BRIDGE: Behavior-Guided Residual Integration with Dual-Frequency Graph Evidence

Zesheng Li
University of the Chinese Academy of Sciences
Beijing, China
2022111515@stu.sufe.edu.cn

Chengchang Pan*
University of the Chinese Academy of Sciences
Beijing, China
166353314@qq.com

Honggang Qi*
University of the Chinese Academy of Sciences
Beijing, China
hgqi@ucas.ac.cn

Abstract

Multimodal recommendation improves item representations by combining visual, textual, and collaborative signals, but stronger cross-view alignment does not always improve ranking. Our diagnostics on Amazon Baby show that direct consistency has an effective range: moderate alignment helps, while stronger alignment suppresses recommendation-specific variation. We also observe a clear spectral split: low-frequency components capture shared structure, whereas higher-frequency components retain more private ranking signal. Based on these findings, we propose BRIDGE, a behavior-guided residual integration framework built on dual-frequency graph evidence. BRIDGE separates the model into three parts: DFGE decomposes graph-smoothed ID, visual, and textual channels into spectral bands; BEN converts training-only co-user overlap into signed candidate evidence; and CRI applies that evidence only inside the base top-K candidate set during training and inference. This design keeps the multimodal backbone and localizes behavior evidence to candidate calibration. Experiments on Amazon Baby, Sports, and Electronics show that BRIDGE reaches 0.1128/0.1262/0.0778 Recall@20 and 0.0525/0.0594/0.0385 NDCG@20, outperforming baselines by up to 7.3% in Recall@20 and 14.9% in NDCG@20. Project materials are available at https://lizesheng13.github.io/bridge/.

CCS Concepts

• **Information systems** → **Recommender systems**; *Multimedia and multimodal retrieval.*

Keywords

Multimodal Recommendation, Dual-Frequency Graph Evidence, Behavior Graph, Candidate Calibration

ACM Reference Format:
Zesheng Li, Chengchang Pan, and Honggang Qi. 2026. BRIDGE: Behavior-Guided Residual Integration with Dual-Frequency Graph Evidence. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799682.3840961

*Corresponding authors.

CIKM '26, November 7–11, 2026, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840961

1 Introduction

Images, titles, descriptions, and brands provide useful side information for recommendation, especially when interaction data are sparse. This has motivated a broad family of multimodal recommenders that combine user-item interactions with visual and textual item features [3, 26–28]. Recent surveys synthesize this literature from representation, modeling, and optimization perspectives [17]. Recent work has improved the representation layer by building item-item graphs from content, denoising multimodal features, or enforcing stronger consistency across views [5, 10, 25, 33, 40]. The common premise is simple: a better multimodal representation should yield a better ranker.

Our diagnostics point to a narrower role for content evidence. Moderate cross-view consistency improves ranking, whereas stronger consistency suppresses recommendation-specific variation. Frequency analysis shows a shared-vs-private split: low-frequency bands agree across views, while higher-frequency bands retain more ranking signal. This distinction is most important inside the small candidate set where ranking is decided. Over-aligning the views blurs the differences between plausible items and near neighbors.

BRIDGE turns these observations into a behavior-guided design. The multimodal encoder defines the base candidate geometry, and co-user behavior authorizes only local residual corrections inside that candidate set. DFGE preserves frequency-aware multimodal evidence, BEN converts training-only co-user overlap into signed behavior support, and CRI applies that support only within the candidate scope. On Amazon Baby, Sports, and Electronics, BRIDGE reaches 0.1128, 0.1262, and 0.0778 Recall@20, respectively, with corresponding NDCG@20 values of 0.0525, 0.0594, and 0.0385.

The deployment setting adds a second constraint. In a retrieval-and-rank pipeline, the shortlist is the decision bottleneck. A content signal that helps representation learning can hurt ranking when it moves every catalog score. BRIDGE keeps the multimodal backbone responsible for the shortlist and lets behavior evidence revise only candidates that the backbone has already made plausible.

The contributions of this work are:

- We identify two diagnostics in multimodal recommendation: direct cross-view consistency has an effective range, and the representation spectrum separates shared low-frequency evidence from more private higher-frequency variation.
- We propose DFGE, a dual-frequency graph evidence encoder that preserves that spectrum with band routing and structural regularization.
- We propose BEN and CRI, which turn training-only co-user overlap into signed behavior evidence and use it only to calibrate the candidate scope used at training and inference.

2 Related Work

2.1 Multimodal Recommendation and Cross-View Alignment

Multimodal recommendation extends collaborative filtering with visual and textual signals. VBPR [3] shows that content can alleviate sparsity, while graph-based methods such as MMGCN [28], GRCN [27], DualGNN [26], and LATTICE [36] organize these signals with user-item and item-item structure. Recent variants add denoising, diffusion, robustness, dual-path routing, retrieval completion, or stronger cross-view consistency [1, 2, 5, 7–12, 19, 20, 25, 30, 37, 39, 40]. Related alignment and fusion work also explores dual-graph attention, hierarchical multi-step fusion, and multimodal large language model coupling [22, 32, 35, 41, 42]. The common theme is to improve representation quality before ranking. BRIDGE differs by asking when content evidence should be trusted by the ranker rather than how aggressively it should be fused.

Most alignment methods use cross-view agreement as a proxy for better ranking. When content matching dominates, it can erase the differences between a plausible item and a near duplicate or popular lookalike. BRIDGE uses the same multimodal evidence to define candidate geometry, then decides whether behavior evidence should change the rank of each candidate. This decision is part of the model rather than a post-hoc reranking rule.

2.2 Frequency-Aware and Behavior-Guided Recommendation

Frequency-aware recommenders separate stable structure from fine-grained variation. SMORE [16], FITMM [33], SSR [34], MSCF-Net [6], DuGRec [13], D-DPDG [29], and Sequences-as-Nodes [22] show that spectral views can organize multimodal or collaborative evidence. These works differ in whether the spectrum is used for denoising, co-frequency enhancement, or explicit band routing. BRIDGE uses the same general intuition but keeps the band decomposition as a representation core rather than a standalone predictor.

That makes BRIDGE different from frequency-aware denoising pipelines. Those models usually use the spectrum to suppress noise or separate stable from unstable components, while BRIDGE uses the decomposition to decide which bands remain shared evidence and which bands are left as residual variation for candidate calibration. The band split is therefore an authorization step, not only a regularizer.

The same point applies to band routing. A decomposition can denoise a representation, but denoising alone does not explain how the output should affect ranking. BRIDGE keeps the spectrum inside the backbone and lets candidate calibration consume only the bands that remain useful after decomposition. This keeps the representation and the calibration rule aligned with one another.

2.3 Behavior-Guided and Candidate-Side Control

Item-neighborhood methods rank with interaction-derived item scores, while linear item–item models learn a global coefficient matrix from the interaction data [23, 24]. These approaches are close to the behavior branch in BRIDGE, but they are normally used as standalone catalog-wide rankers. BRIDGE instead treats raw co-occurrence and EASE as candidate-side residual controls and applies signed behavior evidence only after the multimodal backbone has selected a top-K candidate set.

In parallel, behavior-guided methods such as BPR [21], LightGCN [4], DRAGON [38], FGCM [15], JBM-Diff [18], and causal-inspired multimodal recommendation [31] use interaction structure to validate or filter content-derived relations. More recent work extends this idea to dual-graph diffusion, modality coherence, and multi-intention or sequential settings [2, 12, 32, 35, 42]. BRIDGE combines these lines, but uses behavior only as a local authorization signal for candidate-level residual calibration rather than as a global fusion or denoising target.

Unlike a post-hoc re-ranker or a catalog-wide item–item score, BRIDGE trains and evaluates its residual under the same detached top-K scope. The base scorer remains responsible for retrieving new positives, while behavior evidence changes only the local ordering of plausible candidates.

Together, these strands motivate BRIDGE's dual-frequency and candidate-calibrated design.

3 Problem Formulation

Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item sets. We observe implicit feedback $\mathcal{R} = \{(u, i) \mid u \in \mathcal{U}, i \in \mathcal{I}\}$, where an edge indicates that user u interacted with item i. Each item also has multimodal content features, including visual features $\mathbf{x}_i^v$ and textual features $\mathbf{x}_i^t$. For a user u, let $\mathcal{H}_u$ be the interacted item set; for an item i, let $\mathcal{U}_i$ be the set of users who interacted with it. The recommendation task is to learn a score $s(u, i)$ and rank unobserved items for each user under full-sort evaluation.

