Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Modality Importance Does Not Translate into Capacity Gains: A Controlled Study of Asymmetric Multimodal Recommendation

Emin Talip Demirkiran

etd@eskisehir.edu.tr

Eskisehir Technical University

Research Article

Keywords: multimodal recommendation, recommender systems, modality importance, representation capacity, reproducibility, graph recommendation

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10998658/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When Modality Importance Does Not Translate into Capacity Gains: A Controlled Study of Asymmetric Multimodal Recommendation

Emin Talip Demirkiran
Department of Computer Engineering, Eskişehir Technical University, Eskişehir, Türkiye
Correspondence: etd@eskisehir.edu.tr

Abstract. Multimodal recommender systems commonly assume that a modality shown to be more informative should receive greater representational capacity, but this inference has rarely been tested as a controlled architectural hypothesis. We examine whether text-prioritized projection width improves a FREEDOM-style frozen/denoised-graph recommender while holding the first-layer projection-weight count fixed. Across three Amazon e-commerce domains and five matched seeds, asymmetric capacity does not provide a reliable accuracy advantage; none of the pre-specified treatment-control comparisons satisfies the locked significance rule. Post-hoc diagnostics clarify why this result should be interpreted narrowly. The capacity reallocation substantially changes text and image projection representations and actively uses the added text units, yet the primary recommendation loss and inference score have no direct gradient path to the projection parameters; influence is mediated through a weak auxiliary objective. In a small subset of runs, checkpoint selection amplifies these local optimization differences and produces endpoint instability. Separately, multimodal utility itself is strongly dataset-dependent: the FREEDOM-style models outperform a LightGCN floor on Clothing but underperform it on Baby and Sports. The findings distinguish modality informativeness from capacity allocation and causal access to the ranking objective. Although limited to one architecture family and three e-commerce domains, the study provides a reproducible diagnostic principle: capacity should be reallocated only after verifying that the manipulated branch has a direct and stable route to the recommendation objective.

Keywords: multimodal recommendation; recommender systems; modality importance; representation capacity; reproducibility; graph recommendation

1 Introduction

Modern recommender systems increasingly combine collaborative signals with visual, textual, audio, and other forms of item content. Multimodal recommendation has consequently developed from simple side-information fusion into a broad family of graph-based, self-supervised, alignment-oriented, and robustness-oriented models [1]. The premise is intuitively attractive: when user-item interactions are sparse, semantic content may supply evidence that collaborative data alone cannot provide. Yet recent work has challenged the stronger assumption that *more modalities are necessarily better*. Large-scale knockout studies show that the usefulness of multimodal content can be highly asymmetric across modalities and architectures [2], [3].

A natural next step is to convert diagnosed modality importance into architectural capacity. If text is consistently more informative than image in a given family of e-commerce recommenders, should the model allocate more representational width to text? This reasoning appears plausible, but it conflates three different questions: whether a modality carries useful signal, whether a wider branch can exploit that signal, and whether changes in that branch have a sufficiently direct optimization path to alter the ranking function. The distinction is especially important in architectures such as FREEDOM [4], where content-derived graphs and collaborative embeddings interact through a specific computation structure rather than through an unrestricted late-fusion network.

We test this distinction directly. Starting from a symmetric FREEDOM-style multimodal recommender, we reallocate projection width toward text while keeping the first-layer projection-weight count exactly matched. The treatment is compared with the symmetric control on Baby, Sports, and Clothing using five matched random seeds and four top-K ranking metrics. A standard LightGCN model [5] is retained as a non-multimodal floor so that differences between multimodal variants are not mistaken for evidence that multimodal information itself is beneficial.

The confirmatory result is negative: the asymmetric treatment does not show a reliable accuracy advantage under the pre-specified decision rule. That result is scientifically informative only if its mechanism and scope are made explicit. We therefore perform a separate post-hoc diagnostic on the saved final checkpoints. It shows that the capacity manipulation is strong at the projection level, but the corrected architecture provides no direct gradient path from either the primary recommendation loss or the inference score to the projection parameters. The intervention reaches collaborative representations only through a heavily down-weighted auxiliary objective. The same diagnostic also reveals a second mechanism: when small training differences alter the discrete early-stopping trajectory, endpoint representations can diverge sharply even though the instantaneous causal channel is weak.

This emphasis on controlled comparison and mechanism follows broader concerns about recommender-system evaluation. Reported improvements are sensitive to baselines, data processing, model selection, and reproducibility choices [6], [7], [8]. Accordingly, the contribution of this study is not a new state-of-the-art model. It is a controlled test of a plausible design inference and a diagnosis of why the inference fails to produce a stable gain in the tested architecture.

The paper makes four contributions. First, it provides a matched-seed test of text-prioritized asymmetric projection capacity under an exactly matched first-layer weight count. Second, it separates representational change from causal access to the ranking objective through checkpoint-level gradient and representation diagnostics. Third, it shows that the value of multimodal information is strongly dataset-dependent relative to a LightGCN floor, with modality-knockout direction mirroring this heterogeneity. Fourth, it documents the implementation checks and repeated experiments following implementation verification required to ensure that the final evidence conforms to the intended architecture, initialization fairness, and Recall@20 model-selection rule.

