Can Complementary Signals Bridge Similarity Islands? Manifold-Augmented Graph Embedding for Multimodal Recommendation

Zhenye Yang[1,2] **Jinpeng Chen**[1,2*] **Wenbo Fu**[1,2] **Jun Ma**[1,2] **Huan Li**[3]
Qimin Zhou[1,2] **Tao Wang**[1,2] **Li Sun**[1] **Senzhang Wang**[4]
[1]School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, Beijing, China
[2]Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education
[3]Zhejiang University, Hangzhou, China [4]Central South University, Changsha, China
{yangzhenye,jpchen,clearmylove,majun,zhouqimin, wangtao005, lsun}@bupt.edu.cn
lihuan.cs@zju.edu.cn, szwang@csu.edu.cn

Abstract

Multimodal Recommender Systems (MMRS) leverage rich item features to mitigate data sparsity, yet similarity-driven graph construction and propagation can favor semantically similar substitutes over functionally related complements. When behavioral associations connect semantically distant items, complementary signals can remain geometrically isolated, yielding fragmented "Similarity Islands" and semantic drift when standard propagation is applied. To overcome these limitations, we propose **MAGE** (Manifold-Augmented Graph Embedding), a framework for personalized multimodal ranking that decouples complementarity modeling into offline geometric warping and online topological propagation. The offline phase enhances item representations by extracting complement-aware residual directions via a Soft Conceptor mechanism and injecting them under a local gap-preserving constraint, incorporating co-purchase evidence while controlling the magnitude of geometric perturbation. Synergistically, the online phase employs an asymmetric dual-stream GNN that treats visual features as stable anchors and augmented textual features as flexible probes, fused through a self-calibrated cross-modal mechanism to reconcile shared semantics and modality-specific residual cues. Extensive experiments on four Amazon datasets show that MAGE yields consistent improvements over strong multimodal recommendation baselines. Together with the supporting analyses in the appendix, these results suggest that geometry-aware augmentation can alleviate the semantic mismatch that arises in complementary recommendation settings. Our code is available at `https://github.com/Alex-bupt/mage`.

1 Introduction

Recommender systems help users navigate large collections of online information (Zhang et al., 2019). While traditional collaborative filtering struggles with persistent *data sparsity* (Natarajan et al., 2020), Multimodal Recommender Systems (MMRS) have emerged as a powerful remedy. By integrating rich side information such as product images and textual descriptions, MMRS construct dense semantic representations for items, substantially improving recommendation quality, particularly in cold-start and long-tail scenarios (Liu et al., 2024; Chen et al., 2017; Lin et al., 2024).

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Graph Neural Networks (GNNs) have become the dominant paradigm in MMRS due to their ability to propagate user preferences through higher-order connectivity (Gao et al., 2023). Models such as MMGCN (Wei et al., 2019), MENTOR (Xu et al., 2025), and GUME (Lin et al., 2024) refine graph structures and multimodal fusion. However, similarity-driven graph construction and propagation favor the **homophily assumption** (Zhu et al., 2020). Low-pass propagation (Chien et al., 2021) encourages representation smoothness across connected nodes, implicitly assuming that "connected items should be similar" (Platonov et al., 2023). This favors *substitutes*, such as two similar phone models, that naturally cluster in the embedding space.

Nevertheless, real-world purchasing behavior is frequently driven by **Complementarity** (e.g., co-purchase patterns), where users seek items that are functionally compatible rather than semantically similar (e.g., a *Charger Plug* for an iPhone) (McAuley et al., 2015). Although complementary recommendation has been studied using behavioral signals and graph-based formulations (Shang et al., 2024; Luo et al., 2024), similarity-driven multimodal approaches can retain a mismatch between behavioral topology and item features even after adding behavioral edges. Such designs become particularly limited in **multimodal settings**, where complementary items often differ drastically in both visual appearance and textual descriptions, giving rise to pronounced **heterophily** (Zheng et al., 2022). In contrast to generic heterophily-aware graph learning, our setting must additionally reconcile modality conflict: co-purchase relations provide imperfect evidence of functional relatedness, while the raw feature space remains organized around semantic similarity. This mismatch yields fragmented "Similarity Islands": substitute items cluster naturally, while behaviorally associated complements remain geometrically separated and difficult to connect through similarity-driven propagation. Empirical statistics and visualizations supporting the existence of these "Similarity Islands" are provided in Appendix C. Uniform smoothing over such heterophilic relations can then introduce **semantic drift** (Rusch et al., 2023), because enforcing smoothness between semantically distant but behaviorally related items may contaminate the seed item's core representation and weaken complementary exploration.

To better address these "Similarity Islands" without exacerbating semantic drift, we argue that multimodal recommendation with complementary signals may benefit from *geometric transformation* in addition to topological augmentation. We therefore propose **MAGE** (**M**anifold-**A**ugmented **G**raph **E**mbedding), a framework that decouples complementary modeling into two tightly coupled components: offline geometric augmentation and online asymmetric propagation. Our task is personalized user–item ranking, rather than recommending complements for a query item alone. At the representation level, MAGE performs offline manifold-based enhancement by extracting complement-aware residual directions via local SVD and Soft Conceptor projection, and injecting them into the textual modality under an explicit local gap-preserving constraint. At the architectural level, MAGE adopts an **asymmetric dual-stream** design (Wang et al., 2020), treating visual features as stable perceptual anchors and augmented textual features as flexible functional probes. A **Self-Calibrated Gate** is further introduced to harmonize these heterogeneous signals, so that complementary information can be incorporated without overwhelming an item's core identity.

i We study the representation–topology mismatch of complementary signals in personalized multimodal recommendation. Building on existing complementarity and heterophily research, we introduce **geometry-aware manifold reshaping** before topological propagation.

ii We propose MAGE, a framework that combines offline geometric augmentation with online asymmetric propagation. In Phase I, complement-aware residuals are injected under an explicit local gap-preserving rule with an adaptive coefficient; in Phase II, the *Self-Calibrated Gate* harmonizes stable visual anchors with augmented textual signals.

iii Experiments on four benchmark datasets show consistent improvements over strong baselines. A targeted hard-complement evaluation, fine-grained ablations, original-text controls, and graph perturbations characterize the contributions of geometric augmentation, modality-specific propagation, and text preprocessing, as well as sensitivity to imperfect graph evidence.

2 Related Work

Multimodal graph recommendation. VBPR integrates visual features with collaborative ranking (He and McAuley, 2016); MMGCN, MGCN, LGMRec, MENTOR, and GUME develop multi-

modal propagation, representation learning, and fusion (Wei et al., 2019; Yu et al., 2023; Guo et al., 2024; Xu et al., 2025; Lin et al., 2024). MAGE builds on this setting but explicitly transforms the textual geometry before assigning similarity and complementary connections to different streams.

Complementarity and heterophily. Prior complementary-item recommendation already models relevance together with dissimilarity (McAuley et al., 2015; Luo et al., 2024; Shang et al., 2024). In particular, SComGNN uses low- and mid-pass spectral filters to capture these two attributes (Luo et al., 2024). MAGE instead targets personalized user–item ranking with multimodal inputs, coupling a local feature transformation to modality-specific propagation. Generic heterophily-aware learning motivates departing from uniform smoothing (Zhu et al., 2020; Zheng et al., 2022; Pei et al., 2020); here the additional issue is how behavioral topology interacts differently with visual and textual features.

3 Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items. Each item $i \in \mathcal{I}$ is associated with multimodal features: a d_v-dimensional visual embedding $\mathbf{x}_{i,v} \in \mathbb{R}^{d_v}$ and a d_t-dimensional textual embedding $\mathbf{x}_{i,t} \in \mathbb{R}^{d_t}$. User–item interactions are given by a binary matrix $\mathbf{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$ with $R_{u,i} = 1$ indicating an observed interaction. We further leverage co-purchase relations $\mathcal{G}_{\text{co}} = (\mathcal{I}, \mathcal{E}_{\text{co}})$, where $(a, b) \in \mathcal{E}_{\text{co}}$ denotes an observed co-purchase association, used as imperfect evidence of complementarity. The graph comes from item metadata, not held-out user–item labels.

Similarity-based propagation encourages connected items to have similar representations. However, edges in $\mathcal{E}_{\text{co}}$ are often heterophilic: complementary items are functionally related but semantically distant in the raw feature space, making naive propagation prone to semantic drift.

Our goal is personalized user–item ranking with score $\hat{y}_{u,i}$. Specifically, we seek a **manifold transformation** $\phi(\cdot)$ that maps $\mathbf{x}_{i,t}$ to an augmented space $\tilde{\mathbf{x}}_{i,t}$, leveraging locality-preserving geometry to capture complementary signals beyond the dominant semantic subspace (He and Niyogi, 2003; Pei et al., 2020). The final model combines $\tilde{\mathbf{x}}_{i,t}$ with the stable visual anchors $\mathbf{x}_{i,v}$ to model both substitute and complementary preferences.

4 Methodology

In this section, we present MAGE, a framework designed to bridge the semantic gap in multimodal recommendation by leveraging complementary signals. The core philosophy of MAGE is to decouple the challenge of modeling **heterophilic complementary signals** into two distinct spaces: (1) the **Geometric Space**, where we resolve feature conflicts offline via manifold learning; and (2) the **Topological Space**, where we propagate complementary signals online via an asymmetric graph neural network. An overview of the proposed MAGE framework is illustrated in Figure 1.

4.1 Modality Initialization: Evidence-Bound Extraction

Raw product text in e-commerce is often noisy and inconsistently formatted, making functional cues such as compatibility constraints and specifications difficult to identify directly. We therefore apply a one-time **LLM-based evidence-bound extraction** step as data preprocessing (Wei et al., 2024), which is instructed to extract explicitly supported functional attributes and remove marketing-heavy content. The processed text is then encoded into a d-dimensional embedding $\mathbf{x}_{i,t}$, which serves as the textual input for subsequent manifold augmentation. The extracted descriptions are frozen before representation learning; no LLM is called during training or inference. Detailed prompts and failure handling are provided in Appendices G and B.3, respectively, and the original-text control is reported in Section 5.5.

4.2 Phase I: Offline Manifold-based Complementarity Injection

Complementary items are often geometrically distant in the raw feature space, making direct aggregation prone to noise and semantic drift. Phase I therefore performs an offline geometric transformation on the textual modality, mapping $\mathbf{x}_{i,t}$ to $\tilde{\mathbf{x}}_{i,t}$ before graph propagation. The key idea is

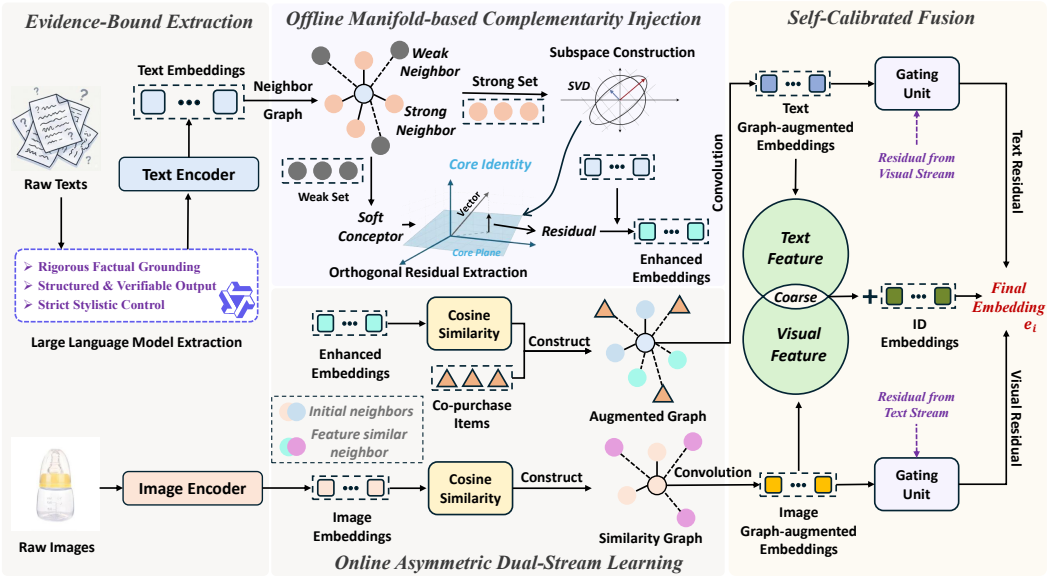


Figure 1: Architecture of the MAGE framework.

to extract a *complement-aware residual direction* from the local co-purchase neighborhood, while constraining its injection so that the original substitute–complement margin is not overly distorted.