The standard training objective is pairwise Bayesian Personalized Ranking (BPR) [21]:

$$\mathcal{L}_{\text{BPR}}(s) = -\sum_{(u,i,j)} \log \sigma(s(u, i) - s(u, j)), \tag{1}$$

where i is an observed positive item and j is a sampled negative item. Existing multimodal recommenders usually define $s(u, i)$ by making collaborative, visual, and textual embeddings consistent or fusing them. In this work, we make the scoring problem more explicit: content-derived signals are allowed to affect ranking only when supported by behavior evidence.

BRIDGE learns a final score of the form

$$s(u, i) = s_{\text{base}}(u, i) + \Delta_{\text{bridge}}(u, i), \tag{2}$$

where the residual correction is behavior-guided and applied only in a controlled candidate scope. The concrete definitions of behavior evidence and the frequency-aware encoder are given in Method.

4 Motivating Observations

BRIDGE is motivated by three empirical observations. Together they suggest a simple design principle: multimodal information should be preserved, but content should be treated as conditional evidence, not as a command.

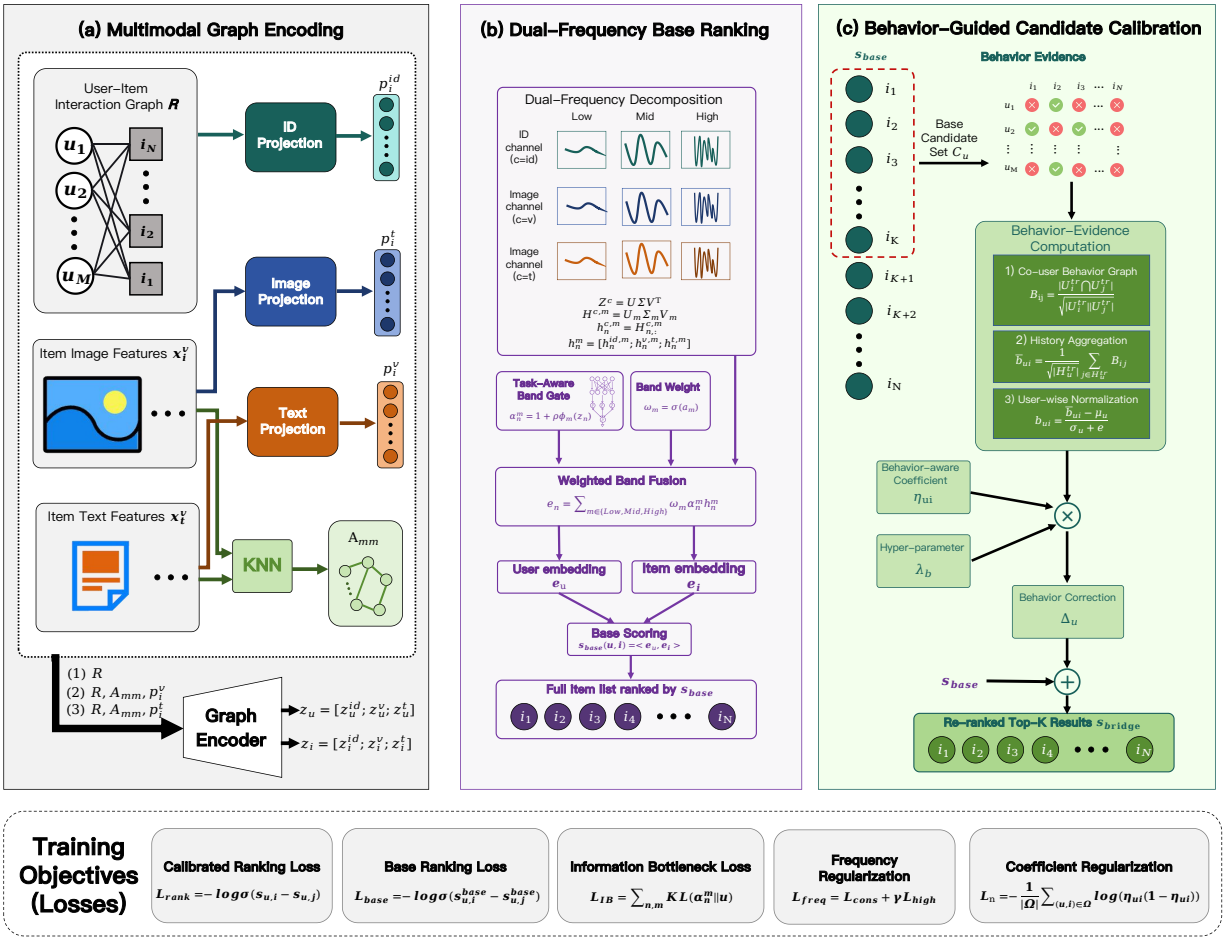

Figure 1: Overview of BRIDGE. DFGE constructs dual-frequency graph evidence; BEN computes normalized behavior support; CRI integrates the residual only inside the selected candidate set.

Table 1: Overall recommendation performance on Amazon Baby, Sports, and Electronics. All methods use the same processed splits, BEIT3 features, and full-sort evaluator. Entries report mean ± standard deviation over five seeds. ** indicates $p < 0.01$ against the strongest baseline under an independent two-sample t-test.

Dataset	Metric	BPR	LightGCN	VBPR	MMGCN	GRCN	DualGNN	SLMRec	LATTICE	BM3	FREEDOM	DiffMM	MMIL	AlignRec	SMORE	FITMM	BRIDGE
Baby	R@10	0.0357	0.0479	0.0418	0.0413	0.0538	0.0507	0.0533	0.0561	0.0573	0.0624	0.0617	0.0670	$0.0674_{\pm 0.0007}$	$0.0680_{\pm 0.0008}$	$0.0716_{\pm 0.0006}$	$\mathbf{0.0752}^{**}_{\pm 0.0004}$
	R@20	0.0575	0.0754	0.0667	0.0649	0.0832	0.0782	0.0788	0.0867	0.0904	0.0985	0.0978	0.1035	$0.1046_{\pm 0.0010}$	$0.1035_{\pm 0.0011}$	$0.1089_{\pm 0.0009}$	$\mathbf{0.1128}^{**}_{\pm 0.0006}$
	N@10	0.0192	0.0257	0.0223	0.0211	0.0285	0.0264	0.0291	0.0305	0.0311	0.0324	0.0321	0.0361	$0.0363_{\pm 0.0004}$	$0.0365_{\pm 0.0005}$	$0.0387_{\pm 0.0005}$	$\mathbf{0.0428}^{**}_{\pm 0.0003}$
	N@20	0.0249	0.0328	0.0287	0.0275	0.0364	0.0335	0.0365	0.0383	0.0395	0.0416	0.0408	0.0455	$0.0458_{\pm 0.0005}$	$0.0457_{\pm 0.0006}$	$0.0484_{\pm 0.0005}$	$\mathbf{0.0525}^{**}_{\pm 0.0003}$
Sports	R@10	0.0432	0.0569	0.0561	0.0394	0.0607	0.0574	0.0667	0.0628	0.0659	0.0705	0.0687	0.0747	$0.0758_{\pm 0.0008}$	$0.0762_{\pm 0.0009}$	$0.0809_{\pm 0.0007}$	$\mathbf{0.0860}^{**}_{\pm 0.0004}$
	R@20	0.0653	0.0864	0.0857	0.0625	0.0928	0.0881	0.0998	0.0961	0.0987	0.1077	0.1035	0.1133	$0.1160_{\pm 0.0012}$	$0.1142_{\pm 0.0013}$	$0.1187_{\pm 0.0010}$	$\mathbf{0.1262}^{**}_{\pm 0.0007}$
	N@10	0.0241	0.0311	0.0305	0.0203	0.0335	0.0316	0.0366	0.0339	0.0357	0.0382	0.0357	0.0405	$0.0414_{\pm 0.0005}$	$0.0408_{\pm 0.0005}$	$0.0441_{\pm 0.0004}$	$\mathbf{0.0490}^{**}_{\pm 0.0002}$
	N@20	0.0298	0.0387	0.0386	0.0266	0.0421	0.0393	0.0454	0.0431	0.0443	0.0478	0.0458	0.0505	$0.0517_{\pm 0.0006}$	$0.0506_{\pm 0.0006}$	$0.0538_{\pm 0.0005}$	$\mathbf{0.0594}^{**}_{\pm 0.0004}$
Electronics	R@10	0.0235	0.0363	0.0293	0.0207	0.0349	0.0363	0.0392	0.0397	0.0437	0.0382	0.0418	0.0456	$0.0472_{\pm 0.0008}$	$0.0474_{\pm 0.0009}$	$0.0503_{\pm 0.0006}$	$\mathbf{0.0549}^{**}_{\pm 0.0002}$
	R@20	0.0367	0.0540	0.0458	0.0331	0.0529	0.0541	0.0581	0.0573	0.0648	0.0588	0.0617	0.0673	$0.0700_{\pm 0.0010}$	$0.0691_{\pm 0.0012}$	$0.0725_{\pm 0.0009}$	$\mathbf{0.0778}^{**}_{\pm 0.0004}$
	N@10	0.0127	0.0204	0.0159	0.0109	0.0195	0.0202	0.0223	0.0224	0.0247	0.0209	0.0224	0.0253	$0.0262_{\pm 0.0004}$	$0.0261_{\pm 0.0004}$	$0.0278_{\pm 0.0003}$	$\mathbf{0.0326}^{**}_{\pm 0.0002}$
	N@20	0.0161	0.0250	0.0202	0.0141	0.0241	0.0248	0.0273	0.0271	0.0302	0.0262	0.0278	0.0309	$0.0321_{\pm 0.0005}$	$0.0318_{\pm 0.0005}$	$0.0335_{\pm 0.0004}$	$\mathbf{0.0385}^{**}_{\pm 0.0003}$