2 Related Work

Foundations and collaborative recommendation. Recommender-system research has evolved from neighborhood and hybrid methods to latent-factor, ranking, and neural formulations [9]. Matrix factorization established a compact latent representation of users and items [10], while implicit-feedback formulations emphasized confidence-weighted learning [11] and pairwise ranking objectives such as BPR [12]. Neural Collaborative Filtering later generalized interaction functions beyond the inner product [13]. These foundations matter here because the tested multimodal branches do not replace collaborative learning; instead, they interact with an ID-based ranking model whose optimization remains central to the final score.

Graph and self-supervised collaborative filtering. NGCF introduced graph propagation over user-item interactions [14], and LightGCN showed that removing feature transformations and nonlinearities can improve both simplicity and accuracy in collaborative graph recommendation [5]. SGL and SimGCL subsequently used self-supervised objectives to improve robustness and representation geometry [15], [16]. This line of work motivates the non-multimodal comparator in our experiments. A multimodal method should not be judged only against another multimodal method; it should also establish that the content channel adds value beyond a strong collaborative graph baseline.

From side information to multimodal graph architectures. Early visual recommenders such as VBPR demonstrated that pretrained image features can improve personalized ranking [17]. MMGCN moved multimodal recommendation toward modality-specific graph propagation [18], followed by architectures that model user-specific modality preference, reduce auxiliary-graph cost, or separate graph views, including DualGNN, BM3, and MGCN [19], [20], [21]. FREEDOM freezes content-derived item-item graphs and denoises the user-item graph, offering an efficient and influential reference design [4]. LGMRec extends local graph learning with global dependencies, while AlignRec addresses misalignment among content, ID, and interaction representations [22], [23]. Together, these approaches show that the usefulness of multimodal information depends not merely on feature availability but on how content is connected to collaborative learning.

Robustness, alignment, and modality-sensitive learning. Recent models increasingly treat multimodal recommendation as an optimization and signal-quality problem rather than a simple fusion problem. Mirror Gradient targets flat minima to improve robustness to noisy multimodal inputs [24]; MENTOR uses multi-level self-supervision while retaining interaction information [25]; SMORE performs spectrum-based modality representation fusion [26]; and COHESION uses dual-stage composite graph fusion [27]. These methods differ substantially in how modality information reaches the final objective. That diversity is directly relevant to our study: a capacity intervention can only matter to the extent that the manipulated representation participates in the learning path that determines recommendation scores.

Do modalities actually help? The most direct motivation for our hypothesis comes from recent diagnostic work. Ye et al. evaluate a broad set of multimodal recommenders and show that whole-modality usefulness does not imply equal usefulness of individual modalities; text is often the stronger non-collaborative signal in the evaluated e-commerce settings [2]. Their IJMIR reproducibility study reinforces the need to test the contribution of multimodal embeddings rather than assume it [3]. Pomo et al. similarly examine whether multimodal systems genuinely exploit semantic content, while Zhou et al. show that modality utility is conditional on domain and setting rather than universal [28], [29]. The literature therefore supports *diagnosing modality importance*, but it does not establish that a fixed a priori reallocation of projection width will improve ranking. That missing link defines the question tested here.

3 Problem Formulation and Method

3.1 Hypothesis and capacity manipulation

Let $x_i^{(v)} \in \mathbb{R}^{d_v}$ and $x_i^{(t)} \in \mathbb{R}^{d_t}$ denote the fixed visual and textual feature vectors of item i. Each modality is projected through a two-layer branch with hidden width h_m and common output dimension $d = 64$:

$$\mathbf{e}_i^{(m)} = W_2^{(m)}\, \phi\left(W_1^{(m)}\mathbf{x}_i^{(m)} + \mathbf{b}_1^{(m)}\right) + \mathbf{b}_2^{(m)}, \qquad m \in \{v, t\}. \tag{1}$$

The fixed source dimensions are $d_v = 4096$ and $d_t = 384$. The control uses $h_v = h_t = 64$. The treatment prioritizes text with $h_v = 58$ and $h_t = 128$. The first projection-layer weight count is therefore identical:

$$4096h_v + 384h_t = 286,720. \tag{2}$$

The hypothesis is that the text-prioritized allocation improves Recall@K and NDCG@K relative to the symmetric allocation on a majority of the tested e-commerce datasets. The pre-specified decision rule requires both a positive treatment-control difference and paired-test significance at $p < 0.01$ on a majority of datasets. The inferential target is therefore narrow: the effect of reallocating hidden width while holding the first-layer input-projection weight count fixed. Output dimension, graph construction, the main ranking objective, training schedule, and evaluation protocol are otherwise unchanged between the two multimodal conditions. Equation (2) should not be interpreted as full-module parameter matching: including biases and second projection layers yields 295,168 parameters in the control and 298,938 in the treatment (+1.277%). Accordingly, H1 tests this operational capacity-allocation hypothesis within the specified architecture; it is not a general claim that textual information is intrinsically more valuable than visual information.

3.2 Recommendation architecture and objective

The final implementation follows the FREEDOM-style separation between content-derived graph construction and collaborative propagation [4]. Raw pretrained visual and textual features construct modality-specific frozen k-nearest-neighbor item graphs, which are combined using a fixed image/text weight of 0.1/0.9. The frozen graph propagates trainable item-ID embeddings rather than the live outputs of Eq. (1). In parallel, a LightGCN-style user-item branch propagates collaborative user and item embeddings [5]. The recommendation score is computed from the resulting user and item representations.

