ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Research Article
https://doi.org/10.1631/ENG.ITEE.2026.0160

Decoupled prompt-guided mixture-of-experts dynamic distillation for multimodal recommendation

Lili WU[1], Renmin ZHANG[1]✉, Bin ZHANG[1], Jincheng ZHANG[3], Xi CHEN[2], Yingjing QIAN[4]

[1]*School of Communication and Electronic Engineering, Jishou University, Jishou 416000, China*
[2]*Tencent Inc., Shenzhen 518057, China*
[3]*Ant Group Co., Ltd., Changsha 410000, China*
[4]*School of Electronic and Information Engineering, Huaihua University, Huaihua 418000, China*

Abstract: Multimodal recommendation aims to enrich preference modeling by leveraging visual and textual features. However, integrating high-dimensional pretrained features introduces substantial computational overhead. While knowledge distillation provides an effective compression strategy, existing frameworks face three intertwined challenges: rank bottlenecks caused by low-dimensional projections, cross-modal interference induced by shared fusion spaces, and optimization instability under static distillation temperatures. To address these issues, we propose ProMoE-DTS, a decoupled prompt-guided mixture-of-experts framework with dynamic temperature scheduling. Using an asymmetric teacher–student architecture, the teacher model leverages modality-aware soft prompts as semantic anchors to route heterogeneous features into parameter-disjoint expert networks, thereby alleviating cross-modal conflicts and resolving the rank bottlenecks. To ensure stable knowledge transfer, a feedback-driven dynamic temperature scheduler adaptively regulates the distillation intensity based on epoch-wise signals. This asymmetric design confines intensive multimodal operations to the offline teacher, leaving the online student model with a highly efficient, pure identifier-based structure. Extensive experiments on three benchmark datasets demonstrate that ProMoE-DTS improves Recall@20 by 2.24%–3.96% over state-of-the-art baselines, while requiring only 3.28%–3.55% of the teacher's parameters.

Key words: Multimodal recommendation; Knowledge distillation (KD); Prompt tuning; Mixture-of-experts (MoE); Dynamic temperature scheduling

1 Introduction

In modern e-commerce, multimodal recommender systems (MMRSs) have emerged as a pivotal technology for personalized preference modeling. By leveraging rich visual and textual item descriptions, these systems significantly boost ranking performance (Qiu et al., 2021; Hu and Zhang, 2025). Early works in this domain primarily focused on extracting shallow modal features, such as raw image pixels or word counts (Geng et al., 2015; He RN and McAuley, 2016). Subsequent research employed attention mechanisms to capture cross-modal correlations and fuse diverse modalities (Xun et al., 2021; Liu F et al., 2023). Recently, leveraging graph neural networks (GNNs) to capture high-order user–item collaborative relations has become the mainstream paradigm, propagating embedding features over historical interaction graphs (Wei YW et al., 2019). To further extract deep item semantics and alleviate data sparsity, modern architectures use powerful pretrained foundation models (Wei W et al., 2023) and actively explore the construction of auxiliary graphs (Mo et al., 2025). Nevertheless, integrating these heavy and high-dimensional pretrained vectors incurs substantial computational and memory overhead during training. To alleviate this heavy footprint, knowledge distillation (KD) is widely adopted to compress cumbersome multimodal models by transferring dark knowledge from a large teacher model to a compact student model (Tang and Wang, 2018; Lee et al., 2019).

However, existing KD frameworks yield suboptimal

✉ Renmin ZHANG, rzhang1981@163.com
Lili WU, https://orcid.org/0009-0001-1004-9874
Renmin ZHANG, https://orcid.org/0000-0002-1019-6735
CLC number: TP391.3

Received: May 26, 2026; Revision accepted: Aug. 4, 2026;
Crosschecked: Aug. 17, 2026

performance in multimodal scenarios due to three fundamental bottlenecks. First, bridging the dimensional gap between pretrained vectors and collaborative filtering (CF) spaces via linear projections triggers a rank bottleneck (Wang J et al., 2021; Nguyen et al., 2024), which restricts semantic expressiveness and exacerbates overfitting. Second, forcing distinct modalities into a shared projection space induces cross-modal destructive interference (Liang et al., 2023; Dai et al., 2024), where heterogeneous modality dependencies degrade into disruptive noise. Finally, existing temperature strategies fail to stabilize pairwise ranking optimization. While static temperatures force a strict trade-off between early-stage gradient explosion and late-stage discriminative degradation (Hinton et al., 2015), transferring classification-based dynamic schedulers fails (Islam et al., 2025) because extreme positive–negative margin fluctuations in pairwise ranking inherently destabilize absolute-logit-based scaling.

To address these issues, we propose ProMoE-DTS, a decoupled prompt-guided mixture-of-experts (ProMoE) framework with dynamic temperature scheduling for multimodal recommendation. Specifically, deployed exclusively on the offline teacher side, the ProMoE module introduces parameter-disjoint and prompt-guided expert spaces to route heterogeneous features independently, effectively alleviating cross-modal collisions. By employing a modality-aware soft prompt-tuning mechanism as semantic anchors to guide this routing process, the teacher projects high-dimensional semantics into a collaborative space, resolving the rank bottleneck. To govern this multi-expert distillation trajectory, we design a feedback-driven dynamic temperature scheduler (DTS) that adaptively scales the temperature based on epoch-wise training errors. This unified scheduler tracks the learning feedback of the student model to secure gradient stability during an initial warm-up phase, while enhancing discriminative power over hard negative samples at later stages. Ultimately, this asymmetric architecture allows the deployed student model to entirely discard the mixture-of-experts (MoE) and prompt layers, offloading heavy multimodal operations. By executing lightweight identifier (ID)-based graph convolutions online, ProMoE-DTS achieves an outstanding equilibrium between high-fidelity recommendation accuracy and improved deployment efficiency.

The main contributions of this work are as follows:

1. We identify a practical challenge in multimodal recommendation, formalizing how to compress high-dimensional pretrained semantics into a compact student model without sustaining heavy parameter and training-stage memory footprints.

2. We propose a decoupled prompt-guided mixture-of-experts distillation framework, ProMoE-DTS, using modality-specific expert spaces and prompt-guided routing to alleviate representation rank bottlenecks and cross-modal optimization conflicts.

3. We design a feedback-driven dynamic temperature distillation strategy tailored to pairwise recommendation learning, enabling stable early-stage optimization and stronger late-stage discrimination over hard negative samples.

4. Extensive experiments on three benchmark Amazon datasets demonstrate that ProMoE-DTS achieves a superior accuracy–efficiency trade-off, delivering extreme parameter compactness and low training memory usage.

2 Related works

2.1 Multimodal recommender systems

The integration of multimodal information in recommender systems has attracted increasing attention for enriching item representations (Wei RX et al., 2025; Lin YF and He, 2026). Early approaches primarily applied attention mechanisms to fuse diverse modalities, as exemplified by attentive collaborative filtering (ACF) (Chen et al., 2017). To capture high-order collaborative connectivity, GNN models such as multimodal graph convolution network (MMGCN) (Wei YW et al., 2019) and hierarchical structure disentanglement (HSD) (Zhang JH et al., 2021) have been widely adopted to incorporate visual and textual signals. Building on these paradigms, recent advances leverage data augmentation strategies (Zhu et al., 2025) or instruction tuning via large language models (LLMs) (Bao et al., 2023) to further align user preferences.

However, a critical limitation remains: these methods rely typically on passive linear projections to map high-dimensional pretrained features into a low-dimensional CF space. This rigid compression substantially restricts their representation capacity, inevitably leading to severe rank bottlenecks (Wei W et al., 2024; Deng et al., 2025). Furthermore, conventional frameworks often overlook the inherent semantic inconsistencies across modalities. Directly fusing these heterogeneous signals introduces irrelevant noise and cross-modal interference, which ultimately disrupts the optimization trajectory of the recommendation model. Consequently, shifting from early online multimodal fusion to an asymmetric knowledge transfer paradigm—where complex multimodal semantics are isolated and distilled into a lightweight ID space—has emerged as a critical necessity for efficient recommendation.

2.2 KD in recommendation

As an established paradigm for model compression, KD has been extensively investigated to inherit essential expertise from cumbersome architectures while optimizing parameter efficiency in recommender systems. Recent advances have introduced various adaptation schemes, including heterogeneous ensembles (Kang et al., 2023), GNN-to-multilayer perceptron (MLP) compression (Tian et al., 2022), and cross-modal or LLM-based knowledge transfer (Huo et al., 2024; Du et al., 2025). Although pioneering frameworks such as collaborative graph diffusion with KD (DiffKD) (Ma et al., 2025) effectively transfer multimodal semantics via graph diffusion, they fundamentally rely on static or empirical temperature configurations. Such rigid settings fail to adaptively reconcile the dynamic trade-off between early-stage optimization stability and late-stage hard negative discrimination (Dong et al., 2023). Moreover, general classification heuristics like DTS (Islam et al., 2025) scale temperatures using absolute logit magnitudes. This mechanism is inherently mismatched for pairwise ranking objectives like the Bayesian personalized ranking (BPR) loss (Rendle et al., 2009). Within highly sparse collaborative spaces (Wang QY et al., 2019), the extreme fluctuations of relative positive–negative margins cause absolute-logit schedulers to trigger gradient vanishing or explosion. Consequently, establishing a feedback-driven

temperature scheduling paradigm explicitly adapted to stabilize pairwise gradient flows remains underexplored. To bridge this gap, our framework moves beyond traditional KD by integrating decoupled distillation and intermediate feature matching, ensuring robust transfer of high-order structural and semantic information.

2.3 Prompt tuning and MoE

Prompt tuning and MoE architectures have achieved remarkable success in parameter-efficient learning and complex relationship modeling. In the context of prompt tuning, pioneering frameworks such as personalized prompt learning for explainable recommendation (PEPLER) (Li et al., 2023) and ClickPrompt (Lin JH et al., 2024) use ID features as soft prompts to adapt LLMs to recommendation tasks. Concurrently, MoE architectures are widely adopted to handle diverse item attributes. For instance, multimodal MoE (MMoE) (Yu et al., 2024) and multi-domain multi-task MoE (M^3oE) recommendation framework (Zhang ZJ et al., 2024) incorporate multi-gate structures to mitigate optimization conflicts in multi-task scenarios, whereas facet-aware MoE (FAME) (Liu MR et al., 2025) leverages MoE components to disentangle multifaceted user interests in sequential behaviors.

Nevertheless, unlike existing prompt mechanisms in recommendation that often serve as textual queries for LLMs—exhibiting high discreteness and lacking deep cross-modal connections (Lan et al., 2025)—our framework redefines prompts as modality-aware semantic anchors. These anchors operate at the architectural bottom to explicitly guide parameter-disjoint feature routing. Furthermore, MoE frameworks in this domain are typically restricted to task-level routing rather than modality-level decoupling, which frequently leads to expert homogenization and entangled optimization trajectories. Crucially, while deploying decoupled expert spaces effectively resolves representation bottlenecks, it inherently introduces massive parameter overhead that prohibits direct online service. To bridge these gaps, combining soft prompt tuning with MoE structures to construct a robust teacher space, followed by compressing this expertise via a feedback-driven dynamic KD strategy, presents a highly promising avenue for efficient multimodal recommendation.

3 Method

In this section, we elaborate on the proposed ProMoE-DTS framework. The overall architecture is illustrated in Fig. 1. Central to our design is a teacher–student ($\mathcal{T}$–$\mathcal{S}$) paradigm, which achieves an optimal trade-off between high-fidelity recommendation accuracy and exceptional parameter efficiency during the training phase.

