions to learn user preferences and item properties. While incorporating multimodal information can explicitly represent these preferences and properties, existing works often use it only as side information, failing to fully leverage its potential. In this paper, we propose MDVT, a model-agnostic approach that constructs multimodal-driven virtual triplets to provide valuable supervision signals, effectively mitigating the data sparsity problem in multimodal recommendation systems. To ensure high-quality virtual triplets, we introduce three tailored warm-up threshold strategies: static, dynamic, and hybrid. The static warm-up threshold strategy exhaustively searches for the optimal number of warm-up epochs but is time-consuming and computationally intensive. The dynamic warm-up threshold strategy adjusts the warm-up period based on loss trends, improving efficiency but potentially missing optimal performance. The hybrid strategy combines both, using the dynamic strategy to find the approximate optimal number of warm-up epochs and then refining it with the static strategy in a narrow hyper-parameter space. Once the warm-up threshold is satisfied, the virtual triplets are used for joint model optimization by our enhanced pair-wise loss function without causing significant gradient skew. Extensive experiments on multiple real-world datasets demonstrate that integrating MDVT into advanced multimodal recommendation models effectively alleviates the data sparsity problem and improves recommendation performance, particularly in sparse data scenarios.

CCS Concepts

• **Information systems → Recommender systems;**.

Keywords

Recommender System, Multimedia, Model-Agnostic, Virtual Triplets

ACM Reference Format:
Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, Yijie Li, Mengran Li, Puzhen Wu, and Edith C. H. Ngai. 2025. MDVT: 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 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3737042

1 Introduction

The rapid development of the internet has led to an information explosion, making recommender systems indispensable for navigating vast amounts of data. Traditional recommender systems rely

[*] Corresponding authors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '25, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737042

on modeling user preferences through historical user-item interactions [6, 9, 30, 31]. However, the data sparsity problem significantly hinders the performance of these systems, as they depend solely on limited historical interactions to implicitly learn user preferences and item properties. Incorporating multimodal information [2, 15]—such as images and textual descriptions—allows for explicit representation of user preferences and item properties, potentially alleviating the data sparsity problem. Several recent works [3, 8, 34] have integrated multimodal content into recommendation models. For example, VBPR [8] extends the matrix factorization framework to incorporate item visual features, while ACF [3] introduces a hierarchically structured attention network to capture user preferences at the component level. Graph Convolutional Networks (GCNs) have also gained attention in this context [7, 27, 28, 39, 44]. Models like MMGCN [28] and GRCN [27] employ GCNs to integrate multimodal information into the message-passing process, enhancing the inference of user and item representations. To further exploit the rich multimodal information, LATTICE [39] and FREEDOM [44] construct item-item graphs to aggregate semantically similar items. LGMRec [7] utilizes hyper-graph structures to learn both global and local representations, capturing complex relationships in multimodal information.

However, existing works typically use multimodal information only as side information to enhance the learning of user preferences, failing to fully leverage its potential. They primarily focus on improving item representations using multimodal content, while user representations are still learned solely from historical interactions. This limitation becomes more pronounced in data sparsity scenarios, where users have limited interaction records.

We propose that the similarity between user and item modality representations can serve as valuable supervision signals beyond explicit user-item interactions. To leverage this insight, we introduce **M**ultimodal-**D**riven **V**irtual **T**riplets (MDVT), a novel, model-agnostic approach that constructs virtual triplets based on multimodal information. These virtual triplets provide informative supervision signals, effectively mitigating the data sparsity problem in multimodal recommendation systems. A key challenge is that, unlike items, users do not have inherent multimodal information in recommendation scenarios. Users' modality representations must be learned from scratch by initializing embeddings randomly and refining them through model optimization. Consequently, the initial similarity between user and item modality representations may not provide high-quality supervision signals. To address this, we introduce three tailored warm-up threshold strategies:

- **Static Warm-up Threshold Strategy**: This strategy exhaustively searches for the optimal number of warm-up epochs, ensuring that user modality representations are sufficiently learned before constructing virtual triplets. While effective, it is time-consuming and computationally intensive due to the thorough hyper-parameter tuning required.
- **Dynamic Warm-up Threshold Strategy**: This strategy adjusts the warm-up period based on the trend of loss changes during training. It improves efficiency by reducing the need for extensive hyper-parameter tuning, automatically determining when user representations are adequately learned. However, it may not always find the optimal number of warm-up epochs compared to the static strategy.
- **Hybrid Warm-up Threshold Strategy**: Combining the strengths of both strategies, the hybrid strategy first employs the dynamic strategy to find the approximate optimal number of warm-up epochs and then applies the static strategy within a narrow hyper-parameter space. This allows for efficient training with a balance between computational cost and performance optimization.

Once the warm-up threshold is satisfied, the virtual triplets are used for joint model optimization through our enhanced pair-wise loss function, enhancing the learning process without causing significant gradient skew [13, 38]. Our MDVT approach is plug-and-play and can be easily integrated into any existing multimodal recommendation model, improving their performance, particularly in data sparsity scenarios. To validate the effectiveness of MDVT, we conducted extensive experiments on multiple real-world datasets adopting various advanced multimodal recommendation models. The results demonstrate that integrating MDVT into these models significantly alleviates the data sparsity problem and improves recommendation performance, especially for users with limited interaction records. Additionally, we would like to highlight the key distinction between our work and prior studies. Our virtual triplets are constructed based on the similarity between dynamically learned user and item representations, which are better aligned with the recommendation task with a sufficient warm-up phase. In contrast, prior works typically construct virtual samples based on the similarity of raw features between items.

2 Preliminary

In this section, we provide an overview of graph collaborative filtering (GCF), the common paradigm of advanced multimodal recommendations, which adopts graph neural network (GNN) into collaborative filtering (CF) with multimodal features. CF tasks usually contains a user set $\mathcal{U} = \{u_1, ..., u_{|\mathcal{U}|}\}$ and an item set $\mathcal{I} = \{i_1, ..., i_{|\mathcal{I}|}\}$. 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. GCF naturally constructs the bipartite graph by user-item interaction matrix $\mathcal{R}$. This graph can be denoted by $\mathcal{G} = (\mathcal{U}, \mathcal{I}, \mathcal{E})$, 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 $(i, u) \in \mathcal{E}$. We random initialize $\mathbf{E}_{u^m} \in \mathbb{R}^{d_m\times|\mathcal{U}|}$ to represent user embedding with modality m. $\mathbf{E}_{i^m} \in \mathbb{R}^{d_m\times|\mathcal{I}|}$ represents item initialized embedding with modality m, which extracted by pre-trained encoders. Here d_m represents the hidden dimensionality. Based on the user-item graph $\mathcal{G}$, GNNs conduct neighbor aggregation to enhance user/item embeddings for extracting high-order user-item collaborative signals. Take the most widely-used GNN backbone LightGCN [9] as an example, the embeddings for user u and item i in the l-th layer are:

$$\mathbf{e}_{u_m}^{(l)} = \frac{1}{d_u} \sum_{j|(u,j)\in\mathcal{E}} \frac{1}{d_j}\mathbf{e}_{j_m}^{(l-1)}, \quad \mathbf{e}_{i_m}^{(l)} = \frac{1}{d_i} \sum_{v|(i,v)\in\mathcal{E}} \frac{1}{d_v}\mathbf{e}_{v_m}^{(l-1)}, \quad (1)$$

where d_* denotes degree of node. $\mathbf{e}_*^{(l)}$ represents node embedding in l-th layer. After L layers of neighbor aggregation, the final representations of modality m for user u and item i as:

$$\bar{\mathbf{e}}_{u_m} = \sum_{l=0}^{L} \mathbf{e}_{u_m}^{l}, \quad \bar{\mathbf{e}}_{i_m} = \sum_{l=0}^{L} \mathbf{e}_{i_m}^{l}. \tag{2}$$

The predicted user-item relation score can be calculated by $\hat{y}_{u,i} = \sum_{m \in \mathcal{M}} (\bar{\mathbf{e}}_{u_m}^{\top} \bar{\mathbf{e}}_{i_m})$. With the prediction scores $\hat{y}_{u,i}$, the GNN models are optimized by minimizing the BPR loss function [20]:

$$\mathcal{L}_{bpr} = \sum_{(u,i^+,i^-) \in \mathcal{D}} - \log(\sigma(\hat{y}_{u,i^+} - \hat{y}_{u,i^-})), \tag{3}$$

where triplet training dataset $\mathcal{D}$ contains all positive user-item pairs $(u, i^+) \in \mathcal{E}$ and sampled negative user-item pairs $(u, i^-) \notin \mathcal{E}$. $\sigma(\cdot)$ denotes activation function. Though the above GCF paradigm achieves state-of-the-art performance in the recommendation field, its performance is limited by scarce interaction records. In light of this, this paper proposes MDVT, which leverages informative and valuable multimodal information to construct virtual training triplets to mitigate the data sparsity problem.

