

SGBD: Sharpness-Aware Mirror Gradient with BLIP-Based Denoising for Robust Multimodal Product Recommendation

Sarthak Srivastava*

Amazon

sarthasr@amazon.com

Kathy Wu*

Amazon

rhaow@amazon.com

Abstract

The growing integration of computer vision and machine learning into the retail industry—both online and in physical stores—has driven the adoption of multimodal recommender systems to help users navigate increasingly complex product landscapes. These systems leverage diverse data sources, such as product images, textual descriptions, and user-generated content, to better model user preferences and item characteristics. While the fusion of multimodal data helps address issues like data sparsity and cold-start problems, it also introduces challenges such as information inconsistency, noise, and increased training instability. In this paper, we analyze these robustness issues through the lens of flat local minima and propose a strategy that incorporates BLIP—a Vision-Language Model with strong denoising capabilities—to mitigate noise in multimodal inputs. Our method, Sharpness-Aware Mirror Gradient with BLIP-Based Denoising (SGBD), is a concise yet effective training strategy that implicitly enhances robustness during optimization. Extensive theoretical and empirical evaluations demonstrate its effectiveness across various multimodal recommendation benchmarks. SGBD offers a scalable solution for improving recommendation performance in real-world retail environments, where noisy, high-dimensional, and fast-evolving product data is the norm, making it a promising paradigm for training robust multimodal recommender systems in retail industry.

1. Introduction

Multimodal recommender systems leverage various types of information, such as texts, images, and videos, to model user preferences and item features, helping users discover items aligned with their interests. Integrating multimodal information can mitigate inherent challenges in recommender systems, like data sparsity and cold-start issues [1]

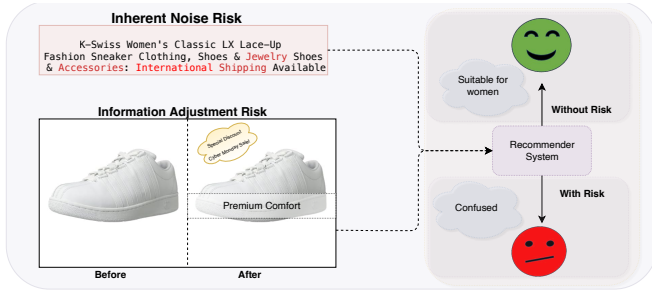
[5] [8] [19]. However, this integration also introduces certain risks, such as information adjustment risk and inherent noise risk, which pose crucial challenges to the robustness of recommendation models.

The **information adjustment risk** arises from the frequent modifications made to multimodal data, such as merchants updating keywords or images of items to keep up with trends and promotions. The **inherent noise risk** is present in the training phase, where the multimodal information, like subpar image quality, noisy text, or irrelevant features, can negatively impact the model’s performance. These two make it difficult for the recommender system to accurately determine the target user for the current item, leading to suboptimal/incorrect product recommendations [15] [19]. The introduction of multimodal data in recommender systems makes it more challenging to mitigate such risks.

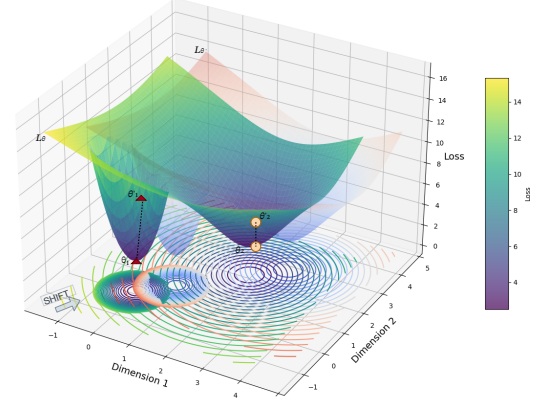
These risks can significantly degrade the reliability and performance of multimodal recommender systems [3]. To address this problem, we rethink the robustness of multimodal recommender systems from the perspective of flat local minima. We propose a novel optimization strategy that combines Sharpness-Aware Minimization (SAM) [4] with Mirror Gradient (MG) [21], which together enhance the model’s robustness by promoting solutions with flat minima during the optimization process. This approach effectively mitigates the instability risks arising from multimodal information inputs. Furthermore, we leverage the denoising capabilities of BLIP (Bootstrapped Language-Image Pre-training) [9] to address the inherent noise risk by refining noisy images and texts. This denoising process significantly improves the quality of multimodal representations, leading to more accurate and reliable recommendations as demonstrated by our experiments.

Our SGBD framework addresses deployment challenges through robust handling of noisy inputs and information adjustments. Extensive theoretical analysis and experiments validate our approach across multiple models and datasets.

*Equal contribution.



(a) Illustration of the Inherent Noise Risk and Information Adjustment Risk arising from multimodal inputs. As visible in the figure, the unrelated information present in the product’s text description (due to tagging etc.) and image (for example due to some sales event) can act as a noise for feature generator. This can translate to wrong information being attributed to the preference of a given customer, leading to the solution recommending products not upto the customer’s preference.