The framework operates through a hierarchical purification and transfer process, synergistically driven by three core components: (1) a graph-driven raw feature initialization module which preserves intrinsic high-dimensional semantics while overcoming modality cold-start; (2) a decoupled ProMoE module that constructs modality-aware soft prompts to resolve representation rank bottlenecks and alleviate cross-modal interference; (3) a DTS designed to adaptively regulate the intensity of distillation alignment from the teacher model ($\mathcal{T}$) to the compact student model ($\mathcal{S}$). For clarity, key mathematical notations are summarized in Table 1.

Table 1 Summary of key mathematical notations

Symbol	Description
$\mathcal{U}, \mathcal{I}, \mathcal{G}$	User set, item set, and interaction graph
$\boldsymbol{A}_{\text{norm}}$	Normalized adjacency matrix
$\boldsymbol{x}_{m,i}, \boldsymbol{p}_{m,i}$	Raw multimodal features and soft prompts
$\boldsymbol{W}_{m,k}, \boldsymbol{W}_{\text{trans},m}$	Expert and transformation matrices
$\boldsymbol{e}^{\text{t}*}, \boldsymbol{e}^{\text{s}*}$	Final teacher/student readout embeddings
$\tau(t), \eta(t)$	Dynamic temperature and adaptive error scale
$\mathcal{L}_{\text{BPR}}, \mathcal{L}_{\text{KD}}(t)$	BPR loss and dynamic KD loss
$\mathcal{L}_{\text{TCKD}}, \mathcal{L}_{\text{NCKD}}$	Target/Non-target class KD losses
$\mathcal{L}_{\text{feat}}, \mathcal{L}_{\text{aux}}$	Feature matching and auxiliary load-balancing losses
$\lambda_1, \lambda_2, \lambda_{\text{p}}$	Feature matching strength, prompt dropout rate, and prompt intensity
α_2, β_m	Global prompt scaling intensity and modality importance weight
μ_1, μ_2	Weighting hyperparameters for the teacher loss objectives

3.1 Problem formulation

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. In the multimodal recommendation scenario, historical interactions are typically formulated as a bipartite graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$, where an edge $e_{u,i} \in \mathcal{E}$ indicates that user u has explicitly interacted with item i and $\mathcal{E}$ denotes the set of observed interaction edges between users and items. Let $\mathcal{I}_u$ represent the set of historical items associated with user u. Beyond the base collaborative ID representations, each item $i \in \mathcal{I}$ is enriched with raw multimodal attributes, specifically, comprising visual and textual features. Given the interaction graph $\mathcal{G}$ and the multimodal feature space, the fundamental objective of our task is to learn a robust mapping function $f(\cdot)$ that estimates the preference score $\hat{y}_{u,i}$ for any unobserved user–item pair (u, i) to generate accurate top-K recommendations. Here, K is a variable representing the cutoff threshold.

3.2 Interaction graph and raw feature initialization

To capture high-order collaborative signals, we formulate the user–item interactions as an attributed bipartite graph:

$$\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E}, X). \tag{1}$$

Here, X denotes the set of multimodal features for all items, defined as $X = \{X_i\}_{i \in \mathcal{I}}$. Each item $i \in \mathcal{I}$ is inherently associated with a set of multimodal features $X_i = \{\boldsymbol{x}_{m,i}\}_{m \in \mathcal{M}}$. Specifically, the raw feature vector $\boldsymbol{x}_{m,i} \in \mathbb{R}^{d_{\text{in},m}}$ encodes visual or textual semantics extracted from pretrained foundation models. $d_{\text{in},m}$ denotes the input feature dimension of modality m. To avert the severe representation rank bottlenecks discussed previously, we strictly preserve the intrinsic high-dimensional structure of these features without applying any passive linear compression at this stage.

Given that users inherently lack explicit multimodal attributes, we leverage the topological connectivity of the

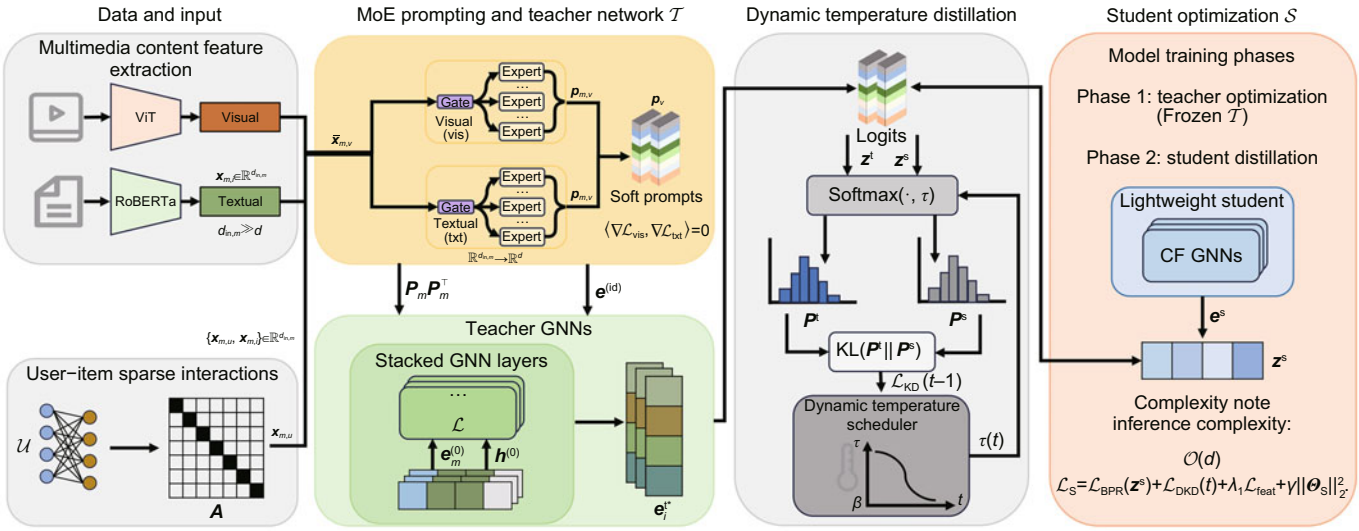

Fig. 1 Overall architecture of the proposed ProMoE-DTS framework. It comprises three interconnected stages: graph-driven raw feature encoding, decoupled ProMoE for representation alignment, and feedback-driven dynamic temperature distillation for structured knowledge transfer. RoBERTa: robustly optimized bidirectional encoder representations from Transformers pretraining approach. ViT: vision Transformer

interaction graph to aggregate raw item semantics for user initialization. The symmetric normalized adjacency matrix $\boldsymbol{A}_{\text{norm}}$, derived from the interaction set $\mathcal{E}$, is defined as

$$\boldsymbol{A}_{\text{norm}} = \boldsymbol{D}^{-0.5}\boldsymbol{A}\boldsymbol{D}^{-0.5}, \tag{2}$$

where $\boldsymbol{A}$ is the binary adjacency matrix and $\boldsymbol{D}$ denotes the corresponding diagonal degree matrix. By executing message passing over this normalized topology, we synthesize the initial multimodal representations for users, thereby mitigating the user-side modality cold-start dilemma:

$$\boldsymbol{x}_{m,u} = \sum_{i \in \mathcal{N}_u} \boldsymbol{A}_{\text{norm},u,i}\boldsymbol{x}_{m,i}, \tag{3}$$

where $\mathcal{N}_u$ denotes the set of historical neighboring items directly interacted with user u in the bipartite interaction graph $\mathcal{G}$.

Ultimately, this graph-driven initialization enables users and items to obtain comprehensive, high-dimensional semantic profiles before entering the subsequent decoupled ProMoE module.

3.3 Decoupled ProMoE

With the high-dimensional multimodal features initialized via the interaction graph, the teacher model must project these heterogeneous signals into a low-dimensional CF space. To accomplish this without succumbing to representation rank bottlenecks or cross-modal interference, we introduce the ProMoE module. This architecture dynamically expands representation capacity and strictly decouples optimization trajectories to alleviate cross-modal conflicts.

3.3.1 Decoupled modality-aware expert routing

To strictly isolate the optimization trajectories across diverse modalities, we construct independent expert libraries for each modality $m \in \mathcal{M}$, formulated as $\Theta_m = \{\boldsymbol{W}_{m,1}, \boldsymbol{W}_{m,2}, \ldots, \boldsymbol{W}_{m,K_m}\}$. Crucially, each expert projection matrix $\boldsymbol{W}_{m,k} \in \mathbb{R}^{d_{\text{in},m} \times d}$ assumes the dual responsibility of performing dimensionality reduction and extracting specialized semantic patterns, bridging the severe dimension gap without relying on rigid and passive linear compression.

Both user and item features are processed symmetrically through this routing mechanism. Let $v \in \mathcal{U} \cup \mathcal{I}$ denote a generic node, associated with its high-dimensional feature $\boldsymbol{x}_{m,v} \in \mathbb{R}^{d_{\text{in},m}}$. To stabilize the gradient flow during routing and prevent numerical explosion, we first apply layer normalization (LayerNorm):

$$\bar{\boldsymbol{x}}_{m,v} = \text{LayerNorm}(\boldsymbol{x}_{m,v}). \tag{4}$$

To ensure differentiability for end-to-end training, $\mathcal{L}_{\text{aux}}$ is computed over the continuous gating probabilities $\boldsymbol{g}_m$ prior to the top-K masking operation, thereby providing a stable gradient signal to update the gating network.

Subsequently, a modality-specific gating network $G_m(\cdot)$ evaluates the routing preferences. To ensure that the gating network receives stable gradients while maintaining strict sparsity during inference, we decouple the routing into dense probability computation and sparse activation. First, the continuous dense router probability is computed as

$$\tilde{\boldsymbol{g}}_m(\bar{\boldsymbol{x}}_{m,v}) = \text{softmax}(\boldsymbol{W}_{\text{g},m}\bar{\boldsymbol{x}}_{m,v} + \boldsymbol{b}_{\text{g},m}), \tag{5}$$

where $\boldsymbol{W}_{\text{g},m} \in \mathbb{R}^{K_m \times d_{\text{in},m}}$ and $\boldsymbol{b}_{\text{g},m} \in \mathbb{R}^{K_m}$ denote the trainable weight matrix and bias vector of the modality-specific gating network $G_m(\cdot)$ for modality m, respectively.

Next, we employ a masked activation strategy to enforce single-expert routing (i.e., hyperparameter $k_{\text{top}} = 1$) via the MaskTopK operator, which retains the largest entry and sets the remaining entries to zero followed by normalization:

$$\boldsymbol{g}_m(\bar{\boldsymbol{x}}_{m,v}) = \text{MaskTopK}(\tilde{\boldsymbol{g}}_m(\bar{\boldsymbol{x}}_{m,v}), k_{\text{top}}). \tag{6}$$

Importantly, the auxiliary load-balancing loss $\mathcal{L}_{\text{aux}}$ is computed on the continuous dense distribution $\tilde{\boldsymbol{g}}_m$ to guarantee end-to-end differentiability, while the sparse output $\boldsymbol{g}_m$ is used for expert activation.

Driven by these gating weights, the modality-specific soft prompt $\boldsymbol{p}_{m,v} \in \mathbb{R}^d$ is generated via a sparse weighted combination of the activated experts:

$$\boldsymbol{p}_{m,v} = \sum_{k=1}^{K_m} g_{m,k}(\bar{\boldsymbol{x}}_{m,v}) \boldsymbol{W}_{m,k}^{\mathrm{T}} \bar{\boldsymbol{x}}_{m,v}, \tag{7}$$

where $g_{m,k}$ denotes the sparse gating weight assigned to the k^{th} expert for modality m, and $k \in \{1, 2, \ldots, K_m\}$ serves as the expert index within the expert library of size K_m.