4.2.1 Local Relative Subspace Characterization

Given a seed item i, we consider its co-purchase neighborhood $\mathcal{N}_{full}(i)$ and compute cosine similarities between the seed embedding $\mathbf{x}_{i,t}$ and its neighbors. Instead of using a global threshold shared across all items, we adopt a **per-seed relative percentile partition**: neighbors in the upper tail of the local similarity distribution form the *Strong Set* $\mathcal{S}_i$, while those in the lower tail form the *Weak Set* $\mathcal{W}_i$. These sets are local geometric proxies for substitute-like and complement-like relations, not verified semantic labels. To further suppress noisy co-purchase relations, we apply an additional similarity floor to weak neighbors, excluding extremely dissimilar items that are unlikely to encode meaningful complementarity.

We then build a local semantic matrix from the seed item and its strong neighbors,

$$\mathbf{C}_i = \begin{bmatrix} \mathbf{x}_{i,t} \\ \{\mathbf{x}_{j,t}\}_{j\in\mathcal{S}_i} \end{bmatrix}, \qquad \mathbf{C}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\top, \tag{1}$$

where $\mathbf{V}_i$ spans the dominant substitute subspace around item i. Rather than performing a hard truncation on singular directions, we use a **Soft Conceptor** to softly characterize this subspace:

$$w_{i,k} = \frac{\sigma_k^2}{\sigma_k^2 + \gamma^{-2}}, \tag{2}$$

where σ_k is the k-th singular value and γ controls the softness of the subspace boundary. Stronger semantic directions receive larger weights and are thus more aggressively suppressed in subsequent residual extraction.

4.2.2 Complement Residual Extraction

We use deviations from the dominant substitute-like subspace as candidate complementary signals. Co-purchase evidence and geometric constraints motivate this proxy, but do not certify that each residual is semantically complementary. Let $\mathbf{V}_i = [\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,r}]$ denote the right singular vectors and $\mathbf{w}_i = [w_{i,1}, \ldots, w_{i,r}]$ the corresponding Soft Conceptor weights. We define the soft projection operator as

$$\mathbf{P}_i = \mathbf{V}_i \mathrm{diag}(\mathbf{w}_i)\mathbf{V}_i^\top, \tag{3}$$

and remove the substitute-dominated component from each weak neighbor. Aggregating these filtered signals yields a consensus complementary direction:

$$\bar{\mathbf{r}}_i = \frac{1}{|\mathcal{W}_i|} \sum_{j \in \mathcal{W}_i} (\mathbf{I} - \mathbf{P}_i)\mathbf{x}_{j,t}. \tag{4}$$

To remove the component aligned with the seed representation, we orthogonalize this consensus direction against the unit-normalized seed embedding:

$$\mathbf{r}_i = \bar{\mathbf{r}}_i - (\bar{\mathbf{r}}_i^\top \mathbf{x}_{i,t})\, \mathbf{x}_{i,t}. \tag{5}$$

The resulting $\mathbf{r}_i$ serves as the final complement-aware residual direction of item i.

4.2.3 Gap-Constrained Adaptive Injection

A major risk of residual enhancement is over-injection: a large update may collapse the local separation between substitute-like and complement-like neighbors. We therefore adapt the injection strength to each seed rather than using a fixed global coefficient. The constraint concerns the strong/weak geometric proxies defined above; it does not assign semantic labels to individual neighbors.

Let $\boldsymbol{\mu}_{\mathcal{S}_i}$ and $\boldsymbol{\mu}_{\mathcal{W}_i}$ be the strong- and weak-set means, and define $\boldsymbol{\Delta}_i = \boldsymbol{\mu}_{\mathcal{S}_i} - \boldsymbol{\mu}_{\mathcal{W}_i}$, $g_i = \mathbf{x}_{i,t}^\top \boldsymbol{\Delta}_i$, and $d_i = \mathbf{r}_i^\top \boldsymbol{\Delta}_i$. Here the two means average the original textual embeddings in their respective sets. The value g_i measures how much more closely the seed aligns with its strong neighbors, while d_i measures how the residual changes this separation. For $g_i > 0$, we set the target gap to $g_i^{\text{target}} = \rho g_i$, where $\rho \in (0, 1)$ is the keep ratio. For an unnormalized update $\mathbf{z}_i(\alpha) = \mathbf{x}_{i,t} + \alpha \mathbf{r}_i$, the requirement is

$$\mathbf{z}_i(\alpha)^\top \boldsymbol{\Delta}_i \geq \rho g_i. \tag{6}$$

Since this update has gap $g_i + \alpha d_i$, for $g_i > 0$ the initial coefficient is

$$\alpha_i = \begin{cases} \min\left(\alpha_{\max}, \dfrac{(1-\rho)g_i}{-d_i + \varepsilon}\right), & d_i < 0, \\ \alpha_{\max}, & d_i \geq 0. \end{cases} \tag{7}$$

For $d_i < 0$, a larger ρ tightens the injection bound, while an active safety cap can leave the coefficient unchanged. For $d_i \geq 0$, the unnormalized gap does not decrease, but the cap still limits perturbation. We then normalize $\tilde{\mathbf{x}}_{i,t} = (\mathbf{x}_{i,t} + \alpha_i \mathbf{r}_i)/\|\mathbf{x}_{i,t} + \alpha_i \mathbf{r}_i\|_2$ and backtrack α_i if $\tilde{\mathbf{x}}_{i,t}^\top \boldsymbol{\Delta}_i < \rho g_i$. The cap and normalized-gap check control injection magnitude; they do not prove semantic correctness. Appendix A.1 gives the expanded derivation, and Section 5.5 isolates the safeguards empirically.

4.3 Phase II: Online Asymmetric Dual-Stream Learning

Because Phase I changes textual features while leaving visual features intact, we use an **Asymmetric Dual-Stream** architecture for complement-aware multimodal recommendation, treating visuals as stable anchors and augmented text as flexible probes, and assigning stability to the visual stream and plasticity to the textual stream.

4.3.1 Asymmetric Encoding Strategy

We differentiate the roles of visual and textual modalities into "Stability" and "Plasticity" to balance identity preservation with complementary exploration:

Visual Anchor (Stability): We treat the visual modality as a stable anchor to preserve the item's perceptual identity. We feed the *raw* visual embeddings $\mathbf{x}_{i,v}$ into a non-linear space projector $\phi_v(\cdot)$, implemented as a multi-layer perceptron (MLP) with Sigmoid activation, to map them into the interaction latent space: $\mathbf{e}_{i,v}^{(0)} = \phi_v(\mathbf{x}_{i,v})$.

Textual Probe (Plasticity): We treat the textual modality as a flexible probe to explore latent complementarity. Instead of raw text, we input the *manifold-augmented* embeddings $\tilde{\mathbf{x}}_{i,t}$ into a separate encoder $\phi_t(\cdot)$ with distinct parameters: $\mathbf{e}_{i,t}^{(0)} = \phi_t(\tilde{\mathbf{x}}_{i,t})$. This effectively injects the offline-learned geometric residuals into the online learning pipeline.

4.3.2 Modal-Asymmetric Topology Construction

To decouple similarity from complementarity, we design two parallel graph convolution tracks with asymmetric connectivity. This design is motivated by our observation that visual features are prone to noise when propagating over complementary edges, whereas augmented textual features are assigned the role of carrying complementary information.

Visual Stream (Homophilic Topology): This stream operates on a purified *Similarity Graph* $\mathcal{G}_{\text{sim}}$, constructed as the intersection of the visual k-NN graph and the textual k-NN graph. An item–item edge is retained only when it appears in both modality-specific nearest-neighbor graphs, so the resulting topology reflects visual–textual consensus without adding edges solely on the basis of co-purchase relations. The GCN on this track acts as a conservative smoother:

$$\mathbf{H}_v^{(l+1)} = \sigma(\tilde{\mathbf{A}}_{\text{sim}}\mathbf{H}_v^{(l)}\mathbf{W}_v^{(l)}), \tag{8}$$

where $\tilde{\mathbf{A}}_{\text{sim}}$ is the normalized adjacency matrix, $\mathbf{H}_v^{(l)}$ is the layer-l visual item representation matrix, $\mathbf{W}_v^{(l)}$ is a learnable transformation, and $\sigma(\cdot)$ is a non-linear activation. This homophilic propagation keeps visual representations grounded in intrinsic category identity without being perturbed by heterophilic complementary edges.

Textual Stream (Hybrid Topology): This stream operates on an *Augmented Complementary Graph* $\mathcal{G}_{\text{aug}}$. Here, we inject the filtered co-purchase edges $\mathcal{E}_{\text{co}}$ into the k-NN structure. To maintain structural consistency and sparsity, we apply a unified truncation strategy where both similarity and complementarity are bounded by the same neighbor count k.

We define a weighted adjacency matrix $\mathbf{A}_{\text{aug}}$ to balance the homophilic similarity signals and the heterophilic complementary residuals. For an edge (i, j), the weight is assigned as follows:

$$\mathbf{A}_{\text{aug}}(i,j) = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{E}_{\text{knn}} \\ \omega_{\text{co}}, & \text{if } (i,j) \in \mathcal{E}_{\text{co}} \text{ and } \text{rank}(j) \le k \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Here, $\mathcal{E}_{\text{knn}}$ denotes the edge set of the k-NN similarity graph, and $\text{rank}(j)$ is the rank of neighbor j among co-purchase candidates of item i sorted by co-purchase strength. The hyperparameter k bounds each edge source separately (their union can contain up to $2k$ neighbors before any symmetrization), while ω_{co} serves as a signal modulator that down-weights complementary edges to prevent heterophilic propagation from overwhelming the item's core semantic identity. The textual GCN then performs a hybrid propagation:

$$\mathbf{H}_t^{(l+1)} = \sigma(\tilde{\mathbf{A}}_{\text{aug}}\mathbf{H}_t^{(l)}\mathbf{W}_t^{(l)}). \tag{10}$$

By propagating the manifold-augmented textual features over this hybrid topology, MAGE diffuses high-order complementary signals while preserving the stability of the item representations.

In parallel, we maintain a collaborative ID backbone by propagating learnable user and item ID embeddings over the normalized user–item graph augmented with the textual item–item topology. The resulting representations, denoted as $\mathbf{e}_{u,\text{id}}$ and $\mathbf{e}_{i,\text{id}}$, serve as the behavioral backbone for fusion and contrastive regularization.

4.4 Self-Calibrated Cross-Modal Fusion

The streams contain shared category information and modality-specific cues. Our disentangle-then-fuse design estimates their common structure and calibrates the residuals before integration. From the final item streams $\mathbf{h}_{i,v}$ and $\mathbf{h}_{i,t}$, attention forms a shared prototype $\mathbf{h}_{i,c} = \alpha_v\mathbf{h}_{i,v} + \alpha_t\mathbf{h}_{i,t}$, where $(\alpha_v, \alpha_t) = \text{Softmax}(\phi_{attn}(\mathbf{h}_{i,v}) \parallel \phi_{attn}(\mathbf{h}_{i,t}))$. The collaborative ID representation modulates each residual: $\mathbf{r}_{i,m} = \sigma_{beh}^{(m)}(\mathbf{e}_{i,\text{id}}) \odot (\mathbf{h}_{i,m} - \mathbf{h}_{i,c})$. This separates deviations from the shared prototype and weights them using collaborative information; these residuals are not guaranteed to be purely complementary semantics. For $m \in \{v, t\}$ and its counterpart $\bar{m}$, the bidirectional gate refines it as

$$\mathbf{g}_{i,m} = \sigma(\text{MLP}_{\bar{m}\to m}(\mathbf{r}_{i,\bar{m}}) + \mathbf{W}_{m\to m}\mathbf{r}_{i,m}), \qquad \tilde{\mathbf{r}}_{i,m} = \mathbf{r}_{i,m} + \beta_g\, \mathbf{r}_{i,m} \odot \mathbf{g}_{i,m}. \tag{11}$$

The gate combines a cross-modal suggestion with self-modal calibration. Residual-style refinement preserves the input signal, with β_g controlling the adjustment. The final item vector combines the ID backbone with the shared and refined signals:

$$\mathbf{e}_i = \mathbf{e}_{i,\mathrm{id}} + \mathbf{e}_{i,\mathrm{fused}}, \qquad \mathbf{e}_{i,\mathrm{fused}} = \tfrac{1}{3}(\mathbf{h}_{i,c} + \tilde{\mathbf{r}}_{i,v} + \tilde{\mathbf{r}}_{i,t}). \tag{12}$$

Users have no raw modality features. Their visual/textual views, $\mathbf{e}^{\mathrm{ext}}_{u,v}$ and $\mathbf{e}^{\mathrm{ext}}_{u,t}$, are induced by propagating item signals through interaction histories. Applying the same prototype/residual/gating rule to these history-induced views yields $\mathbf{e}_{u,\mathrm{fused}}$, giving $\mathbf{e}_u = \mathbf{e}_{u,\mathrm{id}} + \mathbf{e}_{u,\mathrm{fused}}$. Appendix A.2 expands the computation.