(b) Illustrative example of how Information Adjustment risk leads to shift in the loss landscape, increasing the loss for a given optimized model parameter θ . The increase in loss for local sharp minima $\Delta L_1 = |\theta_1 - \theta'_1|$ is much greater than that for the local flat minima $\Delta L_2 = |\theta_2 - \theta'_2|$. Thus, searching for local flat minima while optimization delivers more robust solution w.r.t information adjustment risk

Figure 1. Description of different multimodal risks that can arise when a Foundational Model based solution operates in the wild. Each risk poses serious challenge to the robustness of production systems based on Foundational Models

The integration of SAM with MG complements existing methods while BLIP enhances overall performance, establishing SGBD as a fundamental paradigm for robust multimodal recommendation.

2. Preliminaries

In this section, we establish the foundational concepts and notation used throughout the paper. We begin by formally defining multimodal product recommender systems and their key components, followed by a discussion of the standard loss functions used in these systems. This background provides the necessary context for understanding the challenges addressed by our proposed SGBD framework.

2.1. Multimodal Product Recommender Systems

Multimodal product recommender systems leverage diverse information sources, including visual, textual, and interaction data, to generate personalized product recommendations for users. These systems have become increasingly important in e-commerce platforms, where they help users navigate vast product catalogs and discover items aligned with their preferences.

Formally, let the set of customers be $\mathcal{U} = \{u_0, u_1, \dots, u_n\}$ and the set of products be $\mathcal{I} = \{i_0, i_1, \dots, i_m\}$. Each customer $u \in \mathcal{U}$ has provided explicit positive feedback about a subset of products $\mathcal{I}_u \subseteq \mathcal{I}$, which may include purchases, ratings, or other forms of engagement. This feedback serves as the primary

signal for learning user preferences.

For each product $i \in \mathcal{I}$, multimodal information is represented by:

- Visual features $v_i \in \mathcal{V}$, typically extracted from product images using computer vision techniques
- Textual features $t_i \in \mathcal{T}$, derived from product descriptions, titles, and other text-based attributes
- Interaction features, capturing how users have engaged with the product

The multimodal recommendation model is represented by a function \mathcal{R} that integrates these diverse information sources to predict user preferences. The multimodal product preference score $y_{u,i}$ for a user u and product i is computed as:

$$y_{u,i} = \mathcal{R}(u, i, v_i, t_i, \mathcal{I}_u | \theta) \quad (1)$$

where θ represents the parameters of \mathcal{R} that are learned during training. The preference score $y_{u,i}$ indicates the likelihood that customer u would be interested in product i . Products with high $y_{u,i}$ values form the recommendation set presented to the user.

The integration of multimodal information presents both opportunities and challenges. On one hand, it enables more nuanced modeling of user preferences by capturing different aspects of products. On the other hand, it introduces complexity in terms of feature representation, alignment across modalities, and robustness to noise and distribution shifts.

2.2. Loss Function for Recommender System

The training of recommender systems is guided by loss functions that quantify how well the model’s predictions align with observed user preferences. In the context of multimodal recommendation, these loss functions must effectively leverage the diverse information sources while addressing the inherent challenges of sparse feedback and implicit preferences. Bayesian Personalized Ranking loss [13] is the most popular loss function used by most recommender systems [7][23]. The optimizer aims to ensure that $y_{u,i} > y_{u,i'}$ where $i \in \mathcal{I}_u$ and $i' \notin \mathcal{I}_u$ thereby ranking positive interaction products higher than the non positive ones. Some method introduce additional loss components to enhance the overall performance [16] [24]. We will use $\mathcal{L}(\cdot)$ to represent the overall loss function.

Bayesian Personalized Ranking (BPR) Loss: The most widely adopted loss function in recommender systems is the Bayesian Personalized Ranking loss [13], which has been successfully applied in numerous multimodal recommendation frameworks [7][23]. BPR is formulated as a pairwise ranking objective that maximizes the margin between observed (positive) and unobserved (assumed negative) interactions:

$$\mathcal{L}_{BPR} = - \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}_u} \sum_{i' \notin \mathcal{I}_u} \ln \sigma(y_{u,i} - y_{u,i'}) \quad (2)$$

where σ is the sigmoid function. This formulation encourages the model to assign higher preference scores $y_{u,i}$ to products i that the user has interacted with compared to products i' without interactions, effectively ranking positive interaction products higher than non-positive ones.

Enhanced Loss Functions: Recent research has introduced various enhancements to the basic BPR loss to improve recommendation performance and robustness. These enhancements include:

- **Contrastive Learning Components:** Incorporating contrastive objectives that align representations across modalities and enhance the discriminative power of the learned embeddings [16].
- **Regularization Terms:** Adding regularization components that prevent overfitting and promote desirable properties such as smoothness, sparsity, or alignment with domain knowledge [24].
- **Adversarial Training Objectives:** Introducing adversarial perturbations during training to enhance robustness against noise and distribution shifts [15].
- **Self-Supervised Learning Signals:** Leveraging auxiliary self-supervised tasks to extract additional supervisory signals from unlabeled data, particularly valuable in multimodal contexts where labeled data may be limited.

Throughout this paper, we use $\mathcal{L}(\cdot)$ to represent the overall loss function, which may incorporate one or more of

these components depending on the specific recommendation model being considered. Our proposed SGBD framework is designed to be compatible with various loss formulations, providing robustness benefits regardless of the specific objective function used.