3 Methodology

In this section, we present our MDVT, a plug-and-play framework, which can improve all existing multimodal recommendation models' performance by constructing virtual training triplets to mitigate the data sparsity problem. The overall framework of our proposed MDVR is illustrated in Figure 1[1]. Our proposed MDVT contains three main components:

- **Multimodal-Driven Virtual Triplets Constructor** (Section 3.1): We construct virtual triplets by the top-n positive and negative items for each user based on fused multimodal representation.
- **Threshold Strategies** (Section 3.2): To ensure the quality of virtual triples, we define three different threshold strategies: 1) **Static**: a heuristic static warm-up threshold strategy, 2) **Dynamic**: a loss-based dynamic warm-up threshold strategy, and 3) **Hybrid**: a hybrid warm-up threshold strategy.
- **Enhanced Pair-wise Loss Function** (Section 3.3): Based on our constructed virtual triplets, we propose a simple yet effective enhanced pair-wise loss function, which can be directly plugged into all multimodal recommendation models.

3.1 Multimodal-Driven Virtual Triplets

Compared to traditional recommendation scenarios, items in multimodal recommendation settings contain rich modality features such as visual and textual information. While most previous multimodal recommendation studies [11, 44] have used multimodal information merely as side information to infer user preferences, recent studies [18, 43] in the explainable recommendation field demonstrate that leveraging multimodal data can explicitly reveal user preferences and item attributes. Inspired by these explainable recommendation approaches, we utilize modality information to provide additional supervision signals to alleviate the data sparsity problem that recommender systems often suffer from. To better plug into different multimodal recommendation models, we construct virtual triplets based on the final aggregated representations. We simplify the final representation aggregation for different multimodal recommendation models as:

$$\bar{\mathbf{e}}_u = \mathrm{F}(\bar{\mathbf{e}}_{u_m} | m \in \mathcal{M}) \quad \bar{\mathbf{e}}_i = \mathrm{F}(\bar{\mathbf{e}}_{i_m} | m \in \mathcal{M}), \tag{4}$$

where $\mathrm{F}(\cdot)$ represents modality representation fusion operation. Here, the fused representation contains valuable multimodal information [41, 44, 45], which can be used to construct model-agnostic multimodal-driven virtual triplets. To calculate the top-n positive and negative representation for each user node to construct virtual triplets, we maintain a user-item representation similarity matrix by cosine similarity, formally:

$$\mathcal{S}_{u,i} = \frac{\bar{\mathbf{e}}_u^{\top} \bar{\mathbf{e}}_\mathbf{i}}{\|\bar{\mathbf{e}}_u\| \|\bar{\mathbf{e}}_\mathbf{i}\|}. \tag{5}$$

Based on the similarity matrix, for each user, we select the most similar n and the least similar n items to construct the virtual triplet:

$$\mathcal{D}_{u,i^+}^{V} = \mathrm{Max_n}(\mathcal{S}_{u,i^*} | i^* \in \mathcal{I}), \quad \mathcal{D}_{u,i^-}^{V} = \mathrm{Max_n}(-\mathcal{S}_{u,i^*} | i^* \in \mathcal{I}), \tag{6}$$

where $\mathrm{Max_n}(\cdot)$ denotes top-n similarity filter operation. For user u, $\mathcal{D}_{u,i^+}^{V}$ and $\mathcal{D}_{u,i^-}^{V}$ contain n positive items and negative items, respectively. Therefore, we construct a new virtual triplet training dataset $\mathcal{D}^V$, which is updated with model optimization. For each triplet with user u in this virtual triplet training dataset $\mathcal{D}^V$ can be expressed as $\mathcal{D}_u^V = (u, \mathcal{D}_{u,i^+}^{V}, \mathcal{D}_{u,i^-}^{V})$.

3.2 Threshold Strategies

High-quality representations are essential for constructing effective virtual triplets. It is worth noting that, unlike items—which naturally possess multimodal information such as images and textual descriptions—users in recommendation scenarios do not inherently have multimodal information. Therefore, user representations are initially randomized and progressively refined during training, the model requires sufficient warm-up epochs to ensure the representations are adequately optimized for constructing high-quality triplets. To this end, we propose three threshold strategies to determine the optimal number of warm-up epochs. These strategies ensure that virtual triplets are incorporated only when the representation quality is sufficient to provide high-quality supervision signals. Specifically, we propose three warm-up threshold strategies: Static Warm-up Threshold (**Static**), Dynamic Warm-up Threshold (**Dynamic**), and Hybrid Warm-up Threshold (**Hybrid**).

3.2.1 *Static Warm-up Threshold Strategy.*

The warm-up epochs required for learning high-quality representations varies across different models and hyper-parameter settings (e.g., learning rate, batch size). Therefore, a simple and effective strategy is to set a predefined threshold $\mathcal{T}_\mathcal{S}$, and our virtual triplets will jointly optimize the multimodal recommendation model after $\mathcal{T}_\mathcal{S}$ epochs training. For our static warm-up threshold strategy, we search the optimal number of warm-up epochs within a manually defined threshold set $\mathcal{S}_\mathcal{T}$. When computational resources are abundant, we can search for the optimal parameters by exhaustively traversing the hyper-parameter space. Conversely, when computational resources are

[1] While Figure 1 only depicts ID embeddings, visual and textual modalities, it is important to note that our MDVT is model-agnostic and can be easily applied to all multimodal recommendation models, regardless of the number and types of modalities involved.

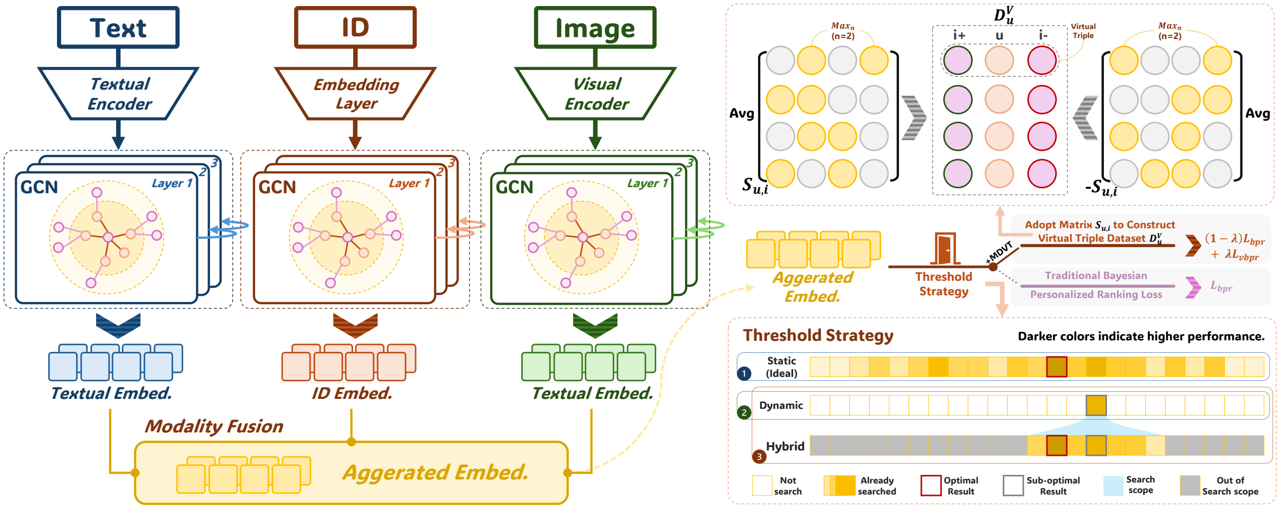


Figure 1: The overall architecture of our proposed MDVT.

limited, we rely on researchers' extensive domain knowledge and experience in model training to define threshold set $\mathcal{S}_{\mathcal{T}}$.

3.2.2 Dynamic Warm-up Threshold Strategy. In scenarios with limited computational resources, exhaustively traversing all hyper-parameters to find the optimal warm-up epoch number is impractical, and researchers may not be familiar with all existing models to manually set thresholds appropriately. Therefore, a dynamic threshold strategy with a lower hyper-parameter tuning cost is necessary. Inspired by numerous studies [1, 13, 21, 29] on the relationship between training loss and model convergence, we propose a dynamic warm-up threshold strategy based on the trend of loss changes. Specifically, we assess whether the model is approaching convergence by comparing the ratio of the loss decrease between the current and previous epochs. When the rate of loss change is low enough, indicating that the model has sufficiently converged and this epoch is an approximate optimal threshold $\mathcal{T}_O$. Then, we adopt virtual triplets to optimize the multimodal recommendation model jointly. We define the loss at epoch t as $\mathcal{L}^t$. Virtual triplets are incorporated into the model optimization process when the rate of loss change falls below a pre-defined hyper-parameter g.

3.2.3 Hybrid Warm-up Threshold Strategy. An effective and satisfactory threshold selection strategy is the hybrid warm-up threshold strategy, which combines the dynamic warm-up threshold strategy and static warm-up threshold strategy. Specifically, it adopts the dynamic warm-up threshold strategy to find an approximate optimal threshold $\mathcal{T}^{cur}$, then adopts the static warm-up threshold strategy within a small scope $[\mathcal{T}^{cur}-s, \mathcal{T}^{cur}+s]$ to find the optimal threshold $\mathcal{T}_O$, where s is the search scope hyper-parameter. This hybrid warm-up threshold strategy allows for a high probability of finding the optimal number of warm-up epochs without searching the entire hyper-parameter space.