4.1 Direct Consistency Has an Effective Range

We run a compact direct-consistency diagnostic on Amazon Baby using the same backbone and split, while varying only the consistency strength. Without direct consistency, Recall@20 is 0.0922. A moderate setting with a contrastive loss weight of 0.01, a similarity weight of 0.3, and UI cosine improves it to 0.1024, showing that cross-view consistency can be useful. However, a stronger setting with a contrastive loss weight of 0.1 and a similarity weight of 1.0 drops Recall@20 to 0.0836, showing that direct consistency has an

effective range. Once this objective dominates, it can suppress private recommendation factors that are not shared by all modalities. We also checked a lighter setting with a contrastive loss weight of 0.001 and a similarity weight of 0.1 under another seed, which reached Recall@20 of 0.0982 and still stayed below the moderate setting.

4.2 Low Frequency Agrees; High Frequency Ranks

Frequency diagnostics show where the useful information sits. Low-frequency components have stronger cross-view agreement, while high-frequency components retain more ranking signal. Both are weaker than the complete representation. Figure 2 shows this comparison. We define these bands by ordering the singular components of graph-smoothed channel matrices; Section 5.1 gives the formal construction. Low frequency captures shared structure, whereas higher-frequency variation carries more local ranking evidence.

4.3 Behavior Evidence Decides When Calibration Helps

Content-derived corrections are not uniformly beneficial. In our diagnostics, global score correction often hurts, while top-K candidate correction can improve ranking. Co-user behavior is most useful on the sparser Amazon domains, so we use it as a local signal for candidate calibration. A candidate supported by co-occurrence can receive a small correction; otherwise, the same content relation may identify a substitute or visual lookalike.

This yields three design requirements for BRIDGE:

(1) Preserve frequency-aware multimodal structure instead of collapsing all signals into an undifferentiated representation.
(2) Use behavior evidence inside the model so that user histories can authorize or reject candidate corrections.
(3) Restrict score calibration to the candidate set, so uncertain evidence cannot perturb the whole item space.

5 Method

Figure 1 summarizes BRIDGE. The method preserves frequency-specific multimodal evidence before using behavior to calibrate the final ranking. DFGE learns the frequency-aware graph representation and base score $s_{\text{base}}(u,i)$. BEN estimates normalized behavior evidence b_{ui} from the training interaction graph. CRI injects that evidence as a signed residual inside the candidate scope, so the base encoder defines the ranking geometry and the behavior branch adjusts local decisions.

5.1 Dual-Frequency Graph Evidence Encoder

Each item has an ID embedding, a visual feature, and a textual feature. The visual and textual features are projected to the same latent dimension as the ID channel. DFGE first applies channel-specific projection, then two rounds of symmetric normalized message passing on the user-item graph, and finally sums the layer states to form each channel embedding. Let $\mathbf{A}_{ui}$ denote the user-item adjacency and $\hat{\mathbf{A}}_{ui} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_{ui}\mathbf{D}^{-\frac{1}{2}}$ its symmetric normalization. For channel c, the input state is $\mathbf{H}^{c,(0)} = [\mathbf{P}_u^c; \mathbf{W}_c\mathbf{x}_i^c]$, where $\mathbf{P}_u^c$ is a learned user preference table and $\mathbf{W}_c$ is a linear projection for that channel; for the ID channel, the second term is the learned item ID embedding. We then apply $L = 2$ rounds of normalized message passing, $\mathbf{H}^{c,(\ell+1)} = \hat{\mathbf{A}}_{ui}\mathbf{H}^{c,(\ell)}$, and read out the channel embedding by summing all layers, $\mathbf{Z}^c = \sum_{\ell=0}^{L}\mathbf{H}^{c,(\ell)}$, for $c \in \{\text{id}, v, t\}$. The item-side representation is refined with a cached multimodal item graph built from top-K cosine neighbors over the content features. When both image and text are available, the cached item graph is the weighted sum $0.1\,\mathbf{A}_v + 0.9\,\mathbf{A}_t$ with $K = 10$. The item graph acts as a second smoothing stage before frequency decomposition, using one propagation layer and keeping the final smoothed item states rather than another layer-sum readout. This produces user and item representations for the ID, visual, and textual channels:

$$\mathbf{z}_u^c, \mathbf{z}_i^c = \text{GraphEnc}_c(\mathcal{R}, \mathbf{A}_{\text{mm}}, \mathbf{x}^c), \quad c \in \{\text{id}, v, t\}. \tag{3}$$

The three channels are then concatenated:

$$\mathbf{z}_u = [\mathbf{z}_u^{\text{id}}; \mathbf{z}_u^v; \mathbf{z}_u^t], \quad \mathbf{z}_i = [\mathbf{z}_i^{\text{id}}; \mathbf{z}_i^v; \mathbf{z}_i^t]. \tag{4}$$

Each channel has dimension 64, so the fused representation dimension is 192.

5.2 Spectral Band Routing

DFGE decomposes each modality channel into M frequency bands before producing the base ranking embedding. For a representation matrix $\mathbf{Z}^c$ of modality c, we compute a singular decomposition $\mathbf{Z}^c = \mathbf{U}\Sigma\mathbf{V}^\top$ and split the singular components into M contiguous bands. The m-th band is:

$$\mathbf{H}^{c,m} = \mathbf{U}_m\Sigma_m\mathbf{V}_m^\top. \tag{5}$$

For each node n and channel c, let $\mathbf{h}_n^{c,m} = (\mathbf{H}^{c,m})_{n,:}$ denote the m-th band slice. The band-specific multimodal state is then

$$\mathbf{h}_n^m = [\mathbf{h}_n^{\text{id},m}; \mathbf{h}_n^{v,m}; \mathbf{h}_n^{t,m}], \quad n \in \mathcal{U} \cup \mathcal{I}. \tag{6}$$

A task-aware band gate then fuses the bands:

$$\alpha_n^m = 1 + \rho \cdot \phi_m(\mathbf{z}_n), \quad \omega_m = \sigma(a_m), \tag{7}$$

$$\mathbf{e}_n = \sum_{m=1}^{M} \omega_m\alpha_n^m\mathbf{h}_n^m. \tag{8}$$

Here $\phi(\mathbf{z}_n) = \sigma(\mathbf{W}_g\mathbf{z}_n + \mathbf{b}_g)$ is a one-layer gate network with output dimension M; in the main setting, $\phi_m(\cdot)$ is the m-th sigmoid output of this linear gate, and ρ is a learned scalar initialized to 0.5. The base score is computed by inner product:

$$s_{\text{base}}(u,i) = \langle \mathbf{e}_u, \mathbf{e}_i \rangle. \tag{9}$$

The frequency decomposition is part of the base ranking representation. Low-frequency bands, in the sense of leading singular directions of the propagated representation spectrum, provide shared graph evidence, whereas higher-frequency bands preserve private ranking variation. CRI later decides whether a candidate score should receive a behavior-authorized residual. The high-frequency bands remain part of the base representation rather than a separate score for every item.

