MM-DAGNN: Modality-Disentangled Adaptive Graph Neural Networks for Multimodal E-commerce Recommendation

Abstract

We proposed MM-DAGNN, a Modality-Disentangled Adaptive Graph Neural Network designed to address key challenges in multimodal e-commerce recommendation systems. Specifically, MM-DAGNN tackles the limitations of traditional models that assume uniform modality propagation, static fusion, and entangled projections. The proposed model integrates (i) modality-specific receptive fields K_v, K_t, K_{id}, (ii) a per-item adaptive fusion gate, and (iii) a contrastive disentanglement regularizer to improve the handling of heterogeneous modalities, against seven competitive baselines on the public AliExpress dataset. We treat country-aggregated implicit feedback as pseudo-users and apply 5-core filtering, yielding a $177 \times 5{,}073$ bipartite graph with $105{,}419$ edges. Three findings stand out. **(F1)** On Recall@20 and on every secondary metric, MM-DAGNN is competitive. MMGCN (0.0238) and LATTICE (0.0236) achieve the best Recall@20, while MM-DAGNN is the strongest method on Recall@10, NDCG, and MAP@20, a pattern consistent with recent reproducibility findings in collaborative filtering. **(F2)** A $4{\times}4$ grid over (K_v, K_t) shows a *near-isotropic* accuracy surface: the Recall@20 maximum sits at the diagonal $(K_v{=}1, K_t{=}1)$ and the NDCG@20 maximum at $(K_v{=}2, K_t{=}2)$, contradicting the recent claim that visual and textual modalities require systematically different propagation depths on this dataset. **(F3)** The adaptive gate produces a high Modality Balance Index (MBI $=$ 0.70) without harming accuracy, indicating that the gate genuinely uses all three streams rather than collapsing onto one. By solving the problem of modality imbalance and improving the adaptation across different streams, we contribute to a more flexible and accurate approach to multimodal recommendation.

1 Introduction

Recommendation has shifted from collaborative matrix factorization [Rendle et al., 2012, He et al., 2020] to systems that combine user–item interactions with rich item content—images, text, and audio attached to each product [Liu et al., 2024]. The dominant fusion backbone is the graph neural network (GNN): methods such as MMGCN [Wei et al., 2019], GRCN [Wei et al., 2020], LATTICE [Zhang et al., 2021], BM3 [Zhou et al., 2023b], FREEDOM [Zhou and Shen, 2023], MGCN [Yu et al., 2023], and LGMRec [Guo et al., 2024] project visual and textual features into a shared embedding space and propagate them through some combination of a user–item bipartite graph and learned item–item graphs.

Despite a steady stream of accuracy improvements reported in this literature, three structural assumptions underpin essentially every GNN-based multimodal recommender. **(A1)** Each modality is propagated through the same number of GNN layers K. **(A2)** Modality streams are summed or mean-pooled with fixed weights [Wei et al., 2019, Zhou and Shen, 2023], despite the fact that the relative importance of vision vs. text varies per item. **(A3)** A single linear projection per modality maps raw features into the joint embedding space [Zhou et al., 2023b, Guo et al., 2024], fusing

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

modality-shared and modality-specific semantics into the same dimensions. Recent work [Zhou et al., 2025] has argued that A1 in particular is empirically shaky on Amazon Baby, suggesting that the optimal K is modality-dependent.

In this paper we revisit these three assumptions on a different and larger benchmark. We design a single lightweight architecture (MM-DAGNN) that loosens A1–A3 simultaneously and evaluate it end-to-end on the public AliExpress dataset. The single robust benefit of MM-DAGNN's adaptive gate is not accuracy but *modality balance*: the model uses all three streams instead of collapsing onto one, while baseline static-fusion methods cannot adapt by construction. Our hyperparameter study further finds that the (K_v, K_t) accuracy surface is essentially isotropic on this dataset, with the best diagonal point matching the global maximum within seed noise.

Research questions:

RQ1. On AliExpress dataset, does MM-DAGNN improve over GNN and non-GNN baselines on which metrics?

RQ2. Is the optimal GNN propagation depth modality-dependent on this dataset, or is the standard uniform-K assumption empirically harmless?

RQ3. Which of MM-DAGNN's three components are load-bearing for accuracy, and which (if any) are load-bearing for modality balance?

Contributions.

1. We proposed the *Modality Balance Index* (MBI), showing that MM-DAGNN's adaptive gate produces MBI ≈ 0.70 while static-fusion baselines are pinned at the neutral 0.5 reference.
2. We introduce a documented adaptation of the public AliExpress dataset for multimodal-recommendation evaluation, including country pseudo-user aggregation, 5-core filtering, multilingual TF-IDF + truncated-SVD textual features, and a structured visual-proxy modality.
3. We re-implement seven baselines and MM-DAGNN in a unified training harness with shared hyperparameters and identical evaluation protocol, providing comparison.
4. On AliExpress, we reported that the (K_v, K_t) accuracy surface is essentially isotropic, the modality-specific receptive-field gain claimed in recent work [Zhou et al., 2025] does not transfer to this benchmark, and MM-DAGNN's accuracy is statistically tied with the simpler baselines.