Analysis. The static warm-up threshold strategy requires comprehensive hyper-parameter tuning because it involves manually selecting the optimal number of warm-up epochs through an exhaustive search of the hyper-parameter space. In scenarios where a full traversal of hyper-parameters is feasible, this strategy can effectively find the optimal number of warm-up epochs, ensuring the model performs at its best. However, this process is time-consuming and computationally intensive due to the high demand for hyper-parameter tuning. In contrast, the dynamic warm-up threshold strategy reduces the need for extensive hyper-parameter tuning by automatically selecting the number of warm-up rounds based on the trend of loss change. This strategy adjusts dynamically to each model's loss change trend, allowing for a more efficient training process with lower hyper-parameter tuning demands. The dynamic warm-up threshold strategy is particularly beneficial when computational resources are limited. However, despite its advantages, the dynamic warm-up threshold strategy may not always find the optimal number of warm-up epochs compared to the static warm-up threshold strategy with full hyper-parameter traversal. Since it does not exhaustively explore the hyper-parameter space, there's a possibility that it might miss the optimal number of warm-up epochs for a given model. Therefore, while the dynamic warm-up threshold strategy improves efficiency and requires less manual tuning, it might sacrifice some performance optimization achievable through the static method. Moreover, the hybrid warm-up threshold strategy combines the advantages of both the static and dynamic warm-up threshold strategies. Specifically, it first adopts the dynamic warm-up threshold strategy to find the approximate optimal number of warm-up epochs. Then, it applies the static warm-up threshold strategy within a small scope, offering the potential for optimal performance. We present the procedure in Appendix A.1.

3.3 Enhanced Pair-wise Loss Function

Once the multimodal recommendation model has learned high-quality representations (when the threshold strategy is satisfied), we adopt the widely used Bayesian Personalized Ranking (BPR) loss on our virtual triplet training dataset to optimize the model:

$$\mathcal{L}_{vbpr} = \sum_{(u,\mathcal{D}^V_{u,i^+},\mathcal{D}^V_{u,i^-})\in\mathcal{D}^V} -\log(\sigma(\bar{\mathbf{e}}_u^\top \hat{\mathbf{e}}_{u,i^+} - \bar{\mathbf{e}}_u^\top \hat{\mathbf{e}}_{u,i^-})), \quad (7)$$

$$\hat{\mathbf{e}}_{u,i^+} = \frac{1}{n}\sum_{i^+\in\mathcal{D}^V_{u,i^+}} \bar{\mathbf{e}}_{i^+} \quad \hat{\mathbf{e}}_{u,i^-} = \frac{1}{n}\sum_{i^-\in\mathcal{D}^V_{u,i^-}} \bar{\mathbf{e}}_{i^-}, \quad (8)$$

Table 1: Statistics of the three 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

where we calculate the representation mean of the similar group and the representation mean of the dissimilar group to get informative representations of the virtual positive item and negative item, respectively. We jointly optimize the model with two loss functions: the BPR loss defined in Eq.3 applied to the training dataset $\mathcal{D}$, and the BPR loss applied to the virtual training dataset $\mathcal{D}^V$ as defined in Eq.7. The final learning loss can be expressed as:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{bpr} + \lambda\mathcal{L}_{vbpr}, \tag{9}$$

where λ regulates the influence of our virtual training loss $\mathcal{L}_{vbpr}$. Note that the users involved in $\mathcal{L}_{bpr}$ and $\mathcal{L}_{vbpr}$ completely overlap, and each training triplet in both cases consists of one positive item and one negative item. Adding $\mathcal{L}_{vbpr}$ alters the loss magnitude, so λ ensures balanced scaling to prevent gradient skew [13, 38], enabling smooth joint optimization.

4 Evaluation

We conduct extensive experiments on MDVT, aiming to answer the following research questions (RQs): **RQ1:** Can MDVT enhance the performance of multimodal recommender systems? **RQ2:** How do the various components in MDVT affect performance enhancement? **RQ3:** What are the effectiveness and costs of different threshold strategies in MDVT? **RQ4:** Can MDVT have a positive impact on the convergence speed? **RQ5:** Can MDVT be compatible with robust training and data augmentation strategies? **RQ6:** How do different warm-up threshold strategies work in practical training? **RQ7:** What is the impact of key hyper-parameters in MDVT?

4.1 Experimental Settings

4.1.1 Datasets. The experiments are conducted on three real-world datasets: Baby, Sports, and Clothing from Amazon [19]. All the datasets comprise textual and visual features in the form of item descriptions and images. To further evaluate the performance of MDVT in scenarios involving multiple modalities, we also conduct experiments on the TikTok dataset [11]. Our data preprocessing methodology follows the approach outlined in MMRec [42]. Table 1 shows the statistics of these datasets. We adopt two widely used metrics to evaluate the performance fairly: Recall@K (R@K) and NDCG@K (N@K). We report the average metrics of all users in the test dataset under both K = 5 and K = 10. We follow the popular evaluation setting [7, 44] with a random data splitting 8:1:1 for training, validation, and testing.

4.1.2 Baselines. We extensively examine the performance of our MDVT across a variety of multimodal recommendation models, including MMGCN [28], SLMRec [23], FREEDOM [44], DRAGON [45], LGMRec [7], and MMSSL [25]. Moreover, we test the compatibility of our MDVT with the adversarial training strategy (AMR [22]) and LLM-based data augmentation strategy (GPT-4o [36]).

4.1.3 Implementation Details. We retain the standard settings for all baselines and fix batch size B to 2048. For MDVT, we apply a grid search on hyper-parameters λ in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, and the value n in Eq. 6 in $\{1, 2, 4, 8\}$. For the static warm-up threshold strategy, we define threshold set as $\mathcal{S}_{\mathcal{T}} = \{0, 5, 10, 20, 40, 80\}$. We perform a grid search in $\{0.1, 0.2, 0.3, 0.4\}$ for hyper-parameter g in dynamic and hybrid warm-up threshold strategies, a grid search in $\{1, 2, 3, 4, 5\}$ for hyper-parameter s in hybrid warm-up threshold strategy. The common optimizer is Adam [12] 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 [36] 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.

4.2 Overall Performance (RQ1)

We evaluate the effectiveness of our MDVT with various warm-up threshold strategies on various models for multimodal recommendation scenarios. From Table 2, we find the following observations: **Observation1: MDVT with all warm-up threshold strategies can enhance the performance of various multimodal recommendation models.** As shown in Table 2, we conduct extensive experiments with MDVT on five multimodal recommendation models across three distinct public datasets. The experimental results demonstrate that all warm-up threshold strategies significantly improve over all baselines across all evaluation metrics. The static warm-up threshold strategy consistently achieves superior results compared to the dynamic warm-up threshold strategy. Moreover, the hybrid warm-up threshold strategy consistently outperforms both the static and dynamic warm-up threshold strategies. In summary, the experimental results validate that leveraging multimodal information to construct virtual triplets can effectively improve recommender performance by mitigating the data sparsity problem. **Observation2: Hybrid warm-up threshold strategy can find ideal warm-up epochs within an affordable hyper-parameter tuning cost.** For all multimodal recommendation models across all datasets, the dynamic warm-up threshold strategy achieves performance improvements comparable to the static warm-up threshold strategy. This indicates that the dynamic warm-up threshold strategy can identify approximately optimal warm-up epochs without requiring extensive hyper-parameter tuning or substantial expert knowledge. Building on this, the hybrid warm-up threshold strategy utilizes the approximately optimal number of warm-up epochs found by the dynamic strategy to adopt the static warm-up threshold strategy within a small scope. Consequently, it finds the ideal number of warm-up epochs closer to the optimal number and achieves superior performance improvements within an affordable hyper-parameter tuning cost.

4.3 Ablation Study (RQ2)

To discern the impact of our MDVT's core components, we conducted an ablation study with various configurations:

Table 2: Performance comparison of baselines with or without MDVT 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. (S), (D), and (H) denote Static, Dynamic, and Hybrid, respectively.

