OpticalRec: Unified Optical Vision-Language Representation for Multimodal Recommendation

Anonymous ACL submission

Abstract

Recent advances in vision-language modeling have substantially improved multimodal encoding, retrieval and reasoning. Yet for multimodal recommendation, effectively encoding vision-language semantics remains a long-standing bottleneck. Mainstream approaches primarily adopt independent encoding of each modality followed by late fusion, inherently inducing cross-modal semantic distortion and subpar recommendation performance. To address this challenge, we propose OpticalRec, the first pixel-space unified encoding paradigm for multimodal collaborative filtering, a fundamental recommendation setting. Instead of separating multimodality, OpticalRec renders structured item image and text into a unified pixel-space to preserve intrinsic vision-language correlations. A single encoder then derives homogeneous item embeddings directly from the unified pixel-space, avoiding semantic distortion caused by late fusion. Extensive experiments across multiple benchmarks and downstream recommendation models demonstrate that OpticalRec consistently outperforms state-of-the-art baselines. Furthermore, theoretical analysis via mutual information and matrix effective rank confirms its efficacy in preserving fine-grained vision-language variances. As a plug-and-play module, OpticalRec integrates seamlessly into existing multimodal collaborative filtering and potentially broader recommendation frameworks without major architectural modifications.

1 Introduction

Recent progress in vision-language modeling and multimodal learning has opened new opportunities for recommender systems (OpenAI, 2026; Li et al., 2024a; Bai et al., 2025). These improvements mainly stem from two aspects: first, textual and visual representations offer rich semantic signals that help characterize user interests and item properties, such as product images and textual descriptions; second, the general representation and reasoning abilities learned during large-scale vision-language pretraining can be effectively transferred to recommendation tasks (Malitesta et al., 2025; Xu et al., 2025). As a result, multimodal representations derived from vision-language models have become a powerful basis for improving recommendation quality, contextual understanding, and personalization in modern recommender systems (Pomo et al., 2025; Ye et al., 2025a).

Despite above mentioned developments of multimodality and their implied transfer-learning capabilities, injecting rich vision-language features into recommender systems continues to pose significant challenges. A fundamental bottleneck lies in the decoupled encoding of each modality followed by lossy late fusion, which distorts cross-modal semantics and weakens the resulting representation. For example, e-commerce items naturally organize images and textual metadata into visually grounded layouts for human browsing. Mainstream multimodal collaborative filtering recommenders independently encode each modality and integrate them through late fusion for downstream recommendation (Xu et al., 2025; Malitesta et al., 2025). Specifically, summation or concatenation-based late fusion is recognized to distort modality-specific representations (Zhou et al., 2025a); Graph-based methods exploit multimodal fusion by similarity graphs, but initial graph construction still falls back to separate encoding with late fusion (Zhang et al., 2021; Zhou and Shen, 2023). Recent works continues to refine downstream graph architectures while overlooking information loss during multimodal encoding and fusion (Ong and Khong, 2025; Yu et al., 2025). Hence, a unified-space encoding paradigm is optimal for multimodal recommendation that satisfies:

1. Multimodal content should be presented in a unified input space that preserves vision-

language layout structures and correlation.

2. The unified input should be encoded once into a homogeneous embedding space, without additional late fusion.

Motivated by these observations and recent advances in vision-language grounding (Bai et al., 2025; OpenGV, 2025), we propose OpticalRec, a lightweight, plug-and-play pixel-space unified encoding framework for multimodal recommendation. Through a one-time rendering process that transforms textual metadata into visual text, OpticalRec preserves the original optical layout of images and text within a unified pixel space, consistent with how users visually browse online products. This also follows a "see-as-users-see" principle: multimodal content should be encoded in the same integrated visual form in which users naturally perceive it. A single vision-language encoder then derives homogeneous multimodal embeddings directly from the unified optical presentation. Such unified encoding preserves fine-grained vision-language correlations often disrupted by separate encoding and late fusion. These benefits translate into consistent improvements across benchmark datasets supported by theoretical analysis of mutual-information and matrix effective rank. We summarize our contribution as follows:

1. We propose OpticalRec, the first pixel-space unified encoding framework for multimodal collaborative filtering that eliminates late fusion.

2. We provide theoretical analysis based on mutual information and matrix effective rank, showing that unified pixel-space encoding preserves richer vision-language correlation with reduced information loss.

3. Experiments across multiple Amazon Review benchmarks and recommendation backbones demonstrate OpticalRec's consistent improvements over strong multimodal baseline encoders.

4. OpticalRec is lightweight and plug-and-play, making it readily adaptable to existing multimodal collaborative filtering frameworks.