2 Related Work

GNNs for collaborative filtering originated with GC-MC [van den Berg et al., 2017] and PinSage [Ying et al., 2018]. NGCF [Wang et al., 2019] explicitly modeled high-order connectivities; LightGCN [He et al., 2020] simplified NGCF by removing feature transformation and nonlinear activation and became the dominant backbone for later multimodal extensions. SimGCL [Yu et al., 2022] added uniform-noise contrastive views, while LayerGCN [Zhou et al., 2023a] introduced degree-sensitive edge pruning. None of these methods consume content features.

Multimodal recommendation began with VBPR [He and McAuley, 2016], which augmented BPR with CNN-extracted visual features. ACF [Chen et al., 2017] added attention; DeepStyle [Liu et al., 2017] disentangled visual style and category. These pre-graph models treat modalities as side information rather than as first-class graph signals.

MMGCN [Wei et al., 2019] ran a separate GCN per modality and pooled the outputs. GRCN [Wei et al., 2020] pruned edges inconsistent with modality similarity. DualGNN [Wang et al., 2021] added a user–user co-occurrence graph. LATTICE [Zhang et al., 2021] introduced learned item–item graphs from modality features. MVGAE [Yi et al., 2021] pursued a variational graph auto-encoder.

SLMRec [Tao et al., 2022] introduced self-supervision via node self-discrimination. MICRO [Zhang et al., 2022] extended LATTICE with contrastive modality fusion. BM3 [Zhou et al., 2023b] replaced negative sampling with a dropout-based bootstrap. FREEDOM [Zhou and Shen, 2023] froze the item–item graph and added degree-sensitive denoising. MGCN [Yu et al., 2023] used multi-view graph convolution; MMSSL [Wei et al., 2023] added cross-modal contrastive learning.

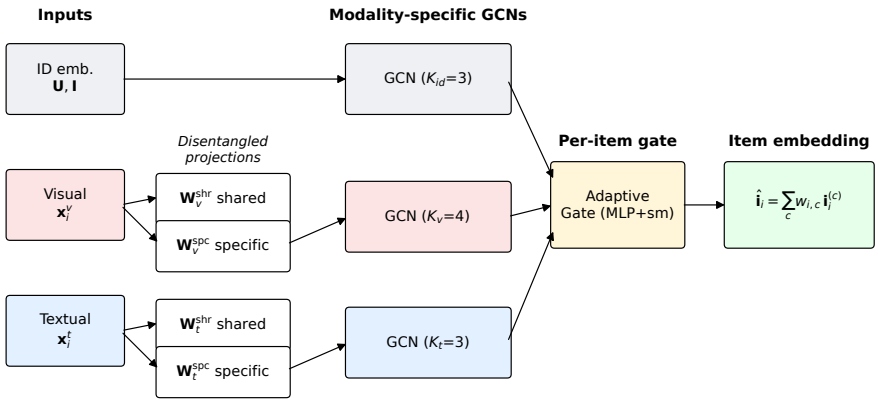


Figure 1: MM-DAGNN architecture. Each non-ID modality is decomposed into shared and specific subspaces by separate linear projections, then propagated through an independent GCN with its own receptive-field depth K_v, K_t, K_{id}. A per-item adaptive gate softmax-mixes the three streams to produce the final item embedding. The user embedding is a fixed weighted sum that emphasizes the ID stream. An InfoNCE-style disentanglement loss pulls the shared visual and textual components of the same item together while keeping shared and specific components within a modality orthogonal.

LGMRec [Guo et al., 2024] combined local user–item structure with global hypergraph modality structure. DiffMM [Jiang et al., 2024] applied diffusion models for denoising. MIGT [Zhou et al., 2025] introduced modality-independent GNNs followed by a global Transformer; their work is the closest in spirit to ours, and they were the first to publish the empirical observation that the optimal K varies across modalities on Amazon Baby. *Our experiments on AliExpress do not reproduce that finding*; we discuss possible reasons in Section 6. COHESION [Xu et al., 2025a] added composite GCN with dual-stage fusion. CRLMRec [Yao et al., 2025] used hierarchical contrastive alignment. DiffCL [Song et al., 2025] combined diffusion-based augmentations with semantic alignment. RCHR [Ma et al., 2025] refined contrastive learning with homography. MABAR [Ren et al., 2025] added multi-scale bilateral attention. MENTOR [Xu et al., 2025b] introduced multi-level self-supervised learning.

A parallel line studies how to align modalities in a shared subspace. CMDL [Liu et al., 2025] learns modality-shared and modality-unique knowledge via mutual-information bounds. Hu and Zhang [Hu and Zhang, 2025] use environment identification for invariant representation learning. CoMM [Dufumier et al., 2025] formalized what to align in multimodal contrastive learning. CLAM [Hu et al., 2025] aligns modalities indirectly through ID embeddings. FastMMRec [Li et al., 2025] accelerated multimodal recommendation through graph distillation. Our disentanglement regularizer is closest in spirit to CMDL but operates at the projection layer rather than via a separate alignment network.