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	0.0331	0.0463	0.0172	0.0231
+MDVT (S)	0.0255*	0.0418*	0.0169*	0.0221*	0.0234*	0.0404*	0.0157*	0.0211*	**0.0148***	**0.0244***	**0.0099***	**0.0126***	0.0381*	0.0539*	0.0201*	0.0268*
Improv.	6.25%	10.58%	5.62%	10.50%	8.33%	9.19%	9.79%	9.33%	13.85%	11.93%	12.50%	14.55%	15.11%	16.41%	16.86%	16.02%
+MDVT (D)	0.0251*	0.0413*	0.0167*	0.0218*	0.0231*	0.0400*	0.0155*	0.0208*	**0.0148***	**0.0244***	**0.0099***	**0.0126***	0.0373*	0.0520*	0.0197*	0.0263*
Improv.	4.58%	9.26%	4.37%	9.00%	6.94%	8.11%	8.39%	7.77%	13.85%	11.93%	12.50%	14.55%	12.69%	12.31%	14.53%	13.85%
+MDVT (H)	**0.0257***	**0.0420***	**0.0170***	**0.0224***	**0.0236***	**0.0406***	**0.0158***	**0.0213***	**0.0148***	**0.0244***	**0.0099***	**0.0126***	**0.0383***	**0.0543***	**0.0203***	**0.0272***
Improv.	7.08%	11.11%	6.25%	12.00%	9.26%	9.73%	10.49%	10.36%	13.85%	11.93%	12.50%	14.55%	15.71%	17.28%	18.02%	17.75%
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	0.0349	0.0503	0.0188	0.0251
+MDVT (S)	0.0357*	0.0560*	0.0239*	0.0319*	0.0458*	0.0705*	0.0307*	0.0388*	0.0317*	0.0493*	0.0214*	0.0267*	0.0384*	0.0550*	0.0208*	0.0276*
Improv.	4.08%	5.86%	5.75%	10.00%	6.76%	6.33%	6.60%	6.30%	8.56%	9.07%	9.18%	8.10%	10.03%	9.34%	10.64%	9.96%
+MDVT (D)	0.0353*	0.0553*	0.0234*	0.0313*	0.0456*	0.0701*	0.0305*	0.0385*	0.0314*	0.0488*	0.0212*	0.0263*	0.0379*	0.0542*	0.0204*	0.0271*
Improv.	2.92%	4.54%	3.54%	7.93%	6.29%	5.73%	5.90%	5.48%	7.53%	7.96%	8.16%	6.48%	8.60%	7.75%	8.51%	7.97%
+MDVT (H)	**0.0359***	**0.0563***	**0.0241***	**0.0323***	**0.0460***	**0.0709***	**0.0309***	**0.0391***	**0.0319***	**0.0497***	**0.0216***	**0.0270***	**0.0393***	**0.0566***	**0.0209***	**0.0284***
Improv.	4.66%	6.43%	6.64%	11.38%	7.23%	6.94%	7.29%	7.12%	9.25%	9.96%	10.20%	9.31%	12.61%	12.52%	11.17%	13.15%
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	0.0399	0.0589	0.0214	0.0295
+MDVT (S)	0.0391*	0.0652*	0.0257*	0.0350*	0.0472*	0.0752*	0.0312*	0.0406*	0.0410*	0.0662*	0.0275*	0.0361*	0.0427*	0.0629*	0.0226*	0.0312*
Improv.	4.55%	3.99%	5.76%	6.06%	5.83%	4.88%	7.22%	5.45%	5.67%	5.25%	7.00%	5.87%	7.02%	6.79%	5.61%	5.76%
+MDVT (D)	0.0387*	0.0648*	0.0253*	0.0347*	0.0469*	0.0750*	0.0310*	0.0403*	0.0405*	0.0655*	0.0270*	0.0354*	0.0420*	0.0622*	0.0223*	0.0309*
Improv.	3.48%	3.35%	4.12%	5.15%	5.16%	4.60%	6.53%	4.68%	4.38%	4.13%	5.06%	3.81%	5.26%	5.60%	4.21%	4.75%
+MDVT (H)	**0.0398***	**0.0662***	**0.0262***	**0.0357***	**0.0476***	**0.0757***	**0.0315***	**0.0410***	**0.0412***	**0.0665***	**0.0277***	**0.0364***	**0.0431***	**0.0629***	**0.0229***	**0.0318***
Improv.	6.42%	5.58%	7.82%	8.18%	6.73%	5.58%	8.25%	6.49%	6.19%	5.72%	7.78%	6.74%	8.02%	6.79%	7.01%	7.80%
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	0.0451	0.0682	0.0244	0.0341
+MDVT (S)	0.0396*	0.0689*	0.0262*	0.0360*	0.0474*	0.0780*	0.0311*	0.0434*	0.0430*	0.0710*	0.0287*	0.0385*	0.0475*	0.0718*	0.0259*	0.0362*
Improv.	4.21%	4.08%	5.22%	4.35%	5.57%	3.72%	5.07%	5.08%	7.23%	5.81%	6.30%	5.48%	5.32%	5.28%	6.15%	6.16%
+MDVT (D)	0.0391*	0.0685*	0.0259*	0.0357*	0.0470*	0.0776*	0.0308*	0.0430*	0.0428*	0.0704*	0.0285*	0.0382*	0.0471*	0.0712*	0.0257*	0.0359*
Improv.	2.89%	3.47%	4.02%	3.48%	4.68%	3.20%	4.05%	4.12%	6.73%	4.92%	5.56%	4.66%	4.43%	4.40%	5.33%	5.28%
+MDVT (H)	**0.0398***	**0.0692***	**0.0264***	**0.0364***	**0.0479***	**0.0788***	**0.0314***	**0.0440***	**0.0432***	**0.0713***	**0.0288***	**0.0387***	**0.0480***	**0.0724***	**0.0262***	**0.0366***
Improv.	4.74%	4.53%	6.02%	5.51%	6.68%	4.79%	6.08%	6.54%	7.73%	6.26%	6.67%	6.03%	6.43%	6.16%	7.38%	7.33%
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	0.0406	0.0610	0.0217	0.0304
+MDVT (S)	0.0416*	0.0656*	0.0280*	0.0359*	0.0474*	0.0769*	0.0311*	0.0417*	0.0409*	0.0619*	0.0274*	0.0335*	0.0431*	0.0641*	0.0231*	0.0320*
Improv.	11.23%	4.79%	12.45%	7.81%	6.28%	6.95%	7.99%	7.75%	10.24%	11.53%	11.38%	10.93%	6.16%	5.08%	6.45%	5.26%
+MDVT (D)	0.0413*	0.0651*	0.0279*	0.0356*	0.0471*	0.0762*	0.0308*	0.0412*	0.0403*	0.0612*	0.0270*	0.0328*	0.0425*	0.0635*	0.0228*	0.0317*
Improv.	10.43%	3.99%	12.05%	6.91%	5.61%	5.98%	6.94%	6.46%	8.63%	10.27%	9.76%	8.61%	4.68%	4.10%	5.07%	4.28%
+MDVT (H)	**0.0417***	**0.0660***	**0.0281***	**0.0360***	**0.0475***	**0.0771***	**0.0312***	**0.0419***	**0.0411***	**0.0622***	**0.0276***	**0.0337***	**0.0435***	**0.0647***	**0.0233***	**0.0324***
Improv.	11.50%	5.43%	12.85%	8.11%	6.50%	7.23%	8.33%	8.27%	10.78%	12.07%	12.20%	11.92%	7.14%	6.07%	7.37%	6.58%
MMSSL	0.0369	0.0613	0.0241	0.0326	0.0451	0.0693	0.0294	0.0369	0.0382	0.0619	0.0253	0.0335	0.0395	0.0575	0.0210	0.0287
+MDVT (S)	0.0396*	0.0655*	0.0257*	0.0348*	0.0478*	0.0732*	0.0312*	0.0389*	0.0402*	0.0648*	0.0267*	0.0351*	0.0423*	0.0613*	0.0224*	0.0305*
Improv.	7.32%	6.85%	6.64%	6.75%	5.99%	5.63%	6.12%	5.42%	5.24%	4.68%	5.53%	4.78%	7.09%	6.61%	6.67%	6.27%
+MDVT (D)	0.0388*	0.0649*	0.0252*	0.0343*	0.0473*	0.0727*	0.0308*	0.0385*	0.0400*	0.0643*	0.0263*	0.0348*	0.0419*	0.0609*	0.0221*	0.0302*
Improv.	5.15%	5.87%	4.56%	5.21%	4.88%	4.91%	4.76%	4.34%	4.71%	3.88%	3.95%	3.88%	6.08%	5.91%	5.24%	5.23%
+MDVT (H)	**0.0404***	**0.0667***	**0.0262***	**0.0354***	**0.0486***	**0.0745***	**0.0317***	**0.0396***	**0.0407***	**0.0657***	**0.0270***	**0.0356***	**0.0429***	**0.0621***	**0.0228***	**0.0310***
Improv.	9.49%	8.81%	8.71%	8.59%	7.76%	7.50%	7.82%	7.32%	6.54%	6.14%	6.72%	6.27%	8.61%	8.00%	8.57%	8.01%

- w/o-Aggr: This configuration removes the representation averaging operation specified in Eq. 8, resulting in an asymmetry between the triplets in $\mathcal{D}^V$ and $\mathcal{D}$. Specifically, in $\mathcal{D}^V$, each user is associated with n triplets, whereas in $\mathcal{D}$, each user has 1 triplet.
- w/o-Scale: This configuration removes the align-scale operation in Eq. 9 by modifying the loss function to $\mathcal{L} = \mathcal{L}_{bpr} + \lambda \mathcal{L}_{vbpr}$. This alteration causes the virtual triplet loss to introduce significant gradient skew when incorporated into the training process.