The main ranking objective is BPR [12]. The projected modality representations in Eq. (1) enter two auxiliary BPR-

Table 1: Dataset characteristics.

Dataset	Users	Items	Interactions	Density (%)	Interactions/item	Share on items ≤ 10 events (%)
Baby	19,445	7,050	160,792	0.1173	22.81	14.23
Sports	35,598	18,357	296,337	0.0453	16.14	24.41
Clothing	39,387	23,033	278,677	0.0307	12.10	36.83

style terms rather than the normal inference-time score. Training minimizes

$$\mathcal{L} = \mathcal{L}_{main} + \lambda_{aux} \left(\mathcal{L}_v + \mathcal{L}_t\right), \qquad \lambda_{aux} = 10^{-4}. \quad (3)$$

Because the auxiliary losses share user representations with the recommender, the modality projections can affect learned collaborative embeddings indirectly during training. They do not, however, constitute a direct scoring branch. This architectural fact becomes central to interpreting the negative result.

3.3 Diagnostic interventions

Two diagnostics are kept conceptually separate from the confirmatory H1 test. First, because directly zeroing the live projection outputs would be causally inert at inference, the modality-knockout diagnostic reconstructs the frozen graph after ablating one raw modality and measures the resulting ranking change. This probes the utility of each modality in the fixed multimodal graph, not the effect of projection width.

Second, after the pre-specified confirmatory analysis was completed, we analyzed the saved best checkpoints for all 15 matched dataset-seed control-treatment pairs. For a common deterministic sample of 2,048 items, we measured projection norms, hidden-unit activity, representation similarity, and normalized Frobenius differences. On a deterministic interaction batch of at most 512 examples, we decomposed gradients from the main BPR objective and the weighted auxiliary objective to projection parameters, user embeddings, and item-ID embeddings. A fixed full-sort score probe was also differentiated with respect to projection parameters. These diagnostics were not pre-specified and are therefore interpreted as explanatory, not confirmatory, evidence.

4 Experimental Design

4.1 Datasets and comparators

We use three public Amazon-review-derived e-commerce domains associated with established multimodal recommendation benchmarks [30]. The processed user-item interactions, text and image feature representations, and fixed per-user 80/10/10 train/validation/test splits are held identical across conditions. Restricting evaluation to three domains from the same benchmark family intentionally controls preprocessing, representation, and splitting conventions, allowing dataset-level differences to be examined within a common e-commerce setting; it does not by itself support cross-domain generalization. Table 1 summarizes the interaction structure. Clothing is the sparsest of the three domains and has the largest share of interactions attached to items with at most ten observations.

The confirmatory comparison is symmetric control versus text-prioritized treatment. An unmodified LightGCN serves as a strong non-multimodal diagnostic floor [5] because it retains collaborative user-item graph propagation while removing the content-derived multimodal component. This makes the secondary comparison a direct test of whether the multimodal machinery adds value beyond a strong collaborative signal on the same data splits and ranking metrics, rather than an exhaustive state-of-the-art benchmark exercise [6]. Control and treatment use two user-item propagation layers; the LightGCN comparator retains its comparator-specific three-layer configuration rather than being weakened solely to match depth. All three conditions are evaluated with Recall@10, Recall@20, NDCG@10, and NDCG@20.

4.2 Training, model selection, and inference

Each condition is trained with seeds 42, 123, 2024, 7, and 999, yielding 45 final runs. Control-treatment inference is paired by dataset and seed. The paired replicate is the unit of the H1 significance test within each dataset: the test therefore characterizes variability from stochastic optimization under a fixed dataset and split, not population-sampling uncertainty across users, datasets, or application domains. Early stopping and best-checkpoint selection use validation Recall@20 with patience 20, exactly matching the pre-specified analysis rule, and the maximum training horizon is 1000 epochs. For each dataset-metric pair, treatment-control differences are tested with a two-sided paired t-test over the five matched seeds using the pre-specified $p < 0.01$ decision threshold. Given only five paired replicates, raw paired differences and confidence intervals are emphasized alongside p-values, consistent with recommendations to avoid interpreting a non-significant test as proof of equality [6].

Post-hoc statistical analyses include Cohen's paired d_z, 95% and 99% confidence intervals, leave-one-seed-out influence checks, equivalence-test sensitivity over several relative margins, and a Bayesian region-of-practical-equivalence sensitivity analysis. These analyses characterize effect magnitude, uncertainty, and sensitivity; they do not redefine or rescue the pre-specified H1 decision. No smallest effect size of interest was specified before final testing, so the equivalence results remain descriptive sensitivity analyses rather than confirmatory evidence.

4.3 Implementation verification and repeated experiments

Verification against the reference architecture and the pre-specified protocol identified three discrepancies relevant to construct validity or matched-seed fairness. First, the initial multimodal graph propagation did not reproduce the intended FREEDOM-style separation between raw-feature graph construction and item-ID propagation. Second, control and treatment initialized shared collaborative embeddings at different points in the random-number sequence, confounding matched-seed comparisons. Third, the training implementation

Table 2: Confirmatory H1 summary. Relative ranges cover the four ranking metrics within each dataset.