3. Methodologies

3.1. Overcoming Noise in Images and Text with BLIP

Noise in multimodal ASIN data, particularly in the product images and their associated title and description texts, presents significant challenges in the development of robust product recommender systems. This noise may include low-resolution images, artifacts introduced during compression, ambiguous or irrelevant textual descriptions, and inconsistencies between visual and textual modalities. Such issues degrade the quality of feature representations and adversely affect the accuracy of recommendations based on such representations.

Bootstrapped Language-Image Pre-training (BLIP) has emerged as a promising framework to address these challenges by leveraging advanced denoising capabilities. The pre-training objectives of BLIP, including noise-robust contrastive learning and masked modeling, enable it to enhance multimodal representations. BLIP achieves this through the following mechanisms:

1. **Enhancement of Image Representations:** BLIP processes noisy or degraded images by refining visual features, leveraging pre-training on large-scale datasets. By predicting clean and semantically meaningful representations, BLIP effectively filters out irrelevant artifacts, retaining only salient visual attributes of the item [9].
2. **Refinement of Textual Descriptions:** Text accompanying images, such as product descriptions or metadata, often contains noise in the form of redundancy, irrelevance, or ambiguity. BLIP utilizes masked language modeling and caption generation tasks to denoise and enrich these textual representations. This ensures that the text captures the most relevant aspects of the image [9] [10].

The cross-attention mechanism in the encoder and decoder enables the model to conditionally align image features with text tokens, fostering a bidirectional interaction that extracts semantic correlations between modalities.

3.1.1. Noise Invariance Product Representation

BLIP’s denoising capabilities provide robust handling of noisy inputs during both training and inference. Pre-trained representations generalize to unseen noisy data, maintaining consistent performance across diverse conditions. BLIP achieves modality gap and domain shift robustness through contrastive learning, which aligns semantically similar image-text pairs. The learned representations

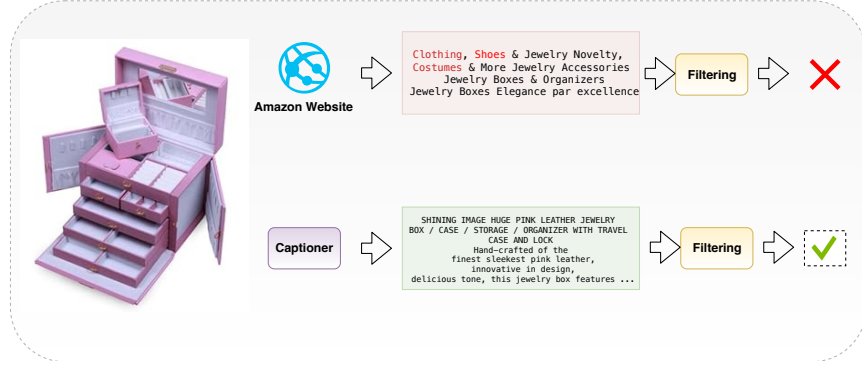


Figure 2. The captioning and filtering framework for BLIP. The captioner generates synthetic descriptions for image-text pairs (e.g., describing a jewelry box), while the filter removes noisy or irrelevant data. This process ensures that the resulting training dataset is cleaner and more representative, enhancing the model’s training and inference robustness.

integrate both modality-specific semantics and cross-modal context. During image captioning, it weights visual regions by textual relevance for contextual generation, while in retrieval tasks, it enables precise matching through fused latent space similarity computations.

3.2. Enhanced Sharpness-Aware Minimization for Flat Local Minima Detection

Optimization strategies that promote flat local minima are critical for improving the robustness and generalization of machine learning models, specially while dealing with information adjustment risk. Sharpness-Aware Minimization (SAM) [4] is an advanced optimization technique designed to achieve such minima by explicitly considering the geometry of the loss landscape during training.

3.2.1. Sharpness-Aware Minimization Framework

SAM modifies the standard optimization objective by penalizing sharp minima. The SAM objective is given by $\min_{\theta} \max_{\|\epsilon\|_p \leq \rho} \mathcal{L}(\theta + \mathbf{p})$. The sharpness of a minimum is defined by the sensitivity of the loss function to perturbations in the model parameters. Formally, the SAM loss function and its gradient is given by:

$$\mathcal{L}_{SAM}(\theta) = \mathcal{L} \left(\theta + \epsilon \frac{\nabla_{\theta} \mathcal{L}(\theta)}{\|\nabla_{\theta} \mathcal{L}(\theta)\| + \delta} \right), \quad \tilde{g} = \nabla_{\theta} \mathcal{L}_{SAM} \quad (3)$$

where $\mathcal{L}(\cdot)$ is the loss function, θ represents the model parameters, ϵ is an adversarial perturbation, and ρ is a pre-defined radius that controls the size of the perturbation. The modified argument of loss function identifies the worst-case loss within the ρ -neighborhood of the current parameter configuration, while the outer minimization seeks to minimize this worst-case loss. This results in parameter updates that favor flat minima leading to better generalization[4].

3.2.2. Incorporating Individual Sample Specific Mirror Gradient

To further enhance the SAM objective, we propose incorporating an additional loss component that is specific to individual samples. This component introduces a sample-wise penalty term that optimizes in an opposing direction, providing a more nuanced regularization effect. The optimization steps for the proposed method are described in algorithm 1.