To regularize the gating network and avert routing collapse—where the model degenerates exclusively using a single expert—we incorporate an auxiliary load-balancing loss $\mathcal{L}_{\mathrm{aux}}$ across the expert layer. For a given mini-batch of nodes $\mathcal{B}$, it forces a uniform routing distribution:

$$\mathcal{L}_{\mathrm{aux}} = \sum_{m=1}^{|\mathcal{M}|} \sum_{k=1}^{K_m} f_{m,k} P_{m,k}, \tag{8}$$

where $f_{m,k} = \frac{1}{|\mathcal{B}|} \sum_{v \in \mathcal{B}} \mathbb{I}(k \in \mathrm{TopK}(\tilde{\boldsymbol{g}}_m))$ denotes the fraction of nodes routed to expert k, and $P_{m,k} = \frac{1}{|\mathcal{B}|} \sum_{v \in \mathcal{B}} \tilde{g}_{m,k}(\bar{\boldsymbol{x}}_{m,v})$ is its mean routing probability. $\mathbb{I}(\cdot)$ denotes the indicator function, which evaluates to 1 if the k^{th} expert is selected among the top-k_{top} activated experts according to the continuous dense gating distribution $\tilde{\boldsymbol{g}}_m$, and 0 otherwise. Finally, the decoupled modality-specific prompts are aggregated into a unified global soft prompt $\boldsymbol{p}_v$ for each node, ensuring comprehensive multimodal representation:

$$\boldsymbol{p}_v = \sum_{m \in \mathcal{M}} \boldsymbol{p}_{m,v}. \tag{9}$$

3.3.2 Prompt-enhanced subspace graph convolution

After generating the soft prompts, the critical step is to seamlessly inject these semantic priors. Observing that items are the authentic source of multimodal semantics, we deploy an asymmetric enhancement strategy. The detailed mechanism of this prompt-guided subspace projection is illustrated in Fig. 2.

Specifically, we design a subspace projection mechanism exclusively for items. Let $\boldsymbol{P}_m \in \mathbb{R}^{|\mathcal{I}| \times d}$ and $\boldsymbol{X}_m \in \mathbb{R}^{|\mathcal{I}| \times d_{\mathrm{in},m}}$ denote the global prompt and raw feature matrices for modality m. Conceptually, as illustrated in the feature smoothing process in Fig. 2, the feature increment $\Delta \boldsymbol{x}_{m,i}$ for a specific item i is intuitively computed by aggregating features from all items $j \in \mathcal{I}$, weighted by their prompt-guided similarity:

$$\Delta \boldsymbol{x}_{m,i} = \sum_{j \in \mathcal{I}} (\boldsymbol{p}_{m,i}^{\mathrm{T}} \boldsymbol{p}_{m,j}) \boldsymbol{x}_{m,j}. \tag{10}$$

However, evaluating this item-wise similarity directly corresponds to computing a dense Gram matrix, denoted as $\boldsymbol{P}_m \boldsymbol{P}_m^{\mathrm{T}}$, which incurs a prohibitive $\mathcal{O}(|\mathcal{I}|^2)$ computational complexity. To strictly avoid this computational bottleneck while preserving mathematical equivalence, we reformulate the operation into a global matrix form and compute the enhanced feature increment matrix $\Delta \boldsymbol{X}_m$ using matrix associativity:

$$\Delta \boldsymbol{X}_m = \boldsymbol{P}_m \left(\boldsymbol{P}_m^{\mathrm{T}} \boldsymbol{X}_m \right). \tag{11}$$

From the resulting global matrix $\Delta \boldsymbol{X}_m$, we extract the corresponding instance-level increment $\Delta \boldsymbol{x}_{m,i}$ to project the enhanced high-dimensional feature into the collaborative space $\mathbb{R}^d$. We use a unified linear transformation $\boldsymbol{W}_{\mathrm{trans},m} \in \mathbb{R}^{d \times d_{\mathrm{in},m}}$ to map the combined feature increment and raw input into the target collaborative dimension:

$$\tilde{\boldsymbol{e}}_{m,i}^{(0)} = \mathrm{Dropout}\left(\boldsymbol{W}_{\mathrm{trans},m} \left(\boldsymbol{x}_{m,i} + \lambda_{\mathrm{p}} \cdot \mathrm{Norm}(\Delta \boldsymbol{x}_{m,i}) \right) \right), \tag{12}$$

where $\tilde{\boldsymbol{e}}_{m,i}^{(0)}$ is the initial projected multimodal feature representation of item i for modality m after prompt-enhanced subspace projection. Dropout$(\cdot)$ denotes the standard dropout regularization operator applied during training to randomly zero out elements with a given probability, preventing over-reliance on specific projection features. λ_{p} controls the prompt intensity, and Norm$(\cdot)$ denotes the ℓ_2 normalization applied to stabilize the magnitude of the injected semantic increments. For users who acquire initial high-dimensional features via graph aggregation as formulated in Eq. (3), we directly project them without prompt-guided enhancement to preserve the asymmetric architecture:

$$\tilde{\boldsymbol{e}}_{m,u}^{(0)} = \mathrm{Dropout}(\boldsymbol{W}_{\mathrm{trans},m} \boldsymbol{x}_{m,u}), \tag{13}$$

where $\tilde{\boldsymbol{e}}_{m,u}^{(0)}$ is the initial projected multimodal feature representation of user u for modality m via direct linear transformation.

Concurrently, let $\boldsymbol{e}_u^{(\mathrm{id})}, \boldsymbol{e}_i^{(\mathrm{id})} \in \mathbb{R}^d$ denote the base collaborative ID embeddings. We inject them with their respective global prompts to initialize the ID states:

$$\begin{cases} \boldsymbol{h}_u^{(0)} = \boldsymbol{e}_u^{(\mathrm{id})} + \alpha_2 \cdot \mathrm{Norm}(\boldsymbol{p}_u), \\ \boldsymbol{h}_i^{(0)} = \boldsymbol{e}_i^{(\mathrm{id})} + \alpha_2 \cdot \mathrm{Norm}(\boldsymbol{p}_i), \end{cases} \tag{14}$$

where $\boldsymbol{h}_u^{(0)}$ and $\boldsymbol{h}_i^{(0)}$ are the initial prompt-enhanced collaborative ID state representations of user u and item i, respectively. $\boldsymbol{p}_u$ and $\boldsymbol{p}_i$ are the unified global soft prompt vectors for user u and item i, respectively. α_2 is a tunable hyperparameter governing the intensity of the injected global prompts. Subsequently, the enhanced ID representations and multimodal features are propagated over the bipartite graph for L layers using $\boldsymbol{A}_{\mathrm{norm}}$:

$$\begin{cases} \boldsymbol{h}_u^{(l)} = \sum_{i \in \mathcal{N}_u} \boldsymbol{A}_{\mathrm{norm},u,i} \boldsymbol{h}_i^{(l-1)}, \\ \tilde{\boldsymbol{e}}_{m,u}^{(l)} = \sum_{i \in \mathcal{N}_u} \boldsymbol{A}_{\mathrm{norm},u,i} \tilde{\boldsymbol{e}}_{m,i}^{(l-1)}, \end{cases} \tag{15}$$

with $\boldsymbol{h}_u^{(l)}$ being the collaborative ID embedding of user u at the l^{th} graph convolution network (GCN) propagation layer and $\tilde{\boldsymbol{e}}_{m,u}^{(l)}$ being the multimodal semantic representation of user u for modality m at the l^{th} GCN propagation layer, and similarly for the item representations $\boldsymbol{h}_i^{(l-1)}$ and $\tilde{\boldsymbol{e}}_{m,i}^{(l-1)}$ by aggregating from connected users. After L layers of propagation, we adopt a late-fusion strategy to obtain the final high-fidelity teacher representations by reading out the ID embeddings and concatenating the multimodal semantics with $\boldsymbol{e}_i^{\mathrm{t}*}$ denoting the final comprehensive output readout embedding of item i generated by the teacher model:

$$\boldsymbol{e}_i^{\mathrm{t}*} = \frac{1}{L+1} \sum_{l=0}^{L} \boldsymbol{h}_i^{(l)} + \sum_{m=1}^{|\mathcal{M}|} \beta_m \cdot \mathrm{Norm}(\tilde{\boldsymbol{e}}_{m,i}^{(L)}), \tag{16}$$

and with the user representation $\boldsymbol{e}_u^{\mathrm{t}*}$ obtained symmetrically, where β_m denotes the modality-specific importance weight.

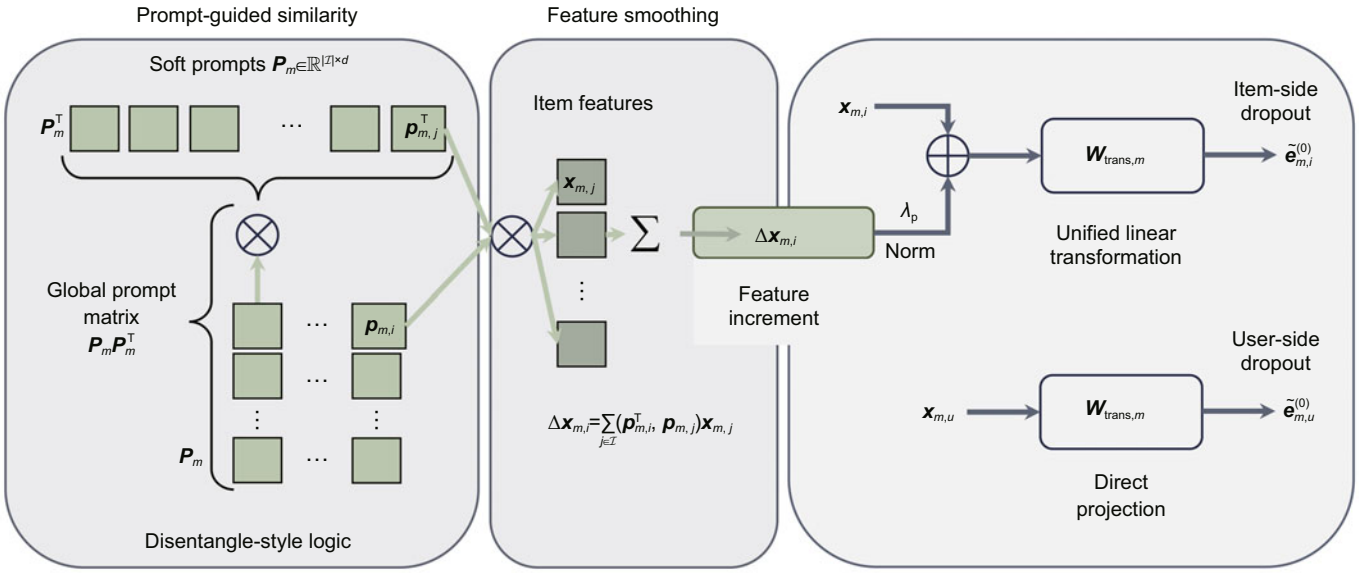

Fig. 2 Detailed architecture of the prompt-enhanced subspace projection in ProMoE. Modality-specific soft prompts first construct prompt-guided item similarities, which are then used for feature smoothing and asymmetric teacher-side projection. $\otimes$ denotes the scalar–vector multiplication weighting each item feature vector $x_{m,j}$ by the prompt-guided similarity scalar ($p_{m,i}^{\mathrm{T}} p_{m,j}$), and $\sum$ denotes the aggregation summation over candidate items $j \in \mathcal{I}$

The final predicted logit for a user–item pair (u, i) in the teacher model is computed via the inner product:

$$z_{u,i}^{\mathrm{t}} = (\boldsymbol{e}_u^{\mathrm{t}*})^{\mathrm{T}} \boldsymbol{e}_i^{\mathrm{t}*}. \tag{17}$$

3.3.3 Theoretical analysis of ProMoE

The theoretical superiority of the ProMoE module stems from two fundamental mathematical properties.