We conduct extensive experiments for our MDVT across five multimodal recommendation models on the Baby dataset for various configurations. The findings presented in Figure 2 clearly demonstrate that our MDVT surpasses all its modified configurations, thereby confirming the essential role each component plays in learning high-quality representations. We believe that the inferior performance of all configurations compared to MDVT is due to the discrepancy in scale between the model loss during epochs when virtual triplets participate in joint optimization and the training loss during the warm-up epochs. This discrepancy leads to gradient skew, which is similar to that observed in multi-task learning [13, 38]. To validate our statement, we further investigate the changes in model loss and recommendation performance throughout the training phase. Specifically, we visualize the training loss and recommendation performance (NDCG@10) for two multimodal recommendation models with our MDVT and its configurations with the static warm-up threshold strategy (20 epoch warm-up) on the Baby dataset. The experimental results presented in Figure 3 support our statement. Specifically, compared with MDVT, all configurations exhibit a significant increase in loss and a noticeable

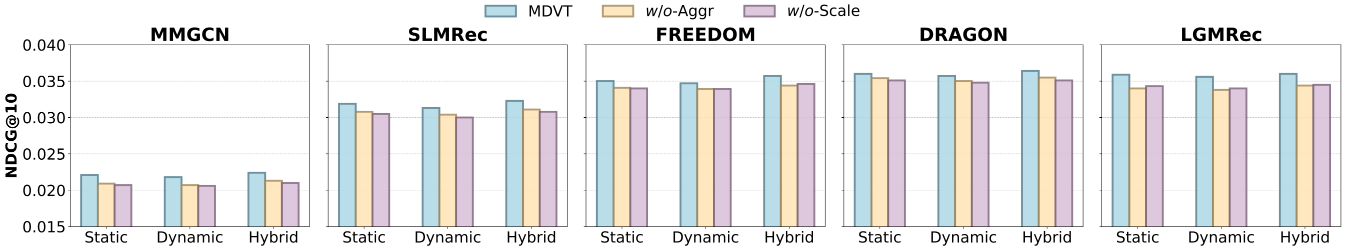


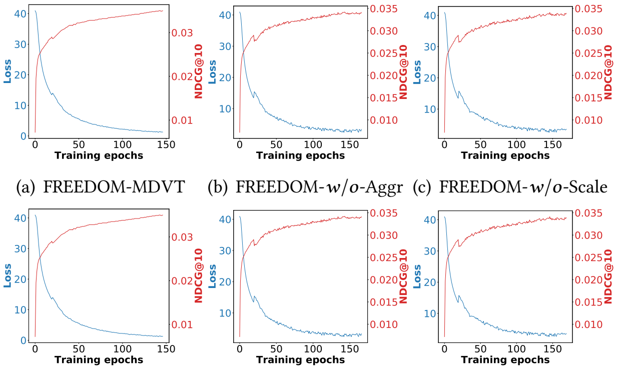
Figure 2: Ablation study on key components of MDVT in terms of NDCG@10.

decrease in performance after 20 warm-up epochs. Furthermore, their subsequent training is more unstable and requires longer training times than MDVT. These observations indicate that our key component design allows virtual triplets to be added to the training process without causing significant gradient skew, thereby maintaining stability and efficiency.

Furthermore, we verify the effect of virtual triples construction. Specifically, we conducted an ablation study with various configurations: a) MDVT$_{ID}$, which builds virtual triples via only ID modality. b) MDVT$_{V}$, which builds virtual triples via only visual modality. c) MDVT$_{T}$, which builds virtual triples via only textual modality. d) MDVT$_{ID-V}$, which builds virtual triples via ID and visual modalities. e) MDVT$_{ID-T}$, which builds virtual triples via ID and textual modalities. f) MDVT$_{V-T}$, which builds virtual triples via both visual and textual modalities. g) MDVT$_{F1}$, which builds virtual triplets directly based on high and low interaction frequencies. g) MDVT$_{F2}$, which top-2n virtual triplets based on representation similarity, selects the top-n triplets according to interaction frequency. All experiments adopt the hybrid strategy. According to experimental results in Table 3, we have the following observations. 1) Models using only visual/textual modalities perform even worse than the original model, as the lack of ID causes the generated virtual triplets to deviate from the recommendation task. 2) The variant using only ID outperforms the original model, demonstrating the dominant role of ID. 3) The performance is further improved when ID is combined with visual/textual modalities. This suggests that the auxiliary modalities provide more modality information. 4) Constructing virtual triplets based on interaction frequency reduces the personalization of recommendations, thereby degrading the overall recommendation performance. 5) MDVT achieves the best performance, attributed to its ability to fuse informative information from multiple modalities to achieve optimal performance.

4.4 Sparsity Study (RQ3)

To evaluate the effectiveness of adopting MDVT in advanced multimodal recommendation models under various data sparsity scenarios, we conduct experiments on subsets of all three datasets with differing sparsity levels. In particular, we compare the performance of three advanced multimodal recommendation models—SLMRec, FREEDOM, and LGMRec—with and without our MDVT. To analyze the effect of data sparsity, we categorize user groups based on their interaction counts in the training set (e.g., the first group consists of users who have interacted with 1-5 items). As illustrated in Figure 4, MDVT consistently enhances the performance of these models across all datasets and sparsity levels, thereby confirming its effectiveness in diverse sparse scenarios. Furthermore, the improvement in recommendation performance of all models with MDVT is particularly significant in sparse scenarios, specifically for users with 1-5 and 6-10 interacted items. We attribute this enhancement to our virtual triplets being especially effective in sparse scenarios.



(a) FREEDOM-MDVT (b) FREEDOM-*w/o*-Aggr (c) FREEDOM-*w/o*-Scale

(d) DRAGON-MDVT (e) DRAGON-*w/o*-Aggr (f) DRAGON-*w/o*-Scale

Figure 3: The learning curve when adopting MDVT and its configurations to optimize the loss on the Baby dataset.

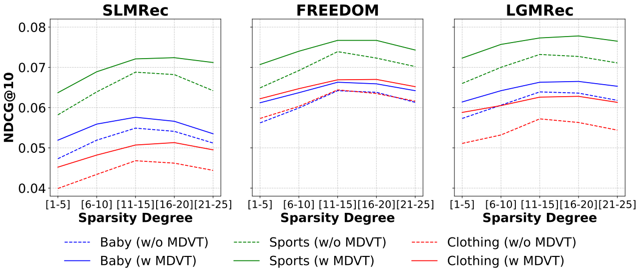


Figure 4: Sparsity study on three advanced multimodal recommendation models across three distinct datasets.

4.5 Convergence Speed (RQ4)

In addition, MDVT helps accelerate model training convergence. We visualized the training loss of three advanced multimodal recommendation models (MMGCN, FREEDOM, and DRAGON) on the Baby dataset. Following previous training settings [44, 45], we used an early stopping strategy with the patience of 20 epochs and set the maximum number of epochs to 1,000. As shown in Figure 5, MDVT effectively improves the convergence speed of all models. We attribute this improvement to the virtual triplets providing informative supervision signals that accelerate model training convergence.

Table 3: Performance comparison for variants on three datasets in terms of Recall@10 (R@10).

Dataset	Model	Original	MDVT$_{ID}$	MDVT$_{V}$	MDVT$_{T}$	MDVT$_{ID-V}$	MDVT$_{ID-T}$	MDVT$_{V-T}$	MDVT$_{F1}$	MDVT$_{F2}$	MDVT
Baby	MMGCN	0.0378	0.0381	0.0360	0.0370	0.0401	0.0412	0.0367	0.0354	0.0410	**0.0420**
	SLMRec	0.0529	0.0533	0.0508	0.0519	0.0545	0.0552	0.0515	0.0493	0.0550	**0.0563**
	FREEDOM	0.0627	0.0631	0.0603	0.0617	0.0641	0.0649	0.0613	0.0589	0.0654	**0.0662**
	DRAGON	0.0662	0.0666	0.0644	0.0653	0.0675	0.0682	0.0649	0.0627	0.0688	**0.0692**
	LGMRec	0.0626	0.0633	0.0602	0.0615	0.0643	0.0652	0.0611	0.0522	0.0652	**0.0660**
Sports	MMGCN	0.0370	0.0375	0.0355	0.0362	0.0393	0.0401	0.0359	0.0341	0.0399	**0.0406**
	SLMRec	0.0663	0.0668	0.0645	0.0653	0.0690	0.0702	0.0651	0.0619	0.0699	**0.0709**
	FREEDOM	0.0717	0.0723	0.0696	0.0707	0.0741	0.0751	0.0703	0.0678	0.0751	**0.0757**
	DRAGON	0.0752	0.0757	0.0736	0.0746	0.0770	0.0778	0.0744	0.0699	0.0773	**0.0788**
	LGMRec	0.0719	0.0724	0.0700	0.0711	0.0749	0.0762	0.0708	0.0675	0.0762	**0.0771**
Clothing	MMGCN	0.0218	0.0225	0.0202	0.0211	0.0232	0.0237	0.0208	0.0200	0.0239	**0.0244**
	SLMRec	0.0452	0.0461	0.0438	0.0446	0.0481	0.0490	0.0445	0.0420	0.0485	**0.0497**
	FREEDOM	0.0629	0.0636	0.0615	0.0623	0.0649	0.0657	0.0621	0.0580	0.0655	**0.0665**
	DRAGON	0.0671	0.0681	0.0653	0.0663	0.0692	0.0701	0.0660	0.0633	0.0704	**0.0713**
	LGMRec	0.0555	0.0559	0.0538	0.0549	0.0596	0.0610	0.0545	0.0518	0.0614	**0.0622**

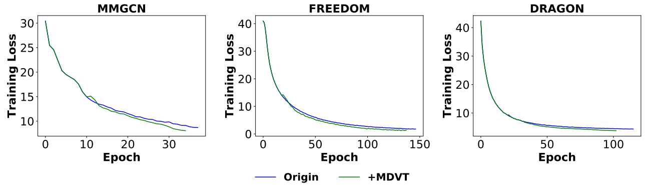


Figure 5: Convergence study on the Baby dataset.