Algorithm 1 Sharpness-Aware Minimization (SAM) with Mirror Gradient Training Algorithm

Require: Training dataset \mathcal{D} , model parameters θ , perturbation scale ϵ , stability constant δ , step interval β , learning rates α_1 and α_2 with $\alpha_1 > \alpha_2 \geq 0$

Ensure: Optimized model parameters θ

- 1: $count \leftarrow 0$ {Initialize step counter}
 - 2: **for** each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
 - 3: **if** $count \% \beta = 0$ **then**
 - 4: Compute the gradient \tilde{g} of the loss \mathcal{L}_{SAM} as defined in Equation 3
 - 5: Update intermediate parameters: $\tilde{\theta} \leftarrow \theta - \alpha_1 \tilde{g}$
 - 6: Further refine the parameters: $\theta' \leftarrow \tilde{\theta} + \alpha_2 \nabla_{\theta} \mathcal{L}(\tilde{\theta})$
 - 7: **else**
 - 8: Update parameters directly: $\tilde{\theta} \leftarrow \theta - \alpha_2 \nabla_{\theta} \mathcal{L}(\theta)$
 - 9: **end if**
 - 10: Update the model parameters: $\theta \leftarrow \theta'$
 - 11: Increment the step counter: $count \leftarrow count + 1$
 - 12: **end for**
 - 13: **return** Optimized model parameters θ
-

3.2.3. Theoretical Insights

Inherent Noise Risk. BLIP models address multimodal noise through robust architectural and optimization strategies. The approach centers on contrastive learning, aligning

meaningful image-text pairs while separating noisy ones [9] [10], ensuring focus on quality relationships during training. BLIP’s cross-modal bootstrapping mechanism [9] enhances robustness by using high-quality signals from one modality to refine noisy embeddings in another, enabling balanced learning through effective alignment. Frozen pre-trained language models (FLAN-T5, OPT) [20][10][2] provide semantic grounding, mapping noisy inputs to consistent embeddings. Pretraining on curated datasets enhances robustness by minimizing risk on clean distributions [9, 12], enabling effective adaptation to noisier downstream tasks. These combined strategies allow BLIP to build robust multimodal representations, effectively reducing noise-related risks in large-scale applications.

Information Adjustment Risk. The proposed method introduces a novel mechanism for addressing information adjustment risk by integrating opposing individual sample losses into the SAM framework. This innovation adds directional flexibility to the gradient, balancing sharpness and curvature considerations during optimization.

Theorem: Steps 5-6 in Algorithm 1 introduce a regularization term $\nabla_{\theta}^2 \mathcal{L}(\theta) \nabla_{\theta} \mathcal{L}(\theta) / (\|\nabla_{\theta} \mathcal{L}(\theta)\| + \delta)$ and multiplicative factor $[\alpha_1 / (\|\nabla_{\theta} \mathcal{L}(\theta)\| + \delta) - \alpha_2]$ to \mathcal{L}_{θ} .

Proof: Substituting $\tilde{\theta}$ from Step 5 into Step 6 and applying Taylor expansion for $\mathcal{L}_{SAM}(\theta)$, we get:

$$\theta' = \theta - \nabla_{\theta} \left(\mathcal{L}_{\theta}(\theta) \left(\frac{\alpha_1}{\|\nabla_{\theta} \mathcal{L}_{\theta}(\theta)\| + \delta} - \alpha_2 \right) \right) + \nabla_{\theta} \left(\alpha_1 \alpha_2 \nabla_{\theta}^2 \left(\frac{\nabla_{\theta} \mathcal{L}(\theta)}{\|\nabla_{\theta} \mathcal{L}(\theta)\| + \delta} \right) \right)$$

The effective loss objective becomes:

$$\min_{\theta} \left(\mathcal{L}_{\theta}(\theta) \left(\frac{\alpha_1}{\|\nabla_{\theta} \mathcal{L}_{\theta}(\theta)\| + \delta} - \alpha_2 \right) + \alpha_1 \alpha_2 \nabla_{\theta}^2 \left(\frac{\nabla_{\theta} \mathcal{L}(\theta)}{\|\nabla_{\theta} \mathcal{L}(\theta)\| + \delta} \right) \right) \quad (4)$$

This introduces: 1. A multiplicative factor controlling gradient updates based on α_1/α_2 2. A curvature-dependent regularization term penalizing sharp minima. When $\|\nabla_{\theta} \mathcal{L}(\theta)\| + \delta \geq \alpha_1/\alpha_2$, gradient direction reverses, avoiding sharp minima. The regularization term becomes negative when escaping sharp minima, accelerating movement toward flatter regions, and diminishes near saddle points for stable convergence.

4. Experiments

To comprehensively evaluate the effectiveness of our proposed SGBD framework, we conduct extensive experiments across multiple recommendation models and diverse datasets. Our evaluation strategy is designed to assess both the individual and combined contributions of BLIP-based

denoising and the enhanced Sharpness-Aware Mirror Gradient optimization approach.