1. Union-of-subspaces formulation. Rather than strictly increasing the rank of a single projection matrix for each instance, ProMoE dynamically expands the representational coverage by assigning different samples to different expert subspaces. Thus, the model behaves as a piecewise linear union-of-subspaces approximator, alleviating the practical information loss and severe low-rank truncation errors caused by a single shared low-rank projection.

2. Parameter-disjoint modality decoupling. Sharing parameters across divergent modalities, such as visual and textual signals, often triggers destructive interference where $\langle \nabla \mathcal{L}_{\mathrm{vis}}, \nabla \mathcal{L}_{\mathrm{txt}} \rangle < 0$. By designing modality-specific independent expert libraries Θ_m, we structurally isolate the optimization trajectories within the routing layers. In the parameter space $\Theta_{\mathrm{MoE}} = \bigcup_{m \in \mathcal{M}} \Theta_m$, the gradients from heterogeneous modalities are applied to disjoint parameters, which substantially reduces destructive interference compared with shared projection layers. Notably, this parameter-level decoupling does not imply full independence, as the modalities are aligned in the downstream collaborative space. Instead, it establishes a structural mechanism to alleviate cross-modal gradient conflicts prior to semantic fusion.

3.4 Dynamic temperature distillation

In the KD process of transferring refined knowledge from the teacher model $\mathcal{T}$ to the lightweight student model $\mathcal{S}$, the temperature scalar τ plays a critical role in smoothing the output probability distribution. However, standard KD typically relies on a fixed static temperature, which introduces an inherent optimization dilemma in complex multimodal recommendation scenarios. Specifically, setting a fixed small τ causes the softmax function to approximate a step characteristic, which drastically increases the Hessian matrix norm early in training, violating Lipschitz continuity and easily triggering gradient explosion. Conversely, setting a fixed large τ to maintain early stability will excessively suppress the gradient magnitude ($\|\nabla\| \propto \tau^{-2}\epsilon \to 0$) at the late training stage as the alignment error shrinks. Here, ∇ denotes the gradient operator with respect to model parameters, and ϵ is a small positive scalar representing the alignment error residue. This prevents the student model from effectively capturing fine-grained top-K ranking differences, leading to premature stagnation.

3.4.1 Dynamic temperature scheduling mechanism

To adapt temperature scaling for pairwise ranking optimization, the dynamic temperature scheduling mechanism regulates distillation intensity over the training epoch t. Unlike classification tasks relying on absolute logits, pairwise ranking optimizes the relative margin between positive and negative items. Let $\Delta z = z_{u,i} - z_{u,j}$ denote the score margin, where $z_{u,i}$ and $z_{u,j}$ denote the predicted preference scores for the observed positive item i and unobserved negative item j, respectively. The gradient of the temperature-scaled BPR loss $\mathcal{L}_{\mathrm{BPR}}$ with respect to this margin is mathematically derived as

$$\frac{\partial \mathcal{L}_{\mathrm{BPR}}}{\partial \Delta z} = -\frac{1}{\tau(t)} \left(1 - \sigma\left(\frac{\Delta z}{\tau(t)}\right)\right), \tag{18}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the standard sigmoid activation function.

In highly sparse implicit feedback graphs, the severe positive–negative imbalance drives Δz to extensive ranges. A naive time-decay or absolute-logit scaling fails under this

gradient mechanism. As τ approaches zero, extreme margins trigger infinite multipliers leading to gradient explosion, whereas an excessively large τ causes the gradient to aggressively vanish.

Therefore, to buffer these gradient fluctuations, we design a sequential feedback loop governed by an adaptive error scaling factor $\eta(t)$. This factor stabilizes the decay amplitude using the moving average of the preceding epoch's decoupled distillation loss $\mathcal{L}_{\mathrm{DKD}}(t-1)$:

$$\eta(t) = \frac{\mathcal{L}_{\mathrm{DKD}}(t-1)}{\mathcal{L}_{\mathrm{DKD}}(t-1) + \epsilon_0}, \tag{19}$$

where ϵ_0 is a small constant that maintains numerical stability. For the initial epoch $t = 1$, $\mathcal{L}_{\mathrm{DKD}}(0)$ is initialized to 1.0. This feedback signal constructs a target temperature trajectory $\tau_{\mathrm{tgt}}(t)$ by integrating with a cosine annealing schedule:

$$\tau_{\mathrm{tgt}}(t) = \tau_{\min} + \frac{\tau_{\max} - \tau_{\min}}{2} \cdot \left[1 + \cos\left(\frac{\pi t}{T_{\max}}\right)\right] \cdot \eta(t), \tag{20}$$

where $\tau_{\min}$ and $\tau_{\max}$ represent the lower and upper bounds of the dynamic distillation temperature, and $T_{\max}$ denotes the maximum number of training epochs. To mitigate temperature oscillations caused by sparse batch interactions, a momentum term ρ updates the current temperature:

$$\tau(t) = \rho\tau(t-1) + (1-\rho)\tau_{\mathrm{tgt}}(t). \tag{21}$$

This dynamic temperature $\tau(t)$ softens the output logit vector $\boldsymbol{z}$ for the teacher and student models. For the i^{th} candidate item, the probability distribution $p_i(\cdot,\cdot)$ is computed as

$$p_i(\boldsymbol{z}, \tau(t)) = \frac{\exp(z_i/\tau(t))}{\sum_{j=1}^{C} \exp(z_j/\tau(t))}, \tag{22}$$

where C specifies the size of the sampled candidate set containing the observed positive item and random unobserved negatives. The dynamic KD loss $\mathcal{L}_{\mathrm{KD}}$ is formulated via the Kullback–Leibler (KL) divergence between the teacher and student distributions:

$$\mathcal{L}_{\mathrm{KD}}(t) = \sum_{i=1}^{C} p_i(\boldsymbol{z}^{\mathrm{t}}, \tau(t)) \ln\left(\frac{p_i(\boldsymbol{z}^{\mathrm{t}}, \tau(t))}{p_i(\boldsymbol{z}^{\mathrm{s}}, \tau(t))}\right), \tag{23}$$

where $\boldsymbol{z}^{\mathrm{t}}$ and $\boldsymbol{z}^{\mathrm{s}}$ represent the teacher's predicted preference logit vector and student's predicted preference logit vector, respectively

3.4.2 Theoretical analysis of stable gradient dynamics

The theoretical advantage of the DTS mechanism lies in its ability to maintain stable gradient dynamics throughout the optimization cycle by reshaping the loss landscape in real time. During the initial training stage, where $t \to 0$, a large $\tau(t)$ mathematically suppresses the Lipschitz constant, formulated as $L_{\mathrm{lip}} \propto 1/\tau^2$, transforming the originally steep nonconvex KL divergence into a smoother approximation. This ensures that the model can safely traverse regions of high local curvature during random initialization. As training progresses to the late stage, where $t \to T_{\max}$, the student model's output gradually approaches that of the teacher model, driving the output residual $\|\boldsymbol{z}^{\mathrm{s}} - \boldsymbol{z}^{\mathrm{t}}\|$ toward zero. At this juncture, the dynamic decay of $\tau(t)$ amplifies the local curvature, counteracting the gradient vanishing effect caused by alignment error reduction. This mechanism enables the student model to learn discriminative features near the minima, achieving a favorable balance between optimization stability and transfer precision.

3.4.3 Decoupled KD and feature matching

Equipped with the dynamically scheduled temperature $\tau(t)$, we reformulate the standard KD by decoupling it into target class KD (TCKD) and non-target class KD (NCKD). This structural decoupling prevents the gradient noise of negative samples from interfering with the alignment of positive interactions.

Specifically, let $p_{\mathrm{T}} = \sum_{i\in\mathcal{P}} p_i(\boldsymbol{z}, \tau(t))$ and $p_{\mathrm{N}} = \sum_{j\in\mathcal{N}} p_j(\boldsymbol{z}, \tau(t))$ denote the cumulative probabilities of the positive target subset $\mathcal{P}$ and negative non-target subset $\mathcal{N}$, respectively. The objective components are formally defined as

$$\mathcal{L}_{\mathrm{TCKD}} = \mathrm{KL}\left([p_{\mathrm{T}}^{\mathrm{t}}, p_{\mathrm{N}}^{\mathrm{t}}] \parallel [p_{\mathrm{T}}^{\mathrm{s}}, p_{\mathrm{N}}^{\mathrm{s}}]\right), \tag{24}$$

$$\mathcal{L}_{\mathrm{NCKD}} = \sum_{j\in\mathcal{N}} \hat{p}_j^{\mathrm{t}} \ln\left(\frac{\hat{p}_j^{\mathrm{t}}}{\hat{p}_j^{\mathrm{s}}}\right), \tag{25}$$

where $\mathcal{L}_{\mathrm{TCKD}}$ focuses on transferring the teacher's relative confidence regarding the interacted positive items, $\mathcal{L}_{\mathrm{NCKD}}$ aligns the latent rankings among the massive unobserved items, and $\hat{p}_j = p_j/p_{\mathrm{N}}$ represents the normalized internal probability distribution among the non-target items. Finally, we aggregate these decoupled components to formulate the overall dynamic KD loss:

$$\mathcal{L}_{\mathrm{DKD}}(t) = \alpha_{\mathrm{KD}}\mathcal{L}_{\mathrm{TCKD}} + \beta_{\mathrm{KD}}\mathcal{L}_{\mathrm{NCKD}}, \tag{26}$$

where α_{KD} and β_{KD} are the weighting parameters balancing the target-class and non-target-class distillation objectives.

Furthermore, to ensure that the compact student model structurally inherits the multimodal geometry, we enforce an aggregated feature-level alignment. By minimizing the mean squared error (MSE) between the student's pure ID embedding and the teacher's multimodal representations, we implicitly constrain the student to capture the semantic centroid of the heterogeneous modalities:

$$\mathcal{L}_{\mathrm{feat}} = \sum_{i\in\mathcal{I}} \sum_{m\in\mathcal{M}} \mathrm{MSE}(\boldsymbol{e}_i^{\mathrm{s}*}, \tilde{\boldsymbol{e}}_{m,i}^{(L)}), \tag{27}$$

where $\boldsymbol{e}_i^{\mathrm{s}*}$ denotes the pure ID representation learned by the student model via light graph convolutional network (LightGCN) propagation, and $\tilde{\boldsymbol{e}}_{m,i}^{(L)}$ denotes the teacher's representations with gradients detached to ensure that only the student model is updated during distillation.

3.5 Model optimization and two-phase training

To construct a unified learning paradigm, we implement a strict two-phase training strategy that effectively isolates the computational burden. The complete optimization procedure is systematically summarized in Algorithm 1.