2 Related Works

2.1 Multimodal Collaborative Filtering

Pretrained language and vision-language models are widely adopted in multimodal collaborative filtering to provide transferable visual and textual item representations (Malitesta et al., 2024; Xu et al., 2025). A major line of work adopts pretrained multimodal embeddings as frozen item features for collaborative filtering (Xin et al., 2019; Wang et al., 2021; Zhou et al., 2025b; Li et al., 2024c; Sheng et al., 2024). Early methods such as VBPR (He and McAuley, 2016) leverage CNN-based image features to model visually driven user preferences. Later methods further introduce textual features and graph neural networks for high-order multimodal interaction modeling (Chen et al., 2017; Zhou, 2023; Yu et al., 2023; Liu et al., 2023; Zhou et al., 2023a). For example, MMGCN (Wei et al., 2019) and GRCN (Wei et al., 2020) extend graph-based recommendation with modality-aware multimodal fusion over interaction graphs. Later developments such as LATTICE (Zhang et al., 2021) and FREEDOM (Zhou and Shen, 2023) construct multimodal similarity graphs from visual and textual embeddings for refined graph-based fusion. Recent methods continues to explore tailored graph structures or investigate improved training objectives for multimodal preference alignment (Zheng et al., 2024; Xu et al., 2025; Liu et al., 2025). LGMRec (Guo et al., 2024) models local and global modality correlations via hypergraphs, while BM3 (Zhou et al., 2023b) introduces self-supervised multimodal representation learning. In contrast, our work shifts the focus from increasingly specialized graph-fusion architectures to unified vision-language encoding that avoids separate modality fusion.

Beyond feature-based multimodal collaborative filtering, recent studies also explore multimodal sequential or generative recommendation (Wang et al., 2024; Ye et al., 2025b; Wang et al., 2025; Yue et al., 2025; Zhai et al., 2025). However, our work focuses on the collaborative filtering setting, where multimodal item representations are directly integrated with collaborative signals, providing a cleaner setup for studying vision-language encoding and fusion.

2.2 Vision-Language Grounding and Unified Encoding

Feature-based multimodal recommendation predominantly follows a separate-encoding paradigm, where visual and textual features are independently extracted and later fused through summation, concatenation, or graph-based content similarity modeling (Ye et al., 2025a; He et al., 2025). This

paradigm aligns with early vision-language representation learning such as CLIP, which learns shared image-text embeddings but still encodes each modality separately. Later models such as Flamingo and LLaVA move toward interleaved vision-language modeling, allowing visual and textual inputs to be jointly processed. More recent models such as Qwen-VL (Bai et al., 2025), InternVL (OpenGV, 2025), and LLaVA model series (Li et al., 2024a,b) further strengthen visual grounding, layout-aware understanding, and unified perception over images, text, and spatial structures. However, how such unified vision-language perception can benefit multimodal collaborative filtering remains unexplored, as user-item collaborative signal is inherently multifaceted. As such, OpticalRec builds a unified optical encoding paradigm to preserve vision-language correlations and improve recommendation performance.

3 Methodologies

3.1 Multimodal Collaborative Filtering Recommendation

We briefly introduce the Multimodal Collaborative Filtering (MMCF) task. Specifically, Let $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$ and $\mathcal{I} = \{i_1, i_2, \ldots, i_N\}$ denote the sets of M users and N items, respectively. Each item i is described by an image $\mathbf{v}_i$ and textual description $\mathbf{t}_i$, its multimodal embedding is derived by applying a multimodal encoder f; user information is usually unavailable given ethical requirements thus derived from interacted items:

$$\mathbf{q}_i = f(\mathbf{v}_i, \mathbf{t}_i), \quad \mathbf{p}_u = \mathcal{A}(\{\mathbf{q}_i\}_{i \in \mathcal{N}_u}) \tag{1}$$

where $\mathcal{N}_u$ denotes item set user u interacted with and $\mathcal{A}(\cdot)$ is an aggregation function such as mean pooling. A recommendation model $\mathcal{F}(\cdot)$ is then applied to transform $\mathbf{p}_u$, $\mathbf{q}_i$, and user-item interaction graph $\mathcal{G}$ to final representations $\mathbf{p}'_u \in \mathbb{R}^d$ and $\mathbf{q}'_i \in \mathbb{R}^d$ bearing user-item preference signal:

$$(\mathbf{p}'_u, \mathbf{q}'_i) = \mathcal{F}(\mathbf{p}_u, \mathbf{q}_i, \mathcal{G}) \tag{2}$$

The final recommendation score, i.e., the likelihood that item i is recommended to user u is defined as their inner-product:

$$y_{ui} = \langle \mathbf{p}'_u, \mathbf{q}'_i \rangle \tag{3}$$

3.2 OpticalRec as Pixel-Space Unified Encoding

In Section 3.1, we have introduced multimodal collaborative filtering (MMCF). Instead of pursuing increasingly tailored designs of downstream recommender $\mathcal{F}(\cdot)$, as in mainstream approaches, we address multimodal recommendation from its source: the encoding of multimodal item content $f(\mathbf{v}_i, \mathbf{t}_i)$, where $\mathbf{v}_i$, $\mathbf{t}_i)$ is image-text input of item i. As discussed in Section 1, commonly adopted independent encoding with late fusion suffers from embedding distortion and feature corruption. We therefore propose OpticalRec, the first unified pixel-space encoding for multimodal recommendation that is late-fusion free and yields richer multimodal embeddings.