4.5 Prediction and Optimization

Given the final user and item representations $\mathbf{e}_u$ and $\mathbf{e}_i$, we predict the preference score by inner product:

$$\hat{y}_{u,i} = \mathbf{e}_u^\top \mathbf{e}_i. \tag{13}$$

We optimize MAGE with a multi-task objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{BPR}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{bm}}\mathcal{L}_{\mathrm{BM}} + \lambda_{\mathrm{info}}\mathcal{L}_{\mathrm{UM}} + \lambda_{\mathrm{reg}}\|\Theta\|_2^2. \tag{14}$$

Here, $\mathcal{L}_{\mathrm{BPR}}$ is the main recommendation objective, while the remaining terms act as regularizers. The Info Loss in the sensitivity analysis is $\mathcal{L}_{\mathrm{UM}}$, weighted by λ_{info}; $\mathcal{L}_{\mathrm{BM}}$ is a separate regularizer.

To reduce modality-wise scale mismatch, we match the moments of the stream embeddings $\mathbf{E}_v$ and $\mathbf{E}_t$. We also align the fused representations with the collaborative backbone:

$$\mathcal{L}_{\mathrm{align}} = |\mu(\mathbf{E}_v) - \mu(\mathbf{E}_t)| + |\sigma^2(\mathbf{E}_v) - \sigma^2(\mathbf{E}_t)|, \tag{15}$$

$$\mathcal{L}_{\mathrm{BM}} = \mathcal{L}_{\mathrm{NCE}}(\mathbf{e}_{u,\mathrm{fused}}, \mathbf{e}_{u,\mathrm{id}}) + \mathcal{L}_{\mathrm{NCE}}(\mathbf{e}_{i,\mathrm{fused}}, \mathbf{e}_{i,\mathrm{id}}). \tag{16}$$

For users, let $\mathbf{e}_{u,\mathrm{expl}} = \frac{1}{2}(\mathbf{e}^{\mathrm{ext}}_{u,v} + \mathbf{e}^{\mathrm{ext}}_{u,t})$ denote the history-induced dual-stream view. The user regularizer combines agreement with this view and consistency between perturbed versions:

$$\begin{aligned}\mathcal{L}_{\mathrm{UM}} = {} & \mathcal{L}_{\mathrm{NCE}}(\mathbf{e}_{u,\mathrm{fused}}, \mathbf{e}_{u,\mathrm{expl}}) \\ & + \mathcal{L}_{\mathrm{NCE}}(\tilde{\mathbf{e}}^{(1)}_{u,\mathrm{fused}}, \tilde{\mathbf{e}}^{(2)}_{u,\mathrm{fused}}) + \mathcal{L}_{\mathrm{NCE}}(\tilde{\mathbf{e}}^{(1)}_{u,\mathrm{expl}}, \tilde{\mathbf{e}}^{(2)}_{u,\mathrm{expl}}).\end{aligned} \tag{17}$$

These terms regularize representation learning alongside BPR; they are not independent complementarity objectives. Appendix A.3 details the perturbation construction.

5 Experiments

5.1 Datasets and Experimental Inputs

We conduct experiments on four public Amazon benchmark datasets: *Baby*, *Sports*, *Clothing*, and *Electronics* (Yu et al., 2023; Xu et al., 2025; Lin et al., 2024). For the proposed pipeline, we preprocess raw product text with the **Evidence-Bound Extraction** pipeline using Qwen3-32B-AWQ (Yang et al., 2025), and encode the resulting descriptions into 512-dimensional textual features. Visual features are extracted as 4096-dimensional embeddings using a pre-trained encoder. The co-purchase graph is constructed from item-level metadata relations and kept separate from the user–item interaction labels used for evaluation.

To distinguish preprocessing effects from architectural gains, we evaluate baseline text replacement and MAGE with original text in Section 5.5. Dataset statistics, implementation settings, and sensitivity analyses are in Appendices B.1, B.2, and F.1.

5.2 Compared Methods

We compare MAGE with representative baselines from three categories: traditional collaborative filtering methods (**MF-BPR** (Rendle et al., 2009), **LightGCN** (He et al., 2020)), multimodal representation learning methods (**VBPR** (He and McAuley, 2016), **SLMRec** (Tao et al., 2023), **BM3** (Zhou et al., 2023), **MGCN** (Yu et al., 2023), **LGMRec** (Guo et al., 2024)), and recent graph structure learning methods that refine multimodal connectivity or cross-modal relations (**MENTOR** (Xu et al., 2025), **GUME** (Lin et al., 2024)).

Table 1: Performance comparison of baselines and MAGE in terms of Recall@K (R@K) and NDCG@K (N@K). The best results are **bolded**, and the second-best are underlined. The improvements of MAGE over the strongest baseline are statistically significant under a paired t-test over five random seeds ($p < 0.01$).

Dataset	Metric	Methods									
		MF-BPR	**LightGCN**	**VBPR**	**SLMRec**	**BM3**	**MGCN**	**LGMRec**	**MENTOR**	**GUME**	**MAGE**
Baby	R@10	0.0357	0.0479	0.0423	0.0540	0.0564	0.0620	0.0639	0.0678	0.0673	**0.0702**
	R@20	0.0575	0.0754	0.0663	0.0810	0.0883	0.0964	0.0989	0.1048	0.1042	**0.1077**
	N@10	0.0192	0.0257	0.0227	0.0285	0.0321	0.0339	0.0337	0.0362	0.0365	**0.0379**
	N@20	0.0249	0.0328	0.0284	0.0357	0.0383	0.0427	0.0430	0.0450	0.0460	**0.0475**
Sports	R@10	0.0432	0.0569	0.0558	0.0676	0.0656	0.0729	0.0719	0.0763	0.0778	**0.0821**
	R@20	0.0653	0.0806	0.0856	0.1076	0.1087	0.1106	0.1068	0.1139	0.1165	**0.1221**
	N@10	0.0241	0.0313	0.0307	0.0410	0.0408	0.0397	0.0387	0.0409	0.0427	**0.0440**
	N@20	0.0298	0.0387	0.0384	0.0462	0.0443	0.0496	0.0477	0.0511	0.0527	**0.0544**
Clothing	R@10	0.0187	0.0340	0.0423	0.0406	0.0482	0.0641	0.0555	0.0668	0.0703	**0.0738**
	R@20	0.0279	0.0526	0.0663	0.0710	0.0743	0.0945	0.0828	0.0989	0.1024	**0.1088**
	N@10	0.0103	0.0188	0.0223	0.0285	0.0231	0.0347	0.0302	0.0360	0.0384	**0.0396**
	N@20	0.0126	0.0236	0.0284	0.0357	0.0281	0.0428	0.0371	0.0441	0.0466	**0.0484**
Electronics	R@10	0.0158	0.0190	0.0229	0.0351	0.0343	0.0339	0.0341	0.0347	0.0355	**0.0383**
	R@20	0.0241	0.0320	0.0337	0.0549	0.0545	0.0520	0.0517	0.0540	0.0551	**0.0587**
	N@10	0.0076	0.0094	0.0113	0.0180	0.0171	0.0181	0.0183	0.0183	0.0178	**0.0200**
	N@20	0.0100	0.0124	0.0140	0.0228	0.0229	0.0227	0.0228	0.0232	0.0232	**0.0252**

5.3 Overall Performance

Table 1 reports full-test Recall and NDCG on four Amazon datasets. MAGE achieves the best reported values on every dataset and metric. Recall@20 improves over the strongest baseline by approximately 2.77%, 4.81%, 6.25%, and 6.53% on Baby, Sports, Clothing, and Electronics, respectively. These results show that the full pipeline improves ranking quality across the four evaluated domains. Compared with recent graph-based multimodal baselines, the gains are consistent with the motivation for transforming item geometry before propagation. The overall table alone cannot identify which design choices cause the improvements; the ablations below separate their contributions. Because standard top-K metrics do not isolate complementarity, we next examine semantically distant, behaviorally connected cases. Similarity-island statistics, long-tail results, and qualitative analyses remain in Appendices C, E.3, and E.2.

5.4 Performance on the Hard Complement Subset

To test the intended setting more directly, we select test positives connected through metadata co-purchase relations to a user's training-history item, with low visual and textual similarity to the selected history anchor. Both similarities must fall in the bottom 30% of the eligible pool; Appendix E.1 specifies the construction. This is a targeted diagnostic because it uses the relation type that motivates MAGE, while the primary benchmark remains the full-test evaluation above.

Table 2 shows improvements over GUME and MENTOR on all 16 dataset–metric combinations. Relative to the stronger of the two baselines for each metric, the mean improvement is 9.87%, with dataset means of 6.08%, 14.00%, 11.71%, and 7.69% on Baby, Sports, Clothing, and Electronics. The results support the ability to rank behaviorally connected but semantically distant items; they do not by themselves isolate which component produces the gain.

A separate training-frequency-based grouping, without semantic-distance selection, also shows higher tail Recall for MAGE than matched GUME on Sports and Clothing (Appendix E.3.1). This complements the targeted hard-subset analysis.

5.5 Ablation Study

We evaluate four block-level removals in Table 3. Removing Phase I retains complementary propagation, while removing complementary topology retains Phase-I features. Both removals reduce the two reported metrics on all four datasets, supporting their cumulative contribution.

Table 2: Performance on the hard complement subset. The stronger of GUME and MENTOR is the reference for the reported 9.87% mean relative gain. Subset construction is a targeted diagnostic, not an independent full-test benchmark.

Dataset	Method	R@10	R@20	N@10	N@20
Baby	GUME	0.1093	0.1591	0.0590	0.0720
	MENTOR	0.1093	0.1599	0.0558	0.0687
	MAGE	**0.1184**	**0.1676**	**0.0628**	**0.0754**
Sports	GUME	0.1751	0.2345	0.0998	0.1150
	MENTOR	0.1616	0.2167	0.0952	0.1093
	MAGE	**0.2030**	**0.2682**	**0.1127**	**0.1297**
Clothing	GUME	0.1596	0.2245	0.0832	0.0996
	MENTOR	0.1411	0.1954	0.0701	0.0836
	MAGE	**0.1762**	**0.2567**	**0.0913**	**0.1119**
Electronics	GUME	0.0720	0.1134	0.0367	0.0470
	MENTOR	0.0704	0.1092	0.0314	0.0425
	MAGE	**0.0762**	**0.1191**	**0.0408**	**0.0511**

Table 3: Ablation results of MAGE.

Variants	**Baby**		**Sports**		**Clothing**		**Electronics**	
	R@20	**N@20**	**R@20**	**N@20**	**R@20**	**N@20**	**R@20**	**N@20**
MAGE	**0.1077**	**0.0475**	**0.1221**	**0.0544**	**0.1088**	**0.0484**	**0.0587**	**0.0252**
w/o Phase I	0.1045	0.0458	0.1198	0.0510	0.1046	0.0472	0.0570	0.0246
w/o Asym. Prop.	0.1070	0.0473	0.1210	0.0521	0.1073	0.0480	0.0580	0.0245
w/o Comp. Topology	0.1061	0.0467	0.1210	0.0520	0.1049	0.0469	0.0552	0.0223
w/o Gate	0.1050	0.0465	0.1211	0.0528	0.1060	0.0475	0.0566	0.0231

Architecture-level controls. **w/o Phase I.** Removing offline augmentation reduces both metrics on all four datasets while retaining complementary topology, supporting a contribution from geometric preprocessing beyond adding behavioral edges. **w/o Asym. Prop.** Symmetric dual-stream propagation also underperforms the complete model, supporting distinct modality roles rather than the use of two streams alone. **w/o Comp. Topology.** Retaining augmented features but removing complementary edges lowers performance, especially on Electronics. Thus, feature augmentation and complementary topology contribute jointly within the evaluated pipeline. **w/o Gate.** Replacing the self-calibrated gate with concatenation reduces performance, supporting adaptive integration of the visual and textual streams. These block-level comparisons motivate finer controls, rather than establishing that every internal operation is necessary.

Simpler residuals and individual safeguards. Table 4 compares Phase-I variants with the remaining configuration and validation-based checkpoint selection fixed. Complete Phase I leads all 40 dataset–variant–metric comparisons in the full table (Appendix D). Mean residuals and hard-SVD truncation remain competitive but are consistently lower. The gap constraint has the largest average contribution among the three removed safeguards, and orthogonalization the smallest. The differences are descriptive and modest; no separate statistical significance or indispensability claim is made for each safeguard.