Phase 1: teacher optimization. The high-capacity teacher model $\mathcal{T}$ is trained to generate precise prompts and high-fidelity multimodal collaborative signals. Its overall loss is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{T}} = {} & \mathcal{L}_{\mathrm{BPR}}(\boldsymbol{z}^{\mathrm{t}}) + \mu_1\mathcal{L}_{\mathrm{BPR}}(\boldsymbol{p}) \\ & + \mu_2 \sum_{m\in\mathcal{M}} \mathcal{L}_{\mathrm{BPR}}(\tilde{\boldsymbol{e}}_m) \\ & + \omega\mathcal{L}_{\mathrm{aux}} + \gamma\|\boldsymbol{\Theta}_{\mathrm{T}}\|_2^2,\end{aligned} \tag{28}$$

Algorithm 1 Two-phase training paradigm of ProMoE-DTS

Require: interaction graph $\mathcal{G}$, multimodal feature set X, maximum number of training epochs $T_{\max}$, hyperparameters ρ, $\tau_{\min}$, and $\tau_{\max}$
Ensure: optimal ID embeddings $\boldsymbol{\Theta}_{\mathrm{S}}$ for the student model
/* Phase 1: teacher optimization */
repeat
  sample a mini-batch of user–item interactions from $\mathcal{G}$
  compute modality-specific soft prompt $\boldsymbol{p}_{m,v}$ and global soft prompt $\boldsymbol{p}_v$ via ProMoE routing (Eqs. (7) and (9))
  execute prompt-enhanced subspace GCN to obtain teacher representations $\boldsymbol{e}_u^{\mathrm{t}\,*}$ and $\boldsymbol{e}_i^{\mathrm{t}*}$ (Eqs. (11)–(16))
  calculate the teacher objective $\mathcal{L}_{\mathrm{T}}$ (Eq. (28)) and update parameters $\boldsymbol{\Theta}_{\mathrm{T}}$ via gradient descent
until convergence
strictly freeze all teacher parameters $\boldsymbol{\Theta}_{\mathrm{T}}$
/* Phase 2: student distillation */
initialize the dynamic temperature $\tau(0)$ and previous distillation loss $\mathcal{L}_{\mathrm{DKD}}(0)$
for epoch $t = 1, 2, \ldots, T_{\max}$ **do**
  calculate adaptive error scale $\eta(t)$ (Eq. (19)), target trajectory (Eq. (20)), and update dynamic temperature $\tau(t)$ (Eq. (21))
  for each mini-batch **do**
    read out pure ID embeddings $\boldsymbol{e}_u^{\mathrm{s}\,*}$ and $\boldsymbol{e}_i^{\mathrm{s}*}$ via compact graph propagation
    compute the decoupled dynamic distillation loss $\mathcal{L}_{\mathrm{DKD}}(t)$ with softened probabilities (Eqs. (24)–(26))
    compute the feature matching loss $\mathcal{L}_{\mathrm{feat}}$ against offline multimodal semantics (Eq. (27))
    calculate the overall student objective $\mathcal{L}_{\mathrm{S}}$ (Eq. (29)) and update parameters $\boldsymbol{\Theta}_{\mathrm{S}}$
  end for
end for
return $\boldsymbol{\Theta}_{\mathrm{S}}$

where $\mathcal{L}_{\mathrm{BPR}}(\cdot)$ denotes the standard pairwise ranking loss computed strictly over the inner-product scores derived from the corresponding representation sets (i.e., the predicted logits $\boldsymbol{z}^{\mathrm{t}}$, soft prompt embeddings $\boldsymbol{p}$, and projected multimodal features $\tilde{\boldsymbol{e}}_m$), $\mathcal{L}_{\mathrm{aux}}$ acts as the ProMoE load-balancing penalty, and γ governs the ℓ_2 regularization. $\boldsymbol{\Theta}_{\mathrm{T}}$ denotes the complete trainable parameter vector of the teacher model $\mathcal{T}$. ω is the balancing hyperparameter governing the auxiliary load-balancing loss $\mathcal{L}_{\mathrm{aux}}$.

Phase 2: student distillation. Once the teacher model converges, its parameters are strictly frozen. The compact student model $\mathcal{S}$ is subsequently trained by combining the fundamental collaborative ranking task with decoupled dynamic distillation:

$$\mathcal{L}_{\mathrm{S}} = \mathcal{L}_{\mathrm{BPR}}(\boldsymbol{z}^{\mathrm{s}}) + \mathcal{L}_{\mathrm{DKD}}(t) + \lambda_1 \mathcal{L}_{\mathrm{feat}} + \gamma \|\boldsymbol{\Theta}_{\mathrm{S}}\|_2^2. \tag{29}$$

During this phase, gradients are back-propagated exclusively to update the student's ID embeddings $\boldsymbol{\Theta}_{\mathrm{S}}$.

Online inference efficiency is paramount for industrial deployability. Current graph-based multimodal recommenders suffer from severe latency and memory bottlenecks during online service due to real-time modality feature reduction, which incurs a complexity of $\mathcal{O}\big(\sum_{m\in\mathcal{M}} |\mathcal{I}| \times d_{\mathrm{in},m} \times d\big)$, and dynamic GNN operations, bounded by $\mathcal{O}(L \times |\mathcal{E}| \times d)$. To overcome this, ProMoE-DTS adopts an asymmetric architecture that strictly confines these intensive operations—including graph convolutions and expert routing, characterized by a complexity of $\mathcal{O}\big(\sum_{m\in\mathcal{M}} |\mathcal{I}| \times K_m \times d_{\mathrm{in},m} \times d\big)$—to the offline teacher model. During online inference, the distilled student model $\mathcal{S}$ completely discards the heavyweight multimodal layers and deep graph structures. By caching the learned representations, the student solely performs highly efficient inner products using pure ID embeddings, drastically reducing the online scoring time complexity to $\mathcal{O}(d)$. This theoretically eliminates the heavy computational dependencies of multimodal GNNs, achieving deployment-friendly low-latency scoring while preserving high-fidelity accuracy.

4 Experiments and analysis

4.1 Experimental settings

4.1.1 Datasets

Following existing protocols (Zhou et al., 2023; Liu YF et al., 2024), we evaluate on three Amazon datasets (McAuley et al., 2015): clothing, sports, and baby. After applying a 5-core filter and excluding items with incomplete modalities, we extract 4096-dimensional visual features and 1024-dimensional textual embeddings via Sentence-BERT (BERT: bidirectional encoder representations from Transformers) (Reimers and Gurevych, 2019). To strictly prevent data leakage, interactions are chronologically partitioned into training, validation, and test sets at an 8:1:1 ratio. Dataset statistics are in Table 2.

4.1.2 Evaluation metrics

Following previous works (Wei YW et al., 2020, 2022), we evaluate the top-K recommendation performance using Recall@K and normalized discounted cumulative gain at K (NDCG@K) for $K \in \{20, 50\}$ (i.e., R@20, R@50, N@20, and N@50). The evaluation adopts the all-ranking strategy by ranking each ground-truth positive item against all unobserved items rather than a sampled subset. To ensure reliability

Table 2 Statistics of the experimental datasets

Dataset	Dimension	Number of users	Number of items	Number of interactions	Sparsity (%)
Clothing	4096 1024	39 387	23 033	278 677	99.969
Sports	4096 1024	35 598	18 357	296 337	99.955
Baby	4096 1024	19 445	7050	160 792	99.883

4096 denotes the 4096-dimensional visual features and 1024 denotes the 1024-dimensional textual embeddings

against initialization variances, all main experiments are independently executed across three specific random seeds, namely, 42, 2024, and 2026. We report the average metrics alongside the corresponding standard deviations. Furthermore, we conduct paired t-tests over user-level metric differences to compute the p-value (p stands for probability), thereby verifying the statistical significance of the improvements.

4.1.3 Implementation details

The ProMoE-DTS framework is implemented via PyTorch, with execution conducted on an NVIDIA GeForce RTX 3090 graphics processing unit (GPU). The collaborative embedding dimensionality is fixed at $d = 64$, whereas raw visual and textual features maintain their original dimensional boundaries and undergo transformation steps within the teacher-side ProMoE module. The layer configurations for the teacher and student GNNs are set to 3 and 1, respectively. The ProMoE module operates with four expert networks alongside a top-1 routing allocation strategy. To modulate the representation compression bounds, the prompt dropout rate is assigned a value of 0.6. During the decoupled KD stage, the weighting parameters α_{KD} and β_{KD} are designated as 0.1 and 0.5, respectively, with the feature matching coefficient λ_1 set to 0.05. For the dynamic temperature scheduler, the boundary parameters are configured to $\tau_{\min} = 0.5$ and $\tau_{\max} = 5.0$. Optimization is performed using the AdamW optimizer (Loshchilov and Hutter, 2019) under a batch size parameter of 1024. The learning rate is evaluated within the numerical range $[10^{-4}, 10^{-2}]$ on the validation split, and the $\mathcal{L}_2$ regularization coefficient γ is sampled from $\{10^{-4}, 10^{-3}, 10^{-2}\}$. Baseline models are deployed using their respective verified source code configurations under equivalent environmental constraints.

4.1.4 Baselines

We compare ProMoE-DTS against nine state-of-the-art baselines, broadly categorized into traditional CF models and MMRS.

CF models:

1. BPR-MF (Rendle et al., 2009): a classic matrix factorization method optimized by BPR loss.

2. Neural graph collaborative filtering (NGCF) (Wang X et al., 2019): a graph-based CF framework that integrates GNNs into a CF framework to explicitly model high-order user–item connectivities.

3. LightGCN (He XN et al., 2020): a simplified GCN-based model that streamlines graph convolutions by removing feature transformations and nonlinear activations.

Multimodal recommender systems:

1. Visual Bayesian personalized ranking (VBPR) (He RN and McAuley, 2016): an extension of traditional matrix factorization that incorporates visual signals.

2. Multimodal graph convolution network (MMGCN) (Wei YW et al., 2019): a multimodal graph neural architecture that constructs modality-specific graphs to guide preference representation learning.

3. Graph-refined convolutional network (GRCN) (Wei YW et al., 2020): a graph-refinement model that adaptively prunes false-positive interaction edges using multimodal features.

4. PromptMM (Wei W et al., 2024): a multimodal KD framework that incorporates prompt-tuning mechanisms to compress cross-modal multi-view representations.

5. DiffKD (Ma et al., 2025): a graph diffusion-based architecture designed to transfer high-order structural and semantic knowledge into compact student networks.

6. HierarchyMD (Jian et al., 2026): a hierarchy-aware multimodal distillation paradigm optimized for fine-grained structural feature matching within collaborative spaces.

4.2 Performance comparisons

The overall recommendation performance is reported in Table 3. From the results, we have the following key observations:

1. ProMoE-DTS yields the highest empirical metrics compared to the selected baselines across the evaluated datasets. Specifically, compared to HierarchyMD (Jian et al., 2026), it registers measured relative increments ranging from 2.24% to 3.96% for R@20 and from 3.12% to 3.95% for N@20. Baselines

Table 3 Overall performance comparison on three datasets

Baseline	Sports				Clothing				Baby			
	R@20	N@20	R@50	N@50	R@20	N@20	R@50	N@50	R@20	N@20	R@50	N@50
BPR-MF	0.0701	0.0315	0.1163	0.0407	0.0329	0.0146	0.0545	0.0189	0.0579	0.0242	0.1086	0.0343
NGCF	0.0648	0.0279	0.1092	0.0367	0.0354	0.0149	0.0612	0.0200	0.0526	0.0219	0.0989	0.0311
LightGCN	0.0814	0.0365	0.1320	0.0466	0.0463	0.0200	0.0751	0.0259	0.0714	0.0298	0.1276	0.0411
VBPR	0.0714	0.0315	0.1172	0.0407	0.0329	0.0147	0.0538	0.0189	0.0596	0.0250	0.1110	0.0353
MMGCN	0.0531	0.0237	0.0953	0.0325	0.0338	0.0147	0.0602	0.0200	0.0603	0.0271	0.1084	0.0377
GRCN	0.0822	0.0375	0.1376	0.0484	0.0623	0.0270	0.1038	0.0353	0.0746	0.0317	0.1310	0.0423
PromptMM	0.0854	0.0392	0.1405	0.0512	0.0645	0.0281	0.1051	0.0362	0.0762	0.0329	0.1331	0.0438
DiffKD	0.0881	0.0410	0.1432	0.0528	0.0662	0.0293	0.1068	0.0374	0.0781	0.0341	0.1350	0.0452
HierarchyMD†	0.0906	0.0423	0.1461	0.0539	0.0681	0.0304	0.1082	0.0385	0.0802	0.0353	0.1369	0.0469
Ours†	**0.0931**	**0.0438**	**0.1488**	**0.0553**	**0.0708**	**0.0316**	**0.1095**	**0.0394**	**0.0820**	**0.0364**	**0.1388**	**0.0484**
p-value	1.60×10^{-6}	5.90×10^{-5}	2.99×10^{-7}	1.11×10^{-6}	1.41×10^{-4}	5.59×10^{-4}	5.00×10^{-6}	1.29×10^{-5}	3.24×10^{-5}	2.96×10^{-6}	7.51×10^{-7}	4.63×10^{-6}

† Standard deviations across three random seeds are omitted from the main grid to maintain layout clarity. The maximum variance is rigorously bounded within ±0.0012 for the proposed method and ±0.0015 for HierarchyMD across all reported metrics. The best results are in bold. The p-values are computed between the proposed method and the strongest baseline

such as PromptMM (Wei W et al., 2024) apply textual prompt tuning on discrete views, and DiffKD (Ma et al., 2025) implements graph diffusion to regulate knowledge transfer, but their frameworks maintain static distillation temperatures during training iterations. The integration of parameter-disjoint expert routing and error-feedback temperature scheduling in ProMoE-DTS scales the loss optimization trajectory, allowing the student model to replicate structural and semantic patterns with less representation truncation.