We now formalize and compare the two paradigms (1) Independent encoding with Late fusion and (2) Early fusion with Unified encoding. Both pipelines can be expressed as Markov chains as current output depends solely on the previous step. Independent encoding process each modality separately with vision and language encoders f_V, f_T into modality-specific vectors $(\mathbf{z}_i^v, \mathbf{z}_i^t)$ before applying a fusion operator g:

$$(\mathbf{v}_i, \mathbf{t}_i) \xrightarrow{f_V, f_T} \mathbf{z}_i^{\text{LF}} \triangleq (\mathbf{z}_i^v, \mathbf{z}_i^t) \xrightarrow{g} \mathbf{q}_i^{\text{LF}} \tag{4}$$

where $\mathbf{z}_i^{\text{LF}}$ denotes output of two independent encoders requiring late fusion and $\mathbf{q}_i^{\text{LF}}$ means late-fused final embedding.

OpticalRec as unified encoding first applies an efficient rendering operation $\mathcal{R}(\cdot)$ that transforms into a homogeneous pixel space $\mathbb{R}^d$ all optical elements such as (1) item image $\mathbf{v}_i$, (2) item textual description $\mathbf{t}_i$, and (3) image-text layout structure $\mathbf{s}_i^{vt}$ following visual organization of product pages:

$$\tilde{\mathbf{v}}_i = \mathcal{R}(\mathbf{v}_i, \mathbf{t}_i, \mathbf{s}_i^{vt}) \tag{5}$$

The rendering operation is performed once offline and introduces negligible overhead during recommendation discussed in Section 6. A rendered item card example is shown in Figure 1. Then, the unified input $\tilde{\mathbf{v}}_i$ is mapped to the final embedding $\mathbf{z}_i^{\text{OR}} = \mathbf{q}_i^{\text{OR}}$ using a single vision-language encoder f_Φ. The whole OpticalRec encoding is illustrated as:

$$(\mathbf{v}_i, \mathbf{t}_i, \mathbf{s}_i^{vt}) \xrightarrow{\mathcal{R}} \tilde{\mathbf{v}}_i \xrightarrow{f_\Phi} \mathbf{z}_i^{\text{OR}} = \mathbf{q}_i^{\text{OR}} \tag{6}$$

As a result, OpticalRec directly produces a unified item representation without post-hoc late fusion,

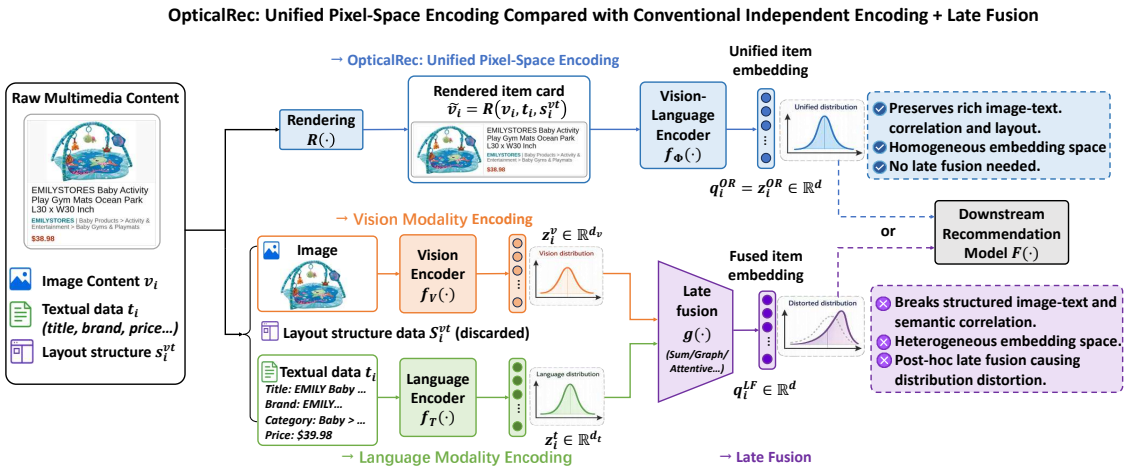

Figure 1: Comparing OpticalRec: Unified Pixel-Space Encoding with Conventional Independent Encoding + Late Fusion. Independent encoding discards multimodal content structure and separately encodes each modality followed by late fusion, causing semantic distortion and feature corruption; OpticalRec efficiently renders item text into the pixel space together with item image, preserving rich content structure and vision-language correlation. Through a unified encoder, final item embedding resides in a homogeneous space favorable for downstream recommendation.

preserving image-text-layout correlations. We further analyze this effect via mutual information gap analysis in Section 3.3.

3.3 Mutual Information Gap between the Unified and Late-Fusion Encoding

In this section, we measure encoded representation informativeness of OpticalRec and late-fusion-based encoding discussed in Section 3.2. Specifically, we compare the information differences in each derived embedding by introducing the mutual information gap. Let $\mathcal{I}_{\text{cross}}^{\text{OR}}$ denote the cross-modal interaction information captured by OpticalRec's unified encoding, and let $\mathcal{D}_{\text{intra}}^{\text{LF}}$ denote the information lost when late fusion operation g compress independently encoded vectors $\mathbf{z}_i^{\text{LF}} \triangleq (\mathbf{z}_i^v, \mathbf{z}_i^t)$ into final late-fused representation $\mathbf{q}_i^{\text{LF}}$. The information gap then decomposes as:

$$I(\mathbf{q}_i^{\text{OR}}; (\mathbf{v}_i, \mathbf{t}_i)) - I(\mathbf{q}_i^{\text{LF}}; (\mathbf{v}_i, \mathbf{t}_i)) = \mathcal{I}_{\text{cross}}^{\text{OR}} + \mathcal{D}_{\text{intra}}^{\text{LF}} \quad (7)$$

Here, $I(X;Y)$ denotes mutual information, measuring how much information X preserves about Y. Intuitively, Equation (7) compares the information retained by OpticalRec and late-fusion representations about the original image-text input; we next justify why both right-hand terms are positive and that their sum is substantial in practice.