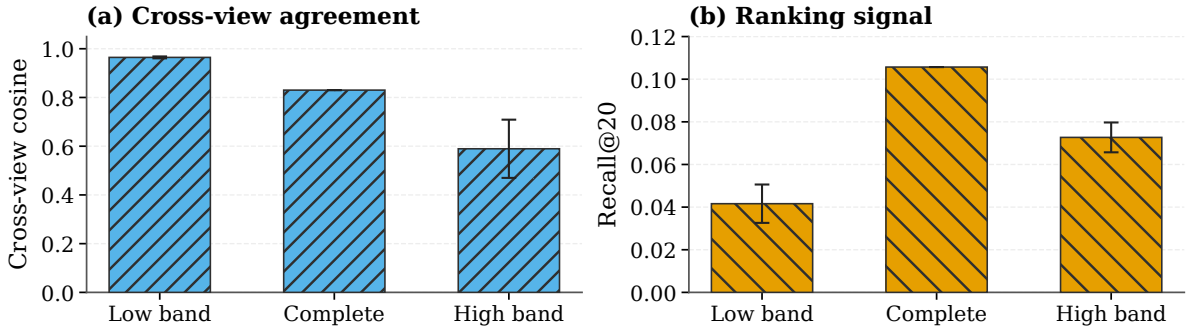

Figure 2: Representation-spectral frequency diagnostic on Amazon Baby. Low-frequency bands have stronger cross-view agreement, whereas high-frequency bands retain more ranking signal; the complete representation performs best. Error bars show the observed ranges.

Interpretation of frequency. Here, frequency means the ordering of representation components after graph smoothing. Let $\mathbf{Z} \in \mathbb{R}^{N\times d}$ be a graph-smoothed channel matrix after message passing. BRIDGE computes $\mathbf{Z} = \mathbf{U}\Sigma\mathbf{V}^\top$, partitions the right singular directions as $\mathbf{V} = [\mathbf{V}_1, \dots, \mathbf{V}_M]$ in descending singular-value order, and reconstructs band evidence by $\mathbf{H}^m = \mathbf{Z}\mathbf{V}_m\mathbf{V}_m^\top$. The leading bands capture dominant post-smoothing variation, while later bands collect residual directions with less shared energy. Because normalized graph propagation is already low-pass, the leading singular bands act as stable shared evidence and the trailing bands retain more local variation. A graph Laplacian basis would order components by graph oscillation, whereas this decomposition orders latent energy.

Channel-wise decomposition. The decomposition is applied to each modality channel before fusion. Each user/item representation is split into 64-dimensional ID, image, and text blocks; each block is decomposed separately and then reassembled into 192-dimensional band representations. With $M = 3$, the 64 singular directions are partitioned into bands of size $22/21/21$ in descending singular-value order. For controlled analysis, we also evaluate non-SVD decomposition operators under the same encoder and fusion capacity. In the equal-capacity no-decomposition control, the band inputs are

$$\mathbf{H}^{c,m} = \frac{1}{M}\mathbf{Z}^c, \quad m = 1, \dots, M. \tag{10}$$

This keeps the same number of band gates, band weights, graph encoders, and fusion parameters while removing spectral separation itself. We also evaluate fixed-basis variants in which $\mathbf{H}^{c,m} = \mathbf{Z}^c\mathbf{Q}_m\mathbf{Q}_m^\top$: Gram uses eigenvectors of $\mathbf{Z}^\top\mathbf{Z}$, DCT uses a deterministic cosine basis, and Random uses a fixed orthogonal basis.

5.3 Behavior Evidence Normalizer

The Behavior Evidence Normalizer (BEN) builds an item-item behavior graph from co-user similarity after the interaction split is fixed. Let $\mathcal{D}_{\text{tr}}$ denote the training interactions only, and let $\mathcal{U}_i^{\text{tr}}$ denote the users who interacted with item i in $\mathcal{D}_{\text{tr}}$. The raw behavior edge between two items is:

$$B_{ij} = \frac{|\mathcal{U}_i^{\text{tr}} \cap \mathcal{U}_j^{\text{tr}}|}{\sqrt{|\mathcal{U}_i^{\text{tr}}|}\sqrt{|\mathcal{U}_j^{\text{tr}}|}}. \tag{11}$$

Only the top K_b neighbors of each item are kept. For a user-item pair (u, i), the raw candidate behavior evidence is aggregated from the user's training history $\mathcal{H}_u^{\text{tr}}$:

$$\tilde{b}_{ui} = \frac{1}{\sqrt{|\mathcal{H}_u^{\text{tr}}|}} \sum_{j\in\mathcal{H}_u^{\text{tr}}} B_{ij}. \tag{12}$$

Validation and test interactions are never used to construct B_{ij}, $\mathcal{H}_u^{\text{tr}}$, item popularity, or user degree features; they are used only for model selection and final evaluation. The reported setting uses cosine co-user similarity, $K_b = 1000$, square-root history normalization, and centered behavior scores:

$$b_{ui} = \frac{\tilde{b}_{ui} - \mu_u}{\sigma_u + \epsilon}. \tag{13}$$

Although B_{ij} and $\tilde{b}_{ui}$ are non-negative, the normalized residual b_{ui} can be positive or negative. Positive values indicate stronger-than-usual behavior support under the user's history, while negative values indicate weaker-than-usual support and can demote a candidate when used in Eq. 20. When the full user-item behavior matrix fits in memory, BRIDGE normalizes each user's complete behavior row once and reuses the same b_{ui} values in sampled training pairs and candidate inference. For larger sparse datasets, b_{ui} is computed on demand; both paths yield the same normalized score and avoid hidden train–test mismatch.

5.4 Candidate Residual Integrator

The Candidate Residual Integrator (CRI) uses BEN evidence as a signed score residual after DFGE has selected a candidate set. The central design choice is the candidate-level scope rather than an unconstrained learned coefficient. The main BRIDGE setting uses a validation-selected residual strength:

$$\Delta_{\text{bridge}}(u, i) = \mathbb{I}[i \in \mathcal{C}_u^{\text{tr}}]\lambda_b b_{ui}. \tag{14}$$

This is the formulation used in the main experiments. We also include a conservative control with lightweight user and item biases β_u, β_i and scalar coefficients a_s, a_b:

$$\eta_{ui} = \sigma(\beta_u + \beta_i + a_s \tanh(s_{\text{base}}(u, i)) + a_b \tanh(b_{ui})). \tag{15}$$

The corresponding candidate score is

$$s_{\text{train}}(u, i) = s_{\text{base}}(u, i) + \mathbb{I}[i \in \mathcal{C}_u^{\text{tr}}]\lambda_b\eta_{ui}b_{ui}. \tag{16}$$

In both cases, the sign of the correction comes from b_{ui}; the coefficient only scales the residual. This distinction matters empirically: the fixed-λ_b calibrator is the primary CRI instantiation, while the conservative coefficient remains a useful control. During training, BRIDGE uses the same candidate scope as inference. For each training user, we first form a detached base-score candidate set:

$$\mathcal{C}_u^{\mathrm{tr}} = \mathrm{TopK}(\mathrm{sg}(s_{\mathrm{base}}(u,\cdot)), K_c), \tag{17}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. This set is computed with the same candidate size K_c used at inference. It is used only as a scope mask, not as an additional supervision target. In the reported setting, the set is recomputed for each mini-batch from detached base embeddings rather than cached across epochs, so the candidate scope tracks the evolving base scorer. Sampled positive and negative items are then scored as:

$$s_{\mathrm{train}}(u,i) = s_{\mathrm{base}}(u,i) + \mathbb{I}[i \in \mathcal{C}_u^{\mathrm{tr}}]\lambda_b b_{ui}. \tag{18}$$

The indicator ensures that the residual branch is active only on candidate items. Items outside $\mathcal{C}_u^{\mathrm{tr}}$ still contribute through $s_{\mathrm{base}}(u,i)$, so the base scorer can learn to move future positives into the candidate set. Gradients are taken through the base score for all sampled pairs and through the calibration coefficient only when the conservative control is used; the top-K membership itself is detached. This candidate-aware objective reduces the train–test scope mismatch of training a residual on globally sampled pairs while applying it only to a top-K candidate set at inference. During inference, BRIDGE first obtains a candidate set from the base multimodal score:

$$\mathcal{C}_u = \mathrm{TopK}(s_{\mathrm{base}}(u,\cdot), K_c). \tag{19}$$

At inference, the final score is:

$$s(u,i) = s_{\mathrm{base}}(u,i) + \mathbb{I}[i \in \mathcal{C}_u]\lambda_b b_{ui}. \tag{20}$$

This top-K restriction is the main operational scope. It lets behavior evidence reorder plausible candidates, but prevents it from rewriting the whole item space. The reported setting uses $K_c = 200$, and λ_b is selected on validation Recall@20 from a small calibration grid. The conservative coefficient shares the same score form and candidate scope, but it is used only in the control variant. BRIDGE therefore always performs behavior-residual candidate calibration, and the calibration strength is selected on validation rather than tuned on test.

Candidate-aware training. A simpler implementation would train Eq. 18 without the indicator and then apply Eq. 20 only inside $\mathcal{C}_u$ at test time. We treat that variant as a reranking baseline and compare against it in the control suite. BRIDGE instead trains the residual under the same scope used at inference, which keeps the residual branch focused on candidate items and leaves the base scorer responsible for moving new positives into the candidate set.