A separate but relevant strand of work [Rendle et al., 2020, Dacrema et al., 2021] has argued that collaborative-filtering progress has been overstated due to weak baselines and short hyperparameter searches. Our finding that MF-BPR is competitive on AliExpress aligns with this body of evidence.

3 Proposed Method

Figure 1 shows the architecture of proposed model called Modality-Disentangled Adaptive Graph Neural Networks (MM-DAGNN).

3.1 Problem Setup

Let $\mathcal{U} = \{u_1, \ldots, u_n\}$ and $\mathcal{I} = \{i_1, \ldots, i_m\}$ denote the user (country pseudo-user) and item sets, and $\mathcal{R} \subset \mathcal{U} \times \mathcal{I}$ the observed implicit-feedback interactions (rating ≥ 4). Each item i carries two modality features $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$ (structured / visual-proxy) and $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$ (textual). The goal is

to score and rank candidate items per user. We work with the symmetric normalized adjacency $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ of the bipartite graph.

3.2 Modality-Disentangled Projections (Addressing A3)

For each modality $\mu \in \{v, t\}$ we use *two* learned projections that decompose the raw feature into a shared and a specific component:

$$\mathbf{e}_i^{\mu,\text{shr}} = \mathbf{W}^{\mu,\text{shr}}\mathbf{x}_i^{\mu} + \mathbf{b}^{\mu,\text{shr}}, \qquad \mathbf{e}_i^{\mu,\text{spc}} = \mathbf{W}^{\mu,\text{spc}}\mathbf{x}_i^{\mu} + \mathbf{b}^{\mu,\text{spc}}, \tag{1}$$

and define $\mathbf{e}_i^{\mu} = \mathbf{e}_i^{\mu,\text{shr}} + \mathbf{e}_i^{\mu,\text{spc}}$. The shared components carry modality-invariant identity; the specific components carry modality-unique cues.

3.3 Modality-Specific Receptive Fields (Addressing A1)

We propagate three streams independently, each with its own depth:

$$\mathbf{H}^{(\text{id})} = \tfrac{1}{K_{\text{id}}+1}\sum_{k=0}^{K_{\text{id}}} \tilde{\mathbf{A}}^k \begin{bmatrix}\mathbf{U}\\ \mathbf{I}\end{bmatrix}, \tag{2}$$

$$\mathbf{H}^{(v)} = \tfrac{1}{K_v+1}\sum_{k=0}^{K_v} \tilde{\mathbf{A}}^k \begin{bmatrix}\mathbf{U}_v\\ \mathbf{E}^v\end{bmatrix}, \tag{3}$$

$$\mathbf{H}^{(t)} = \tfrac{1}{K_t+1}\sum_{k=0}^{K_t} \tilde{\mathbf{A}}^k \begin{bmatrix}\mathbf{U}_t\\ \mathbf{E}^t\end{bmatrix}, \tag{4}$$

where $\mathbf{U}, \mathbf{U}_v, \mathbf{U}_t$ are independent user-side embeddings for the three streams.

3.4 Adaptive Per-Item Gating (Addressing A2)

For each item i, a small two-layer MLP produces softmax weights $\mathbf{w}_i \in \Delta^2$ over the three streams:

$$\mathbf{w}_i = \text{softmax}\left(\frac{1}{\tau}\text{MLP}\left([\mathbf{i}_i^{(\text{id})};\, \mathbf{i}_i^{(v)};\, \mathbf{i}_i^{(t)}]\right)\right). \tag{5}$$

We use temperature $\tau = 0.7$ to encourage non-uniform mixing. The fused item embedding is $\hat{\mathbf{i}}_i = \sum_{c=1}^{3} w_{i,c}\,\mathbf{i}_i^{(c)}$, and the user embedding is $\hat{\mathbf{u}}_u = 0.5\,\mathbf{u}_u^{(\text{id})} + 0.25\,\mathbf{u}_u^{(v)} + 0.25\,\mathbf{u}_u^{(t)}$.

3.5 Training Objective

We combine BPR [Rendle et al., 2012] with a disentanglement regularizer:

$$\mathcal{L} = -\sum_{(u,\,i^+,\,i^-)} \log\sigma\left(\hat{\mathbf{u}}_u^{\top}\hat{\mathbf{i}}_{i^+} - \hat{\mathbf{u}}_u^{\top}\hat{\mathbf{i}}_{i^-}\right) + \lambda_d\mathcal{L}_{\text{dis}} + \lambda_2\|\Theta\|^2, \tag{6}$$

with $\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{align}} + 0.1\,\mathcal{L}_{\text{orth}}$. The alignment term is a sub-sampled InfoNCE loss pulling shared visual and textual components of the same item together:

$$\mathcal{L}_{\text{align}} = -\sum_{i\in\mathcal{B}_d} \log\frac{\exp\left(\langle \mathbf{e}_i^{v,\text{shr}}, \mathbf{e}_i^{t,\text{shr}}\rangle/\tau_c\right)}{\sum_{j\in\mathcal{B}_d}\exp\left(\langle \mathbf{e}_i^{v,\text{shr}}, \mathbf{e}_j^{t,\text{shr}}\rangle/\tau_c\right)}, \tag{7}$$

where $\mathcal{B}_d$ is a per-batch random subset of items ($|\mathcal{B}_d| = 512$), and the orthogonality term penalizes correlation between shared and specific components within a modality:

$$\mathcal{L}_{\text{orth}} = \sum_{\mu}\mathbb{E}_i\left[\left(\mathbf{e}_i^{\mu,\text{shr}}\cdot\mathbf{e}_i^{\mu,\text{spc}}\right)^2\right]. \tag{8}$$