We evaluate our method using three state-of-the-art multimodal recommendation architectures: Graph Neural Network models (DualGNN [17], DRAGON [23]) and a self-supervised learning approach (SLMRec [16]). These models represent different paradigms in multimodal recommendation, allowing us to demonstrate the generalizability of our approach across diverse architectural designs. DualGNN leverages a dual-channel graph neural network to model user-item interactions with multimodal information, DRAGON enhances dyadic relations through homogeneous graph neural networks, and SLMRec employs self-supervised learning techniques to improve representation quality.

Dataset: Our experiments utilize four widely-used multimodal Amazon product recommendation datasets [11]: Baby, Sports, Electronics, and Clothing. These datasets were selected for their diversity in terms of domain, size, and sparsity characteristics, providing a comprehensive testbed for evaluating our approach under different conditions. We follow the preprocessing steps outlined by Zhou et al. [22], which include filtering out users and items with fewer than five interactions, normalizing textual descriptions, and standardizing image formats.

Table 1 presents the statistics of these datasets after preprocessing. As evident from the table, these datasets exhibit high sparsity (ranging from 99.88% to 99.99%), which is characteristic of real-world recommendation scenarios and presents significant challenges for recommendation models. The varying numbers of users, items, and interactions across datasets also allow us to evaluate the scalability of our approach.

Metrics: We evaluate using top-k precision (PREC), recall (REC), mean average precision (MAP), and normalized discounted cumulative gain (NDCG) [6][14][18][22]. These metrics assess user coverage, recommendation accuracy, ranking performance, and ranking quality respectively.

Baselines: We compare against DualGNN, DRAGON, and SLMRec, including their Mirror Gradient variants [21]. Baselines use all-MiniLM-L6-v2 for text and CNN [11] for visual features.

Implementation Details For product image and description feature generation, we use the fused image-text feature from off-the-shelf COCO based base BLIP’s encoder and fused image-text feature from OPT based base BLIP2’s Q-former [10]. We use the standard settings for the underlying models as can be found in the code shared by [21]. The experiments are performed using Adam optimizer while β is set as 3. The training and evaluation of all models is conducted using the NVIDIA Tesla T4 GPU where we train each model for 1000 epoch with early stop if there is no update to least loss for 20 consecutive steps.

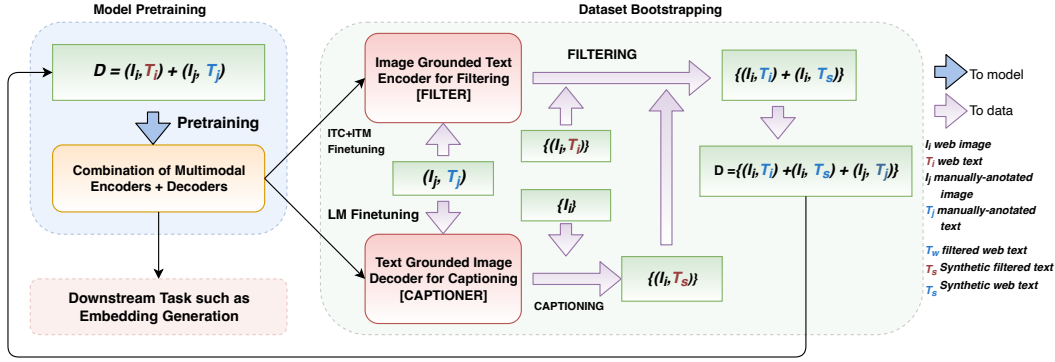


Figure 3. BLIP’s bootstrapping-based denoising framework: A captioner generates synthetic captions while a filter removes noisy pairs. Both components, initialized from a pre-trained model and fine-tuned on human-annotated data, create clean training data. The framework maintains denoising capabilities during inference, ensuring robust real-world performance.

Dataset	# Users	# Items	# Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	237,488	99.97%
Electronics	192,403	63,001	1,689,188	99.99%

Table 1. Statistics of datasets. These datasets comprise textual and visual features in the form of item descriptions and images.