Text preprocessing and transfer. Passing original non-LLM text embeddings through the same MAGE architecture changes the four metrics by -0.45% to -1.07% on Baby and -1.03% to -1.69% on Clothing. Sports changes by $+0.25\%$ to $+1.85\%$, but uses longer validation patience to avoid premature stopping; it is a convergence diagnostic, not a strictly matched superiority claim (Table 9). Moreover, replacing only baseline text features with LLM-derived features produces mixed results, whereas offline augmentation improves all 24 reported Sports/Clothing metrics across

Table 4: Fine-grained Phase-I controls. Remaining settings are matched within each dataset. Full four-metric results are in Table 8.

	Sports		Clothing	
Variant	R@20	N@20	R@20	N@20
Complete Phase I	0.1221	0.0544	0.1088	0.0484
Mean residual	0.1211	0.0540	0.1074	0.0478
Hard-SVD residual	0.1206	0.0537	0.1072	0.0477
Without orthogonalization	0.1218	0.0543	0.1084	0.0482
Without gap constraint	0.1204	0.0536	0.1067	0.0475
Without normalized-gap backtracking	0.1209	0.0539	0.1080	0.0480

GUME, MGCN, and MENTOR (Table 7). Mixed Baby transfer results limit generalization. Together, these controls support geometric preprocessing beyond text refinement alone, without establishing universal transfer or ruling out simpler alternatives.

5.6 Robustness and Efficiency

Dropping or source-degree-preserving rewiring of 30% of eligible co-purchase edges changes Sports/Clothing metrics by at most 1.96% in magnitude. Each perturbed graph is used to recompute Phase-I features and complementary topology and retrain the model; the interaction graph, similarity graph, and evaluation set remain fixed (Appendix F.2). These tests assess the reliability of imperfect co-purchase evidence under the specified corruptions.

Once embeddings are available, Phase I takes 1.89/4.39/4.36 seconds on Baby/Sports/Clothing with peak memory below 0.4 GB. These costs exclude text extraction and embedding generation. The gate adds 12,448 parameters relative to same-input GUME; no LLM is invoked during training or inference. Appendix F.3 gives the scaling analysis.

In Phase II, MAGE follows the standard efficiency profile of lightweight GCN-based recommenders: sparse graph convolution dominates the cost, while the self-calibrated gate only involves low-dimensional linear operations and element-wise products.

6 Conclusion

In this paper, we proposed MAGE, a multimodal recommendation framework that incorporates heterophilic complementary signals through offline manifold augmentation and online asymmetric graph propagation. Experiments on real-world datasets show that MAGE consistently improves over strong baselines. Additional analyses indicate that MAGE better organizes functionally related yet semantically distant items, highlighting the potential of manifold-aware representation learning for complement-aware multimodal recommendation.

Limitations. Co-purchase edges are imperfect evidence, and geometric safeguards do not certify the semantic meaning of individual residuals. The tested 30% corruptions do not establish robustness to arbitrary graph noise, extreme sparsity, or temporal distribution shift. The hard-complement subset is a targeted diagnostic; transfer is mixed on Baby, and the Sports raw-text control changes validation patience. The design combines existing operators, with modest cumulative gains from individual safeguards. Simpler alternatives and broader deployment evaluation remain open directions.

Societal considerations. Complement-aware ranking may improve discovery, but co-purchase signals can also reflect popularity and commercial exposure biases. Offline accuracy does not establish user benefit or fair exposure; deployment should monitor these effects and the reliability of extracted product claims.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (No. 62572075), in part by the Beijing Natural Science Foundation (No. L233034, No. L253004) and in part by Fundamental Research Funds for the Beijing University of Posts and Telecommunications (No. 2025TSQY01).

References

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 *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344, 2017.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, 1(1):1–51, 2023.

Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. LGMRec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8454–8462, 2024.

Ruining He and Julian McAuley. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 639–648, 2020.

Xiaofei He and Partha Niyogi. Locality preserving projections. *Advances in Neural Information Processing Systems*, 16, 2003.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Guojiao Lin, Meng Zhen, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. GUME: Graphs and user modalities enhancement for long-tail multimodal recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1400–1409, 2024.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17, 2024.

Haitong Luo, Xuying Meng, Suhang Wang, Hanyun Cao, Weiyao Zhang, Yequan Wang, and Yujun Zhang. Spectral-based graph neural networks for complementary item recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8868–8876, 2024.

Sichun Luo, Chen Ma, Yuanzhang Xiao, and Linqi Song. Improving long-tail item recommendation with graph augmentation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1707–1716, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2015.

Senthilselvan Natarajan, Subramaniyaswamy Vairavasundaram, Sivaramakrishnan Natarajan, and Amir H Gandomi. Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data. *Expert Systems with Applications*, 149:113248, 2020.

Hongbin Pei, Bingzhe Wei, Kevin Chen Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond. *Advances in Neural Information Processing Systems*, 36:523–548, 2023.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009.

T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Jin Shang, Yang Jiao, Chenghuan Guo, Minghao Sun, Yan Gao, Jia Liu, Michinari Momma, Itetsu Taru, and Yi Sun. Transitivity-encoded graph attention networks for complementary item recommendations. In *2024 IEEE International Conference on Data Mining (ICDM)*, pages 430–439. IEEE, 2024.

Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116, 2023. doi: 10.1109/TMM.2022.3187556.

Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1001–1010, 2020.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. LLMRec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 806–815, 2024.

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 *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1437–1445, 2019.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C. H. Ngai. MENTOR: Multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12908–12917, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6576–6585, 2023.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1):1–38, 2019.

Xin Zheng, Yi Wang, Yixin Liu, Ming Li, Miao Zhang, Di Jin, Philip S Yu, and Shirui Pan. Graph neural networks for graphs with heterophily: A survey. *arXiv preprint arXiv:2202.07082*, 2022.

Xin Zhou. MMRec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–2, 2023.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 845–854, 2023.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in Neural Information Processing Systems*, 33:7793–7804, 2020.

A Supplementary Method Details

A.1 Gap-Constrained Injection Derivation

A major risk of residual enhancement is over-injection: while stronger perturbations may introduce complementary information, they may also collapse the local semantic separation between substitute-like and complement-like neighbors. To address this issue, we do not use a fixed global injection coefficient. Instead, we compute an **adaptive injection strength** from a local substitute–complement gap constraint.

Let

$$\boldsymbol{\mu}_{\mathcal{S}_i} = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \mathbf{x}_{j,t}, \qquad \boldsymbol{\mu}_{\mathcal{W}_i} = \frac{1}{|\mathcal{W}_i|} \sum_{j \in \mathcal{W}_i} \mathbf{x}_{j,t}, \tag{18}$$

and define the original local substitute–complement gap as

$$g_i^{\text{before}} = \mathbf{x}_{i,t}^{\top} \left(\boldsymbol{\mu}_{\mathcal{S}_i} - \boldsymbol{\mu}_{\mathcal{W}_i} \right). \tag{19}$$

For $g_i^{\text{before}} > 0$, we define the target gap as $g_i^{\text{target}} = \rho g_i^{\text{before}}$ and require the enhanced representation to preserve at least this fraction of the original margin:

$$g_i^{\text{after}} \geq \rho \, g_i^{\text{before}}, \tag{20}$$

where $\rho \in (0, 1)$ is the keep ratio. The derivation below assumes $g_i^{\text{before}} > 0$.

Using the unnormalized injected representation

$$\mathbf{z}_i(\alpha) = \mathbf{x}_{i,t} + \alpha \mathbf{r}_i, \tag{21}$$

the unnormalized post-injection gap can be written as

$$g_i^{\text{unnorm}} = \mathbf{z}_i(\alpha)^{\top} \left(\boldsymbol{\mu}_{\mathcal{S}_i} - \boldsymbol{\mu}_{\mathcal{W}_i} \right) = g_i^{\text{before}} + \alpha d_i, \tag{22}$$

where

$$d_i = \mathbf{r}_i^{\top} \left(\boldsymbol{\mu}_{\mathcal{S}_i} - \boldsymbol{\mu}_{\mathcal{W}_i} \right) \tag{23}$$

measures how the residual direction changes the local gap. If $d_i \geq 0$, the unnormalized gap does not decrease, so we allow the capped injection. If $d_i < 0$, the constraint gives $\alpha \leq (1 - \rho) g_i^{\text{before}} / (-d_i)$. We use the numerically stabilized bound

$$\alpha_i^{\text{gap}} = \frac{(1 - \rho) \, g_i^{\text{before}}}{-d_i + \varepsilon}, \tag{24}$$

where ε is a small constant for numerical stability. In practice, we use

$$\alpha_i = \min(\alpha_{\max}, \alpha_i^{\text{gap}}) \quad \text{if } d_i < 0, \qquad \alpha_i = \alpha_{\max} \quad \text{if } d_i \geq 0, \tag{25}$$

with the fixed safety cap $\alpha_{\max} = 0.35$ in the public implementation. For fixed $g_i^{\text{before}} > 0$ and $d_i < 0$, increasing ρ makes this initial coefficient non-increasing: the gap bound tightens, but the coefficient remains 0.35 while the cap is active. A smaller ρ permits more gap reduction; it does not necessarily produce a larger actual update.

The final enhanced textual representation is then

$$\tilde{\mathbf{x}}_{i,t} = \frac{\mathbf{x}_{i,t} + \alpha_i \mathbf{r}_i}{\|\mathbf{x}_{i,t} + \alpha_i \mathbf{r}_i\|_2}. \tag{26}$$

Since the final representation is ℓ_2-normalized for cosine-based downstream learning, the unnormalized bound alone does not ensure the final gap condition. We further verify the normalized gap $g_i^{\text{after}} = \tilde{\mathbf{x}}_{i,t}^{\top} (\boldsymbol{\mu}_{\mathcal{S}_i} - \boldsymbol{\mu}_{\mathcal{W}_i})$ and backtrack α_i whenever it falls below g_i^{target}.

Phase I differs from standard manifold augmentation in two aspects. First, complementarity is extracted *relatively* from each seed's local geometry through the strong/weak partition, rather than by imposing a global notion of similarity. Second, the residual is injected under an explicit *gap-preserving rule* with post-normalization verification, keeping complementary enhancement bounded by the local substitute–complement structure. As a result, Phase I introduces complement-aware directions into the textual space while controlling the risk of semantic drift before online propagation.

A.2 Self-Calibrated Fusion Details

Let $\mathbf{h}_{i,v}$ and $\mathbf{h}_{i,t}$ denote the final-layer item representations from the visual and textual streams. Since these two streams contain both shared category semantics and modality-specific complementary cues, we adopt a disentangle-then-fuse strategy that separates common structure from complementary residuals before integration.

We first extract a shared coarse-grained prototype by attention:

$$\alpha_v, \alpha_t = \mathrm{Softmax}\left(\phi_{attn}(\mathbf{h}_{i,v}) \parallel \phi_{attn}(\mathbf{h}_{i,t})\right), \qquad \mathbf{h}_{i,\mathrm{coarse}} = \alpha_v \mathbf{h}_{i,v} + \alpha_t \mathbf{h}_{i,t}. \tag{27}$$

This prototype captures the category-level identity jointly supported by both modalities.

To isolate complementary information, we subtract the shared prototype from each modality and modulate the resulting residuals using the collaborative representation:

$$\mathbf{r}_{i,m} = \sigma_{beh}^{(m)}(\mathbf{e}_{i,\mathrm{id}}) \odot \left(\mathbf{h}_{i,m} - \mathbf{h}_{i,\mathrm{coarse}}\right), \qquad m \in \{v, t\}, \tag{28}$$

where $\mathbf{e}_{i,\mathrm{id}}$ is the collaborative ID representation after graph propagation.

We then refine these residuals through a bidirectional self-calibrated gate. For each modality m, the gate combines cross-modal suggestion from the counterpart modality $\bar{m}$ and self-modal calibration:

$$\mathbf{g}_{i,m} = \sigma\left(\mathrm{MLP}_{\bar{m}\to m}(\mathbf{r}_{i,\bar{m}}) + \mathbf{W}_{m\to m}\mathbf{r}_{i,m}\right), \qquad m \in \{v, t\}, \tag{29}$$

where $\bar{m}$ denotes the counterpart modality of m. Instead of directly masking the residuals, we use a residual-style refinement:

$$\tilde{\mathbf{r}}_{i,m} = \mathbf{r}_{i,m} + \beta_g \left(\mathbf{r}_{i,m} \odot \mathbf{g}_{i,m}\right), \qquad m \in \{v, t\}. \tag{30}$$

This residual update retains the modality-specific signal and scales it using learned cross-modal and self-modal gates.