Dataset	Relative change	p-value range	$p < 0.01$	Interpretation
Baby	−18.07% to −17.98%	0.391–0.404	0/4	Influential valid checkpoint trajectory; no stable directional claim
Sports	−0.095% to +0.116%	0.358–0.997	0/4	Near-zero mean differences
Clothing	−0.067% to +0.226%	0.523–0.915	0/4	Near-zero mean differences

used validation NDCG@20 for early stopping and checkpoint selection although the pre-specified analysis plan specified Recall@20.

The correction strategy was conservative: each change restored a design element that had already been specified rather than introducing a performance-driven tuning choice. Results affected by the first two discrepancies were excluded and the corresponding experiments were repeated. The stopping-metric discrepancy was discovered after a superseded set of test outcomes had been inspected; however, Recall@20 was restored because it was fixed in the earlier analysis plan, not selected from alternative rules after observing the results. The full 45-run experiment was therefore repeated under Recall@20 model selection, and all confirmatory and secondary analyses reported here use only this final validated set. Earlier results are retained solely as reproducibility provenance and are not part of the evidential basis of the paper, consistent with concerns about reproducibility and baseline discipline in recommender-system evaluation [7], [8].

5 Results

5.1 Asymmetric capacity does not provide a reliable accuracy gain

Figure 1 reports the paired treatment-control effects for all 12 pre-specified dataset-metric comparisons. None reaches the locked $p < 0.01$ threshold. On Sports and Clothing, mean relative differences are small, ranging from -0.095% to +0.226%. Baby shows a much larger negative mean of approximately 18% across metrics, but its confidence intervals are wide and all four tests remain non-significant ($p = 0.391$–0.404).

The Baby mean requires particular care. The treatment run at seed 123 selects epoch 1 as its best Recall@20 checkpoint, whereas the paired control selects epoch 250. This run is valid under the model-selection rule and shows no numerical failure, so it remains in the confirmatory analysis. A post-hoc leave-one-seed-out check shows that removing seed 123 changes the Baby treatment effect to approximately +0.7% to +1.3% depending on the metric, but still does not produce a significant result. Thus, the evidence supports neither a reliable benefit nor a stable harmful effect of asymmetric capacity on Baby; it identifies sensitivity to model selection.

The post-hoc equivalence analysis is consistent with near-zero effects on Sports and Clothing but not with a global equivalence claim. At a relative +/-1% margin, all four metrics on Sports and all four on Clothing satisfy paired TOST, whereas none on Baby does. Because that margin was selected after the experiment, these results describe practical-effect sensitivity rather than establish confirmatory equivalence.

5.2 Multimodal utility is strongly dataset-dependent

The non-multimodal floor changes the interpretation of the study. Averaged across the four ranking metrics, the symmetric multimodal control is about 15.4% below LightGCN on Baby and 24.1% below on Sports, but 43.0% above it on Clothing. The treatment shows the same Sports/Clothing direction. Table 3 summarizes the comparisons; the unstable Baby treatment trajectory again inflates variance and magnitude.

The graph-level modality knockout provides a compatible descriptive pattern. Removing text improves Recall@10 on Baby and Sports but reduces it on Clothing; image removal has effects close to zero. Clothing is also the sparsest and most item-cold dataset in Table 1. With only three datasets, these observations do not identify a dataset-level causal mechanism. They do, however, show that the assumption "multimodal is better" is not transportable even across closely related Amazon domains, consistent with recent critical evaluations of multimodal recommendation [2], [28], [29].

5.3 The capacity intervention is strong locally but weakly connected to the ranking objective

The post-hoc checkpoint diagnostic rules out a simple explanation that the treatment failed because the wider text branch was inactive. The treatment changes visual projection representations by roughly 77%–97% and text projections by 115%–125% in normalized Frobenius distance, while treatment text hidden-unit activity is approximately 99%–100%. Thus, the manipulated capacity is used and the representation-level intervention is substantial.

The structural gradient result is nevertheless unequivocal: for all 15 matched dataset-seed probes, both the main recommendation loss and the direct score produce exactly zero gradient with respect to projection parameters, while the auxiliary loss produces a non-zero projection gradient. Moreover, the weighted auxiliary objective is only about 0.017%–0.020% of the main loss. Its contribution to the user-embedding gradient is about 0.20%–0.60%, and to the item-embedding gradient about 0.08%–0.27%. These quantities confirm an auxiliary-only, weak instantaneous route from projection capacity to the collaborative model.

Weak instantaneous influence does not imply identical final checkpoints. Twelve of 15 control-treatment pairs select the same best epoch; in those pairs, shared-embedding differences remain small. Three pairs select different epochs and show much larger endpoint divergence. The most extreme case is Baby seed 123, where the best epochs differ by 249 and the user/item embedding differences approach 96%/92%. Across the 15 pairs, absolute best-epoch difference is positively associated with endpoint embedding divergence (post-hoc Spearman $\rho = 0.694$, $p = 0.0041$).