2. Multimodal recommendation architectures generally report higher accuracy scores than CF baselines, such as BPR-MF (Rendle et al., 2009) and NGCF (Wang X et al., 2019), which derive representations solely from user–item historical interaction graphs. These differences in metrics indicate that using explicit user–item connection topologies alone is prone to performance limitations under sparse interaction densities. Incorporating supplementary pretrained visual and textual features via collaborative feature mapping introduces additional item properties that counteract graph-level data sparsity.

3. Among the multimodal baselines, a relative performance stratification is observed between early unified linear projection methods, such as VBPR (He RN and McAuley, 2016) and MMGCN (Wei YW et al., 2019), and subsequent structure-refinement or distillation models, including GRCN (Wei YW et al., 2020), PromptMM (Wei W et al., 2024), DiffKD (Ma et al., 2025), and HierarchyMD (Jian et al., 2026). This trend indicates that compressing high-dimensional semantic matrices into collaborative subspaces via rigid linear layers restricts representation capacity. Conversely, decoupling heterogeneous modality spaces and incorporating continuous gradient adjustments provide more structured cross-modal alignment.

4.3 Ablation study

We conduct an ablation study to evaluate the effectiveness of various components in our proposed framework. We design the following variants:

1. Without (w/o)-Prompt: removes the prompt-guided similarity enhancement and ID prompt injection, while keeping the modality-specific expert projections.

2. w/o-MoE: replaces the expert routing mechanism with a single shared modality-specific projection matrix, while preserving the remaining distillation pipeline.

3. w/o-DTS: replaces the dynamic temperature scheduler with a fixed static temperature $\tau = 1.0$.

Table 4 reports the ablation results. From the table, we observe that removing either the DTS or the prompt module leads to the most severe performance degradation depending on the dataset. Specifically, the absence of DTS causes the largest drop in the sports and baby datasets, indicating that dynamically reshaping the loss landscape is crucial for balancing early-stage convergence and late-stage discriminative transfer. Meanwhile, excluding the prompt module substantially decreases performance (especially on the clothing dataset), highlighting its necessity in bridging the semantic gap and mitigating representation bottlenecks. Furthermore, disabling the MoE module decreases recommendation accuracy across all scenarios, confirming that isolating optimization trajectories via decoupled experts effectively alleviates cross-modal collisions. Overall, each component contributes synergistically to the superior performance of ProMoE-DTS.

Table 4 Model performance of the ablation study

Dataset	Metric	R@20	N@20
Sports	w/o-Prompt	0.0801	0.0375
	w/o-DTS	0.0763	0.0362
	w/o-MoE	0.0871	0.0405
	ProMoE-DTS	**0.0931**	**0.0438**
Clothing	w/o-Prompt	0.0549	0.0248
	w/o-DTS	0.0569	0.0255
	w/o-MoE	0.0643	0.0277
	ProMoE-DTS	**0.0708**	**0.0316**
Baby	w/o-Prompt	0.0709	0.0309
	w/o-DTS	0.0695	0.0301
	w/o-MoE	0.0728	0.0314
	ProMoE-DTS	**0.0820**	**0.0364**

The best results are in bold

4.4 Parameter sensitivity analysis

In this subsection, we investigate the impact of key architectural and optimization hyperparameters on the performance of ProMoE-DTS. Specifically, we evaluate the depth of GNN layers (L_{T} and L_{S}), the representation dimensionality (d_{T} and d_{S}), the feature matching strength (λ_1), the number of MoE experts (n), and the prompt dropout rate (λ_2).

4.4.1 Impact of the depth of GNN layers

Fig. 3a illustrates the recommendation performance with varying GNN layer depths for the teacher (L_{T}) and student (L_{S}). Optimal performance is achieved at $L_{\mathrm{T}} = 3$ and $L_{\mathrm{S}} = 1$. The teacher's performance improves as L_{T} increases to 3, which is essential for encoding high-order collaborative signals, but degrades at $L_{\mathrm{T}} = 4$ due to the over-smoothing issue. Crucially, the student network consistently peaks at a shallow depth of $L_{\mathrm{S}} = 1$. This validates our asymmetric distillation strategy: the lightweight student directly inherits highly refined high-order knowledge, eliminating the need for deep graph convolutions during deployment.

4.4.2 Impact of the representation dimensionality

Fig. 3b presents the sensitivity to the hidden representation dimensionality. As the teacher's capacity (d_{T}) increases to 128, the overall performance improves. More importantly, we observe a distinct dimension decoupling phenomenon. While the symmetric configuration of $d = 64$ yields highly competitive results, the student achieves a marginally higher peak performance at a significantly lower dimensionality of $d_{\mathrm{S}} = 32$ when supervised by a 128-dimensional teacher model. This demonstrates that a low-dimensional student model can effectively absorb condensed multimodal knowledge while naturally resisting overfitting, maintaining robust performance with reduced computational overhead.

4.4.3 Impact of the feature matching strength

Regarding the sensitivity of the feature matching intensity, Figs. 4a and 4b show that the performance on the sports

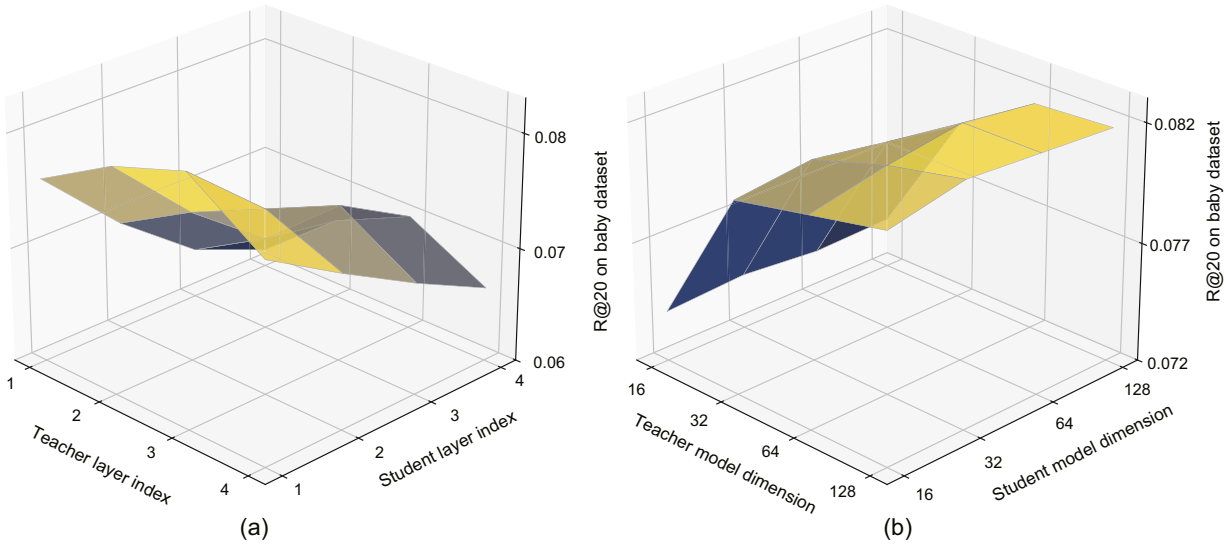

Fig. 3 Architectural parameter sensitivity analysis of ProMoE-DTS: (a) the optimal layer depth combinations between the teacher and student models; (b) the dimension decoupling phenomenon where low-dimensional student model maintains competitive performance

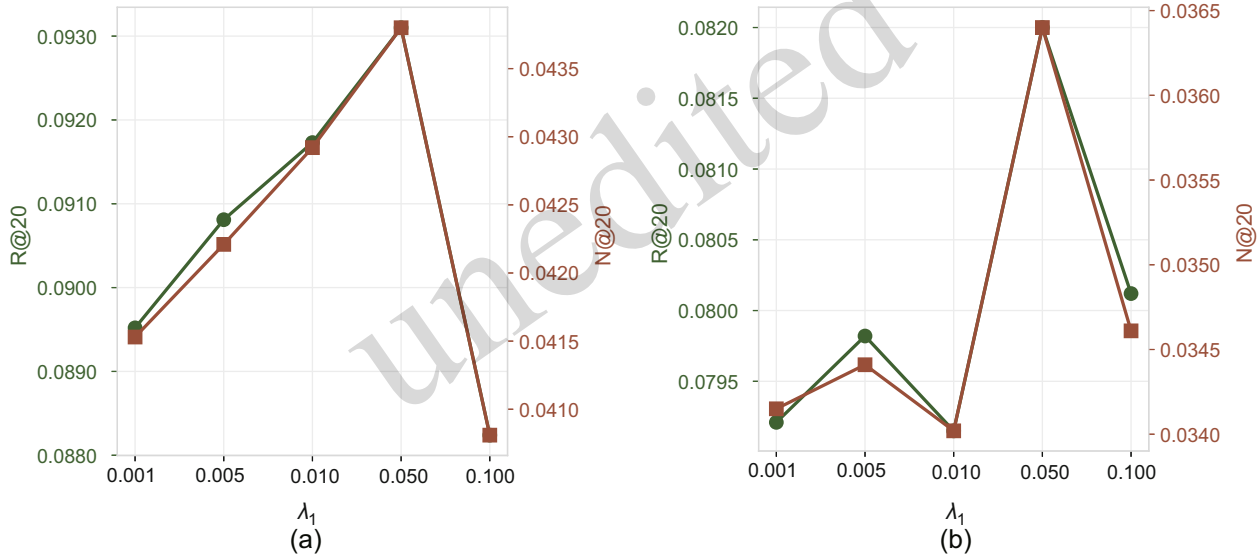

Fig. 4 Performance with different feature matching strengths λ_1 on the sports (a) and baby (b) datasets

and baby datasets reaches its peak at $\lambda_1 = 0.050$, beyond which the performance drops as λ_1 further increases to 0.100. This trajectory stems from the mathematical interaction between graph aggregation and the pairwise BPR optimization loop. Because dense pretrained multimodal vectors possess substantially higher absolute magnitudes than sparse collaborative embeddings, an excessive λ_1 over-amplifies semantic signals during graph propagation, thereby overshadowing collaborative structural preferences. Consequently, the inner-product predictions become biased toward global semantic similarities, which compresses the positive–negative score margin $(\hat{y}_{ui} - \hat{y}_{uj})$ and reduces gradient informativeness. Therefore, $\lambda_1 = 0.05$ functions as an optimal threshold that balances high-dimensional feature matching with collaborative representation learning.