The fused item representation is then obtained by combining the coarse-grained prototype with the two refined residuals:

$$\mathbf{e}_{i,\mathrm{fused}} = \frac{1}{3}\left(\mathbf{h}_{i,\mathrm{coarse}} + \tilde{\mathbf{r}}_{i,v} + \tilde{\mathbf{r}}_{i,t}\right), \qquad \mathbf{e}_i = \mathbf{e}_{i,\mathrm{id}} + \mathbf{e}_{i,\mathrm{fused}}. \tag{31}$$

For users, we do not assume raw visual or textual inputs. Instead, we derive modality-conditioned user views by propagating visual and textual item signals over the user–item interaction graph. The resulting history-induced views, denoted as $\mathbf{e}_{u,v}^{\mathrm{ext}}$ and $\mathbf{e}_{u,t}^{\mathrm{ext}}$, are fused with the same coarse/residual/gating mechanism to obtain $\mathbf{e}_{u,\mathrm{fused}}$. The final user representation is

$$\mathbf{e}_u = \mathbf{e}_{u,\mathrm{id}} + \mathbf{e}_{u,\mathrm{fused}}. \tag{32}$$

Thus, item fusion operates on multimodal item streams, while user fusion operates on modality-conditioned interaction histories.

A.3 Auxiliary Objectives

To reduce modality-wise scale mismatch, we use a lightweight moment-matching term

$$\mathcal{L}_{\mathrm{align}} = |\mu(\mathbf{E}_v) - \mu(\mathbf{E}_t)| + \left|\sigma^2(\mathbf{E}_v) - \sigma^2(\mathbf{E}_t)\right|, \tag{33}$$

where $\mathbf{E}_v$ and $\mathbf{E}_t$ denote the visual and textual stream embeddings.

To stabilize complementary augmentation, we align the fused multimodal representations with the collaborative backbone:

$$\mathcal{L}_{BM} = \mathcal{L}_{\mathrm{NCE}}(\mathbf{e}_{u,\mathrm{fused}}, \mathbf{e}_{u,\mathrm{id}}) + \mathcal{L}_{\mathrm{NCE}}(\mathbf{e}_{i,\mathrm{fused}}, \mathbf{e}_{i,\mathrm{id}}). \tag{34}$$

For users, we additionally define a history-induced dual-stream representation

$$\mathbf{e}_{u,\mathrm{expl}} = \frac{1}{2}\left(\mathbf{e}_{u,v}^{\mathrm{ext}} + \mathbf{e}_{u,t}^{\mathrm{ext}}\right), \tag{35}$$

and regularize the fused user representation against it:

$$\mathcal{L}_{\mathrm{UM}} = \mathcal{L}_{\mathrm{NCE}}(\mathbf{e}_{u,\mathrm{fused}}, \mathbf{e}_{u,\mathrm{expl}}) + \mathcal{L}_{\mathrm{NCE}}(\tilde{\mathbf{e}}_{u,\mathrm{fused}}^{(1)}, \tilde{\mathbf{e}}_{u,\mathrm{fused}}^{(2)}) + \mathcal{L}_{\mathrm{NCE}}(\tilde{\mathbf{e}}_{u,\mathrm{expl}}^{(1)}, \tilde{\mathbf{e}}_{u,\mathrm{expl}}^{(2)}), \tag{36}$$

where perturbed views are generated by

$$\tilde{\mathbf{e}} = \mathbf{e} + \eta\, \mathrm{sign}(\mathbf{e}) \odot \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_2 + \varepsilon}. \tag{37}$$

B Supplementary Experimental Setup

B.1 Dataset Statistics

Table 5 summarizes the basic statistics of the four benchmark datasets used in our experiments, including the numbers of users, items, interactions, sparsity, and the average number of co-purchase neighbors per item.

Table 5: Statistics of the experimental datasets.

Dataset	User	Item	Inter.	Sparsity	Avg. Co-Pur.
Baby	19,445	7,050	160,792	99.88%	23.9
Sports	35,598	18,357	296,337	99.95%	14.0
Clothing	39,387	23,033	278,677	99.97%	13.7
Electronics	192,403	63,001	1,689,188	99.99%	25.6

B.2 Implementation Details

We implemented MAGE and baselines using the MMRec (Zhou, 2023) framework on an NVIDIA RTX 3090 24GB GPU, fixing the embedding size to 64 and optimizing with Adam (Kingma and Ba, 2015) using a learning rate of 10^{-3}. In **Phase I**, we construct the strong and weak neighbor sets using a per-seed relative percentile partition on local cosine similarities, and further apply a minimum similarity filter to suppress noisy weak co-purchase neighbors. We use a Soft Conceptor aperture of $\gamma = 12.0$. The residual injection coefficient is not tuned as a fixed global scalar; instead, it is adaptively determined by a local gap-preservation constraint and further bounded by a safety cap ($\alpha_{\max} = 0.35$ in the public implementation). In **Phase II**, the loss weights λ_{align} and λ_{info} were tuned in $\{10^{-4}, \ldots, 10^{-1}\}$. The contrastive temperature τ_{bm} was fixed at 0.2 on all four datasets; τ_{um} was fixed at 0.1 on Baby and Sports and 0.2 on Clothing and Electronics. The GCN depths were set to 3 for user–item interaction and 1 for item–item propagation. Following our asymmetric design, we set a unified neighbor count $k = 10$ for both similarity and complementary edges, with the complementary edge modulator fixed at $\omega_{\text{co}} = 0.4$.

B.3 Frozen Text Preprocessing and Failure Handling

Qwen3-32B-AWQ is queried once per item through a local OpenAI-compatible endpoint, using the fixed prompt and schema in Appendix G. The configuration uses temperature 0.1, maximum output length 900 tokens, request timeout 120 seconds, and 48 asynchronous workers. Neither top-p nor a decoding seed is explicitly set; they follow serving-backend defaults. The generated descriptions are frozen before representation learning so downstream runs use the same features. Freezing outputs supports downstream reproducibility but does not establish generation invariance across backends.

Responses are requested in JSON-object mode and checked for the seven required top-level fields and their container/value types. Request or parsing failures within an attempt are retried once; a complete generation that still fails validation is regenerated once. Persistently invalid items are excluded from normal preprocessing output and recorded in a separate item-level JSONL file with identifier, failure reason, and truncated diagnostics. This statement concerns preprocessing output, not an unreported change to the evaluation population.

Processed records retain assertions and their evidence spans, rejected claims, model-reported violations, and failure records. The prompts are provided in Appendix G; the public repository linked in the abstract provides the released code and data-access instructions.

C Empirical Evidence for Similarity Islands

A key motivation of MAGE is that complementary items are often not naturally close in the raw embedding space, even when they are behaviorally associated through co-purchase relations. To directly examine this phenomenon, we compare **substitute** item pairs and **complementary** item

pairs in terms of cosine similarity, and visualize their distributions in the original text embedding space.

C.1 Pairwise Similarity Statistics

Table 6 reports the average cosine similarity of substitute and complementary item pairs under both textual and visual modalities. Across all datasets, substitute pairs are consistently much more similar than complementary pairs, with clear gaps in both modalities. For example, on *Baby*, the average text similarity drops from 0.7524 for substitutes to 0.3748 for complements, while the average visual similarity drops from 0.5614 to 0.2044. Similar trends hold on *Sports*, *Clothing*, and *Electronics*. These statistics provide direct empirical evidence that complementary items are systematically more dispersed in the raw multimodal space, rather than being naturally grouped with semantically similar substitutes.

Table 6: Pairwise cosine similarity of substitute and complementary item pairs. The reported gap is defined as the difference between the average substitute similarity and the average complementary similarity. Larger gaps indicate stronger "Similarity Islands" effects.

Dataset	**Modality**	**Substitute** ($\mu \pm \sigma$)	**Complement** ($\mu \pm \sigma$)	**Gap**
Baby	Text	0.7524 ± 0.1099	0.3748 ± 0.0806	0.3776
Baby	Visual	0.5614 ± 0.1350	0.2044 ± 0.0719	0.3570
Sports	Text	0.7288 ± 0.1154	0.3640 ± 0.0848	0.3648
Sports	Visual	0.5894 ± 0.1672	0.1885 ± 0.0857	0.4009
Clothing	Text	0.7319 ± 0.1241	0.4652 ± 0.1109	0.2667
Clothing	Visual	0.6306 ± 0.1640	0.3087 ± 0.1366	0.3219
Elec	Text	0.8754 ± 0.0740	0.6154 ± 0.0985	0.2600
Elec	Visual	0.5537 ± 0.2487	0.2014 ± 0.1056	0.3522

C.2 Text-Space Visualization

To complement the numerical statistics, Figure 2 visualizes the raw *text* embedding space using t-SNE. In each plot, blue circles denote seed items, orange triangles denote substitutes, and red crosses denote complementary items. A consistent pattern can be observed across datasets: substitutes tend to remain closer to the seed items and occupy denser local neighborhoods, whereas complementary items are more scattered and often distributed across separate regions of the space. This effect is especially visible on *Baby* and *Sports*, and remains noticeable on *Clothing* and *Electronics* despite their higher textual similarity among complements.

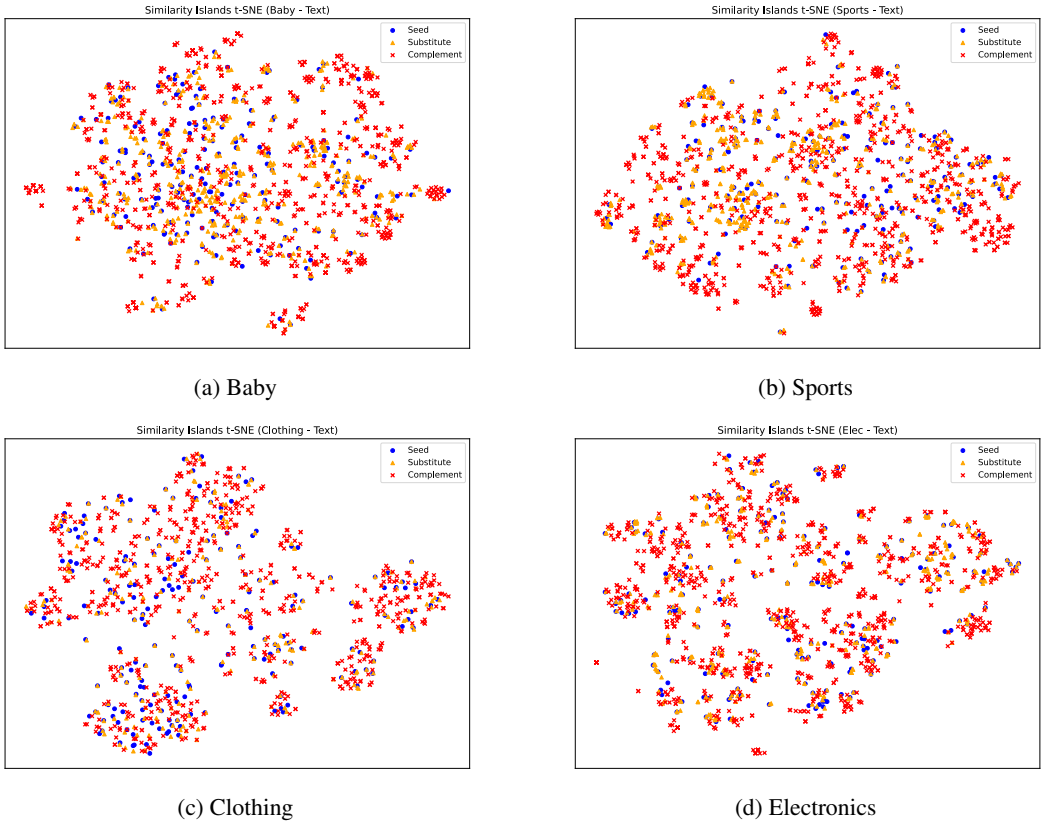
Together, the statistics in Table 6 and the visualization in Figure 2 support the core motivation of our method: complementary recommendation is not merely a matter of adding more edges in the graph, because the raw embedding space itself is organized around semantic similarity. This fragmented geometry forms what we call **Similarity Islands**, making it difficult for standard homophily-based propagation to recover complementary relations without introducing semantic drift.

D Generalization and Source of Performance Gains

D.1 Transferability Across Existing Multimodal Recommenders

A natural question is whether the proposed offline manifold augmentation is intrinsically tied to MAGE, or can serve as a plug-and-play representation enhancement for existing multimodal recommenders. To answer this, we inject our augmented item embeddings into three strong baselines (**GUME**, **MGCN**, and **MENTOR**) without changing their architectures, graph construction, or training objectives. Concretely, we only replace the original item embeddings and keep all other settings fixed.