5.5 Training Objective

BRIDGE optimizes the calibrated BPR loss:

$$\mathcal{L}_{\mathrm{rank}} = -\sum_{(u,i,j)} \log \sigma(s_{\mathrm{train}}(u,i) - s_{\mathrm{train}}(u,j)). \tag{21}$$

An auxiliary BPR loss on the base score keeps the backbone predictive:

$$\mathcal{L}_{\mathrm{base}} = -\sum_{(u,i,j)} \log \sigma(s_{\mathrm{base}}(u,i) - s_{\mathrm{base}}(u,j)). \tag{22}$$

This auxiliary loss matters under candidate-aware training: when a positive item is not yet in $\mathcal{C}_u^{\mathrm{tr}}$, the residual branch is inactive, and the base loss supplies the learning signal that can move the item into later candidate sets. The frequency gates are regularized by two objectives: an information-bottleneck surrogate that controls gate expansion, and a frequency structural regularizer that constrains the decomposed bands. Let $\gamma_{n,m}$ denote the frequency gate value for node n and band m, and define $\delta_{n,m} = \max(\gamma_{n,m} - 1, 0)$ under the positive-gate setting. The information-bottleneck surrogate is:

$$\mathcal{L}_{\mathrm{IB}} = \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}}\left[\alpha_{\mathrm{IB}}\|\delta_n\|_2^2 + \mu_{\mathrm{IB}}\|\delta_n\|_2 \sum_{m=1}^{M}\max(\delta_{n,m} - \phi_+, 0)\right], \tag{23}$$

where $\mathcal{N}$ includes the user and item nodes used by the encoder. This term penalizes unnecessary expansion of the frequency gates: if a node can be represented without strongly amplifying a particular frequency band, the gate is encouraged to stay close to one.

For a training mini-batch $\mathcal{B} = \{(u_b, i_b, j_b)\}_{b=1}^{B}$, BRIDGE computes the frequency structural regularizer from the user bands $\mathbf{h}_{u_b}^m$ and the positive-item bands $\mathbf{h}_{i_b}^m$. Let $\tilde{\mathbf{h}} = \mathbf{h}/(\|\mathbf{h}\|_2 + \epsilon)$ denote the normalized band vector. The regularizer has two parts:

$$\mathcal{L}_{\mathrm{freq}} = \mathcal{L}_{\mathrm{dec}} + \mathcal{L}_{\mathrm{disc}}, \tag{24}$$

where

$$\mathcal{L}_{\mathrm{dec}} = \frac{1}{BM(M-1)}\sum_{b=1}^{B}\sum_{m\neq\ell}\Big[\cos^2(\tilde{\mathbf{h}}_{u_b}^m, \tilde{\mathbf{h}}_{u_b}^\ell) + \cos^2(\tilde{\mathbf{h}}_{i_b}^m, \tilde{\mathbf{h}}_{i_b}^\ell)\Big], \tag{25}$$

and

$$\mathcal{L}_{\mathrm{disc}} = -\frac{1}{B|\mathcal{M}_h|}\sum_{m\in\mathcal{M}_h}\sum_{b=1}^{B}\log\frac{\exp(\langle\tilde{\mathbf{h}}_{u_b}^m, \tilde{\mathbf{h}}_{i_b}^m\rangle/\tau)}{\sum_{r=1}^{B}\exp(\langle\tilde{\mathbf{h}}_{u_b}^m, \tilde{\mathbf{h}}_{i_r}^m\rangle/\tau)}. \tag{26}$$

Here $\mathcal{M}_h$ denotes the higher-frequency bands. The decorrelation term suppresses redundant overlap between different bands, while the high-frequency discrimination term keeps the higher-frequency bands ranking-aware by contrasting each positive pair against other positive items in the same batch. This regularizer is structural rather than a strict orthogonality penalty: it preserves a stable band geometry while keeping the informative bands discriminative. The IB surrogate controls how much each node may amplify frequency bands, whereas the frequency structural regularizer controls the geometry of the resulting band embeddings. When the conservative coefficient control is enabled, we add a soft constraint on the average gate value:

$$\mathcal{L}_{\eta} = \left(\mathbb{E}_{(u,i)}[\eta_{ui}] - \tau_{\eta}\right)^2. \tag{27}$$

For the main fixed-λ_b BRIDGE setting, $\mathcal{L}_\eta = 0$. The final objective is:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{rank}} + \lambda_{\text{base}}\mathcal{L}_{\text{base}} + \lambda_{\text{IB}}\mathcal{L}_{\text{IB}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} \\ &+ \lambda_\eta \mathcal{L}_\eta.\end{aligned} \tag{28}$$

In the reported fixed-λ_b setting, the active coefficients are $\lambda_{\text{base}} = 0.2$, $\lambda_{\text{IB}} = 1.0$, $\lambda_{\text{freq}} = 0.001$, and $\lambda_\eta = 0$. The conservative coefficient control uses $\lambda_\eta = 0.001$ and gate target $\tau_\eta = 0.5$. The candidate-calibration strength follows the same five-point validation grid. This objective makes the method explicit: DFGE learns the frequency-organized multimodal representation, and CRI integrates BEN evidence under the same candidate scope used by inference.

6 Experiments

6.1 Experimental Setup

We evaluate BRIDGE on Amazon Baby, Sports, and Electronics from the Amazon review benchmark [14] under the common MMRec-style implicit-feedback split. Baby has 19,445 users, 7,050 items, and 160,792 interactions. Sports has 35,598 users, 18,357 items, and 296,337 interactions. Electronics has 192,403 users, 63,001 items, and 1,689,188 interactions. We use BEIT3-derived multimodal features and report Recall@10, Recall@20, NDCG@10, and NDCG@20. All baselines in Table 1 use the same processed interaction splits, BEIT3 feature files, and full-sort evaluation protocol. Unless otherwise stated, reported numbers use seed 2020.

For all three datasets, the behavior graph is constructed strictly from the training split. Each interaction file carries a training, validation, or test partition marker. We first build $\mathcal{D}_{\text{tr}}$ from the training partition and compute co-user edges, user histories, item popularity, and user degree from this training matrix only. Validation and test interactions are not used in the behavior graph. Baby contains 118,551 training interactions, 20,559 validation interactions, and 21,682 test interactions. Sports contains 218,409 training interactions, 37,899 validation interactions, and 40,029 test interactions. Electronics contains 1,254,441 training interactions, 211,296 validation interactions, and 223,451 test interactions. After filtering, validation and test contain no cold-start users.

All BRIDGE main results use candidate-aware training. For each training mini-batch, the model computes the detached base top-K_c candidate set of each involved user and masks the behavior residual outside this set, as defined in Eq. 18. At test time, the same K_c-restricted residual is applied to the base top-K_c list in Eq. 20. Training uses pairwise BPR with this same candidate mask, so the residual is learned in the scope where it is used.

The reported hyperparameters follow the dataset and model configuration files. We use a 64-dimensional embedding size, two graph layers, and three frequency bands. The item graph uses 10 neighbors. The learning rate is 10^{-4}, the regularization weight is 10^{-3}, the IB weight is 1.0, the behavior top-K is 1000, the behavior weight is 0.4, the behavior evaluation top-K is 200, and the train candidate top-K is 200. The behavior graph uses cosine similarity, sum-sqrt aggregation, and z-score normalization. Training uses Adam for 300 epochs with batch sizes 2048 and 4096, zero weight decay, and seed 2020. These settings are shared by the main tables, the ablation suite, and the calibration sweeps. The experiment suite is organized to match the method: Table 1 compares overall ranking quality, Tables 2 and 3 isolate the residual and spectral controls, and Table 4 checks whether the gains come from a head-item boost or from a broader redistribution of ranking mass.

6.2 Main Results

Table 1 follows the common overall comparison format and places BRIDGE in the rightmost column. The baseline set spans pairwise ranking (BPR [21]) and graph collaborative models (LightGCN [4], MMGCN [28], GRCN [27], DualGNN [26], and LATTICE [36]). It also includes multimodal variants (VBPR [3], SLMRec [25], BM3 [40], FREEDOM [39], DiffMM [5], MMIL [32], AlignRec [10], SMORE [16], and FITMM [33]). The comparison therefore covers classical collaborative ranking and recent multimodal systems under the same data pipeline. We select the checkpoint and candidate-calibration setting by validation Recall@20. The validation grid sets λ_b to 0.1, 0.2, 0.4, 0.6, or 0.8, and also evaluates a coefficient-controlled variant. The test set is evaluated only after this selection. On the three reported datasets, the selected BRIDGE setting is the fixed-λ_b candidate calibration rather than the conservative coefficient control. For the recent strong baselines and BRIDGE, we report mean and standard deviation across five seeds because these methods define the competitive margin. The remaining baselines are included as single-run protocol controls under the same processed data and feature setting. Hyperparameters are chosen on validation Recall@20 using the ranges provided by the corresponding implementations or MMRec-style configuration files. Table 1 shows that BRIDGE improves every metric on every dataset. The Recall gains are consistent, but the NDCG gains are larger, which means the method mainly sharpens the order near the top of the list rather than only recovering more positives at larger cutoffs. The margin is most pronounced on Electronics, where sparsity and popularity skew make a global residual less reliable. This matches the design of BRIDGE: DFGE sets the candidate geometry, and BEN/CRI refines only the items that already look plausible under the multimodal backbone. Broken down by dataset, Baby shows the smallest but still stable gain, Sports shows the largest balanced improvement across Recall and NDCG, and Electronics shows the clearest benefit from candidate-level calibration. This pattern is consistent with the data regime: as the interaction graph becomes sparser, the model benefits more from a restricted residual than from a global score adjustment.