The cross-modal term $\mathcal{I}_{\text{cross}}^{\text{OR}} > 0$. Let $\mathbf{x}_i = (\mathbf{v}_i, \mathbf{t}_i)$ denotes the original image-text input. Remember $\mathbf{s}_i^{vt}$ denotes the vision-language layout structure defined in Section 3.1. We define the cross-modal gain as the extra information preserved by the unified representation over the late-fusion intermediate embedding:

$$\mathcal{I}_{\text{cross}}^{\text{OR}} \triangleq I(\mathbf{q}_i^{\text{OR}}; \mathbf{x}_i) - I(\mathbf{z}_i^{\text{LF}}; \mathbf{x}_i). \quad (8)$$

Let $\mathbf{c}_i$ denote cross-modal information contained in $\mathbf{s}_i^{vt}$, Since late fusion independently encodes $\mathbf{v}_i$ and $\mathbf{t}_i$, such information is not fully recoverable from $\mathbf{z}_i^{\text{LF}}$:

$$I(\mathbf{c}_i; \mathbf{x}_i \mid \mathbf{z}_i^{\text{LF}}) > 0. \quad (9)$$

In contrast, OpticalRec exposes $\mathbf{c}_i$ implied by $\mathbf{s}_i^{vt}$ to a single vision-language encoder through unified pixel-space rendering. For the resulting non-degenerate cross-modal channel, $I(\mathbf{q}_i^{\text{OR}}; \mathbf{c}_i \mid \mathbf{z}_i^{\text{LF}}) > 0$, and thus $\mathcal{I}_{\text{cross}}^{\text{OR}} > 0$.

The distortion term $\mathcal{D}_{\text{intra}}^{\text{LF}} > 0$. Formally the distortion term decomposes into:

$$\mathcal{D}_{\text{intra}}^{\text{LF}} \triangleq I(\mathbf{z}_i^{\text{LF}}; \mathbf{x}_i) - I(\mathbf{q}_i^{\text{LF}}; \mathbf{x}_i). \quad (10)$$

where the right-hand side denotes information retained before fusion subtracts that retained after fusion. Since late fusion is defined as $\mathbf{q}_i^{\text{LF}} = g(\mathbf{z}_i^{\text{LF}})$, we have the Markov chain $\mathbf{x}_i \to \mathbf{z}_i^{\text{LF}} \to \mathbf{q}_i^{\text{LF}}$. By the data processing inequality (DPI) (Cover and Thomas, 2006):

$$I(\mathbf{q}_i^{\text{LF}}; \mathbf{x}_i) \leq I(\mathbf{z}_i^{\text{LF}}; \mathbf{x}_i), \quad (11)$$

and thus $\mathcal{D}_{\text{intra}}^{\text{LF}} \geq 0$. Equality would require late fusion $g(\cdot)$ be strictly lossless compression, which

does not hold for our multimodal recommendation settings. Hence $\mathcal{D}_{\text{intra}}^{\text{LF}} > 0$ for standard lossy late fusion.

Empirical significance. While the mutual information terms in Equation (7) are difficult to estimate directly for high-dimensional LVLM representations, unified encoding enables a large cross-modal interaction budget. Specifically, image patch tokens and text glyph tokens are processed in the same attention sequence, giving

$$\mathcal{B}_{\text{cross}}^{\text{OR}} \approx L \cdot H \cdot |\mathcal{V}| \cdot |\mathcal{T}| \cdot d_h, \tag{12}$$

where L, H, and d_h are the number of layers, heads, and per-head dimension. For Qwen3-VL-7B with L=32, H=32, d_h=128, and $|\mathcal{V}| \approx |\mathcal{T}| \approx 256$, this yields $\mathcal{B}_{\text{cross}}^{\text{OR}} \approx 8.6 \times 10^9$ potential cross-modal interactions. This suggests that the gap in Equation (7) is practically substantial regarding representation capacity. To further accurately quantify the information gap and data variances in OpticalRec and Late-fusion encoding, we provide one more theoretical analysis in Section 3.4 via matrix effective-rank $erank(\cdot)$.

3.4 Theoretical Analysis via Matrix Effective Rank

In this section, we further quantify the information variances carried in OpticalRec and late-fusion-based embedding. One natural direction is to compute rank-based metrics for embedding matrices due to the fact that in linear algebra and information theory (Bromiley et al., 2004), matrix rank $0 \leq \text{rank}(\mathbf{X}) \leq \min(m, n)$, where $\mathbf{X} \in \mathbb{R}^{m \times n}$, is commonly utilized to represent the number of effective dimension used to carry data variances.