This combination of evidence suggests a two-stage inter-

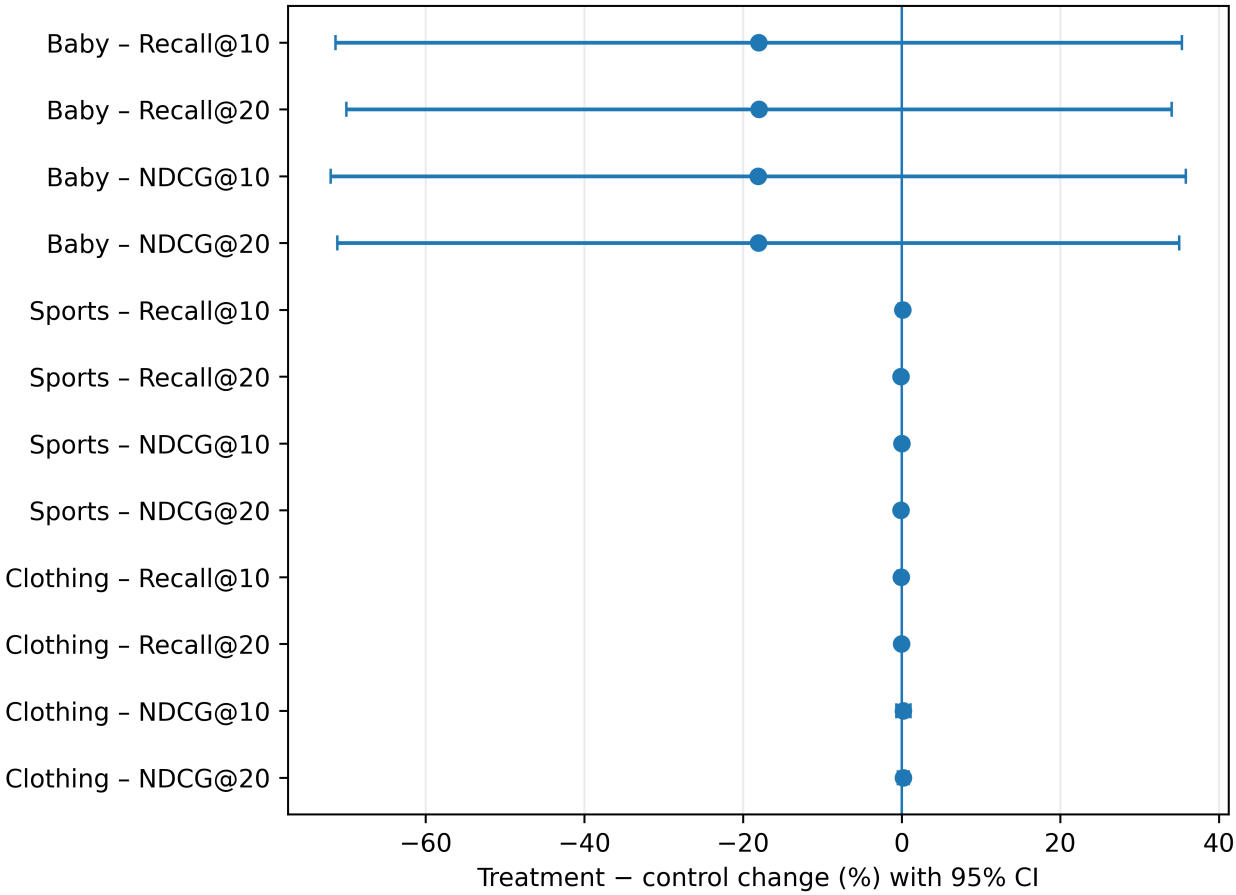


Figure 1: Treatment-control relative changes and 95% confidence intervals for the 12 pre-specified H1 comparisons. Positive values favor text-prioritized asymmetric capacity.

Table 3: Multimodal conditions relative to the LightGCN floor. Values are mean relative differences across Recall@10, Recall@20, NDCG@10, and NDCG@20.

Dataset	Symmetric control vs LightGCN	Significant metrics	Asymmetric treatment vs LightGCN	Significant metrics
Baby	−15.38%	4/4	−30.64%	0/4
Sports	−24.08%	4/4	−24.09%	4/4
Clothing	+43.01%	4/4	+43.14%	4/4

pretation. The architectural channel from projection capacity to the ranking objective is weak at any given training state, but discrete model selection can amplify small optimization perturbations when validation trajectories cross different early-stopping boundaries.

6 Discussion

6.1 Modality informativeness is not a capacity-allocation rule

The central negative result exposes a distinction that is easy to overlook in multimodal model design. A modality knockout asks whether removing a signal damages the system. Projection-width reallocation asks whether a larger parameterized transformation of that signal improves the learned ranking function. These are not equivalent interventions. Prior diagnostic work correctly shows that text can be more informative than image in many e-commerce settings [2], [3], but that observation alone does not imply that text should receive a larger projection branch.

Our results provide a concrete counterexample to that inference. The treatment substantially changes the modality-specific representation and uses its additional capacity, yet the downstream ranking metrics do not improve reliably. The implication is not that modality-aware capacity is generally useless. Rather, capacity should be treated as an architectural resource whose value depends on where it sits in the computational graph and which objective can train it. Modality importance is evidence about *signal value*; capacity allocation additionally requires evidence about *optimization access*.

6.2 Causal access matters more than nominal width

The gradient diagnostic clarifies why simply increasing the asymmetry further would be an incomplete response to the null result. The primary ranking loss cannot directly update the projection module, and the inference score does not consume its output. The projection branch acts through a small auxiliary term that perturbs shared user representations during training. Under this topology, a larger branch may produce very different

Table 4: Post-hoc causal-channel diagnostic (means across five seeds).