As shown in Table 7, offline augmentation yields clear and consistent gains on *Sports* and *Clothing* across all three backbones, indicating that the proposed geometric enhancement provides a transferable inductive bias for modeling **heterophilic complementarity**. These improvements suggest that



(a) Baby (b) Sports

(c) Clothing (d) Electronics

Figure 2: t-SNE visualization of the raw text embedding space for four datasets. Blue circles denote seed items, orange triangles denote substitutes, and red crosses denote complementary items. Across datasets, substitutes remain more locally concentrated around seed items, whereas complementary items are more dispersed, illustrating the "Similarity Islands" phenomenon.

the benefit of our method is not restricted to the specific architecture of MAGE, but can generalize to existing multimodal recommenders as a lightweight plug-and-play enhancement.

On *Baby*, however, the gains are marginal or occasionally negative, indicating that the benefit of transferring augmented features depends on the dataset and backbone. Differences in effective neighborhood coverage or co-purchase noise are possible explanations, but these factors were not isolated in this experiment. The results therefore do not identify sparsity as the cause of the mixed gains.

D.2 LLM Text Refinement Alone Is Not Sufficient

Table 7 also allows us to separate the effect of **LLM-enhanced text embeddings** from that of **offline manifold augmentation**. Replacing only the textual feature file with LLM-enhanced embeddings does not produce consistent gains across models and datasets. Although **GUME** benefits slightly in some settings, the improvements are not stable across backbones; for **MGCN** and **MENTOR**, the LLM-text replacement is often marginal or even negative. This shows that better text-derived inputs alone are insufficient to resolve heterophily.

By contrast, **w/ Offline Aug** delivers clearer and more consistent improvements, especially on *Sports* and *Clothing*. This suggests that the main gain does not come from text refinement itself, but from explicitly reshaping the feature geometry so that semantically distant yet functionally complementary items become easier to connect during downstream graph propagation.

This observation is also consistent with the control design in Table 7: feature replacement alone does not provide the same level of improvement as geometry-aware augmentation combined with graph propagation. What matters is the combination of complementary-aware geometric correction and graph-based propagation, rather than stronger text preprocessing by itself. We leave adaptive,

noise-aware injection that automatically calibrates augmentation strength under extreme sparsity as future work.

Table 7: Generality analysis of offline manifold augmentation and LLM-enhanced text embeddings on existing multimodal recommenders. **Original** denotes the original backbone; **w/ LLM Text Emb** replaces only the textual feature file with LLM-enhanced embeddings; **w/ Offline Aug** replaces the original item embeddings with our offline manifold-augmented ones, while keeping all other settings unchanged.

Model	Setting	Baby				Sports				Clothing			
		R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
GUME	Original	0.0673	**0.1042**	0.0365	**0.0460**	0.0778	0.1165	0.0427	0.0527	0.0703	0.1024	0.0384	0.0466
	w/ LLM Text Emb	0.0671	0.1015	0.0362	0.0451	0.0787	0.1176	**0.0433**	**0.0537**	0.0707	0.1033	0.0385	0.0464
	w/ Offline Aug	**0.0674**	0.1034	**0.0366**	0.0458	**0.0796**	**0.1197**	**0.0433**	0.0536	**0.0712**	**0.1046**	**0.0387**	**0.0471**
MGCN	Original	0.0620	**0.0964**	0.0339	**0.0427**	0.0729	0.1106	0.0397	0.0496	0.0641	0.0945	0.0347	0.0428
	w/ LLM Text Emb	0.0611	0.0947	0.0334	0.0419	0.0722	0.1075	0.0396	0.0486	0.0624	0.0921	0.0335	0.0410
	w/ Offline Aug	**0.0621**	0.0956	**0.0340**	0.0423	**0.0764**	**0.1153**	**0.0417**	**0.0518**	**0.0700**	**0.1018**	**0.0377**	**0.0458**
MENTOR	Original	**0.0678**	**0.1048**	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0668	0.0989	0.0360	0.0441
	w/ LLM Text Emb	0.0653	0.1029	**0.0363**	**0.0460**	0.0756	0.1137	0.0409	0.0504	0.0685	0.1013	0.0372	0.0456
	w/ Offline Aug	0.0671	0.1030	0.0361	0.0450	**0.0787**	**0.1160**	**0.0426**	**0.0522**	**0.0711**	**0.1044**	**0.0387**	**0.0471**

D.3 Fine-Grained Phase-I Ablation

These controls use the default configuration, matched seed settings, and validation-selected checkpoints. They are reported descriptively; no separate significance claim is made for individual safeguards.

Table 8: Complete fine-grained Phase-I controls.

Dataset	Variant	R@10	R@20	N@10	N@20
Clothing	Complete	0.0738	0.1088	0.0396	0.0484
	Mean residual	0.0729	0.1074	0.0391	0.0478
	Hard SVD	0.0727	0.1072	0.0390	0.0477
	Without orthogonalization	0.0735	0.1084	0.0395	0.0482
	Without gap constraint	0.0724	0.1067	0.0388	0.0475
	Without backtracking	0.0732	0.1080	0.0393	0.0480
Sports	Complete	0.0821	0.1221	0.0440	0.0544
	Mean residual	0.0815	0.1211	0.0437	0.0540
	Hard SVD	0.0811	0.1206	0.0435	0.0537
	Without orthogonalization	0.0819	0.1218	0.0439	0.0543
	Without gap constraint	0.0809	0.1204	0.0434	0.0536
	Without backtracking	0.0813	0.1209	0.0436	0.0539

D.4 Original-Text Dependency Control

Table 9: Relative change (%) when MAGE uses original non-LLM text. Sports uses longer validation patience and is not a strictly matched comparison.

Dataset	R@10	R@20	N@10	N@20
Baby	-0.73	-0.45	-1.07	-0.86
Sports	$+1.63$	$+0.25$	$+1.85$	$+0.86$
Clothing	-1.23	-1.09	-1.69	-1.03

The same Phase-I transformation and Phase-II architecture are applied to original text embeddings. Checkpoints remain validation-selected. This control tests dependence on extracted text; it does not establish equivalence between preprocessing choices or isolate convergence effects on Sports.

E Complementarity and Long-Tail Diagnostics

E.1 Hard Complement Subset Evaluation

Although MAGE is designed for general user–item recommendation rather than pure complementary-item retrieval, its central motivation is to better exploit complementary signals when they are behaviorally useful but semantically distant. To directly examine this property, we construct a **hard complement subset** from the test set.

For each test positive pair (u, j), we first collect the user's training-history items $\mathcal{H}_u$ and require that at least one item $i \in \mathcal{H}_u$ is connected to j through a co-purchase relation. The co-purchase graph is built from the union of *also_bought* and *bought_together* relations and treated as undirected. Among all co-purchased history anchors, we select the most semantically distant one according to the average of textual and visual cosine similarities:

$$i^* = \arg \min_{i \in \mathcal{H}_u \cap \mathcal{N}_{\mathrm{co}}(j)} \frac{1}{2} \left(\mathrm{sim}_t(i, j) + \mathrm{sim}_v(i, j) \right). \tag{38}$$

We then retain only those eligible test positives whose textual and visual similarities to i^* both fall into the bottom 30% of the eligible pool. This subset therefore focuses on cases that are behaviorally connected through co-purchase but semantically distant in both modalities. Because it uses the relation type that motivates MAGE, it is a targeted diagnostic rather than a distribution-independent benchmark; the primary evaluation remains the original full-ranking test set.

The corresponding results are reported in the main text (Table 2). The subset is defined for diagnostic evaluation; held-out user–item labels are not used to construct the metadata co-purchase graph.

E.2 Qualitative Analysis: Visualizing the Functional Manifold

To qualitatively examine whether MAGE bridges the semantic gap in complementary recommendation, we visualize the item embedding space of the *Sports* dataset using t-SNE (Maaten and Hinton, 2008).

Setup. We select a *Kershaw T-Tool* as the target item and consider two types of related items: **functional complements** (co-purchased but semantically dissimilar, e.g., folding knives) and **hard negatives** (textually similar but functionally unrelated, e.g., screwdrivers). Figure 3 compares the embedding distributions produced by GUME and MAGE.

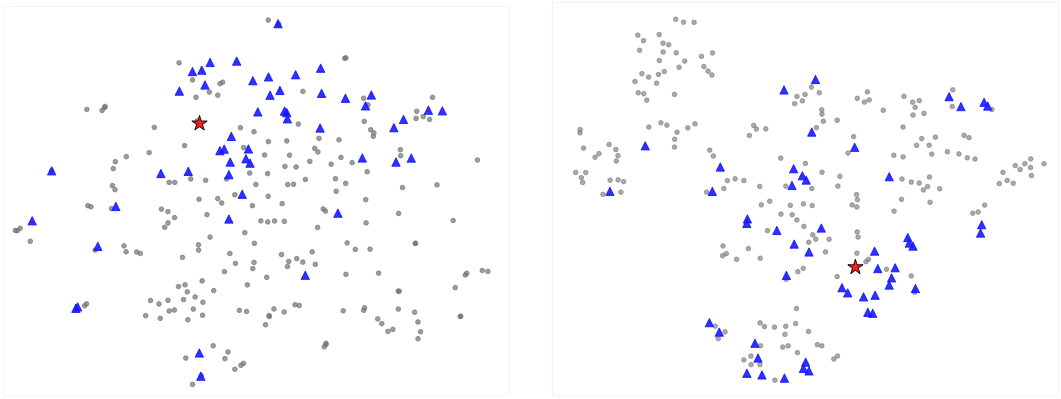
Observations. In the selected visualization in Figure 3, the co-purchased items labeled as functional complements are more dispersed around the target under GUME. Under MAGE, several of these items appear closer to the target, while the selected semantic hard negatives lie farther away. This qualitative pattern is consistent with the intended local reorganization of item representations. However, a selected two-dimensional projection does not independently verify functional compatibility or establish the same behavior across the full item population.

E.2.1 Additional Geometric Case

A supplementary case was selected using a fixed 75th-percentile threshold on Phase-I geometric gain, without inspecting product titles or test-ranking outcomes. For "ILBE Main Pack USMC Generation 2 With Assault Pack," 9 of 10 observed top-10 co-purchase neighbors become closer after Phase I. Their median text-neighborhood rank improves from 3,872 to 1,441, and mean cosine similarity increases from 0.3442 to 0.4662. Selection on geometric gain makes this an illustration of the intended mechanism rather than an unbiased estimate of average improvement. The Kershaw T-Tool visualization in Appendix E.2 provides a complementary qualitative example.

E.3 Impact on Long-Tail Items

To examine MAGE's performance across item popularity levels, we performed a stratified evaluation on the Clothing dataset, following prior work (Lin et al., 2024). We partitioned items into five groups based on interaction frequency, where Group 1 represents the top 20% (Head) and Groups 2–5 represent the long-tail (Tail), following the Pareto Principle (Luo et al., 2023). Figure 4 reports the performance of MAGE compared with the strongest baselines, GUME and MENTOR.



(a) Baseline (GUME) (b) Ours (MAGE)

Figure 3: t-SNE visualization of item embeddings. **Red star**: target item (*T-Tool*); **blue triangles**: functional complements (*knives*); **gray dots**: semantic hard negatives (*other tools*). Compared with the scattered distribution in GUME, MAGE aggregates functional complements into a more cohesive local neighborhood around the target item.

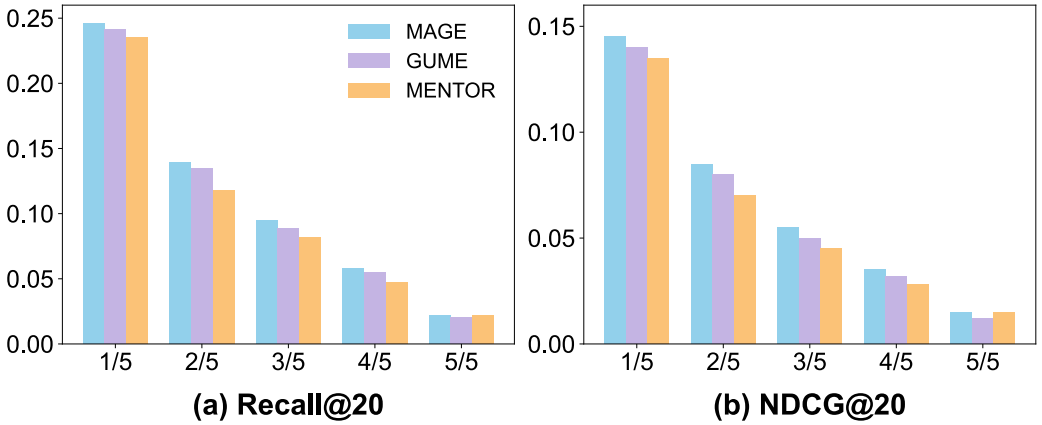


Figure 4: Performance comparison across item popularity groups on the Clothing dataset. Group 1 denotes Head items, while Groups 2–5 correspond to progressively sparser Tail items.