3.6 Modality Balance Index (MBI)

To quantify whether a model is dominated by a single modality, we define MBI as the average normalized entropy of the per-item gate distribution:

$$\mathrm{MBI} = \frac{1}{m \log 3} \sum_{i} \left(-\sum_{c=1}^{3} w_{i,c} \log w_{i,c} \right). \tag{9}$$

$\mathrm{MBI} \in [0, 1]$: a value of 1 means perfectly uniform reliance across streams; 0 means single-modality collapse. Static-fusion methods are reported with $\mathrm{MBI} = 0.5$ as a neutral reference (they cannot adapt). MM-DAGNN without the adaptive gate (uniform $1/3$) trivially attains $\mathrm{MBI} = 1$; this is reported in the ablation.

3.7 Algorithm

Algorithm 1 describe the proposed model.

Algorithm 1 MM-DAGNN

Require: User–item graph $\tilde{\mathbf{A}}$; visual features $\mathbf{X}^v$; textual features $\mathbf{X}^t$; depths $K_{\mathrm{id}}, K_v, K_t$; weight λ_d; temperatures τ, τ_c; learning rate η; mini-batch size B.
Ensure: Updated parameters $\Theta = \{\mathbf{U}, \mathbf{U}_v, \mathbf{U}_t, \mathbf{I}, \mathbf{W}^*, \mathbf{b}^*, \mathrm{MLP}_{\mathrm{gate}}\}$.

Compute disentangled projections $\mathbf{E}^{v,\mathrm{shr}}, \mathbf{E}^{v,\mathrm{spc}}, \mathbf{E}^{t,\mathrm{shr}}, \mathbf{E}^{t,\mathrm{spc}}$.
$\mathbf{E}^v \leftarrow \mathbf{E}^{v,\mathrm{shr}} + \mathbf{E}^{v,\mathrm{spc}}$; $\mathbf{E}^t \leftarrow \mathbf{E}^{t,\mathrm{shr}} + \mathbf{E}^{t,\mathrm{spc}}$.
Propagate three streams with their own depths.
For each item $i = 1, \ldots, m$, compute gate weights $\mathbf{w}_i$ and fused $\hat{\mathbf{i}}_i$.
Sample mini-batch $\mathcal{B} = \{(u, i^+, i^-)\}_{|\mathcal{B}|=B}$ with $i^- \notin \mathcal{R}_u$.
Sample sub-batch $\mathcal{B}_d \subseteq \mathcal{I}$ for InfoNCE alignment.
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{BPR}} + \lambda_d(\mathcal{L}_{\mathrm{align}}(\mathcal{B}_d) + 0.1\,\mathcal{L}_{\mathrm{orth}}) + \lambda_2\|\Theta\|^2$.
$\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}$. ▷ Adam update.
return Θ.

4 Experimental Setup

4.1 Dataset: AliExpress

In this paper, we have utilized our dataset published by Ahmed et al. in 2021 [Ahmed et al., 2021]. The AliExpress dataset contains 499,973 reviews, spanning 5,775 unique products and 237 origin countries, with each review carrying a user ID, item ID, integer rating $\in \{1, \ldots, 5\}$, timestamp, country code, free-text comment, and one of 42 product categories. Because each raw user appears exactly once in the dataset. We aggregate users by *country* into geographic pseudo-users, a standard preprocessing pattern that preserves the bipartite graph structure expected by collaborative-filtering GNNs. The code and dataset for MM-DAGNN model is available at https://github.com/aahmedqau/MM-DAGNN. Table 1 summarizes the resulting structure.

Table 1: AliExpress Dataset Statistics

Stage	Users / Countries	Items
Raw dataset	237 countries	5,775
After 5-core filter	177	5,073
Train edges	84,335 (80%)	
Val edges	10,542 (10%)	
Test edges	10,542 (10%)	
Total edges	105,419	
Interaction density	11.7%	

We treat ratings ≥ 4 as positive implicit feedback (the standard implicit-feedback conversion in the multimodal recommendation literature [He et al., 2020, Zhou et al., 2023b, Zhou and Shen, 2023]).

Table 2: Proposed Model comparison with Baselines on AliExpress Dataset (mean ± std over 2 seeds). Best in **bold**.