Table 4: Performance comparison for strategies on three datasets under Recall@5 (R@5) and NDCG@5 (N@5). +M, +A, and +G denote +MDVT, +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
	+M	0.0257	0.0170	0.0236	0.0158	0.0148	0.0099
	+M+A	0.0260	0.0172	0.0239	0.0160	0.0150	0.0101
	+M+G	0.0266	0.0176	0.0244	0.0163	0.0155	0.0104
	+M+A+G	**0.0268**	**0.0177**	**0.0246**	**0.0164**	**0.0157**	**0.0105**
LGMRec	origin	0.0374	0.0249	0.0446	0.0288	0.0371	0.0246
	+M	0.0417	0.0281	0.0475	0.0312	0.0411	0.0276
	+M+A	0.0420	0.0282	0.0477	0.0313	0.0414	0.0278
	+M+G	0.0427	0.0288	0.0486	0.0318	0.0419	0.0281
	+M+A+G	**0.0431**	**0.0291**	**0.0488**	**0.0320**	**0.0421**	**0.0283**

4.6 Compatibility with Robust Training and Data Augmentation Strategies (RQ5)

Existing studies enhance the robustness of multimodal recommendations by adversarial training strategy [14, 22] and data augmentation method [10, 17, 26]. Therefore, we further evaluate the compatibility of our MDVT with the adversarial training strategy (AMR [22]) and LLM-based data augmentation strategy (GPT-4o [36]). We conducted extensive experiments based on two multimodal recommendation models across three public datasets. Table 4 shows that combining MDVT with both AMR and GPT-4o can further improve model performance. Additionally, GPT-4o outperforms AMR on all datasets, which we attribute to GPT-4o's ability to reduce the inherent gap between visual and textual information of items. Notably, simultaneously using both AMR and GPT-4o achieves more satisfactory performance than adopting either one alone.

4.7 Mechanism for Threshold Strategies (RQ6)

We further explore the practical training with these three warm-up threshold strategies. We conduct the hyper-parameter search for two advanced models (MMGCN and FREEDOM) with these three strategies on the Baby dataset. For the static warm-up threshold strategy, we follow the search range introduced in Section 4.1.3. For the dynamic warm-up threshold strategy, we set $g = 0.1$ (as shown in Section 4.8, $g = 0.1$ or 0.2 can be applied to all datasets). For the hybrid warm-up threshold strategy, we first applied the dynamic strategy with $g = 0.1$ to estimate the approximately optimal warm-up epochs $\mathcal{T}^{cur}$. Then we adopt the static strategy within the range $[\mathcal{T}^{cur} - s, \mathcal{T}^{cur} + s]$, where $s = 2$. As shown in Figure 6, the optimal warm-up epochs for all three strategies are within a similar range. Moreover, the hybrid strategy combines the advantages of both static and dynamic strategies, achieving satisfactory performance with available hyper-parameter adjustment overhead.

4.8 Hyper-parameter Analysis (RQ7)

We evaluate the impact of the key hyper-parameters (λ, n, g, and s) on MDVT's performance across three Amazon datasets in terms of Recall@10. For the hyper-parameters λ and n, we conduct analyses based on the hybrid warm-up threshold strategy, as these hyper-parameters are not related to the choice of warm-up threshold strategy, and the hybrid warm-up threshold strategy has demonstrated superior performance over the static and dynamic warm-up threshold strategies. For the hyper-parameter g, which is contained in both the dynamic and hybrid warm-up threshold strategies, we provide analyses based on these two strategies. Similarly, for the hyper-parameter s, we provide analysis based on the hybrid warm-up threshold strategy, which is the only strategy that contains s.

Hyper-parameter λ and n: From Figure 7 and Figure 8, we have the following observations. For FREEDOM, DRAGON, and LGMRec across all datasets, the optimal hyper-parameters λ and n are 0.2 and 2, respectively. In contrast, for MMGCN and SLMRec, the optimal hyper-parameters are higher, with $\lambda = 0.2$ and $n = 4$ across all datasets. We attribute this phenomenon to the lower baseline performance of MMGCN and SLMRec compared to the other models. These models are more affected by the data sparsity problem

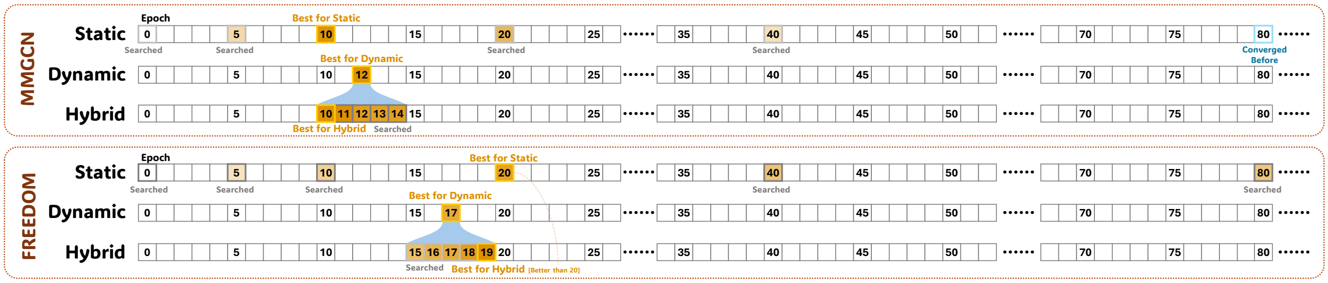


Figure 6: Mechanics of all three warm-up threshold strategies for two advanced models on the Baby dataset.

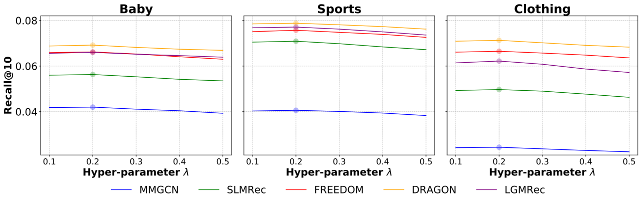


Figure 7: Performance *w.r.t.* hyper-parameter λ.

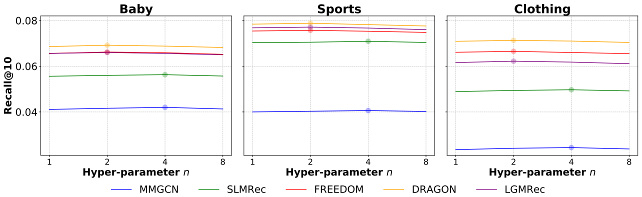


Figure 8: Performance *w.r.t.* hyper-parameter n.

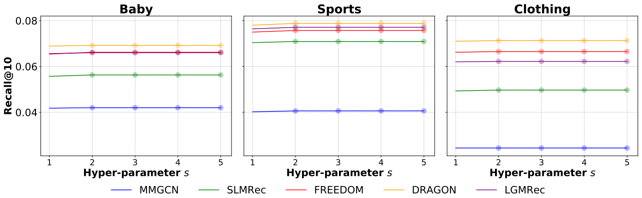


Figure 9: Performance *w.r.t.* hyper-parameter s.

Table 5: Analysis for hyper-parameter g for dynamic and hybrid strategies based on four multimodal recommendation models across all datasets in terms of Recall@10.