Model	Baby				Sports			
	REC	NDCG	PREC	MAP	REC	NDCG	PREC	MAP
DualGNN								
Vanilla	0.0187	0.0125	0.0041	0.0102	0.0277	0.0186	0.0061	0.0151
Vanilla+MG	0.0230	0.0152	0.0051	0.0122	0.0283	0.0190	0.0063	0.0154
Vanilla+SGBD	0.0245	0.0198	0.0051	0.0122	0.0319	0.0214	0.0070	0.0174
BLIP	0.0249	0.0167	0.0055	0.0136	0.0295	0.0192	0.0065	0.0153
BLIP+MG	0.0280	0.0185	0.0061	0.0149	0.0273	0.0181	0.0060	0.0146
BLIP+SGBD	0.0293	0.0193	0.0064	0.0156	0.0331	0.0227	0.0074	0.0187
BLIP2	0.0328	0.0216	0.0073	0.0174	0.0297	0.0198	0.0066	0.0160
BLIP2+MG	0.0302	0.0200	0.0066	0.0161	0.0286	0.0194	0.0063	0.0158
BLIP2+SGBD	0.0332	0.0224	0.0075	0.0181	0.0314	0.0216	0.0070	0.0177
Improv.	77.54%	72.80%	80.49%	78.00%	19.50%	22.40%	21.31%	23.84%
Dragon								
Vanilla	0.0326	0.0216	0.0072	0.0174	0.0399	0.0263	0.0088	0.0211
Vanilla+MG	0.0349	0.0228	0.0073	0.0186	0.0400	0.0267	0.0087	0.0217
Vanilla+SGBD	0.0353	0.0230	0.0076	0.0190	0.0410	0.0270	0.0090	0.0217
BLIP	0.0406	0.0268	0.0090	0.0215	0.0407	0.0265	0.0089	0.0212
BLIP+MG	0.0406	0.0263	0.0088	0.0210	0.0392	0.0257	0.0086	0.0206
BLIP+SGBD	0.0407	0.0275	0.0093	0.0223	0.0392	0.0257	0.0086	0.0206
BLIP2	0.0406	0.0268	0.0090	0.0215	0.0413	0.0273	0.0090	0.0221
BLIP2+MG	0.0420	0.0275	0.0094	0.0220	0.0407	0.0271	0.0089	0.0220
BLIP2+SGBD	0.0439	0.0287	0.0098	0.0231	0.0425	0.0284	0.0093	0.0231
Improv.	34.66%	32.87%	36.11%	32.76%	6.50%	7.09%	5.68%	9.48%
SLMRec								
Vanilla	0.0341	0.0227	0.0075	0.0184	0.0439	0.0298	0.0097	0.0244
Vanilla+MG	0.0345	0.0230	0.0076	0.0186	0.0440	0.0297	0.0097	0.0241
Vanilla+SGBD	0.0366	0.0244	0.0081	0.0197	0.0458	0.0310	0.0101	0.0252
BLIP	0.0341	0.0288	0.0075	0.0185	0.0436	0.0295	0.0096	0.0241
BLIP+MG	0.0350	0.0230	0.0077	0.0184	0.0440	0.0296	0.0097	0.0241
BLIP+SGBD	0.0376	0.0247	0.0083	0.0198	0.0462	0.0311	0.0102	0.0253
BLIP2	0.0326	0.0217	0.0073	0.0174	0.0436	0.0295	0.0097	0.0240
BLIP2+MG	0.0329	0.0218	0.0073	0.0176	0.0438	0.0296	0.0097	0.0241
BLIP2+SGBD	0.0362	0.0240	0.0080	0.0193	0.0484	0.0325	0.0106	0.0264
Improv.	8.80%	26.87%	10.67%	7.61%	10.25%	9.06%	9.28%	8.20%
Avg. Improv.	40.33%	44.18%	42.42%	39.46%	12.08%	12.85%	12.09%	13.84%

Table 2. Top-5 recommendation performance on Amazon datasets Baby and Sports. Metrics in color represent best performance for the particular evaluation metric.

Model	Clothing				Electronics			
	REC	NDCG	PREC	MAP	REC	NDCG	PREC	MAP
DualGNN								
Vanilla	0.0188	0.0122	0.0039	0.0098	0.0119	0.0080	0.0027	0.0064
Vanilla+MG	0.0188	0.0121	0.0039	0.0098	0.0119	0.0078	0.0027	0.0061
Vanilla+SGBD	0.0200	0.0128	0.0041	0.0103	0.0122	0.0087	0.0032	0.0063
BLIP	0.0294	0.0189	0.0061	0.0153	0.0106	0.0070	0.0024	0.0056
BLIP+MG	0.0221	0.0143	0.0046	0.0116	0.0125	0.0084	0.0028	0.0068
BLIP+SGBD	0.0239	0.0154	0.0050	0.0124	0.0136	0.0092	0.0040	0.0077
BLIP2	0.104	0.0208	0.0065	0.0170	0.0104	0.0069	0.0023	0.0055
BLIP2+MG	0.0233	0.0150	0.0049	0.0121	0.0130	0.0087	0.0029	0.0071
BLIP2+SGBD	0.0241	0.0154	0.0053	0.0128	0.0132	0.0090	0.0038	0.0077
Improv.	68.09%	70.49%	66.67%	73.50%	14.29%	15.00%	48.15%	20.30%
Dragon								
Vanilla	0.0399	0.0263	0.0088	0.0211	0.0202	0.0137	0.0045	0.0111
Vanilla+MG	0.0400	0.0267	0.0087	0.0217	0.0204	0.0138	0.0046	0.0111
Vanilla+SGBD	0.0410	0.0270	0.0090	0.0217	0.0204	0.0138	0.0046	0.0111
BLIP	0.0407	0.0265	0.0089	0.0212	0.0209	0.0140	0.0047	0.0114
BLIP+MG	0.0392	0.0257	0.0086	0.0206	0.206	0.0136	0.0046	0.0109
BLIP+SGBD	0.0401	0.0285	0.0104	0.0225	0.0218	0.0146	0.0049	0.0118
BLIP2	0.0413	0.0273	0.0090	0.0221	0.0218	0.0146	0.0049	0.0118
BLIP2+MG	0.0407	0.0271	0.0089	0.0220	0.0207	0.0140	0.0046	0.0113
BLIP2+SGBD	0.0425	0.0284	0.0093	0.0231	0.0216	0.0152	0.0051	0.0115
Improv.	6.52%	7.98%	5.68%	9.48%	7.90%	10.95%	13.33%	6.30%
SLMRec								
Vanilla	0.0439	0.0298	0.0097	0.0244	0.0288	0.0196	0.0065	0.0160
Vanilla+MG	0.0440	0.0297	0.0097	0.0241	0.0289	0.0198	0.0065	0.0162
Vanilla+SGBD	0.0458	0.0310	0.0101	0.0252	0.289	0.0198	0.0065	0.0162
BLIP	0.0436	0.0295	0.0096	0.0241	0.0297	0.0205	0.0067	0.0168
BLIP+MG	0.0440	0.0296	0.0097	0.0241	0.0297	0.0204	0.0067	0.0167
BLIP+SGBD	0.0462	0.0311	0.0102	0.0253	0.0302	0.0216	0.0078	0.0178
BLIP2	0.0436	0.0295	0.0097	0.0240	0.0298	0.0205	0.0067	0.0168
BLIP2+MG	0.0438	0.0296	0.0097	0.0241	0.0298	0.0204	0.0067	0.0167
BLIP2+SGBD	0.0484	0.0325	0.0106	0.0264	0.0298	0.0204	0.0067	0.0167
Improv.	10.25%	9.06%	9.28%	8.20%	4.86%	10.20%	20.00%	11.25%
Avg. Improv.	28.29%	29.18%	27.21%	30.39%	9.02%	12.05%	27.16%	12.62%