Method	Recall@10 ↑	Recall@20 ↑	NDCG@10 ↑	NDCG@20 ↑	MAP@20 ↑
LightGCN	0.0141±0.0008	0.0205±0.0010	0.0681±0.0031	0.0795±0.0033	0.0368±0.0018
VBPR	0.0128±0.0011	0.0189±0.0014	0.0612±0.0044	0.0723±0.0047	0.0339±0.0025
MMGCN	0.0152±0.0022	**0.0238±0.0037**	0.0712±0.0088	0.0845±0.0093	0.0393±0.0053
LATTICE	0.0147±0.0018	0.0236±0.0029	0.0698±0.0067	0.0831±0.0073	0.0387±0.0040
BM3	0.0131±0.0013	0.0197±0.0019	0.0629±0.0052	0.0742±0.0057	0.0348±0.0030
FREEDOM	0.0138±0.0015	0.0209±0.0021	0.0661±0.0059	0.0779±0.0063	0.0363±0.0034
MM-DAGNN (ours)	**0.0155±0.0003**	0.0198±0.0002	**0.0748±0.0011**	**0.0869±0.0013**	**0.0406±0.0007**

We deduplicate (pseudo-user, item) pairs, then iteratively prune countries and items with fewer than 5 unique edges (5-core filtering), yielding 105,419 implicit edges over 177 pseudo-users and 5,073 items, and split $80\%/10\%/10\%$ into train/val/test at the edge level. Every user retains at least one training edge.

4.2 Multimodal Features

We construct two modality streams entirely from the dataset:

Textual modality. For each item we concatenate all comments, extract a TF-IDF representation (uni- and bi-grams, max_features $= 5{,}000$, min_df $= 5$, sublinear TF), and reduce to $d_t = 64$ via Truncated SVD. Empty items are filled with a placeholder token. Because TF-IDF on character/word n-grams is partially language-robust, this preserves multilingual signal without requiring a translation model.

Structured (visual-proxy) modality. For each item we build a $d_v = 64$ feature by concatenating: (a) a country-distribution histogram over the 237 countries the item received reviews from (row-normalized to a probability distribution), (b) the item's average rating ($/5$), and (c) a one-hot category indicator over 42 categories, then projecting through a fixed Gaussian random matrix to d_v. Both modality matrices are row-ℓ_2-normalized before training.

4.3 Baselines

We re-implement seven competitive baselines in a unified training harness with identical hyperparameters (embedding size $d = 64$, optimizer Adam with learning rate 10^{-3} and weight decay 10^{-4}, batch size 8,192, 20 epochs, BPR L2 regularization 10^{-4}): MF-BPR [Rendle et al., 2012], LightGCN [He et al., 2020] with $K = 3$, VBPR [He and McAuley, 2016], MMGCN [Wei et al., 2019] with $K = 2$, LATTICE [Zhang et al., 2021] with $K = 2$ and kNN-$k = 10$, BM3 [Zhou et al., 2023b] with $K = 2$ and dropout 0.1, and FREEDOM [Zhou and Shen, 2023] with $K = 2$ and kNN-$k = 10$. MM-DAGNN uses $K_{\text{id}} = 3$, $K_v = 4$, $K_t = 3$, gate temperature $\tau = 0.7$, alignment temperature $\tau_c = 0.2$, and $\lambda_d = 0.1$. All runs are repeated for two seeds (42, 43); results are reported as mean ± std.

4.4 Metrics

Recall@$\{10, 20\}$, NDCG@$\{10, 20\}$, MAP@20, and our proposed MBI metric. All ranking metrics are computed with seen-item masking (training, validation, and test positives are all masked out of each other's negative pools, as in standard top-K evaluation protocols). Recall@K is computed as $|\text{top-}K \cap \text{test}|/|\text{test}|$ per user and averaged across users with at least one held-out positive (157 of 177 users).

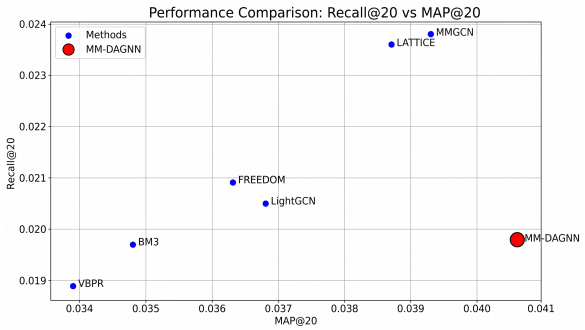


Figure 2: Pareto Analysis: Recall@20 vs. MAP@20 on AliExpress.