Different from the matrix algebraic/true rank, the effective rank $erank(\cdot)$ down-weights near-zero singular values and therefore better reflects representation concentration or collapse (Roy and Vetterli, 2007; Su, 2025). Consider a real-valued matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ which is the item embedding matrix/multimodal feature matrix in our setting, we first conduct singular value decomposition of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^\top, \quad \mathbf{U} \in \mathbb{R}^{m \times r},\ \mathbf{\Sigma} \in \mathbb{R}^{r \times r},\ \mathbf{V} \in \mathbb{R}^{n \times r} \tag{13}$$

We assume $n = r$ for simplicity, then $n = r = \text{rank}(\mathbf{X})$. Let $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_n)$ denote the diagonal matrix of singular values, where $\sigma_i \geq 0$, by normalization we can further construct a probability distribution over all singular values as

Table 1: Statistics of the datasets.

Datasets	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{R}\rvert$	Sparsity (%)
Baby	108,897	17,836	330,771	99.983%
Pets	461,828	61,060	1,772,584	99.993%
Clothing	398,796	75,616	2,109,975	99.993%

$$p_i = \frac{\sigma_i}{\sum_{j=1}^{n} \sigma_j} \tag{14}$$

Based on the constructed probability p_i, we can further compute the Shannon entropy as

$$\text{H(p)} = -\sum_{i=1}^{n} p_i \log p_i \tag{15}$$

$\text{erank}(\mathbf{X})$ can be further defined as the exponential of Shannon entropy, satisfying one desired math property of $\text{erank}(\mathbf{X})$ range discussed later:

$$\text{erank}(\mathbf{X}) = \exp\left(-\sum_{i=1}^{n} p_i \log p_i\right) \tag{16}$$

As mentioned earlier and introduced in (Roy and Vetterli, 2007), the above definition satisfies one desired math property of $\text{erank}(\mathbf{X})$ range:

Property 1. *Let* $\mathbf{p} = (p_1, \ldots, p_n)$ *be a probability distribution induced by the singular values, and let* $\text{erank}(\mathbf{X}) = \exp(H(\mathbf{p}))$*, where* $H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$*, then the effective rank satisfies:*

$$1 \leq \text{erank}(\mathbf{X}) \leq n.$$

where $\text{erank}(\mathbf{X}) = 1$ *if and only if* $\mathbf{p}$ *follows a one-hot distribution; and* $\text{erank}(\mathbf{X}) = n$ *if* $\mathbf{p}$ *follows a uniform distribution.*

Thus, $\text{erank}(\mathbf{X})$ bears more sensitivity than true rank and effectively measures how many singular directions are effectively used by the embedding matrix: smaller values indicate representation collapse, while larger values indicate richer information spread. We report effective rank for OpticalRec and baseline late-fusion embeddings in Section 5.2.

4 Experimental Setup

4.1 Datasets

For evaluating our models and fair comparison, we select three most commonly used benchmarks by multimodal collaborative filtering recommendation from *Amazon.com* known for real-word

sparsity, namely *Baby Products* (Baby), *Pet Supplies* (Pets), and *Clothing, Shoes & Jewelry* (Clothing) (Hou et al., 2024). For preprocessing, we follow (Malitesta et al., 2025; Pomo et al., 2025) to construct the input sequence in chronological order and apply 5-core filtering to exclude users and items with less than five-time appearances. For textural feature selection, we choose *title*, *price*, *brand* and *categories*; For visual features, we use *photos* of the items. We also filter out items without above metadata. Detailed statistics of the datasets are reported in table Table 1 including users (#User), items (#Item), interactions (#Inter.) and dataset sparsity in percentages.

4.2 Baseline Methods

- **VBPR** (He and McAuley, 2016) is an early multimodal collaborative filtering method that extends BPR with CNN-based visual features to model visually driven user preferences.
- **LATTICE** (Zhang et al., 2021) constructs modality-aware item-item similarity graphs for each modality and applies graph convolutions to inject high-order item affinities into item representations for downstream collaborative filtering.
- **FREEDOM** (Zhou and Shen, 2023) freezes the item-item graph and applies degree-sensitive edge pruning to denoise the user-item interaction graph, improving robustness over LATTICE's dynamic graph learning.
- **BM3** (Zhou et al., 2023b) proposes a self-supervised multimodal framework that uses dropout-based contrastive views, eliminating the need for negative sampling and auxiliary graph augmentations.
- **LGMRec** (Guo et al., 2024) captures both local and global modality correlations via hypergraph structures, modeling fine-grained cross-modal interactions for multimodal preference alignment.

4.3 Implementations

For training OpticalRec and all baseline models, we utilize AdamW optimizer with learning rate of 1e-3/1e-4 with maximum epochs of 500. Validation is performed per epoch and the training is stopped once validation performance does not improve for 10 epochs. The model with best validation performance is saved for testing and metrics report. For hyperparameters, we find (1) embedding/model size, the number of Graph Neural Network layers, dropout rate and weight decay be the most sensitive ones for model performance. Specifically, we grid-search the embedding/model size in [64, 128, 256, 512, 1024, 2048], the number of Graph Neural Network layers in [1,2,4,8], dropout rate from [0.1,...,0.8] on a 0.1-stride and weight decay from [1e-6, 1e-4, 1e-2]. For OpticalRec rendering operation, we tune the resolution among [256, 544, 1024], we analyze the efficiency of rendering and resolution tuning process in Section 6. For validation and test, we adopt three most commonly utilized metrics (Malitesta et al., 2025; Pomo et al., 2025) as NDCG@20 and Recall@20 and H@20 (Hit Ratio), typical for recommendation algorithm evaluation.