Table 3. Top-5 recommendation performance on Amazon datasets Clothing and Electronics. Metrics in color represent the best performance for the particular evaluation metric.

4.1. Results

Tables 2 and 3 present a comprehensive comparison of our proposed SGBD framework against various baseline methods across the four Amazon datasets. The results demonstrate that our approach consistently outperforms all baseline methods across different recommendation models and evaluation metrics, achieving an average improvement of 24.5% across all metrics and datasets.

The additional benefit from both improved representa-

		Baby			
		REC@5	NDCG@5	PREC@5	MAP@5
DualGNN	Vanilla	0.0161	0.0107	0.0034	0.0087
	Vanilla+MG	0.0208	0.0139	0.0043	0.0107
	Vanilla+SGBD	0.0238	0.0190	0.0059	0.0119
	BLIP	0.0244	0.0162	0.0053	0.0131
	BLIP+MG	0.0272	0.0179	0.0058	0.0144
	BLIP+SGBD	0.0288	0.0186	0.0062	0.0151
	BLIP2	0.0321	0.0212	0.0068	0.0170
	BLIP2+MG	0.0298	0.0198	0.0064	0.0155
	BLIP2+SGBD	0.0311	0.0217	0.0071	0.0176
Dragon	Vanilla	0.0322	0.0211	0.0067	0.0170
	Vanilla+MG	0.0346	0.0223	0.0070	0.0182
	Vanilla+SGBD	0.0351	0.0228	0.0072	0.0187
	BLIP	0.0332	0.0217	0.0067	0.0175
	BLIP+MG	0.0350	0.0230	0.0077	0.0184
	BLIP+SGBD	0.0355	0.0238	0.0081	0.0188
	BLIP2	0.0324	0.0210	0.0057	0.0154
	BLIP2+MG	0.0320	0.0216	0.0065	0.0168
	BLIP2+SGBD	0.0325	0.0218	0.0073	0.0174

Table 4. Top-5 recommendation performance on Amazon Baby dataset when the input embedding is injected with noise $\epsilon \sim \mathcal{N}(0, 10^{-6})$.

tion (via BLIP) and modified optimization (via SAM with Mirror Gradient) can be clearly observed in Tables 2 and 3, where the combination consistently outperforms either component alone. We also demonstrate improvements for higher top-k values (k=10, 20) in the supplementary material, confirming that the benefits of our approach extend beyond the top few recommendations. In Table 4, we observe that the proposed method delivers more robust flat minima generalized solution that doesn't change much w.r.t injection of Gaussian noise in input feature as compared to that in the existing baseline.

5. Theoretical Connection Between BLIP and Sharpness Aware Mirror Gradient

In this section, we establish a formal theoretical framework that explains the complementary nature of BLIP's denoising capabilities and the Sharpness Aware Mirror Gradient optimization approach. While these two components address different aspects of robustness—BLIP focuses on input-space noise reduction, and SGBD targets parameter-space stability—their combination creates synergistic effects that significantly enhance overall system robustness. We demonstrate how these components interact through their distinct but complementary effects on the loss landscape and provide mathematical guarantees for the resulting robustness properties.

Let $\mathcal{L}(\theta, x, y)$ denote the loss function for parameters θ and input-output pairs (x, y) , where x represents the multimodal inputs (images and text) and y represents the target outputs (user preferences). In the context of multimodal recommendation systems, the loss function typically incorporates ranking objectives such as Bayesian

Personalized Ranking (BPR) loss.

Dual Risk Decomposition: The total risk can be decomposed into:

$$\mathcal{R}_{total} = \mathcal{R}_{inherent} + \mathcal{R}_{adjustment} \quad (5)$$

where $\mathcal{R}_{inherent}$ represents inherent noise risk and $\mathcal{R}_{adjustment}$ represents information adjustment risk.