The results show favorable ranking performance for MAGE across the displayed popularity groups, with gains in several tail groups as well as the Head group. These observations indicate that the measured benefits are not confined to popular items. They are consistent with the motivation to use co-purchase information for less frequently observed items, but do not isolate the mechanism responsible for the gains. In particular, group-level ranking metrics do not establish that popular-item representations or their relative ordering remain unchanged.

E.3.1 Matched Popularity-Stratified Diagnostic

Head and tail groups are fixed using training interaction frequency, without semantic-distance filtering or changes to the full-ranking candidate set. Table 10 reports the available tail Recall measurements from this diagnostic, separately from the hard-complement subset.

F Sensitivity, Robustness, and Efficiency

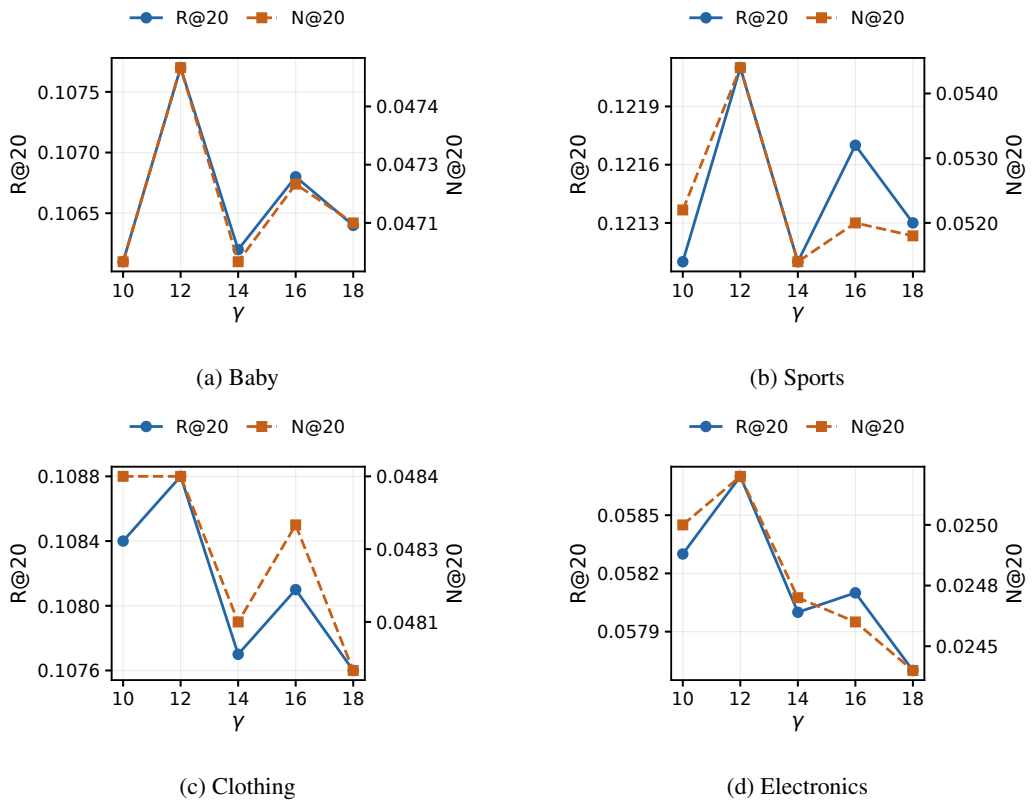
F.1 Hyper-Parameter Sensitivity Analysis

We further examine the sensitivity of MAGE with respect to four groups of hyper-parameters: (*i*) the Soft Conceptor aperture γ, which controls the strength of substitute-subspace filtering in Phase I; (*ii*)

Table 10: Tail Recall under the matched popularity-stratified evaluation.

Dataset	Method	R@10	R@20
Sports	GUME	0.0129	0.0282
	MAGE	0.0152	0.0320
Clothing	GUME	0.0161	0.0322
	MAGE	0.0210	0.0430

the Phase-I keep ratio ρ, which determines how much local structural separation is preserved during the augmentation process; (*iii*) the auxiliary loss weights used in online optimization, namely the Info Loss and Align Loss coefficients; and (*iv*) the complementary neighborhood size k, which controls topological sparsity in Phase II.



(a) Baby

(b) Sports

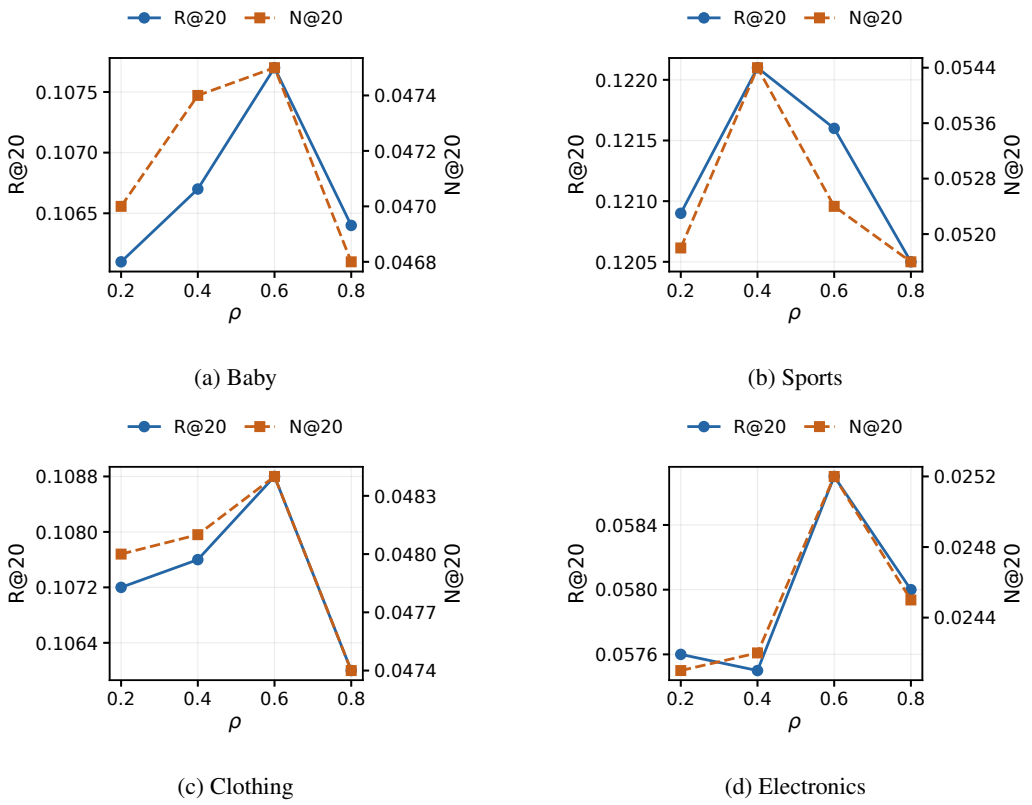
(c) Clothing

(d) Electronics

Figure 5: **Sensitivity to the Soft Conceptor aperture γ.** The best or near-best results occur in an intermediate regime.

Sensitivity to the Soft Conceptor aperture γ. Figure 5 shows the sensitivity of MAGE to the Soft Conceptor aperture γ on all four datasets. A consistent pattern emerges across Baby, Sports, Clothing, and Electronics: performance is strongest in an intermediate regime, with $\gamma = 12$ giving the best or near-best results in all cases. When γ is too small, the Soft Conceptor filtering is too weak to sufficiently suppress the substitute-dominated subspace. When γ becomes too large, the filtering becomes overly aggressive and may remove useful semantic information, leading to mild but consistent degradation. This indicates that Phase I benefits from *moderate subspace filtering* rather than aggressive geometric perturbation.

Sensitivity to the keep ratio ρ. Figure 6 reports the effect of the keep ratio ρ, which controls how much of the local structural gap is preserved during offline augmentation. In the displayed sweeps, both metrics peak at $\rho = 0.4$ on Sports and at $\rho = 0.6$ on Baby, Clothing, and Electronics. A larger keep ratio requires stricter gap preservation and tightens the initial injection bound when the



(a) Baby

(b) Sports

(c) Clothing

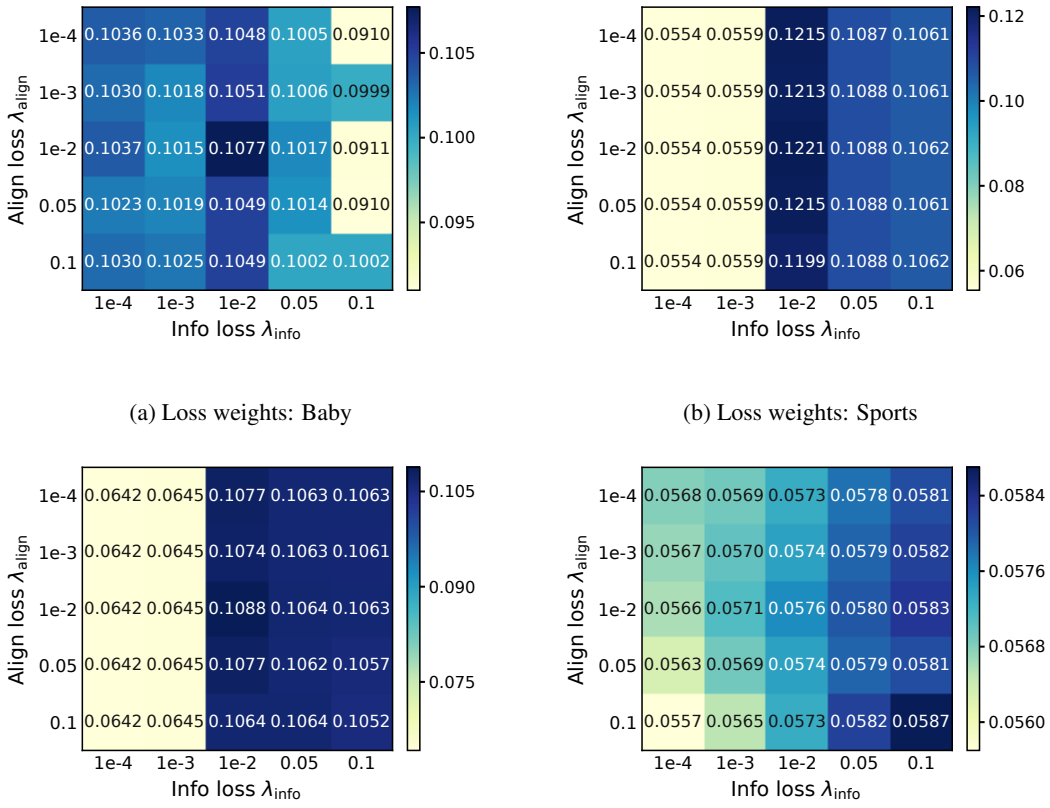
(d) Electronics

Figure 6: **Sensitivity to the Phase-I keep ratio ρ.** A larger keep ratio requires stricter local gap preservation.

original gap is positive and the residual would reduce it; a smaller ratio relaxes this bound. However, changing ρ need not change the coefficient while the 0.35 safety cap is active, and the normalized-gap check may further reduce the update. The curves therefore show dataset-dependent performance under this constraint, rather than a monotonic relationship between ρ, injection magnitude, and ranking quality.

Sensitivity to auxiliary loss weights. Figures 7a–7d report the interaction between the Info Loss and Align Loss coefficients on all four datasets. Here, the Info Loss coefficient λ_{info} weights $\mathcal{L}_{\text{UM}}$ and corresponds to `um_loss` in the implementation. The Align Loss coefficient λ_{align} weights $\mathcal{L}_{\text{align}}$ and corresponds to `vt_loss`. The temperature parameters are held at the dataset-specific values stated in the caption throughout these sweeps. The best displayed results on Baby, Sports, and Clothing occur at $\lambda_{\text{align}} = \lambda_{\text{info}} = 10^{-2}$. Electronics differs: its highest displayed value occurs at $\lambda_{\text{align}} = \lambda_{\text{info}} = 10^{-1}$, the upper boundary of the tested range. The preferred regularization strength is therefore dataset-dependent; the sweeps do not support a universal claim that larger coefficients reduce performance. The boundary optimum on Electronics also leaves behavior beyond the tested range unresolved. These observations motivate selecting the coefficients per dataset rather than imposing a single optimum across all four datasets.

Sensitivity to topological sparsity in Phase II. Figure 8 reports the effect of the complementary neighborhood size k. Despite differences in sparsity across datasets, the same qualitative trend is observed: performance peaks at a relatively small neighborhood and then declines as k increases. In particular, $k = 10$ provides the most stable trade-off and remains consistently near-optimal. This result highlights a key distinction between homophilic and heterophilic propagation: while similarity graphs may benefit from denser smoothing, complementary graphs are much more noise-sensitive and therefore require sparse, precise propagation.