Models	Datasets	Baby		Sports		Clothing	
	Strategies	Dynamic	Hybrid	Dynamic	Hybrid	Dynamic	Hybrid
MMGCN	$g = 0.1$	0.0411	0.0416	0.0397	0.0404	0.0240	0.0242
	$g = 0.2$	**0.0413**	**0.0420**	**0.0400**	**0.0406**	**0.0244**	**0.0244**
	$g = 0.3$	0.0401	0.0404	0.0390	0.0393	0.0231	0.0236
	$g = 0.4$	0.0382	0.0388	0.0373	0.0379	0.0220	0.0225
	$g = 0.5$	0.0365	0.0368	0.0359	0.0362	0.0213	0.0217
FREEDOM	$g = 0.1$	**0.0648**	**0.0662**	**0.0750**	**0.0757**	0.0653	0.0663
	$g = 0.2$	0.0643	0.0655	0.0747	0.0752	**0.0655**	**0.0655**
	$g = 0.3$	0.0635	0.0643	0.0736	0.0743	0.0641	0.0647
	$g = 0.4$	0.0622	0.0628	0.0720	0.0726	0.0626	0.0630
	$g = 0.5$	0.0610	0.0618	0.0703	0.0708	0.0618	0.0620
DRAGON	$g = 0.1$	0.0680	0.0688	**0.0776**	**0.0788**	0.0701	0.0708
	$g = 0.2$	**0.0685**	**0.0692**	0.0774	0.0785	**0.0704**	**0.0713**
	$g = 0.3$	0.0674	0.0679	0.0765	0.0773	0.0692	0.0699
	$g = 0.4$	0.0663	0.0669	0.0757	0.0769	0.0678	0.0678
	$g = 0.5$	0.0648	0.0655	0.0742	0.0750	0.0653	0.0661
LGMRec	$g = 0.1$	0.0648	0.0655	0.0758	0.0767	**0.0612**	**0.0622**
	$g = 0.2$	**0.0651**	**0.0660**	**0.0762**	**0.0771**	0.0610	0.0618
	$g = 0.3$	0.0638	0.0643	0.0745	0.0753	0.0593	0.0599
	$g = 0.4$	0.0625	0.0627	0.0724	0.0728	0.0570	0.0577
	$g = 0.5$	0.0618	0.0621	0.0702	0.0708	0.0548	0.0553

and thus require a larger value of n to fully leverage virtual triplets for performance enhancement.

Hyper-parameter g and s: From Table 5 and Figure 9, we have the following observations. For the hyper-parameter g, values of 0.1 and 0.2 are recommended for all five advanced multimodal recommendation models across all datasets. For the hyper-parameter s, a value of 2 is sufficient to find the optimal number of warm-up epochs for all models. Therefore, we conclude that the hybrid warm-up threshold strategy can achieve satisfactory enhancements without incurring high hyper-parameter tuning costs.

5 Related Work and Model Details

Due to page limitations, we review recent works and their contributions in Appendix A.2. Moreover, we provide details for all utilized models in Appendix A.3. More discussions are in Appendix A.4.

6 Conclusion

In this paper, we aimed to mitigate the data sparsity problem in multimodal recommendation systems by leveraging multimodal information more effectively. We propose a novel, model-agnostic approach called MDVT, which constructs **M**ultimodal-**D**riven **V**irtual **T**riplets to provide valuable supervision signals for model training. To ensure the high quality of these virtual triplets, we introduce three different warm-up threshold strategies tailored to fit various real-world scenarios. Once the warm-up threshold is satisfied, the virtual triplets are used for joint model optimization, enhancing the learning process without causing significant gradient skew. MDVT is model-agnostic and can be easily integrated into any multimodal recommendation model. Extensive experiments on multiple real-world datasets across various advanced models demonstrated the effectiveness of MDVT. These results confirm that leveraging virtual triplets can significantly improve recommendation performance by alleviating the data sparsity problem. In future work, we aim to develop representation enhancement techniques to improve the quality of virtual triplets, thereby enhancing supervision signals and boosting overall model performance.

Acknowledgments

This work was supported by the Hong Kong UGC General Research Fund no. 17203320 and 17209822, and the project grants from the HKU-SCF FinTech Academy.

References

[1] Léon Bottou, Frank E Curtis, and Jorge Nocedal. 2018. Optimization methods for large-scale machine learning. *SIAM review* 60, 2 (2018), 223–311.

[2] Jiaben Chen, Xin Yan, Yihang Chen, Siyuan Cen, Qinwei Ma, Haoyu Zhen, Kaizhi Qian, Lie Lu, and Chuang Gan. 2024. RapVerse: Coherent Vocals and Whole-Body Motions Generations from Text. *arXiv preprint arXiv:2405.20336* (2024).

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 335–344.

[4] 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.

[5] Zheyu Chen, Jinfeng Xu, and Haibo Hu. 2025. Don't Lose Yourself: Boosting Multimodal Recommendation via Reducing Node-neighbor Discrepancy in Graph Convolutional Network. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[6] Zheyu Chen, Jinfeng Xu, Yutong Wei, and Ziyue Peng. 2025. Squeeze and Excitation: A Weighted Graph Contrastive Learning for Collaborative Filtering. *arXiv preprint arXiv:2504.04443* (2025).

[7] 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.

[8] 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.

[9] 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.

[10] 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).

[11] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. (2024).

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems* 31 (2018).

[14] Ruirui 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.

[15] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, et al. 2025. Towards Understanding Camera Motions in Any Video. *arXiv preprint arXiv:2504.15376* (2025).

[16] 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.

[17] 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).

[18] Qiyao Ma, Xubin Ren, and Chao Huang. 2024. XRec: Large Language Models for Explainable Recommendation. *arXiv preprint arXiv:2406.02377* (2024).

[19] 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.

[20] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[21] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. PMLR, 1139–1147.

[22] 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.

[23] 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).

[24] 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).

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. 2017. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems* 30 (2017).

[30] 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.

[31] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Wei Wang, Xiping Hu, and Edith C-H Ngai. 2024. FourierKAN-GCF: Fourier Kolmogorov-Arnold Network–An Effective and Efficient Feature Transformation for Graph Collaborative Filtering. *arXiv preprint arXiv:2406.01034* (2024).

[32] 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. *arXiv preprint arXiv:2504.04452* (2025).

[33] 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.

[34] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2025. A Survey on Multimodal Recommender Systems: Recent Advances and Future Directions. *arXiv preprint arXiv:2502.15711* (2025).

[35] Guipeng Xv, Chen Lin, Wanxian Guan, Jinping Gou, Xubin Li, Hongbo Deng, Jian Xu, and Bo Zheng. 2023. E-commerce search via content collaborative graph neural network. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. 2885–2897.

[36] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.

[37] 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.

[38] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems* 33 (2020), 5824–5836.

[39] 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.

[40] 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.

[41] Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*. 1–2.

[42] Xin Zhou. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497* (2023).

[43] Xin Zhou and Chunyan Miao. 2024. Disentangled Graph Variational Auto-Encoder for Multimodal Recommendation With Interpretability. *IEEE Transactions on Multimedia* (2024).

[44] 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.

[45] 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.

Algorithm 1: Procedure with Threshold Strategies

```
Input: Strategy Type $type$, Threshold Set $\mathcal{S}_{\mathcal{T}}$, Pre-defined
       Hyper-parameter $g$ and $s$;
Output: Optimal warm-up epochs $\mathcal{T}_O$;
Auxiliary Markings: $\mathcal{P}^{bar} = 0$, $\mathcal{T}^{cur}$, and $flag$ = False;
while $flag$ = True do
    Initialize model parameters $\Theta$ and $\mathcal{L}^p = 0$;
    if $type$ = Dynamic or Hybrid then
        $\mathcal{T} = 0$, $f$ = False;
        while not converged do
            $\mathcal{T} = \mathcal{T} + 1$;
            Get fused representations via Eq 1,2, and 4;
            Calculate BPR loss $\mathcal{L}_{\text{bpr}}$ via Eq 3;
            if $\mathcal{L}^p \neq 0$, $f$ = False, and $\frac{\mathcal{L}^{bpr} - \mathcal{L}^p}{\mathcal{L}^p} \leq g$ then
                Update $\mathcal{T}^{cur} = \mathcal{T}$, $f$ = True;
            end
            if $f$ = True then
                Get virtual triplet dataset $\mathcal{D}^V$ via Eq 5-6;
                Calculate virtual loss $\mathcal{L}_{\text{vbpr}}$ via Eq 7-8;
            end
            Calculate final loss $\mathcal{L}$ with Eq 9;
            Update $\mathcal{L}^{prev}$ by current loss $\mathcal{L}$;
            Calculate the gradient of loss $\nabla_\Theta \mathcal{L}$;
            Update $\Theta$ by gradient $\nabla_\Theta \mathcal{L}$ with optimizer;
        end
        Test model performance $\mathcal{P}$;
        Update $\mathcal{P}^{bar} = \mathcal{P}$, $\mathcal{S}_{\mathcal{T}} = [\mathcal{T}^{cur} - s, \mathcal{T}^{cur} + s]$;
    end
    if $type$ = Static or Hybrid then
        for $\mathcal{T}_S \in \mathcal{S}_{\mathcal{T}}$ do
            $\mathcal{T} = 0$;
            while not converged do
                $\mathcal{T} = \mathcal{T} + 1$;
                Get fused representations via Eq 1,2, and 4;
                Calculate BPR loss $\mathcal{L}_{\text{bpr}}$ via Eq 3;
                if $\mathcal{T} >= \mathcal{T}_S$ then
                    Get virtual triplet dataset $\mathcal{D}^V$ via Eq 5-6;
                    Calculate virtual loss $\mathcal{L}_{\text{vbpr}}$ via Eq 7-8;
                end
                Calculate final loss $\mathcal{L}$ with Eq 9;
                Update $\mathcal{L}^{prev}$ by current loss $\mathcal{L}$;
                Calculate the gradient of loss $\nabla_\Theta \mathcal{L}$;
                Update $\Theta$ by gradient $\nabla_\Theta \mathcal{L}$ with optimizer;
            end
            Test model performance $\mathcal{P}$;
            if $\mathcal{P} > \mathcal{P}^{bar}$ then
                Update $\mathcal{P}^{bar} = \mathcal{P}$, $\mathcal{T}^{cur} = \mathcal{T}_S$;
            end
        end
    end
    Update $\mathcal{T}_O = \mathcal{T}^{cur}$;
end
```

A Appendix

A.1 Algorithm and Validation

We present the procedure in Algorithm 1, which integrates our model-agnostic MDVT into the multimodal recommendation model with three different warm-up threshold strategies. Note that hyper-parameters g and s can be easily defined by 0.1 and 2, respectively, to achieve ideal performance for all models across all datasets, which are validated empirically in Section 4.8 and Section 4.7.