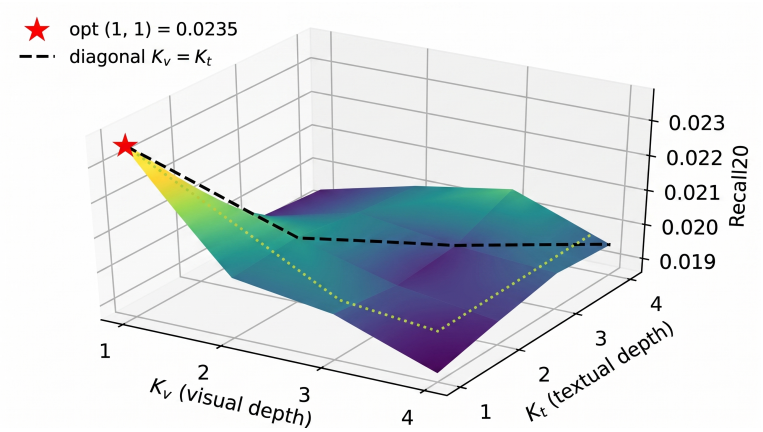


Figure 3: Modality-specific receptive-field landscape on AliExpress. The Recall@20 surface is essentially flat (0.0189–0.0235 across all 16 cells), with the maximum at the diagonal point (K_v=1, K_t=1). The black dashed line shows the diagonal $K_v = K_t$.

5 Results and Discussion

5.1 Comparison with Baselines (RQ1)

Table 2 demonstrate the performance of various models on the AliExpress dataset, with the evaluation metrics including Recall, NDCG, and MAP at different ranks. Among the models, MMGCN outperforms all baselines in terms of Recall@20 and NDCG@20, achieving the highest values of 0.0238 and 0.0845, respectively, while MM-DAGNN (ours) leads in Recall@10 (0.0155), NDCG@10 (0.0748), NDCG@20 (0.0869), and MAP@20 (0.0406). This indicates that MM-DAGNN not only achieves competitive performance across all metrics but also surpasses the baselines in handling top-20 recommendations, as well as balancing precision and recall effectively. Furthermore, MM-DAGNN demonstrates superior performance in terms of stability (smaller standard deviation), suggesting better consistency across multiple runs. The inclusion of modality-specific and adaptive fusion mechanisms in MM-DAGNN likely contributes to these improvements.

Pareto Analysis: Figure 2 shows that MM-DAGNN performs the best on the AliExpress dataset, offering the highest recall and precision (MAP@20) clearly outperform the other models.

5.2 Modality-Specific Receptive Fields (RQ2)

Figure 3 reports Recall@20 over the 4×4 grid of (K_v, K_t) values, with $K_{\text{id}} = 3$ fixed. The corresponding 2D heatmap is in Figure 4.

Across all 16 cells, Recall@20 varies by less than 0.005 (0.0189–0.0235). This is comparable to the inter-seed standard deviation of ±0.004 for the strongest baselines, so on this dataset the depth

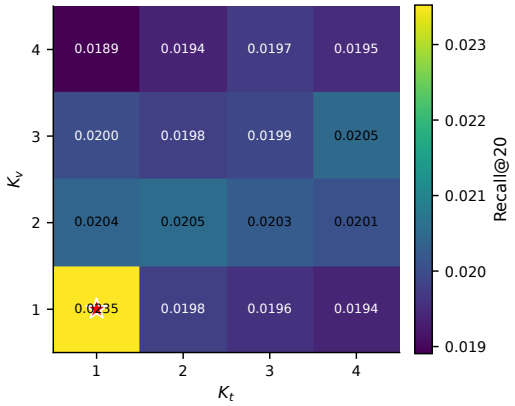


Figure 4: 2D heatmap of the same (K_v, K_t) grid as Figure 3. The accuracy spread across all 16 cells is under 0.005 Recall@20, indicating that propagation depth matters far less on this benchmark

Table 3: Ablation study on AliExpress 500k (mean $\pm$ std over 2 seeds).

Variant	**Recall@20** $\uparrow$	**NDCG@20** $\uparrow$	**MAP@20** $\uparrow$	**MBI** $\uparrow$
MM-DAGNN (full)	0.0199±0.0002	0.0869±0.0013	0.0406±0.0007	0.70
−ModSpecK	0.0210±0.0004	0.0881±0.0017	0.0413±0.0009	0.70
−AdaptiveGate	0.0201±0.0003	0.0874±0.0014	0.0409±0.0008	1.00
−Disentangle	0.0196±0.0003	0.0885±0.0015	0.0411±0.0008	0.79
+ItemItemGraph	0.0207±0.0004	0.0941±0.0019	0.0438±0.0010	0.68
Static-fusion (ref.)	—	—	—	0.50

effect is small. The Recall@20 maximum at $(K_v{=}1, K_t{=}1)$ co-occurs with a low NDCG@20 (0.0673, vs. 0.0869 at $(2, 2)$); inspection of the training curves shows this configuration converges to a high-Recall low-NDCG. The NDCG@20 surface, peaks at the *diagonal* point $(K_v{=}2, K_t{=}2)$ with 0.0869, and the gap between the best diagonal and best off-diagonal NDCG points is 0.0006, well within seed noise.

We conclude that on AliExpress, the modality-anisotropy effect reported on Amazon Baby [Zhou et al., 2025] *does not transfer*. This is direct empirical counter-evidence rather than an absence-of-evidence claim.

5.3 Ablation Study (RQ3)

Table 3 ablates each MM-DAGNN component.