Dataset	Visual proj. diff.	Text proj. diff.	Weighted aux/main loss	Aux/main user-grad	Aux/main item-grad	Same-epoch user emb. diff.
Baby	89.5%	124.6%	0.0179%	0.231%	0.095%	0.53%
Sports	97.3%	122.0%	0.0180%	0.494%	0.219%	0.89%
Clothing	76.9%	114.9%	0.0174%	0.355%	0.182%	1.67%

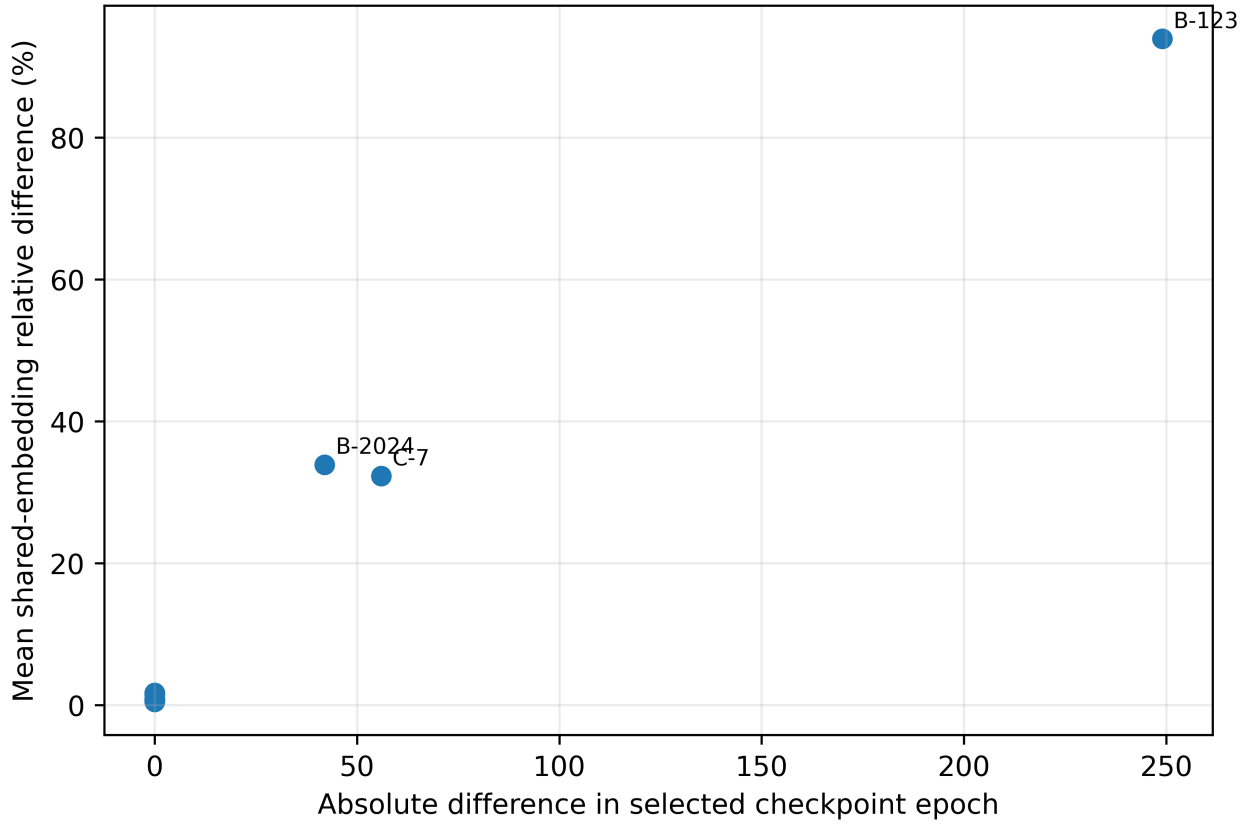


Figure 2: Absolute difference between selected control and treatment checkpoint epochs versus mean relative difference in shared user/item embeddings. Labels identify the three pairs with different selected endpoints.

local representations without exerting a proportionate force on the parameters that dominate recommendation.

This observation has a practical consequence for architecture design. Before performing capacity sweeps, neural architecture search, or modality-specific width tuning, researchers should trace the gradient route from the manipulated component to the target score. A branch that receives only an auxiliary signal may require a different experimental question: not "how wide should the branch be?" but "how directly should modality information participate in scoring or representation updates?" Architectures such as AlignRec, MENTOR, and other alignment-oriented systems make this coupling more explicit [23], [25], suggesting a more appropriate setting for future tests of capacity allocation.

6.3 Model selection can amplify weak training differences

The Baby seed-123 trajectory reveals a second lesson. A weak local causal channel can still matter indirectly when model selection is discrete. Small differences in validation trajectories may change which checkpoint survives early stopping; once different epochs are selected, endpoint parameters can be far apart. This explains why the Baby aggregate appears dramatically negative despite the near-zero treatment-control differences on Sports and Clothing and despite the weak instantaneous gradient pathway.