6.3 Diagnostics

At the validation-selected Baby checkpoint, the pair coefficient mean is 0.7365, the behavior correction absolute mean is 0.6886, the low-item norm mean is 6.3819, the high-item norm mean is 2.7333, and the low-high item cosine is 0.0560. These values show that the two frequency components capture different signals and that the residual branch acts locally. The low-frequency band carries stronger cross-view agreement and larger energy, while the high-frequency band is less aligned but remains informative. Their near-zero cosine indicates that BRIDGE can use the low bands as shared structure and the high bands as local variation.

6.4 Additional Controls and Sensitivity

Tables 2–4 report the controls that separate the roles of DFGE, BEN, and CRI. All controls use the same splits, features, and full-sort protocol. The control suite answers three questions. First, do BEN and CRI need behavior evidence and candidate restriction, or can a generic residual work? Second, does DFGE need a spectral decomposition, or is equal-capacity routing enough? Third, does the residual mainly move score mass away from popular items, or does it improve ranking across strata? Tables 2–4 provide the quantitative controls.

Table 2 isolates the behavior branch, multimodal encoder, and spectral objectives on Baby. Panel (b) compares BRIDGE with alternative residual rules on all three datasets, separating the score itself from the way it is applied.

The first row block in Table 2(a) shows that BEN and CRI are tied together. Removing behavior evidence hurts, but keeping the same evidence and relaxing the scope is still worse than the full candidate-aware model. Global correction recovers part of the drop but remains below BRIDGE because the residual then reaches items outside the base candidate set. The residual helps when it stays local.

The content block in Table 2(a) shows that the item graph carries the largest share of the multimodal gain. Removing the item graph causes a sharp drop, while removing image or text alone produces smaller losses, which means the two modalities are complementary rather than interchangeable. The spectral block then separates architecture from regularization. Disabling the IB surrogate or the frequency regularizer hurts performance, but Gram, DCT, and random bases stay close to one another. The result points to a stable band split as the important factor, rather than a particular basis.

Table 2(b) compares the residual against simpler baselines. Raw co-occurrence is the weakest control, EASE improves it, and the conservative coefficient closes part of the gap. BRIDGE still wins on all three datasets, with the largest margin on Electronics. The comparison shows that BEN is more than an item-item prior. Its normalization and candidate scope matter when the interaction matrix is sparse and the residual must discriminate among near ties.

Table 3 extends the same check across datasets and fusion rules. The cross-dataset block repeats the Baby ordering on Sports and Electronics, so the mechanism is not tied to one domain density. The learned-fusion block then compares a fixed residual with learned scalar mixing, an MLP gate, and an EASE-style residual. The learned variants remain close, but none surpass the fixed BRIDGE setting. A larger fusion head adds little beyond the candidate-restricted residual.

Table 4 tests whether the gain comes from favoring head items. BRIDGE raises Recall@20 in every popularity stratum while reducing head exposure, and the effect is strongest on Sports and Electronics, where the base model is more skewed toward popular items. The conservative coefficient variant keeps more of the original exposure pattern and usually trails the fixed residual, making the fixed calibration rule the stronger operating point.

Across these controls, the same pattern appears. Removing behavior evidence or candidate restriction hurts, and global correction is weaker than candidate-level calibration. Raw co-occurrence and EASE do not match BRIDGE's normalized residual, so BEN is more than an ItemKNN-style lookup. The item graph and multimodal content shape candidate geometry, while the spectral objectives stabilize the calibrated score. Table 4 shows the output effect: BRIDGE raises mid, tail, and cold recall while lowering head exposure.

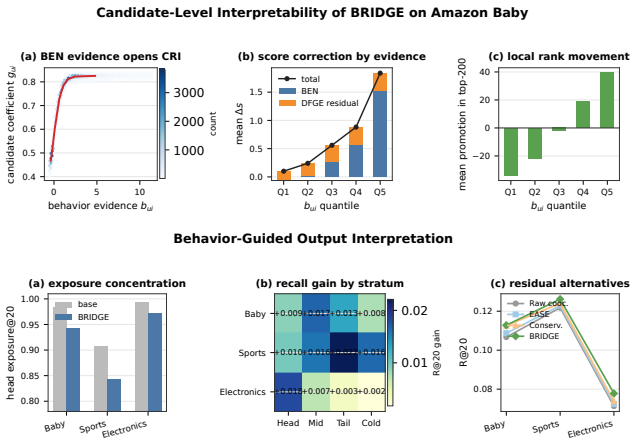

Figure 3: Candidate-level interpretability controls for BRIDGE. The upper row shows BEN evidence, score corrections, and rank changes; the lower row shows exposure shifts, stratum-level recall, and residual alternatives.

7 Discussion and Limitations

The controls separate the two roles in BRIDGE. Candidate restriction keeps the residual from distorting the base score: removing the top-K scope or applying the correction globally weakens ranking. Raw co-occurrence, EASE, and learned fusion controls also remain below the full model. The popularity results add an output-level view: the residual improves all item strata while reducing head exposure, without an explicit anti-popularity objective.

7.1 Practical Implications

DFGE builds the candidate manifold, BEN converts co-user overlap into signed evidence, and CRI decides where that evidence may alter the score. Keeping these roles separate is what makes the residual useful. When the same evidence acts globally, it competes with the multimodal backbone and the improvement shrinks. Candidate restriction keeps the residual tied to the ranking decision.

The popularity diagnostics follow the same division of labor. Head items are easier to reconstruct from sparse co-occurrence, while mid, tail, and cold items depend more on cross-modal structure. BRIDGE uses no explicit anti-popularity penalty. It preserves head accuracy and sends more ranking mass to less exposed items that pass the candidate gate. The exposure shift in Table 4 follows from this localization.

At inference, BRIDGE applies a residual pass to a cached candidate list. The item graph is precomputed, the behavior graph uses training interactions only, and the residual changes scores inside

Table 2: Core Baby ablations and cross-dataset residual alternatives. All rows use the same splits, BEIT3 features, candidate restriction, and full-sort evaluation protocol.

(a) DFGE, BEN/CRI, and frequency ablations on Amazon Baby

Module	Variant	R@10	R@20	N@10	N@20
Full	BRIDGE	**0.0752**	**0.1128**	**0.0428**	**0.0525**
BEN/CRI	w/o behavior evidence	0.0682	0.1016	0.0368	0.0453
BEN/CRI	w/o top-K calibration	0.0651	0.1025	0.0354	0.0450
BEN/CRI	global correction	0.0745	0.1092	0.0419	0.0507
DFGE-Content	w/o item graph	0.0535	0.0824	0.0282	0.0356
DFGE-Content	w/o image feature	0.0735	0.1089	0.0415	0.0506
DFGE-Content	w/o text feature	0.0750	0.1089	0.0422	0.0509
DFGE-Content	w/o content features	0.0608	0.0894	0.0340	0.0414
DFGE-Spectral	w/o IB surrogate	0.0709	0.1031	0.0404	0.0487
DFGE-Spectral	w/o freq. structural reg.	0.0747	0.1107	0.0424	0.0517
DFGE-Spectral	w/o both freq. objectives	0.0712	0.1023	0.0407	0.0487
DFGE-Spectral	w/o freq. decomp. (equal cap.)	0.0746	0.1110	0.0425	0.0519
DFGE-Spectral	Gram decomposition	0.0752	0.1117	0.0425	0.0519
DFGE-Spectral	DCT decomposition	0.0751	0.1089	0.0427	0.0514
DFGE-Spectral	Random orthogonal decomp.	0.0750	0.1113	0.0427	0.0519

(b) Candidate-level alternative residual controls

Dataset	Metric	Raw cooc.	EASE	Conserv. coeff.	BRIDGE
Baby	R@20	0.1067	0.1087	0.1126	**0.1128**
	N@20	0.0478	0.0494	0.0519	**0.0525**
Sports	R@20	0.1218	0.1229	0.1240	**0.1262**
	N@20	0.0553	0.0571	0.0589	**0.0594**
Electronics	R@20	0.0713	0.0724	0.0738	**0.0778**
	N@20	0.0338	0.0344	0.0351	**0.0385**

Table 3: Candidate-side controls on Amazon Baby, Sports, and Electronics. To keep the control study in the main text, we report the primary Recall@20 and NDCG@20 metrics.