5 Experimental Results

5.1 Main Performance Analysis

OpticalRec delivers consistent gains over the strongest baseline across all backbones and datasets. (1) On average, OpticalRec surpasses the best competing variant by 2.12% across all models, metrics, and datasets. The improvement is most pronounced on Pets, with an average gain of 3.21% and the largest single gain of 5.82% on NDCG@20 under FREEDOM. Baby and Clothing yield more moderate but still consistent improvements of 1.65% and 1.51%, respectively. (2) OpticalRec achieves larger gains on ranking-sensitive metrics. Its average improvement on NDCG@20 reaches 2.41%, consistently exceeding the 1.97% gain on Recall@20, across all settings. This indicates that OpticalRec improves not only item retrieval but also top-ranked ordering, benefiting from richer cross-modal semantics encoded through pixel-space unified representation. (3) Graph-based backbones benefit more significantly from OpticalRec. Graph-based recommenders achieve larger average gains, with 2.29% for LATTICE, 2.88% for FREEDOM, and 2.29% for LGMRec, compared with 1.09% for VBPR. This suggests that higher-quality multimodal item embeddings provided by OpticalRec amplify the propagation of semantics along user-item interaction graphs, leading to compounded improvements in graph-based preference modeling.

Table 2: Performance comparison on three Amazon benchmarks. **Bold** = best per group per column; underline = second best. OpticalRec constantly outperforms strong baseline encoding methods in most cases across datasets and models.

Model/Dataset	Baby			Pets			Clothing		
VBPR	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**
+ Sum	0.0791	0.0348	0.0838	0.0981	0.0465	0.1072	0.0434	0.0201	0.0517
+ Concat	0.0791	0.0355	0.0840	**0.1000**	0.0473	0.1091	0.0454	0.0212	0.0536
+ Prompt	0.0796	0.0355	0.0843	0.0994	0.0474	0.1093	0.0453	0.0212	0.0540
+ **OpticalRec**	**0.0808**	**0.0359**	**0.0857**	0.0997	**0.0477**	**0.1097**	**0.0460**	**0.0215**	**0.0551**
LATTICE	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**
+ Sum	0.0784	0.0345	0.0832	0.1036	0.0484	0.1137	0.0498	0.0229	0.0579
+ Concat	0.0762	0.0332	0.0808	0.1026	0.0474	0.1121	0.0468	0.0219	0.0557
+ Graph	0.0770	0.0340	0.0815	0.1042	0.0489	0.1135	**0.0509**	0.0224	0.0573
+ Prompt	0.0779	0.0338	0.0825	0.1046	0.0484	0.1142	0.0504	0.0229	0.0597
+ **OpticalRec**	**0.0803**	**0.0357**	**0.0849**	**0.1084**	**0.0513**	**0.1181**	0.0507	**0.0230**	**0.0601**
FREEDOM	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**
+ Sum	0.0820	0.0358	0.0865	0.0974	0.0447	0.1069	0.0416	0.0182	0.0493
+ Concat	0.0815	0.0358	0.0862	0.0956	0.0428	0.1036	0.0402	0.0171	0.0481
+ Graph	0.0819	0.0356	0.0865	0.0957	0.0428	0.1037	0.0404	0.0173	0.0482
+ Prompt	0.0825	0.0358	0.0870	0.0972	0.0439	0.1056	0.0421	0.0186	0.0501
+ **OpticalRec**	**0.0834**	**0.0363**	**0.0882**	**0.1020**	**0.0473**	**0.1107**	**0.0429**	**0.0194**	**0.0510**
BM3	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**
+ Sum	0.0765	0.0336	0.0812	0.0893	0.0399	0.0970	0.0380	0.0164	0.0452
+ Concat	0.0777	0.0341	0.0825	0.0876	0.0391	0.0951	0.0382	0.0165	0.0455
+ Supervised	0.0792	0.0347	0.0841	0.0912	0.0408	0.0991	0.0389	0.0168	0.0463
+ Prompt	0.0797	0.0350	0.0846	0.0906	0.0405	0.0984	0.0392	0.0169	0.0467
+ **OpticalRec**	**0.0807**	**0.0354**	**0.0860**	**0.0947**	**0.0421**	**0.1024**	**0.0397**	**0.0171**	**0.0475**
LGMRec	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**	**R@20**	**N@20**	**H@20**
+ Sum	0.0804	0.0353	0.0853	0.0944	0.0427	0.1023	0.0395	0.0170	0.0474
+ Concat	0.0797	0.0350	0.0846	0.0929	0.0420	0.1006	0.0391	0.0168	0.0469
+ Attentive	0.0818	0.0357	0.0868	0.0963	0.0435	0.1043	0.0404	0.0174	0.0485
+ Prompt	0.0823	0.0359	0.0876	0.0972	0.0439	0.1053	0.0403	0.0173	0.0484
+ **OpticalRec**	**0.0835**	**0.0366**	**0.0887**	**0.1010**	**0.0454**	**0.1092**	**0.0412**	**0.0177**	**0.0491**

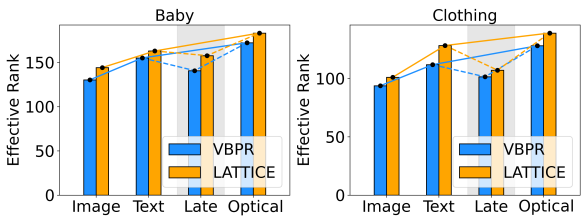

Figure 2: Effective rank comparison among single-modality encoding, Late Fusion, and OpticalRec. Grey regions highlight the effective-rank drop caused by Late Fusion, while OpticalRec consistently ranks highest.