The appropriate response is not to delete the influential run. It is protocol-valid and therefore part of the estimand defined by the training and selection procedure. Removing it after observing the result would change the analysis population. Instead, the run should be retained while its influence is disclosed. More broadly, multimodal recommendation studies with small seed counts should report checkpoint-selection behavior when apparently large mean effects are driven by one training trajectory. Evaluation guidance has long emphasized that model-selection choices are part of the experimental system rather than a neutral post-processing detail [6], [7].

6.4 Multimodal value is conditional on the dataset

The LightGCN comparisons show that the more consequential empirical question may be whether multimodal content helps at all in a particular domain. The multimodal model is strongly favorable on Clothing yet unfavorable on Sports and, in mean performance, Baby. The text-knockout direction follows the same qualitative pattern: removing text helps on Baby and Sports but hurts on Clothing. The sparsest dataset is also the

one where multimodal content is most beneficial.

This pattern is compatible with the view that content becomes more useful when collaborative evidence is scarce, but three datasets are insufficient to identify that relationship. Feature quality, domain semantics, catalog composition, missingness, redundancy between text and image, and the interaction distribution can all contribute. Recent work similarly argues that multimodal gains depend on architecture and context rather than following a universal rule [28], [29]. The practical lesson is therefore comparative: every multimodal experiment should include a strong non-multimodal floor and report modality-specific diagnostics before attributing gains to multimodal fusion.

7 Limitations

The study is deliberately narrow. First, it evaluates one corrected FREEDOM-style architecture family on three Amazon e-commerce domains. The results do not generalize automatically to graph-free, gating-based, adversarial, short-video, or foundation-model-based multimodal recommenders. Second, five seeds per cell provide a matched design but limited power for distributional claims; the influential Baby trajectory demonstrates why confidence intervals and per-seed inspection remain important. Third, the causal-channel, equivalence, Bayesian, and influence analyses were designed after the confirmatory result and must remain explanatory. Fourth, the manipulated branch is connected to the ranking model only through an auxiliary objective. The experiment therefore tests asymmetric capacity *under that auxiliary-only topology*, not asymmetric capacity as a universal design principle. Fifth, the feature encoders and graph-construction settings are fixed; alternative encoders, learned modality gates, or direct scoring paths could change the result. Finally, the dataset-level sparsity/text-utility interpretation is based on only three domains and should be evaluated on a larger, pre-specified collection before being presented as a mechanism. These constraints reinforce broader recommendations to treat recommender-system conclusions as conditional on data, baselines, and evaluation design [6], [8].

A further reproducibility limitation is procedural. The protocol materials were hash-locked locally before execution, but not deposited in an independent preregistration registry, and the original code state was not tied to a version-control commit. The final evidence is instead supported by preserved configurations, experiment metadata records, checkpoint metadata, implementation-correction records, tests, and result files. These materials make the reported analysis auditable, but prospective external registration and commit-level provenance would strengthen a future study.

8 Conclusion

A modality can be important without benefiting from a wider projection branch. In a controlled FREEDOM-style multimodal recommender, text-prioritized asymmetric capacity did not produce a reliable accuracy advantage over a symmetric control under a matched first-layer weight count. The explanation is architectural rather than merely statistical: the manipulated projections change substantially and use the added capacity, but they have no direct gradient path from the primary ranking loss or test-time score. Their influence is mediated through a weak auxiliary objective, while discrete checkpoint selection can occasionally amplify small training differences into large endpoint divergence.

The non-multimodal comparison adds a broader result. Multimodal information is strongly beneficial on Clothing but detrimental in mean performance on Baby and Sports, and text-knockout direction mirrors that dataset dependence. These findings separate three questions that are often conflated: whether a modality is informative, whether the model allocates capacity to it, and whether the architecture gives that capacity a direct and stable route to affect ranking. Future capacity-allocation studies should begin by establishing the third condition rather than assuming it from the first.

Statements and Declarations

Ethics approval and consent to participate. Not applicable. The study uses previously released, de-identified benchmark data and does not recruit human participants or deploy recommendations to live users.

Consent for publication. Not applicable.

Data availability. The experiments use public Amazon-derived benchmark data in the processed form commonly used by the evaluated multimodal-recommendation architecture. Dataset redistribution is subject to the original source and benchmark licenses. The exact split and data-integrity metadata used for the reported experiments are preserved with the study materials.

Code and materials availability. Configuration files, analysis scripts, experiment metadata records, statistical summaries, and the implementation-correction records have been retained in a reproducibility package. Public release of code and derived research materials should follow the applicable source-code and dataset licenses; materials can be supplied for peer review.

Author contributions. Emin Talip Demirkiran is the sole author and takes responsibility for the conception and design of the study, analysis and interpretation of the experimental results, and preparation and revision of the manuscript.

Funding. The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests. The author declares no competing interests.

References

[1] Q. Liu *et al.*, "Multimodal recommender systems: A survey," *ACM Computing Surveys*, vol. 57, no. 2, 2025, doi: 10.1145/3695461.

[2] Y. Ye, J. Fu, Y. Song, K. Zheng, and J. M. Jose, "Are multimodal embeddings truly beneficial for recommendation? A deep dive into whole vs. Individual modalities," in *Advances in information retrieval: 48th european conference on information retrieval (ECIR 2026), proceedings, part III*, in Lecture notes in computer science, vol. 16485. Springer Nature Switzerland, 2026, pp. 66–81. doi: 10.1007/978-3-032-21324-2_5.