(a) Cross-dataset ablations

Dataset	Variant	R@20	N@20
Sports	**BRIDGE**	**0.1262**	**0.0594**
	conserv. coeff.	0.1217	0.0572
	w/o behavior evidence	0.1162	0.0515
	w/o top-K calibration	0.1127	0.0493
	global correction	0.1137	0.0542
	w/o calib. coeff.	0.1194	0.0567
	w/o behavior coeff.	0.1236	0.0581
	w/o image feature	0.1208	0.0567
	w/o text feature	0.1213	0.0572
	w/o content features	0.1048	0.0505
	w/o item graph	0.1010	0.0471
Electronics	**BRIDGE**	**0.0778**	**0.0385**
	conserv. coeff.	0.0747	0.0354
	w/o calib. coeff.	0.0755	0.0359
	w/o behavior evidence	0.0656	0.0298
	w/o top-K calibration	0.0647	0.0292
	global correction	0.0642	0.0303
	w/o image feature	0.0648	0.0293

(b) Learned fusion controls

Dataset	Variant	R@20	N@20
Baby	base encoder	0.1017	0.0455
	fixed λ mix	0.1128	0.0520
	learned λ_u	0.1126	0.0519
	MLP gate	0.1115	0.0512
	EASE residual	0.1087	0.0494
	BRIDGE	**0.1128**	**0.0525**
Sports	base encoder	0.1141	0.0502
	BRIDGE	**0.1262**	0.0594
	learned λ_u	0.1260	**0.0595**
	MLP gate	0.1257	0.0581
	conserv. coeff.	0.1240	0.0589

Table 4: Popularity-stratified diagnostic and exposure track. Head, Mid, Tail, and Cold report Recall@20 within each item-popularity stratum; Head Exp. is the top-20 exposure share assigned to head items.

Dataset	Variant	R@20	Head	Mid	Tail	Cold	Head Exp.
Baby	base	0.1017	0.1649	0.0094	0.0059	0.0031	0.9842
	mix	0.1110	0.1705	0.0270	0.0183	0.0109	0.9356
	logic	0.1113	0.1710	0.0270	0.0189	0.0109	0.9358
	BRIDGE (fixed λ)	**0.1128**	**0.1739**	0.0261	**0.0189**	**0.0109**	0.9426
Sports	base	0.1141	0.1888	0.0338	0.0217	0.0119	0.9073
	BRIDGE (fixed λ)	**0.1262**	**0.1986**	**0.0497**	**0.0438**	**0.0283**	0.8429
	conservative coeff. variant	0.1240	0.1926	0.0452	0.0381	0.0231	0.8658
Electronics	base	0.0637	0.0933	0.0027	0.0012	0.0006	0.9939
	BRIDGE (fixed λ)	**0.0778**	**0.1115**	**0.0099**	**0.0045**	**0.0024**	0.9726
	conservative coeff. variant	0.0738	0.1067	0.0067	0.0029	0.0016	0.9858

the shortlist. This fits a standard retrieval-and-rank pipeline. Candidate quality sets the upper bound: a relevant item that never enters the shortlist cannot be recovered by calibration.

Noise in co-user neighborhoods blurs BEN when overlap is sparse or interaction bursts are highly local. The conservative coefficient limits the correction in these cases. The fixed candidate-weighted

residual is stronger once the backbone produces a reasonable shortlist, which is the setting evaluated in the benchmark.

7.2 Calibration Behavior

The validation sweep gives the same qualitative behavior across the coefficient settings. The fixed-λ_b setting wins on all three datasets, while the conservative control stays close. The residual is a local correction, so a small coefficient has little effect and a large one competes with the backbone. The five-point grid selects the middle range without a large search.

The same pattern appears across the frequency design. The low-frequency component carries the stable cross-view structure, while the higher-frequency bands capture sparse and item-specific variation. The IB surrogate and the frequency regularizer are both needed because they serve different roles: one controls how much each node can route into the bands, the other preserves a useful band geometry once the routing has happened. The ablations show that dropping either term weakens the ranking signal, while changing the basis has only a small effect. The decomposition is structural and largely independent of the basis.

Removing frequency decomposition gives 0.1110 Recall@20 on Baby versus 0.1128 for BRIDGE, while Gram, DCT, and random orthogonal bases remain close. The gain therefore comes from separating shared low-frequency evidence from more private higher-frequency variation; band routing and its regularizers provide a smaller complementary gain by shaping the candidate geometry used by BEN/CRI.

7.3 Scope and Reproducibility

BRIDGE operates at the ranking stage of a two-stage recommender. The multimodal backbone builds the base top-K list, which defines the candidate universe, and BEN/CRI changes only the order inside that list. Recall captures candidate coverage, whereas NDCG also reflects the position of a relevant item in the local ranking. The larger NDCG gains in Table 1 arise from advancing plausible items within the list, especially on the sparser Electronics graph.

Figure 3 makes the mechanism visible at the candidate level. Positive BEN values produce larger score adjustments and generally move the associated items upward in the top-200 order. The dataset-level diagnostics follow the same pattern: Table 4 reports higher Recall@20 in the mid, tail, and cold strata alongside lower head-item exposure on Sports and Electronics. The residual reallocates positions among candidates that have already passed the multimodal gate.

The fixed residual keeps this calibration step transparent. The base representation supplies the retrieval signal, while normalized co-user evidence resolves close scores. Learned scalar mixing and the MLP gate remain competitive in Table 3, yet the fixed rule gives the strongest overall balance across the reported datasets. A five-point validation grid sets its strength, keeping the final ranking step easy to inspect and reproduce.

The ablations also map the design to deployment choices: candidate size sets recall headroom, band count controls spectral resolution, and residual strength controls how far behavior evidence can move a candidate. Baby shows a modest calibration lift, Sports benefits from denser co-user overlap, and Electronics gains the most NDCG from resolving sparse local ties. The residual therefore complements the backbone rather than replacing retrieval.

BRIDGE is most useful when the backbone supplies a semantically plausible shortlist and local ordering is uncertain. Sparse overlap and highly local interaction bursts make BEN less precise, and the residual affects only items retained by the base candidate list. The study covers three Amazon domains with one BEIT3 feature type. The item and behavior graphs are cached and CRI evaluates only the shortlist; end-to-end latency and throughput are not evaluated.

The public repository includes the model and evaluation code, dataset format documentation, configurations, and run scripts for all three benchmarks. Every experiment uses the same feature cache, split manifest, and full-sort evaluator. Behavior evidence is constructed from the training split, and the residual strength is selected on validation data. These settings keep the ranking protocol aligned across all compared methods.

The release also records candidate size, behavior top-K, band count, random seed, feature paths, preprocessing scripts, and evaluation commands. Keeping these choices in the configuration files makes it possible to regenerate the tables with the same cached graphs and feature files, while varying one component at a time.

8 Conclusion

BRIDGE separates candidate retrieval from local ranking. DFGE builds the multimodal candidate geometry, BEN turns co-user overlap into signed evidence, and CRI uses that evidence only to reorder candidates already admitted by the backbone. Across Baby, Sports, and Electronics, this division improves both Recall and NDCG, with the largest gains in NDCG. Multimodal evidence finds plausible items; behavior evidence resolves their close ordering.

The ablations support this division. BRIDGE changes local order without changing the candidate pool, which explains the larger NDCG gains and lower head-item exposure on Sports and Electronics.

The fixed residual works because broad preference matching stays with the base score, while signed behavior evidence is reserved for close alternatives. Its influence is concentrated near the top of the ranked list, where small ordering changes matter most.

The tuning rule follows directly: use the backbone for candidate coverage and apply the residual only after the shortlist is reliable. This separation makes changes easier to diagnose: recall changes point to candidate formation, whereas ranking changes point to local calibration.

This separation also explains the metric pattern: Recall changes only when candidate coverage changes, whereas NDCG can improve when a relevant item is moved upward within the same shortlist. The reported gains therefore measure local ranking quality in addition to retrieval quality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62271466.

GenAI Usage Disclosure

The authors used generative AI tools for language editing, LaTeX formatting, drafting assistance, and figure preparation. The authors developed and verified all technical content, experiments, analyses, and conclusions and take responsibility for the final manuscript.