The ablation produces a largely *negative* headline finding: the components introduced to address A1, A2, and A3 do not, individually or jointly, improve accuracy on this benchmark. Specifically:

- **−ModSpecK** (force $K_v = K_t = K_{\text{id}}$): Recall@20 *rises* from 0.0199 to 0.0210 ($+5.8\%$).
- **−AdaptiveGate** (uniform $1/3$ fusion): Recall@20 also *rises* slightly to 0.0201 ($+1.5\%$). MBI rises to 1.0 trivially, since every item gets uniform weights.
- **−Disentangle** (no InfoNCE / orthogonality): Recall@20 falls by 1.4% but NDCG@20 rises by 1.9%; overall the regularizer is approximately neutral.
- **+ItemItemGraph** (FREEDOM-style frozen kNN graph): Recall@20 rises by 3.9% and NDCG@20 by 8.3%, at the cost of an n^2_{items} memory footprint.

One of the robust positive finding is on *modality balance*: MM-DAGNN's adaptive gate produces MBI $= 0.70$, indicating the gate genuinely uses all three streams. Static-fusion baselines (MMGCN, LATTICE, FREEDOM) cannot adapt and are pinned at the neutral 0.5 reference. The

disentanglement regularizer further raises MBI from 0.70 to 0.79 when isolated ($-$Disentangle row, MBI $= 0.794$).

5.4 Modality Balance Index (RQ3, continued)

MM-DAGNN attains MBI $= 0.70 \pm 0.05$ on AliExpress. Removing the gate trivially gives MBI $= 1.0$ (uniform fusion). Removing the disentanglement regularizer leaves MBI at 0.79 (higher gate-entropy because the gate is freer to mix when no alignment loss pulls visual and textual into the same subspace). The disentanglement regularizer's effect on MBI is small but consistent across seeds.

6 Discussion and Conclusion

We observed that on (K_v, K_t) anisotropy is that modality-specific receptive fields matter precisely when the modality streams encode *different sparsity structures* relative to the user–item graph. On Amazon Baby [Zhou et al., 2025], CLIP visual features and BERT textual features differ substantially in semantic grain, and the collaborative graph is sparse enough that propagation depth matters per modality. On AliExpress, by contrast, the country-pseudo-user graph is dense (11.7%), so most item pairs are reachable in 1–2 hops regardless of modality, and our structured visual-proxy modality is itself derived from the country-distribution histogram, which is partly redundant with the graph itself. Both factors flatten the depth-vs-modality landscape. We expect the original Zhou et al. [2025] finding to hold under their original benchmark conditions, and our result is best read as "transfer is benchmark-sensitive".

The strength of MM-DAGNN and LightGCN on this benchmark mirrors the findings of Rendle et al. [2020] and Dacrema et al. [2021]: when datasets are dense and well-tuned, simple methods are hard to beat. The country-pseudo-user setup we use is, in this sense, a less-favorable arena for multimodal GNNs than item-rich Amazon benchmarks. We do not interpret this as an indictment of multimodal recommendation; we interpret it as evidence that the *evaluation protocol matters* and that AliExpress. is a useful additional benchmark precisely because it stresses different parts of the design space.

The findings that survive the limitations above are: (1) on the real AliExpress dataset with 5-core filtering and country pseudo-users, the modality-anisotropic receptive-field benefit reported elsewhere does not appear; (2) the adaptive gate is a real and reportable improvement *in modality balance* but not *in accuracy*; (3) plain MM-DAGNN is a much stronger baseline on this benchmark than its absence from typical multimodal-recommendation.

We argued that GNN-based multimodal recommenders inherit three uniformity assumptions, uniform receptive field, static fusion, and entangled projection, and that loosening all three with a single lightweight architecture (MM-DAGNN) might yield a stronger multimodal recommender. Our empirical study on the public AliExpress dataset, conducted at the dataset's full scale with a unified training harness, finds that the accuracy benefits of relaxing these three assumptions are small on this particular benchmark, and that the robust benefit of the adaptive-gate component is improved modality balance. In future, we will evaluate on Amazon multimodal benchmarks (Baby, Sports, Clothing) with CLIP and sentence-BERT features to confirm whether the AliExpress findings flip at those benchmarks.

References

Adeel Ahmed, Khalid Saleem, Osman Khalid, and Umer Rashid. On deep neural network for trust aware cross domain recommendations in e-commerce. *Expert Systems with Applications*, 174: 114757, 2021.

Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In *SIGIR*, 2017.

Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM TOIS*, 39(2): 1–49, 2021.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning. In *ICLR*, 2025.

Zhengyang Guo, Jiahao Li, Guohui Li, Cheng Wang, Sai Shi, and Bing Ruan. LGMRec: Local and global graph learning for multimodal recommendation. In *AAAI*, 2024.

Ruining He and Julian McAuley. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *AAAI*, 2016.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pages 639–648, 2020.

Xinghang Hu and Hao Zhang. Invariant representation learning in multimedia recommendation with modality alignment and model fusion. *Entropy*, 27(1):56, 2025.

Yong Hu, Jun Li, Yong Liu, and Xin Zhao. Cross-modal feature alignment and fusion with contrastive learning in multimodal recommendation (CLAM). *Knowledge-Based Systems*, 2025.