[3] Y. Ye, J. Fu, Y. Song, K. Zheng, and J. M. Jose, "A reproducibility study of multimodal embeddings for recommender systems," *International Journal of Multimedia Information Retrieval*, vol. 15, no. 2, 2026, doi: 10.1007/s13735-026-00408-9.

[4] X. Zhou and Z. Shen, "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation," in *Proceedings of the 31st ACM international conference on multimedia*, ACM, 2023, pp. 935–943. doi: 10.1145/3581783.3611943.

[5] X. He, K. Deng, X. Wang, Y. Li, Y.-D. Zhang, and M. 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, ACM, 2020, pp. 639–648. doi: 10.1145/3397271.3401063.

[6] E. Zangerle and C. Bauer, "Evaluating recommender systems: Survey and framework," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, 2022, doi: 10.1145/3556536.

[7] M. F. Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? A worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM conference on recommender systems*, ACM, 2019, pp. 101–109. doi: 10.1145/3298689.3347058.

[8] J. Beel, C. Breitinger, S. Langer, A. Lommatzsch, and B. Gipp, "Towards reproducibility in recommender-systems research," *User Modeling and User-Adapted Interaction*, vol. 26, no. 1, pp. 69–101, 2016, doi: 10.1007/s11257-016-9174-x.

[9] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005, doi: 10.1109/TKDE.2005.99.

[10] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009, doi: 10.1109/MC.2009.263.

[11] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 eighth IEEE international conference on data mining*, IEEE, 2008, pp. 263–272. doi: 10.1109/ICDM.2008.22.

[12] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, AUAI Press, 2009, pp. 452–461.

[13] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, International World Wide Web Conferences Steering Committee, 2017, pp. 173–182. doi: 10.1145/3038912.3052569.

[14] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, ACM, 2019, pp. 165–174. doi: 10.1145/3331184.3331267.

[15] J. Wu *et al.*, "Self-supervised graph learning for recommendation," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, ACM, 2021, pp. 726–735. doi: 10.1145/3404835.3462862.

[16] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, and Q. V. H. Nguyen, "Are graph augmentations necessary?: Simple graph contrastive learning for recommendation," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, ACM, 2022, pp. 1294–1303. doi: 10.1145/3477495.3531937.

[17] R. He and J. McAuley, "VBPR: Visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI conference on artificial intelligence*, 2016. doi: 10.1609/aaai.v30i1.9973.

[18] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM international conference on multimedia*, ACM, 2019, pp. 1437–1445. doi: 10.1145/3343031.3351034.

[19] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, "DualGNN: Dual graph neural network for multimedia recommendation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1074–1084, 2023, doi: 10.1109/TMM.2021.3138298.

[20] X. Zhou *et al.*, "Bootstrap latent representations for multi-modal recommendation," in *Proceedings of the ACM web conference 2023*, ACM, 2023, pp. 845–854. doi: 10.1145/3543507.3583251.

[21] P. Yu, Z. Tan, G. Lu, and B.-K. Bao, "Multi-view graph convolutional network for multimedia recommendation," in *Proceedings of the 31st ACM international conference on multimedia*, ACM, 2023, pp. 6576–6585. doi: 10.1145/3581783.3613915.

[22] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "LGMRec: Local and global graph learning for multimodal recommendation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 8454–8462, 2024, doi: 10.1609/aaai.v38i8.28688.

[23] Y. Liu *et al.*, "AlignRec: Aligning and training in multimodal recommendations," in *Proceedings of the 33rd ACM international conference on information and knowledge management*, ACM, 2024, pp. 1503–1512. doi: 10.1145/3627673.3679626.

[24] S. Zhong, Z. Huang, D. Li, W. Wen, J. Qin, and L. Lin, "Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima," in *Proceedings of the ACM web conference 2024*, ACM, 2024, pp. 3700–3711. doi: 10.1145/3589334.3645553.

[25] J. Xu, Z. Chen, S. Yang, J. Li, H. Wang, and E. C. H. Ngai, "MENTOR: Multi-level self-supervised learning for multimodal recommendation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, pp. 12908–12917, 2025, doi: 10.1609/aaai.v39i12.33408.

[26] R. K. Ong and A. W. H. Khong, "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*, ACM, 2025, pp. 773–781. doi: 10.1145/3701551.3703561.

[27] J. Xu, Z. Chen, W. Wang, X. Hu, S.-W. Kim, and E. C. H. Ngai, "COHESION: Composite graph convolutional network with dual-stage fusion for multimodal recommendation," in *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*, ACM, 2025, pp. 1830–1839. doi: 10.1145/3726302.3729927.

[28] C. Pomo, M. Attimonelli, D. Danese, F. Narducci, and T. D. Noia, "Do recommender systems really leverage multimodal content? A comprehensive analysis on multimodal representations for recommendation," in *Proceedings of the 34th ACM international conference on information and knowledge management*, ACM, 2025, pp. 2377–2387. doi: 10.1145/3746252.3761398.

[29] H. Zhou, Y. Zhang, A. Sun, and Z. Shen, "Does multimodality improve recommender systems as expected? A critical analysis and future directions." 2025. doi: 10.48550/arXiv.2508.05377.

[30] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, ACM, 2015, pp. 43–52. doi: 10.1145/2766462.2767755.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryIJMIRFinal.docx