References

[1] Ji Dai, Quan Fang, Jun Hu, Desheng Cai, Yang Yang, and Can Zhao. 2026. Cross-Modal Attention Network with Dual Graph Learning in Multimodal Recommendation. *arXiv preprint arXiv:2601.11151* (2026).

[2] Ziyuan Guo, Jie Guo, Zhenghao Chen, Bin Song, and Fei Richard Yu. 2025. IGDMRec: Behavior Conditioned Item Graph Diffusion for Multimodal Recommendation. *arXiv preprint arXiv:2512.19983* (2025).

[3] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI '16)*. 144–150. doi:10.1609/aaai.v30i1.9973

[4] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 639–648. doi:10.1145/3397271.3401063

[5] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 7591–7599. doi:10.1145/3664647.3681498

[6] Chenghao Li, Wei Zhou, Yihao Zhang, Jiahao Hu, Huayi Shen, and Junhao Wen. 2026. MSCF-Net: Multi-scale Frequency Denoising and Co-frequency Enhancement Network for Multimodal Recommendation. *Expert Systems with Applications* 309 (2026), 131159. doi:10.1016/j.eswa.2026.131159

[7] Yuan Li, Jun Hu, Jiaxin Jiang, Bryan Hooi, and Bingsheng He. 2026. Robust Multimodal Recommendation via Graph Retrieval-Enhanced Modality Completion. *arXiv preprint arXiv:2605.00670* (2026).

[8] Yuecheng Li, Hengwei Ju, Zeyu Song, Wei Yang, Chi Lu, Peng Jiang, and Kun Gai. 2026. RecGOAT: Graph Optimal Adaptive Transport for LLM-Enhanced Multimodal Recommendation with Dual Semantic Alignment. *arXiv preprint arXiv:2602.00682* (2026).

[9] Yang Li, Qi'Ao Zhao, Chen Lin, Jinsong Su, and Zhilin Zhang. 2024. Who To Align With: Feedback-Oriented Multi-Modal Alignment in Recommendation Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 667–676. doi:10.1145/3626772.3657701

[10] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. 2024. AlignRec: Aligning and Training in Multimodal Recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. 1503–1512. doi:10.1145/3627673.3679626

[11] Zihao Liu and Wen Qu. 2025. DSGRec: Dual-path Selection Graph for Multimodal Recommendation. *PeerJ Computer Science* 11 (2025), e2779. doi:10.7717/peerj-cs.2779

[12] Haokai Ma, Yimeng Yang, Lei Meng, Ruobing Xie, and Xiangxu Meng. 2024. Multimodal Conditioned Diffusion Model for Recommendation. In *Companion Proceedings of the ACM Web Conference 2024*. 1733–1740. doi:10.1145/3589335.3651956

[13] Hongjian Ma, Yan Zhang, Yahui Zhou, Bing Yang, Dunhui Yu, and Zhifei Li. 2026. Let Two Graphs Talk: Self-Supervised Dual-Graph Reconstruction for Multimodal Recommendation. *Information Fusion* 133 (2026), 104299. doi:10.1016/j.inffus.2026.104299

[14] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. 43–52. doi:10.1145/2766462.2767755

[15] Feng Mo, Lin Xiao, Qiya Song, Xieping Gao, Wenzhuo Song, and Shoujin Wang. 2025. FGCM: Modality-Behavior Fusion Model Integrated with Graph Contrastive Learning for Multimodal Recommendation. *IEEE Multimedia* (2025). doi:10.1109/MMUL.2025.3542757

[16] Rongqing Kenneth Ong and Andy W. H. Khong. 2025. Spectrum-based Modality Representation Fusion Graph Convolutional Network for Multimodal Recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25)*. 773–781. doi:10.1145/3701551.3703561

[17] Lin Pan, Zhiqiang Pan, Fei Cai, and Honghui Chen. 2026. Multimodal Recommender Systems: A Survey of Representation, Modeling, and Optimization. *Information Fusion* 128 (2026), 103991. doi:10.1016/j.inffus.2025.103991

[18] Xiangchen Pan and Wei Wei. 2026. Joint Behavior-Guided and Modality-Coherence Conditional Graph Diffusion Denoising for Multi-Modal Recommendation. *arXiv preprint arXiv:2604.03654* (2026).

[19] Yuchao Ping, Shuqin Wang, Ziyi Yang, Bugui He, Nan Zhou, and Yongquan Dong. 2025. DDRec: Dual Denoising Multimodal Graph Recommendation. *IEEE Transactions on Computational Social Systems* 12, 3 (2025), 1100–1114. doi:10.1109/TCSS.2024.3490801

[20] Yuxin Qi, Quan Zhang, Xi Lin, Xiu Su, Jiani Zhu, Jingyu Wang, and Jianhua Li. 2025. Seeing Beyond Noise: Joint Graph Structure Evaluation and Denoising for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12461–12469. doi:10.1609/aaai.v39i12.33358

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*. 452–461.

[22] Bucher Sahyouni, Matthew Vowels, Liqun Chen, and Simon Hadfield. 2026. Sequences as Nodes for Contrastive Multimodal Graph Recommendation. *arXiv preprint arXiv:2602.07208* (2026).

[23] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*. 285–295. doi:10.1145/371920.372071

[24] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. 3251–3257. doi:10.1145/3308558.3313710

[25] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2023. Self-Supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia* 25 (2023), 5107–5116. doi:10.1109/TMM.2022.3187556

[26] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Transactions on Multimedia* 25 (2021), 1074–1084. doi:10.1109/TMM.2021.3138298

[27] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. 3541–3549. doi:10.1145/3394171.3413556

[28] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. 1437–1445. doi:10.1145/3343031.3351034

[29] Jun Wu, Yu Zheng, Tianfeng Zhang, Shilong Jing, Jinyu Liu, Shuai Guo, and Fang Deng. 2026. D-DPDG: Diffusion-based Dual-Graph Attention with Dual-Path Feature Extraction for Multimodal Recommendation. *Journal of Intelligent Information Systems* 64, 2 (2026). doi:10.1007/s10844-025-01014-7

[30] Yuhan Xiu and Xiangrong Tong. 2026. Dual-layer Cross-modal Alignment Recommendation Based on the Diffusion Model. *Information Fusion* 125 (2026), 103472. doi:10.1016/j.inffus.2025.103472

[31] Jie Yang, Chenyang Gu, and Zixuan Liu. 2025. Causal Inspired Multi Modal Recommendation. *arXiv preprint arXiv:2510.12325* (2025).

[32] Wei Yang and Qingchen Yang. 2024. Multimodal-aware Multi-intention Learning for Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 5663–5672. doi:10.1145/3664647.3681412

[33] Wei Yang, Rui Zhong, Yiqun Chen, Shixuan Li, Heng Ping, Chi Lu, and Peng Jiang. 2025. FITMM: Adaptive Frequency-Aware Multimodal Recommendation via Information-Theoretic Representation Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. 6193–6202. doi:10.1145/3746027.3755540

[34] Wei Yang, Rui Zhong, Yiqun Chen, Chi Lu, and Peng Jiang. 2025. Structured Spectral Reasoning for Frequency-Adaptive Multimodal Recommendation. In *Advances in Neural Information Processing Systems (NeurIPS '25)*.

[35] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025. Harnessing Multimodal Large Language Models for Multimodal Sequential Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 12 (2025), 13069–13077. doi:10.1609/aaai.v39i12.33426

[36] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. 3872–3880. doi:10.1145/3474085.3475259

[37] Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. 2024. Mirror Gradient: Towards Robust Multimodal Recommender Systems via Exploring Flat Local Minima. *arXiv preprint arXiv:2402.11262* (2024).

[38] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023. Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation. In *Proceedings of the European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications)*. 3123–3130. doi:10.3233/FAIA230631

[39] Xin Zhou and Zhiqi Shen. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. 935–943. doi:10.1145/

3581783.3611943

[40] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multi-Modal Recommendation. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. 845–854. doi:10.1145/3543507.3583251

[41] Yan Zhou, Jie Guo, Hao Sun, Bin Song, and Fei Richard Yu. 2023. Attention-Guided Multi-Step Fusion: A Hierarchical Fusion Network for Multimodal Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. 1816–1820. doi:10.1145/3539618.3591950

[42] Haodong Zhu, Chenhao Yang, Hongchan Li, Zhongchuan Sun, Yajuan Cui, Yanpei Liu, and Liang Zhu. 2026. Beyond Feature Concatenation: Mutual Information-Driven Fusion for Multimodal Sequential Recommendation. *Knowledge-Based Systems* 341 (2026), 115778. doi:10.1016/j.knosys.2026.115778