We validate the effectiveness of these three threshold strategies in Section 4.2. Furthermore, we provide an in-depth analysis of the mechanism for these three threshold strategies in Section 4.7. Our code link can be found in the footnote 3.

A.2 Related Work

Recent studies incorporate multimodal information to mitigate the data sparsity problem in recommendation systems. Pioneering this approach, VBPR [8] leverages visual content as side information in matrix factorization [20], utilizing item images to enhance recommendations. Building upon this foundation, subsequent works [4, 5, 16, 37] further integrate both visual and textual modalities to enrich item representations and improve performance. Advancements in graph-based methods introduce new avenues for multimodal recommendations. MMGCN [28] is the first to integrate Graph Convolutional Networks (GCNs) to extract modality-specific features from user-item interactions. To explicitly capture commonalities in user preferences and item relationships, models like DualGNN [24] and LATTICE [39] leverage user-user and item-item graphs, respectively. Building on LATTICE, FREEDOM [44] further stabilizes representations by freezing the item semantic graph and reducing noise in the user-item bipartite graph. Recently, self-supervised learning and inter-modal relationships have been explored to enhance recommendation systems. MMSSL [25] and MENTOR [33] employ contrastive self-supervised learning to align modalities with collaborative signals, improving performance without extensive labelled data. Additionally, BM3 [45] investigates inter-modal relationships to boost recommendation accuracy and modality fusion quality. Furthermore, LGMRec [7] leverages hypergraph to capture complex global and local relationships in multimodal information. COHESION [32] design a tailored dual-stage fusion to boost multimodal recommendation performance.

While existing works typically employ multimodal information only as side information to model user preferences, we propose leveraging the similarity between user and item modality representations, which can also provide valuable supervision signals beyond explicit user-item interactions. We construct virtual triplets based on multimodal information to capitalize on this, providing informative supervision signals to mitigate the data sparsity problem.

A.3 Models

In this section, we introduce the five advanced multimodal models used for evaluation:

- **MMGCN** [28] 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.

Table 6: Performance comparison for variants on the Baby dataset in terms of NDCG@10.

Baby	MMGCN	SLMRec	FREEDOM	DRAGON	LGMRec
MDVT	**0.0224**	**0.0323**	**0.0357**	**0.0364**	**0.0360**
MDVT *w* p *w/o* n	0.0212 (0.6)	0.0315 (0.6)	0.0354 (0.6)	0.0360 (0.6)	0.0358 (0.6)
MDVT *w* p *w/o* w,n	0.0200 (0.9)	0.0287 (0.9)	0.0337 (0.7)	0.0353 (0.7)	0.0339 (0.7)
Original	0.0200	0.0290	0.0330	0.0345	0.0333

Table 7: Performance comparison for variants on the Baby dataset in terms of NDCG@10.

Baby	MMGCN	SLMRec	FREEDOM	DRAGON	LGMRec
MDVT	0.0224	0.0323	**0.0357**	**0.0364**	**0.0360**
MDVT (*w* p)	0.0218 (0.9, 4)	0.0318 (0.9, 4)	0.0355 (0.7, 2)	**0.0364** (0.7, 2)	0.0358 (0.7, 2)
MDVT+ (*w* p)	**0.0227** (0.9, [1,4])	**0.0329** (0.9, [1,4])	**0.0357** (0.7, [2,2])	**0.0364** (0.7, [2,2])	**0.0360** (0.7, [2,2])
Original	0.0200	0.0290	0.0330	0.0345	0.0333

- **SLMRec** [23] leverages a self-supervised learning framework for multimodal recommendations by establishing a tailored node self-discrimination task, which reveals hidden multimodal patterns.
- **FREEDOM** [44] 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** [40] leverages heterogeneous and homogeneous graphs to learn high-quality user/item representations.
- **LGMRec** [7] integrates local embeddings, which capture fine-grained topological embeddings, with global embeddings considering hypergraph dependencies among items.
- **MMSSL** [25] combines modality-aware adversarial training with cross-modal contrastive learning to learn both cross-modality and modality-specific features.

A.4 More Discussion

Inspired by CC-GCN [35], we introduced a predefined threshold $\mathcal{T}$ to filter virtual triplets, hypothesizing that $\mathcal{T}$ acts as a dynamic warm-up strategy. Specifically, when the model is undertrained, user-item similarities are low, and few virtual triplets are constructed. To test this, we designed two MDVT variants: MDVT (*w* p *w/o* n) with a warm-up phase and MDVT (*w* p *w/o* w,n) without it, both using $\mathcal{T}$ instead of the top-*n* strategy. NDCG@10 results on the Baby dataset (search space $\mathcal{T} \in 0.5, 0.6, 0.7, 0.8, 0.9$) are reported, with best-performing $\mathcal{T}$ values in parentheses. According to experimental results in Table 6, we observed that the MDVT variant without a warm-up phase (MDVT (*w* p *w/o* w,n)) led to negative optimization effects on SLMRec and showed less improvement on other models compared to MDVT (*w* p *w/o* n). Additionally, this variant required a higher threshold $\mathcal{T}$ to mitigate these issues. To further investigate, we analyzed the changes in similarity between user and item representations during the optimization process. We identified the following reasons: during early training epochs, the BPR loss fails to fully establish a user-item representation space, leading to disordered similarities and repeated selection of incorrect high-similarity items as virtual triplets, disrupting representation learning. A higher $\mathcal{T}$ mitigates this issue by reducing the impact of such errors. Advanced models like FREEDOM, DRAGON, and LGMRec construct user-item representations more effectively within 5-10 epochs, achieving minor improvements even without a warm-up phase. However, the MDVT variant with the warm-up phase (*w* p *w/o* n) partially avoids these issues but still underperforms compared to MDVT. Further analysis showed that for users with sparse interactions (1-3 records), the warm-up phase left few or no items meeting the $\mathcal{T}$ threshold, worsening popularity bias. Lowering $\mathcal{T}$ addressed this for sparse users but generated excessive virtual triplets for dense users, negatively affecting performance. Furthermore, we considered combining the top-*n* strategy with the predefined threshold $\mathcal{T}$ to avoid generating excessive virtual triplets for users with dense interactions. We designed the following two variants: MDVT (*w* p) and MDVT+ (*w* p). The former employs the top-*n* strategy to limit the number of virtual triplets constructed for users with dense interactions. The latter extends this method by replacing top-*n* with an interval [n1, n2], ensuring that all users with sparse interactions can generate at least n1 virtual triplets, while users with dense interactions generate no more than n2 virtual triplets. We report the NDCG@10 results on the Baby dataset. For n1, the hyperparameter search range is $\{0, 1, 2\}$, and for n or n2, the search range is $\{1, 2, 4, 8\}$. In the table below, the numbers in parentheses indicate the optimal hyper-parameters selected for ($\mathcal{T}$, n) or ($\mathcal{T}$, [n1, n2]). Based on the experimental results, we summarize the following observations: the performance of MDVT (*w* p) is inferior to both MDVT+ (*w* p) and MDVT, which validates our earlier finding that 'for most users with sparse interactions (1-3 interaction records), after the warm-up phase, there are either no items or only a few items that satisfy the similarity threshold $\mathcal{T}$, thereby exacerbating the popularity bias.' Moreover, MDVT+ (*w* p) is equivalent to MDVT on advanced models such as FREEDOM, DRAGON, and LGMRec. Additionally, it slightly outperforms MDVT on MMGCN and SLMRec. MMGCN and SLMRec have relatively weaker modeling capabilities, introducing noise for users with sparse interactions. Combining the predefined threshold $\mathcal{T}$ and top-*n* strategy mitigates this issue.

In conclusion, our findings are as follows: 1) The predefined threshold $\mathcal{T}$ alone is insufficient for satisfactory MDVT performance due to fundamental differences from CC-GCN. While CC-GCN uses content-based similarity to construct virtual samples, MDVT dynamically builds virtual triplets based on evolving user-item representation similarities during optimization. 2) Combining the top-*n* strategy with $\mathcal{T}$ outperforms MDVT on weaker models and is equivalent for advanced models. However, it requires an additional top-*n* interval to ensure sufficient virtual triplets for users with sparse interactions and demands careful tuning of $\mathcal{T}$.