4.4.4 Impact of the number of experts

Regarding expert configurations, Fig. 5a shows that the performance on the sports dataset reaches its highest value at $n = 8$ but suffers a performance drop when further scaling to $n = 16$, indicating parameter redundancy on data with lower semantic complexity. Fig. 5b illustrates metrics on the baby dataset scaling continuously up to $n = 16$, where higher feature diversity benefits from expanded capacity within multi-expert subspaces.

However, expanding the expert pool amplifies computational overhead. Table 5 profiles the performance, training cost, and routing distribution on the sports dataset. The configuration with $n = 1$ represents a shared network baseline without routing mechanisms. Scaling from $n = 4$ to $n = 8$

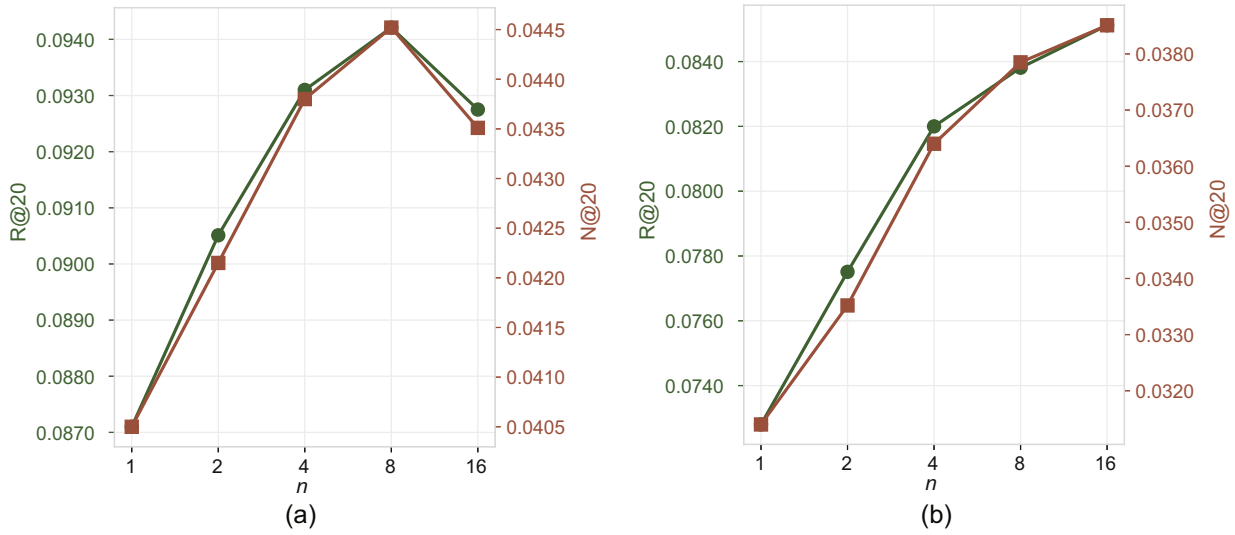

Fig. 5 Performance with varying numbers of experts for the decoupled MoE on the sports (a) and baby (b) datasets

Table 5 Performance, computational cost, and routing entropy with varying numbers of experts on the sports dataset

n	R@20	Time (s)	Memory (GB)	Entropy
1	0.0871	8.5	1.62	–
4	0.0931	10.2	1.99	0.96
8	0.0942	18.5	3.85	0.92
16	0.0927	35.1	7.50	0.88

"–" indicates that the routing entropy is not applicable for the single-network baseline ($n = 1$) because no routing or gating mechanism is involved

roughly doubles the memory footprint and per-epoch training time while yielding merely marginal accuracy gains. Furthermore, to address potential routing collapse, we evaluate the expert utilization balance using normalized routing entropy. An entropy value approaching 1.0 indicates a perfectly uniform routing distribution across experts. Tracing the utilization rates for $n = 4$ reveals a well-balanced allocation of 38.1%, 15.2%, 26.4%, and 20.3% across the experts, achieving an entropy of 0.96. Notably, as the expert pool expands to $n = 16$, the routing entropy degrades to 0.88 alongside a decline in recommendation accuracy. This indicates an increased risk of allocation imbalance and parameter redundancy. While $n = 8$ marginally improves recommendation accuracy, a standardized selection of $n = 4$ secures the Pareto optimal trade-off encompassing deployment efficiency, modeling capacity, and routing robustness.

4.4.5 Impact of the prompt dropout rate

Regarding the parameter sensitivity of the prompt dropout probability, Fig. 6a illustrates the evaluation metrics under varying dropout rates λ_2 on the sports dataset, where the measured performance tracks a stable distribution across different allocations and peaks at $\lambda_2 = 0.6$. Symmetrically, Fig. 6b displays the corresponding evaluation profile on the baby dataset, demonstrating an equivalent metric maximum at the $\lambda_2 = 0.6$ threshold. This shared empirical trajectory across distinct collaborative environments indicates that a moderate dropout allocation operates as a regularization

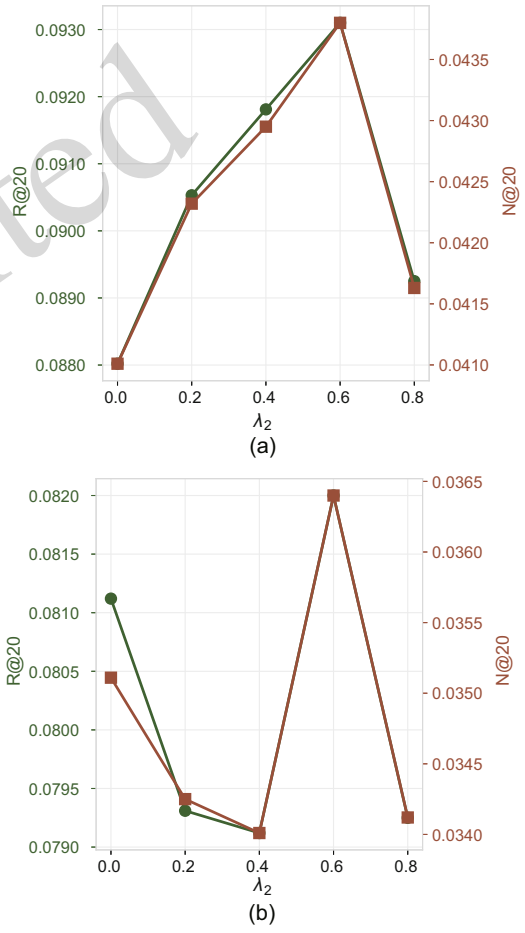

Fig. 6 Performance with different prompt dropout rates λ_2 on the sports (a) and baby (b) datasets

component, restricting the teacher architecture from establishing over-dependence on specific multimodal prompt matrices. Consequently, this constraint bounds the risk of representation overfitting under severe interaction sparsity and stabilizes the generalization parameters of the distilled knowledge.

4.4.6 Impact of temperature boundaries

The dynamic temperature scheduler relies on the boundary parameters $\tau_{\min}$ and $\tau_{\max}$ to constrain the adaptation range. Table 6 profiles the model performance under varying boundary combinations on the sports dataset to validate the default configuration. Setting $\tau_{\max}$ to an excessively high value of 10.0 over-smooths the early-stage logits, delaying convergence and yielding suboptimal representation learning. Conversely, configuring a restricted upper bound of 2.0 fails to provide sufficient gradient stabilization during the initial data-sparse phase. Furthermore, setting $\tau_{\min}$ bounded away from zero is critical. Approaching zero with a $\tau_{\min}$ of 0.1 sharply amplifies local curvature, triggering late-stage gradient explosion as the output residual approaches zero. Consequently, the combination of 0.5 and 5.0 establishes the most robust optimization trajectory.

Table 6 Performance with varying temperature boundaries on the sports dataset

$\tau_{\min}$	$\tau_{\max}$	R@20	N@20
0.1	5.0	0.0895	0.0418
0.5	2.0	0.0912	0.0425
0.5	10.0	0.0905	0.0420
0.5	5.0	**0.0931**	**0.0438**

The best results are in bold

4.5 Efficiency analysis

Table 7 quantifies the computational metrics of ProMoE-DTS alongside the teacher architecture and HierarchyMD. The measurements indicate that HierarchyMD introduces higher runtime resource demands, producing an out-of-memory (OOM) fault on the 24 GB hardware configuration during the execution of its dense topology layers, whereas ProMoE-DTS operates with a lower memory allocation of 2.38 GB. Furthermore, ProMoE-DTS registers a restricted parameter size, consuming between 3.28% and 3.55% of the teacher model's total parameter allocation (representing a scale of 3.470×10^6 to 4.010×10^6). These evaluated scores indicate that the prompt-based distillation protocol compresses multimodal representations with less computational resource consumption than graph-dependent baselines that rely on explicit adjacent matrix scaling.

4.6 In-depth component analysis

To evaluate temperature scaling in pairwise ranking, we compare the proposed mechanism against a static temperature baseline ($\tau = 1.0$), cosine annealing omitting loss feedback, and absolute loss-magnitude scaling (Islam et al., 2025) on the sports dataset. As Table 8 indicates, the direct application of absolute loss scaling degrades performance under the highly sparse pairwise setting due to gradient instability. The proposed feedback-driven mechanism overcomes this limitation by maintaining optimization stability, thereby consistently outperforming the baselines.

Table 8 Performance comparison of temperature scheduling strategies

Dataset	Strategy	R@20	N@20
Sports	Static	0.0763	0.0362
	Cosine annealing	0.0892	0.0416
	Loss-magnitude	0.0875	0.0408
	ProMoE-DTS	**0.0931**	**0.0438**

The best results are in bold

Furthermore, to investigate whether MSE causes semantic averaging compared to fine-grained contrastive learning, we substitute the alignment objective with an InfoNCE cross-modal loss on the baby dataset. Table 9 demonstrates that the contrastive variant decreases accuracy and amplifies computational latency. This performance degradation occurs because randomly sampled non-target items frequently act as false negatives. Pushing the student identifier embedding away from these latent positive items damages the collaborative structure. Conversely, the proposed MSE strategy functions as a parameter-free regularizer pulling the student toward the semantic centroid, efficiently averting false-negative distortion.

Table 9 Performance comparison of alignment strategies on the baby dataset

Alignment	R@20	N@20	Relative time
InfoNCE	0.0805	0.0351	1.45×
MSE	**0.0820**	**0.0364**	**1.00**×

Better results are in bold

Table 7 Efficiency comparison of different models in terms of training time (per epoch), peak GPU memory usage, and parameter size

Dataset	Model	Time (s)	Memory (GB)	Number of parameters	Ratio (%)
Sports	Teacher	18.30	2.95	9.777×10^7	–
	HierarchyMD	61.00	18.24	2.406×10^7	24.61
	ProMoE-DTS	10.20	1.99	3.470×10^6	3.55
Clothing	Teacher	24.88	3.42	1.223×10^8	–
	HierarchyMD†	45.10	37.69	9.839×10^7	80.45
	ProMoE-DTS	13.90	2.38	4.010×10^6	3.28

† HierarchyMD encounters an OOM error on the default RTX 3090 GPU (24 GB); its results are obtained using an NVIDIA A100 (80 GB). The ratio column denotes the parameter count of the corresponding model relative to the teacher model. "–" in the ratio column indicates that the teacher model serves as the benchmark (100%), against which student parameter compression ratios are calculated

5 Conclusions

This paper addresses the critical bottlenecks in multimodal recommendation, specifically, the representation rank bottlenecks, destructive cross-modal interference, and the optimization instability inherent in static distillation strategies. To overcome these challenges, we propose ProMoE-DTS, a novel asymmetric KD framework. Specifically, the ProMoE module leverages modality-specific expert subspaces for robust union-of-subspaces approximation and independent expert libraries for the decoupled optimization of heterogeneous features, thereby alleviating cross-modal conflicts. Concurrently, the DTS adaptively regulates the distillation intensity via a sequential error-feedback loop to maintain stable gradient dynamics. Extensive empirical results demonstrate that ProMoE-DTS significantly outperforms state-of-the-art baselines. By condensing heavy multimodal graph operations into a highly compact ID-based architecture—using only 3.28%–3.55% of the teacher's parameters—our framework achieves a favorable balance between high-fidelity recommendation accuracy and deployment efficiency in resource-constrained scenarios. Future research will explore incorporating user-side multimodal features and mining higher-order latent structural relationships to generalize this asymmetric distillation paradigm.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62061017), the Natural Science Foundation of Hunan Province (No. 2025JJ70624), the Education Commission of Hunan Province (Nos. 24A0380 and 23B0729), and the Graduate Innovation Project of Jishou University (No. Jdy25009).