(a) Loss weights: Baby

(b) Loss weights: Sports

(c) Loss weights: Clothing

(d) Loss weights: Electronics

Figure 7: **Sensitivity to online optimization.** Joint sensitivity of the Info Loss (λ_{info}, UM) and Align Loss (λ_{align}, VT) coefficients. Cell values and colors represent Recall@20. We fix $\tau_{bm} = 0.2$ on all datasets and $\tau_{um} = 0.1$ on Baby/Sports and 0.2 on Clothing/Electronics.

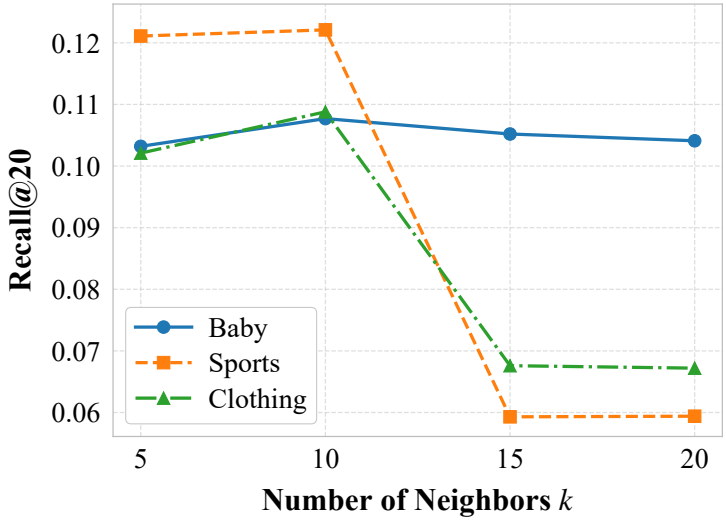


Figure 8: **Sensitivity to topological sparsity in Phase II.** The complementary neighborhood size k controls the sparsity of heterophilic propagation, and a relatively small neighborhood yields the most stable performance.

Taken together, these sweeps characterize the sensitivity of geometric augmentation, auxiliary regularization, and neighborhood size within the tested ranges. They motivate dataset-specific choices of augmentation and regularization strengths, together with explicit gap control and neighborhood truncation. The preferred settings depend on both the parameter and the dataset, as illustrated by the different loss-weight optima on Electronics. These results do not establish a single optimal configuration across datasets or beyond the evaluated ranges.

F.2 Robustness to Co-purchase Graph Corruption

With perturbation seed 999, we independently drop or rewire 30% of eligible co-purchase edges; rewiring preserves source degree and total edge count. For each graph, we recompute Phase-I partitions and residuals, reconstruct complementary topology, and retrain from scratch with validation-only checkpoint selection. The interaction graph, similarity graph, and full-ranking evaluation remain fixed. Table 11 shows that relative metric changes remain within $\pm 1.96\%$. This supports stability under the tested corruptions, not arbitrary noise or extreme sparsity.

Table 11: Relative change (%) from the matched unperturbed configuration under 30% graph corruption.

Dataset	Corruption	R@10	R@20	N@10	N@20
Sports	Dropout	-0.25	$+0.75$	$+1.72$	$+1.96$
Sports	Target rewiring	$+0.50$	$+1.25$	$+1.67$	$+1.51$
Clothing	Dropout	-0.91	-0.83	-0.76	-0.41
Clothing	Target rewiring	-1.07	-0.79	-1.27	-1.03

F.3 Complexity and Offline Efficiency

Manifold-based methods are often criticized for poor scalability. MAGE remains efficient because both its geometric augmentation and graph propagation are localized.

Theoretical complexity. In Phase I, offline manifold augmentation is performed independently for each item. For item i, SVD is applied only to a local matrix of size $(|\mathcal{S}_i| + 1) \times d$, where $|\mathcal{S}_i|$ is bounded by a small neighborhood size and d is fixed. The dominant per-item cost comes from local similarity computation, thin-SVD, and weak-neighbor projection:

$$O(\bar{k}d + r^2 d + |\mathcal{W}_i| r d), \tag{39}$$

where $\bar{k}$ is the average co-purchase neighborhood size and $r \leq |\mathcal{S}_i| + 1$ is the local SVD rank. Since these quantities are locally bounded in practice, the overall offline cost scales approximately linearly with the number of items N. This phase is executed once as preprocessing and introduces no overhead during training or inference.

CPU-only benchmark. We further benchmark the Phase-I preprocessing on a CPU-only Apple Silicon machine using the three Amazon datasets. Phase-I preprocessing from existing embeddings, including its I/O but excluding LLM extraction and embedding generation, takes **1.89s**, **4.39s**, and **4.36s** on Baby, Sports, and Clothing, respectively, with peak memory below **0.4GB**. The median per-seed latency is only **0.16–0.20 ms**. A scalability test that clones the Baby graph from $1\times$ to $10\times$ shows near-linear growth, with per-seed latency increasing only from 0.170 ms to 0.205 ms. These results indicate that Phase I is a one-shot, CPU-friendly preprocessing module, rather than a computational bottleneck.

G Prompt Design for Evidence-Bound Text Preprocessing

To improve reproducibility, we provide the main prompt components used in our evidence-bound preprocessing pipeline. Consistent with the discussion in Section 4.1, the LLM is used only as a lightweight preprocessing tool to extract explicit functional attributes, normalize heterogeneous expressions, and remove marketing-heavy noise. It is not used to freely rewrite or semantically enrich the original text.

G.1 System Instruction

Prompt 1: System Instruction

```
You are an evidence-bound information extractor and normalizer.

HARD RULES:
Only extract facts directly supported by verbatim snippets (<=12 words) from TITLE,
     DESCRIPTION, BRAND, or CATEGORIES; if no evidence, leave the field null or []. Do
     not invent or imply facts.

Use canonical product types {water_bottle, hydration_pack, backpack, sports_bag,
     yoga_mat, resistance_band, jump_rope, fitness_equipment, strength_training,
     recovery_tool, ball, goal_net, protective_gear, helmet, sport_eyewear, swim_gear,
     racquet_sport, golf_accessory, fishing_gear, cycling_accessory, wheel_sports,
     camping_gear, trekking_pole, lighting_outdoor, cooler, sports_towel, sports_socks,
     sportswear, dance_apparel, other}; assign based on clear textual cues, otherwise
     default to "other".

For potentially regulated items (e.g., "knife", "tactical knife", "spring assist", "
     pepper spray", "stun gun", "ammunition"):
- set product_type to "other"
- add a violation tag like "restricted_category: knife"
- keep copy strictly neutral (no marketing or purchase encouragement)

List fields must always be arrays. Do not assign age_range unless an exact numeric or
     stage is stated. Normalize units (oz -> ml; inch -> cm) when explicit; if a backpack
     /pack capacity in liters is explicit (e.g., "20L"), map to capacity_l (do not infer)
     .

De-duplicate short values. Structured text must only use the extracted assertions, avoid
      marketing superlatives unless explicitly present, and maintain grammatical number
     consistency. Keep neutral tone; strictly neutral for regulated items.

STYLE:
- Keep outputs concise and concrete. No fluff.
- Avoid unverifiable attributes unless verbatim.
- Output VALID JSON only (no code fences).
```

G.2 Output Schema Constraints

Prompt 2: Output Schema Constraints

```
Output JSON with this schema:
{
  "normalized": {
    "product_type": "string|null",
    "materials": ["string", ...],
    "capacity_ml": number|null,
    "capacity_l": number|null,
    "age_range": "string|null",
    "compatibility": ["string", ...],
    "safety_claims": ["string", ...],
    "made_in": "string|null",
    "intended_use": ["string", ...]
  },
  "key_points": ["bullet 1", "bullet 2", "..."],
  "assertions": [
    {
      "claim": "string",
      "field": "product_type|materials|capacity_ml|capacity_l|age_range|compatibility|
     safety_claims|made_in|intended_use|other",
```

```
      "evidence": "verbatim snippet from input"
    }
  ],
  "structured_text": "string",
  "evidence": {
    "product_type": ["..."],
    "materials": ["..."],
    "capacity_ml": ["..."],
    "capacity_l": ["..."],
    "age_range": ["..."],
    "compatibility": ["..."],
    "safety_claims": ["..."],
    "made_in": ["..."],
    "intended_use": ["..."]
  },
  "rejected_claims": ["string", "..."],
  "violations": ["string", "..."]
}

CONSTRAINTS:
- If a field has no evidence, keep it null/[] and do NOT surface it in ‘assertions‘ or ‘
    structured_text‘.
- Each assertion must quote its evidence verbatim (short phrases OK).
- If there is any uncertainty, prefer to omit and record it in ‘rejected_claims‘.
- For regulated items (e.g., knives), do not include encouraging language; keep it purely
     descriptive using only evidence.
```

Discussion. The two prompt components jointly enforce three properties required by our preprocessing pipeline: evidence grounding, schema-level normalization, and neutral noise filtering. In this way, the LLM serves only as a lightweight extractor/normalizer before feature encoding, rather than as a generator for open-ended text enhancement.

H Licenses and Terms of Use of Existing Assets

Our experiments use several existing public assets, including benchmark datasets, open-source codebases, and pre-trained models.

Datasets. We use four public Amazon benchmark datasets (*Baby*, *Sports*, *Clothing*, and *Electronics*) obtained from the public collection released by McAuley and collaborators.[2] We use these datasets only for academic research and follow their publicly stated access conditions and terms of use.

Code framework. Our implementation is built on the MMRec framework (Zhou, 2023). We use the public implementation released by its authors and follow the license and usage terms provided in the corresponding repository.

Pre-trained models. For evidence-bound preprocessing, we use Qwen3-32B-AWQ (Yang et al., 2025). This model is used only as a feature extractor or preprocessing tool, and we follow the corresponding license and terms of use provided by its official release.

License compliance. We credit the original creators of all existing assets used in this work and use them in accordance with their licenses and stated research-use conditions. We do not claim ownership of these external assets.

[2] http://jmcauley.ucsd.edu/data/amazon/links.html

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main motivation, methodological contributions, and empirical scope of the paper. These claims are consistent with the reported experimental results, including the overall performance comparison and the supporting analyses in the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion, including the dependence on co-purchase relation quality, possible degradation under noisy or extremely sparse co-purchase graphs, and the additional preprocessing/graph-construction costs introduced by the two-stage design.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical theorems or proofs. Our contributions are primarily methodological and empirical, focusing on architecture design, geometric augmentation, and experimental validation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the model architecture, optimization objective, datasets, feature preparation pipeline, evaluation protocol, and ablation settings in the main paper, with additional implementation details deferred to the appendix. We also provide a public code repository at `https://github.com/Alex-bupt/mage`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public code repository is available at `https://github.com/Alex-bupt/mage`, and the datasets used in the paper are publicly available benchmark datasets. The supplemental material and appendix describe the experimental setup and implementation details needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, feature preparation procedure, baseline methods, optimization objective, and evaluation metrics in the core text, while detailed implementation settings and hyper-parameter configurations are provided in the appendix. Together, these details are sufficient to understand and interpret the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report statistical significance information for the main experimental claims using repeated runs and paired significance testing. Specifically, the paper includes significance analysis over multiple random seeds for representative comparisons, and the corresponding testing protocol is described in the experimental discussion.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the main compute environment used for experiments, including the GPU type and memory configuration, in the appendix implementation details. These resources are sufficient to understand the scale of the reported experiments and to reproduce the setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research described in this paper conforms to the NeurIPS Code of Ethics. The work uses public benchmark datasets and does not involve human subjects or sensitive private data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential benefits and risks of improved complementary recommendation. On the positive side, better multimodal recommendation may improve retrieval quality and user experience; on the negative side, such systems may amplify existing behavioral biases, create filter effects, or reinforce commercially skewed exposure if deployed without proper monitoring and safeguards.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models, scraped sensitive datasets, or other assets that would require special misuse safeguards beyond standard code release and citation of public benchmark data sources.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the public datasets, code framework, and pre-trained model used in this work, and we explicitly state their sources and licenses or terms of use in the appendix. The external model is used only for preprocessing or feature extraction under its official usage conditions.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, model family, or benchmark as a standalone released asset. The main contribution is a new recommendation method evaluated on existing public datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing public benchmark datasets.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or any direct research with human subjects, and therefore IRB approval or an equivalent review process was not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper uses an LLM-based evidence-bound extraction step as a non-standard preprocessing component of the method. Its role, scope, and prompt design are described in the methodology and appendix, where we clarify that the LLM is used to extract and normalize structured functional information rather than to generate the final recommendation outputs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.