5.2 Embedding Quality via Matrix Effective Rank

We compute the matrix effective rank $erank(\cdot)$ to evaluate the informativeness of item embeddings produced by Image-only encoding, Text-only encoding, Late Fusion, and OpticalRec. Late Fusion denotes concatenation-based fusion for VBPR and graph-based fusion for LATTICE, following their default designs. Results are shown in Figure 2 and we observe:

1. OpticalRec yields the most informative embedding space. OpticalRec achieves the highest average effective rank of 155.6, improving over late fusion by 22.81% and over the strongest single-modality baseline, Text-only encoding, by 11.54%. This suggests that unified optical encoding preserves the richest vision-language variances in the embedding space.

2. Late Fusion suffers from dimensional collapse. We also noted that Late Fusion obtains a lower average effective rank of 126.7, falling 9.18%

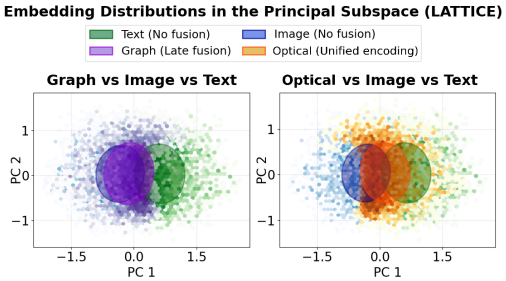

Figure 3: PCA visualization of embedding distributions with 50% confidence ellipses under the graph-based backbone. OpticalRec (Orange) forms a unified distribution bridging Image and Text representations, while Late Fusion (Purple) remains biased toward a single modality.

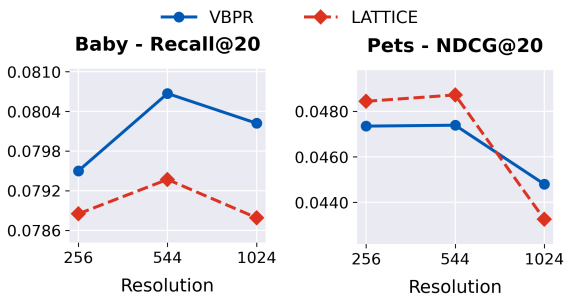

Figure 4: Performance under different Item Card resolutions. The 544×544 performs the best and is robust across datasets/models for efficient deployment.

below the strongest single-modality baseline. This indicates that fusing independently encoded modalities can distort, rather than enrich, the multimodal embedding space.

5.3 Cross-Modal Correlation Visualization

To further examine the embedding geometry behind effective rank, we visualize PCA-projected item representations of Image-only, Text-only, Late Fusion, and OpticalRec using hexbin density plots with 50% confidence ellipses. As shown in Figure 3, OpticalRec forms a unified manifold between image and text representations, suggesting stronger vision-language alignment. In contrast, Late Fusion is biased toward a single modality, i.e., Text in VBPR and Image in LATTICE, indicating that post-hoc fusion struggles to balance heterogeneous modality distributions.

6 Ablation Study

6.1 Rendering Resolution and Efficiency

We evaluate the impact of rendering resolution (256×256), (544×544), and (1024×1024 pixel) on the rendered Item Card. As shown in Figure 4, 544×544 achieves the best overall performance, while both lower and higher resolutions lead to slight degradation. This suggests that medium-resolution rendering already preserves sufficient visual-language semantics, avoiding unnecessary high-resolution vision-language-model inference and reducing resolution tuning cost.

We report the one-time offline rendering overhead in Table 3. Even for the largest Clothing dataset, rendering 75,616 items takes 57 minutes with 8 workers and 10.1 GB storage, indicating practical overhead on modern machines. Together with the resolution robustness in Figure 4, OpticalRec does not require repeated high-resolution rendering.

Table 3: Rendering efficiency as one-time offline preprocessing. Speed denotes overall throughput in cards per second.

Dataset	Items	Speed	Workers	Time (min)	Storage (GB)
Baby	17,836	22	8	13.5	2.4
Pets	61,060	22	8	46.0	8.1
Clothing	75,616	22	8	57.0	10.1

Table 4: Ablation study on Item Card components. "Card-Text" and "Card-Image" denote uni-modal variants rendered from metadata or product images only.

	Method	Baby			Pets		
		R@20	N@20	H@20	R@20	N@20	H@20
VBPR	Text	0.0794	0.0354	0.0841	0.1001	0.0473	0.1092
	Image	0.0781	0.0350	0.0827	0.0997	0.0470	0.1088
	Full	**0.0808**	**0.0359**	**0.0857**	**0.0997**	**0.0477**	**0.1097**
	Imp.(%)	+1.76%	+1.41%	+1.90%	-0.40%	+0.85%	+0.46%
LATTICE	Text	0.0772	0.0332	0.0819	0.1027	0.0474	0.1122
	Image	0.0756	0.0328	0.0801	0.1031	0.0478	0.1125
	Full	**0.0803**	**0.0357**	**0.0849**	**0.1084**	**0.0513**	**0.1181**
	Imp.(%)	+4.02%	+7.53%	+3.66%	+5.14%	+7.32%	+4.98%

6.2 Rendering Components

As shown in Table 4, the full Item Card outperforms text-only or image-only rendering by 3.22% on average, showing the benefit of preserving visual and textual semantics in the same optical space.