Author contributions

Lili WU designed the research and drafted the paper. Bin ZHANG processed the data. Jincheng ZHANG and Xi CHEN helped review the paper. Renmin ZHANG proposed the core technical methodology, guided the experimental design, and supervised the project. Yingjing QIAN co-supervised the project.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in the Amazon review dataset repository at https://nijianmo.github.io/amazon/index.html.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used DeepSeek to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Bao KQ, Zhang JZ, Zhang Y, et al., 2023. TALLRec: an effective and efficient tuning framework to align large language model with recommendation. Proc 17th ACM Conf on Recommender Systems, p.1007-1014. https://doi.org/10.1145/3604915.3608857

Chen JY, Zhang HW, He XN, et al., 2017. Attentive collaborative filtering: multimedia recommendation with item- and component-level attention. Proc 40th Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.335-344. https://doi.org/10.1145/3077136.3080797

Dai BQ, Du ZC, Zhu JM, et al., 2024. UniEmbedding: learning universal multi-modal multi-domain item embeddings via user-view contrastive learning. Proc 33rd ACM Int Conf on Information and Knowledge Management, p.4446-4453. https://doi.org/10.1145/3627673.3680098

Deng JX, Wang SY, Cai K, et al., 2025. OneRec: unifying retrieve and rank with generative recommender and iterative preference alignment. https://arxiv.org/abs/2502.18965

Dong XM, Huang O, Thulasiraman P, et al., 2023. Improved knowledge distillation via teacher assistants for sentiment analysis. Proc IEEE Symp Series on Computational Intelligence, p.300-305. https://doi.org/10.1109/SSCI52147.2023.10371965

Du YP, Sun Z, Wang ZY, et al., 2025. Active large language model-based knowledge distillation for session-based recommendation. Proc 39th AAAI Conf on Artificial Intelligence, p.11607-11615. https://doi.org/10.1609/aaai.v39i11.33263

Geng X, Zhang HW, Bian JW, et al., 2015. Learning image and user features for recommendation in social networks. Proc IEEE Int Conf on Computer Vision, p.4274-4282. https://doi.org/10.1109/ICCV.2015.486

He RN, McAuley J, 2016. VBPR: visual Bayesian personalized ranking from implicit feedback. Proc 30th AAAI Conf on Artificial Intelligence, p.144-150. https://doi.org/10.1609/aaai.v30i1.9973

He XN, Deng K, Wang X, et al., 2020. LightGCN: simplifying and powering graph convolution network for recommendation. Proc 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.639-648. https://doi.org/10.1145/3397271.3401063

Hinton G, Vinyals O, Dean J, 2015. Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531

Hu XH, Zhang HT, 2025. Invariant representation learning in multimedia recommendation with modality alignment and model fusion. *Entropy*, 27(1):56. https://doi.org/10.3390/e27010056

Huo FS, Xu WC, Guo JC, et al., 2024. C^2KD: bridging the modality gap for cross-modal knowledge distillation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16006-16015. https://doi.org/10.1109/CVPR52733.2024.01515

Islam SU, Ahad JI, Rahman F, et al., 2025. Dynamic temperature scheduler for knowledge distillation. https://arxiv.org/abs/2511.13767

Jian M, Wang T, Yang MJ, et al., 2026. Hierarchy-aware multimodal distillation for recommendation. *IEEE Trans Multim*, 28:2279-2290. https://doi.org/10.1109/TMM.2026.3651049

Kang S, Kweon W, Lee D, et al., 2023. Distillation from heterogeneous models for top-K recommendation. Proc ACM Web Conf, p.801-811. https://doi.org/10.1145/3543507.3583209

Lan PX, Xu HY, Yang EN, et al., 2025. Efficient and effective prompt tuning via prompt decomposition and compressed outer product. Proc Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, p.4406-4421. https://doi.org/10.18653/v1/2025.naacl-long.225

Lee JW, Choi M, Lee J, et al., 2019. Collaborative distillation for top-N recommendation. Proc IEEE Int Conf on Data Mining, p.369-378. https://doi.org/10.1109/ICDM.2019.00047

Li L, Zhang YF, Chen L, 2023. Personalized prompt learning for explainable recommendation. *ACM Trans Inform Syst*, 41(4):103. https://doi.org/10.1145/3580488

Liang JH, Zhao XY, Li MY, et al., 2023. MMMLP: multi-modal multilayer perceptron for sequential recommendations. Proc ACM Web Conf, p.1109-1117. https://doi.org/10.1145/3543507.3583378

Lin JH, Chen B, Wang HY, et al., 2024. ClickPrompt: CTR models are strong prompt generators for adapting language models to CTR prediction. Proc ACM Web Conf, p.3319-3330. https://doi.org/10.1145/3589334.3645396

Lin YF, He JY, 2026. A reinforcement learning framework for multi-modal recommendation explanation generation. *Authorea*. https://doi.org/10.22541/au.177368713.30806120/v1

Liu F, Chen HL, Cheng ZY, et al., 2023. Disentangled multimodal representation learning for recommendation. *IEEE Trans Multim*, 25:7149-7159. https://doi.org/10.1109/TMM.2022.3217449

Liu MR, Zhang SX, Long C, 2025. Facet-aware multi-head mixture-of-experts model for sequential recommendation. Proc 18th ACM Int Conf on Web Search and Data Mining, p.127-135. https://doi.org/10.1145/3701551.3703552

Liu YF, Zhang KN, Ren XY, et al., 2024. AlignRec: aligning and training in multimodal recommendations. Proc 33rd ACM Int Conf on Information and Knowledge Management, p.1503-1512. https://doi.org/10.1145/3627673.3679626

Loshchilov I, Hutter F, 2019. Decoupled weight decay regularization. https://doi.org/10.48550/arXiv.1711.05101

Ma WY, Xia HB, Liu Y, 2025. DiffKD: collaborative graph diffusion with knowledge distillation for multimodal recommendation. *J Intell Inform Syst*, 63(5):1487-1510. https://doi.org/10.1007/s10844-025-00946-4

McAuley J, Targett C, Shi QF, et al., 2015. Image-based recommendations on styles and substitutes. Proc 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.43-52. https://doi.org/10.1145/2766462.2767755

Mo F, Xiao L, Song QY, et al., 2025. FGCM: modality-behavior fusion model integrated with graph contrastive learning for multimodal recommendation. *IEEE Multim*, 32(3):27-37. https://doi.org/10.1109/MMUL.2025.3542757

Nguyen NH, Nguyen TA, Nguyen T, et al., 2024. Towards efficient communication and secure federated recommendation system via low-rank training. Proc ACM Web Conf, p.3940-3951. https://doi.org/10.1145/3589334.3645702

Qiu RH, Wang S, Chen Z, et al., 2021. CausalRec: causal inference for visual debiasing in visually-aware recommendation. Proc 29th ACM Int Conf on Multimedia, p.3844-3852. https://doi.org/10.1145/3474085.3475266

Reimers N, Gurevych I, 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proc Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing, p.3982-3992. https://doi.org/10.18653/v1/D19-1410

Rendle S, Freudenthaler C, Gantner Z, et al., 2009. BPR: Bayesian personalized ranking from implicit feedback. Proc 25th Conf on Uncertainty in Artificial Intelligence, p.452-461.

Tang JX, Wang K, 2018. Ranking distillation: learning compact ranking models with high performance for recommender system. Proc 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, p.2289-2298. https://doi.org/10.1145/3219819.3220021

Tian YJ, Zhang CX, Guo ZC, et al., 2022. NOSMOG: learning noise-robust and structure-aware MLPs on graphs. https://doi.org/10.48550/arXiv.2208.10010

Wang J, Zhu L, Dai T, et al., 2021. Low-rank and sparse matrix factorization with prior relations for recommender systems. *Appl Intell*, 51(6):3435-3449. https://doi.org/10.1007/s10489-020-02023-5

Wang QY, Yin HZ, Wang H, et al., 2019. Enhancing collaborative filtering with generative augmentation. Proc 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, p.548-556. https://doi.org/10.1145/3292500.3330873

Wang X, He XN, Wang M, et al., 2019. Neural graph collaborative filtering. Proc 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.165-174. https://doi.org/10.1145/3331184.3331267

Wei RX, Lan JM, Li KK, et al., 2025. MFDB: multimodal feature fusion and dynamic behavior modeling for interactive recommendation systems. *Knowl-Based Syst*, 326:114047. https://doi.org/10.1016/j.knosys.2025.114047

Wei W, Huang C, Xia LH, et al., 2023. Multi-modal self-supervised learning for recommendation. Proc ACM Web Conf, p.790-800. https://doi.org/10.1145/3543507.3583206

Wei W, Tang JB, Xia LH, et al., 2024. PromptMM: multi-modal knowledge distillation for recommendation with prompt-tuning. Proc ACM Web Conf, p.3217-3228. https://doi.org/10.1145/3589334.3645359

Wei YW, Wang X, Nie LQ, et al., 2019. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. Proc 27th ACM Int Conf on Multimedia, p.1437-1445. https://doi.org/10.1145/3343031.3351034

Wei YW, Wang X, Nie LQ, et al., 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. Proc 28th ACM Int Conf on Multimedia, p.3541-3549. https://doi.org/10.1145/3394171.3413556

Wei YW, Wang X, He XN, et al., 2022. Hierarchical user intent graph network for multimedia recommendation. *IEEE Trans Multim*, 24:2701-2712. https://doi.org/10.1109/TMM.2021.3088307

Xun JH, Zhang SY, Zhao Z, et al., 2021. Why do we click: visual impression-aware news recommendation. Proc 29th ACM Int Conf on Multimedia, p.3881-3890. https://doi.org/10.1145/3474085.3475514

Yu HF, Qi ZY, Jang LK, et al., 2024. MMoE: enhancing multimodal models with mixtures of multimodal interaction experts. Proc Conf on Empirical Methods in Natural Language Processing, p.10006-10030. https://doi.org/10.18653/v1/2024.emnlp-main.558

Zhang JH, Zhu YQ, Liu Q, et al., 2021. Mining latent structures for multimedia recommendation. Proc 29th ACM Int Conf on Multimedia, p.3872-3880. https://doi.org/10.1145/3474085.3475259

Zhang ZJ, Liu SC, Yu JA, et al., 2024. M^{3}oE: multi-domain multi-task mixture-of-experts recommendation framework. Proc 47th Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.893-902. https://doi.org/10.1145/3626772.3657686

Zhou X, Zhou HY, Liu Y, et al., 2023. Bootstrap latent representations for multi-modal recommendation. Proc ACM Web Conf, p.845-854. https://doi.org/10.1145/3543507.3583251

Zhu NJ, Ren YQ, Liu Y, et al., 2025. MMGCL: multi-scale and multi-channel graph contrastive learning for flight anomaly detection. *Knowl-Based Syst*, 329:114275. https://doi.org/10.1016/j.knosys.2025.114275