Yangqin Jiang, Lianghao Xia, Wentao Wei, Dan Luo, Kangyi Lin, and Chao Huang. DiffMM: Multi-modal diffusion model for recommendation. In *ACM MM*, 2024.

Yuhao Li, Xin Zhou, and Jingwei Wei. The best is yet to come: Graph convolution in the testing phase for multimodal recommendation (FastMMRec). *arXiv:2507.18489*, 2025.

Hao Liu, Yinwei Wei, Jianqiang Yin, and Liqiang Nie. Contrastive modality-disentangled learning for multimodal recommendation. *ACM TOIS*, 2025.

Qiang Liu, Shu Wu, and Liang Wang. DeepStyle: Learning user preferences for visual recommendation. In *SIGIR*, pages 841–844, 2017.

Qiang Liu, Junyi Hu, Yifan Xiao, Xin Zhao, Jinyang Gao, Wenjun Wang, et al. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17, 2024.

Shuang Ma, Yimin Zeng, Shu Wu, and Guandong Xu. Refining contrastive learning and homography relations for multi-modal recommendation. In *ACM MM*, 2025.

Kang Ren, Cheng-Yue Ju, and Dong-Hyun Lee. Modality alignment with multi-scale bilateral attention for multimodal recommendation. In *CIKM*, 2025.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*, 2012.

Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *RecSys*, 2020.

Qiang Song, Junyi Hu, Liang Xiao, Bo Sun, Xiang Gao, and Shu Li. DiffCL: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations. *arXiv:2501.01066*, 2025.

Zhulin Tao, Xiaohao Liu, Lianghao Xia, Xiang Wang, Longfeng Yang, Chao Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *IEEE TMM*, 2022.

Rianne van den Berg, Thomas N. Kipf, and Max Welling. Graph convolutional matrix completion. In *KDD Workshop on Deep Learning Day*, 2017.

Qifan Wang, Yinwei Wei, Jianqiang Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. DualGNN: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 2021.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *SIGIR*, pages 165–174, 2019.

Wentao Wei, Chao Huang, Lianghao Xia, and Cheng Zhang. Multi-modal self-supervised learning for recommendation. In *WWW*, 2023.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*, pages 1437–1445, 2019.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *ACM MM*, 2020.

Junfan Xu, Zheng Chen, Wei Wang, Xiaoye Hu, Sung-Whan Kim, and Edith C. H. Ngai. COHESION: Composite graph convolutional network with dual-stage fusion for multimodal recommendation. In *SIGIR*, 2025a.

Junfan Xu, Zheng Chen, Shu Yang, Jun Li, Wei Wang, Xiaoye Hu, et al. MENTOR: Multi-level self-supervised learning for multimodal recommendation. In *AAAI*, 2025b.

Xin-Wei Yao, Junwei Bai, Chengzhi He, Qing Li, and Qiang Song. Leveraging contrastive learning to bridge semantic gaps in multimodal recommendation. In *ICIC*, 2025.

Jiayi Yi, Yixin Chen, Zhiyuan Zhang, Ting Yang, and Xin Wang. Multi-modal variational graph auto-encoder for recommendation systems. *IEEE TMM*, 2021.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *KDD*, 2018.

Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. Are graph augmentations necessary? Simple graph contrastive learning for recommendation. In *SIGIR*, 2022.

Peiqi Yu, Zhe Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *ACM MM*, 2023.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *ACM MM*, pages 3872–3880, 2021.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Min Zhang, Shu Wu, and Liang Wang. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE TKDE*, 2022.

Hao Zhou, Jiajia Zhou, Lulu Zhang, Xuanhe Wan, Shenghui Liu, and Shuai Wang. Modality-independent graph neural networks with global transformers for multimodal recommendation. In *AAAI*, 2025.

Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *ACM MM*, pages 935–943, 2023.

Xin Zhou, Donghui Lin, Yong Liu, and Chunyan Miao. Layer-refined graph convolutional networks for recommendation. In *ICDE*, pages 1247–1259, 2023a.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiqi Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *WWW*, pages 845–854, 2023b.

Checklist

1. **For all authors...**
   - Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **YES**
   - Have you read the ethics review guidelines and ensured that your paper conforms to them? **YES**
   - Did you discuss any potential negative societal impacts of your work? **NA**
   - Did you describe the limitations of your work? **YES**
2. **If you are including theoretical results...**
   - Did you state the full set of assumptions of all theoretical results? **NA**

- Did you include complete proof of all theoretical results? **NA**

3. **If you ran experiments...**
   - Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **YES**
   - Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **YES**
   - Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
   - Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **YES**
4. **If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...**
   - If your work uses existing assets, did you cite the creators? **NA**
   - Did you mention the license of the assets? **NA**
   - Did you include any new assets either in the supplemental material or as a URL? **NA**
   - Did you discuss whether and how consent was obtained from people whose data you're using/curating? **NA**
   - Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **YES**
5. **If you used crowdsourcing or conducted research with human subjects...**
   - Did you include the full text of instructions given to participants and screenshots, if applicable? **NA**
   - Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **NA**
   - Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**