Limitations

This work studies multimodal content encoding for CF under controlled settings, rather than exhausting all architectures, fine-tuning strategies, or proprietary VLMs. OpticalRec also requires one-time offline rendering, which is practical in our efficiency study but may need further optimization for extremely large or frequently updated catalogs.

References

Shuai Bai, Yuxuan Cai, and 1 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

PA Bromiley, NA Thacker, and E Bouhova-Thacker. 2004. Shannon entropy, renyi entropy, and information. *Statistics and Inf. Series (2004-004)*, 9(2004):2–8.

Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 335–344.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. John Wiley & Sons, Hoboken, NJ.

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

Ruining He and Julian McAuley. 2016. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Yicheng He, Zhou Kaiyu, Haoyue Bai, Fengbin Zhu, and Yonghui Yang. 2025. Understanding embedding scaling in collaborative filtering. *arXiv preprint arXiv:2509.15709*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *Preprint*, arXiv:2403.03952.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Guohui Li, Li Zou, Zhiying Deng, and Qi Chen. 2024c. Neighborhood-enhanced multimodal collaborative filtering for item cold start recommendation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7815–7819. IEEE.

Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. Megcf: Multimodal entity graph collaborative filtering for personalized recommendation. *ACM transactions on information systems*, 41(2):1–27.

Shiqin Liu, Chaozhuo Li, Minjun Zhao, Litian Zhang, and Jiajun Bu. 2025. Modalsync: Synchronizing user behavior with multimodal features for multimodal pre-training recommendation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2278–2287.

Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. 2025. Formalizing multimedia recommendation through multimodal deep learning. *ACM Transactions on Recommender Systems*, 3(3):1–33.

Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D Malliaros. 2024. Do we really need to drop items with missing modalities in multimodal recommendation? In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3943–3948.

Rongqing Kenneth Ong and Andy WH 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*, pages 773–781.

OpenAI. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

OpenGV. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Claudio Pomo, Matteo Attimonelli, Danilo Danese, Fedelucio Narducci, and Tommaso Di Noia. 2025. 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*, pages 2377–2387.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. 2024. Language representations can be what recommenders need: Findings and potentials. *arXiv preprint arXiv:2407.05441*.

Jianlin Su. 2025. Effective rank construction and computation.

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*, 25:1074–1084.

Yueqi Wang, Zhenrui Yue, Huimin Zeng, Dong Wang, and Julian McAuley. 2024. Train once, deploy anywhere: Matryoshka representation learning for multimodal recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13461–13472.

Yuxiang Wang, Xin Shi, and Xueqing Zhao. 2025. Mllm4rec: multimodal information enhancing llm for sequential recommendation. *Journal of Intelligent Information Systems*, 63(3):745–761.

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*, pages 3541–3549.

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*, pages 1437–1445.

Xin Xin, Xiangnan He, Yongfeng Zhang, Yongdong Zhang, and Joemon Jose. 2019. Relational collaborative filtering: Modeling multiple item relations for recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 125–134.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2025. A survey on multimodal recommender systems: Recent advances and future directions. *arXiv preprint arXiv:2502.15711*.

Yu Ye, Junchen Fu, Yu Song, Kaiwen Zheng, and Joemon M Jose. 2025a. Are multimodal embeddings truly beneficial for recommendation? a deep dive into whole vs. individual modalities. *arXiv preprint arXiv:2508.07399*.

Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025b. Harnessing multimodal large language models for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13069–13077.

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

Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2025. Mind individual information! principal graph learning for multimedia recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 13096–13105.

Zhenrui Yue, Huimin Zeng, Yueqi Wang, Julian McAuley, and Dong Wang. 2025. Preference-optimized retrieval and ranking for efficient multimodal recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3692–3703.

Jianyang Zhai, Zi-Feng Mai, Chang-Dong Wang, Feidiao Yang, Xiawu Zheng, Hui Li, and Yonghong Tian. 2025. Multimodal quantitative language for generative recommendation. In *The Thirteenth International Conference on Learning Representations*.

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*, pages 3872–3880.

Bowen Zheng, Junjie Zhang, Hongyu Lu, Yu Chen, Ming Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Enhancing graph contrastive learning with reliable and informative augmentation for recommendation. *arXiv preprint arXiv:2409.05633*.

Hongyu Zhou, Yinan Zhang, Aixin Sun, and Zhiqi Shen. 2025a. Does multimodality improve recommender systems as expected? a critical analysis and future directions. *Preprint*, arXiv:2508.05377.

Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023a. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. *arXiv preprint arXiv:2301.12097*.

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

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*, pages 935–943.

Xin Zhou, Xiaoxiong Zhang, Dusit Niyato, and Zhiqi Shen. 2025b. Learning item representations directly from multimodal features for effective recommendation. *arXiv preprint arXiv:2505.04960*.

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