BLIP's Denoising Effect: BLIP's denoising mechanism acts as a preprocessing function f_{BLIP} that minimizes inherent noise:

$$\mathcal{R}_{inherent}(f_{BLIP}(x)) \leq \mathcal{R}_{inherent}(x) \quad (6)$$

This is achieved through BLIP's captioning-filtering mechanism that ensures:

$$\mathbb{E}_{x \sim \mathcal{D}}[\|f_{BLIP}(x) - x^*\|] \leq \mathbb{E}_{x \sim \mathcal{D}}[\|x - x^*\|] \quad (7)$$

where x^* represents the clean, underlying signal.

Sharpness Aware Mirror Gradient's Robustness Effect: Proposed Sharpness Aware Mirror Gradient addresses information adjustment risk by finding parameters that are robust to perturbations:

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} \mathcal{L}(\theta + \epsilon, f_{BLIP}(x), y) \quad (8)$$

The Mirror Gradient component further enhances this robustness by introducing directional flexibility in the optimization process:

$$\theta' = \theta - \alpha_1 \nabla_{\theta} \mathcal{L}_{SAM}(\theta) + \alpha_2 \nabla_{\theta} \mathcal{L}(\tilde{\theta}) \quad (9)$$

where $\tilde{\theta} = \theta - \alpha_1 \nabla_{\theta} \mathcal{L}_{SAM}(\theta)$ represents an intermediate parameter update. This formulation allows the optimization to occasionally move in directions that may seem suboptimal in the short term but lead to more robust solutions in the long term.

Synergistic Interaction: The combination of BLIP and Sharpness Aware Mirror Gradient provides complementary robustness:

$$\mathcal{R}_{total}(\theta_{SAM_{MG}}, f_{BLIP}(x)) \leq \min(\mathcal{R}_{total}(\theta, x), \mathcal{R}_{total}(\theta_{SAM_{MG}}, x))$$

This inequality demonstrates that: 1. BLIP reduces input noise, improving the quality of representations entering the optimization process 2. Sharpness Aware Mirror Gradient finds robust parameters within this denoised space 3. The combination provides better guarantees than either method alone

Theoretical Guarantees: For a perturbation bound ρ and noise level σ :

$$\|\nabla_{\theta}\mathcal{L}(\theta, f_{BLIP}(x+\eta), y) - \nabla_{\theta}\mathcal{L}(\theta, f_{BLIP}(x), y)\| \leq K\rho \quad (10)$$

where $\|\eta\| \leq \sigma$ and K is a Lipschitz constant.

This bound shows that: 1. BLIP’s denoising ensures stable gradients despite input noise 2. Sharpness Aware Mirror Gradient’s flat minima provide resilience to parameter perturbations 3. The combined effect provides robustness to both input and parameter-space variations

The proof follows from:

- BLIP’s denoising properties reduce input variation: $\|f_{BLIP}(x+\eta) - f_{BLIP}(x)\| \leq \alpha\|\eta\|$ for some $\alpha < 1$, due to its contraction mapping behavior.
- Sharpness Aware Mirror Gradient’s flat minima ensure: $\|\nabla_{\theta}^2\mathcal{L}(\theta, x, y)\| \leq \beta$ for some bounded β , limiting the curvature of the loss landscape around the optimized parameters.
- By the mean value theorem and the properties of Lipschitz continuous functions, we can establish that:

$$\begin{aligned} \|\nabla_{\theta}\mathcal{L}(\theta, f_{BLIP}(x+\eta), y) - \nabla_{\theta}\mathcal{L}(\theta, f_{BLIP}(x), y)\| \\ \leq L\|f_{BLIP}(x+\eta) - f_{BLIP}(x)\| \leq L\alpha\|\eta\| \leq L\alpha\sigma \end{aligned}$$

where L is the Lipschitz constant of the gradient with respect to the input.

- Similarly, for parameter perturbations δ with $\|\delta\| \leq \rho$:

$$\|\nabla_{\theta}\mathcal{L}(\theta + \delta, x, y) - \nabla_{\theta}\mathcal{L}(\theta, x, y)\| \leq \beta\|\delta\| \leq \beta\rho$$

- Combining these bounds and setting $K = L\alpha\beta$ yields the final result.

This comprehensive theoretical framework establishes that while BLIP and Sharpness Aware Mirror Gradient operate on different aspects of the robustness problem (input space vs. parameter space), their combination provides multiplicative benefits for overall system robustness. These theoretical guarantees explain the empirical success of our approach and provide a solid foundation for understanding why SGBD significantly improves the efficacy and reliability of multimodal recommendation systems deployed in production environments.

6. Discussion

This work advances recommender system training by integrating BLIP’s noise-robust representations with enhanced sharpness-aware optimization. BLIP delivers high-quality multimodal embeddings through cross-modal bootstrapping and curated pretraining, enabling noise-tolerant representations for accurate recommendations. Our approach

leverages cross-attention fused embeddings over individual image-text embeddings due to superior downstream performance (detailed in supplementary). The SGBD framework guides optimization toward flat local minima for improved generalization and robustness. By penalizing sharp minima and facilitating escape from suboptimal solutions, it maintains stable optimization despite noisy gradients and complex loss landscapes.

Empirical results show 24.5% average improvement across metrics (REC, PREC, MAP, NDCG) for top-5 recommendations under BPR loss, demonstrating the effectiveness of combining robust multimodal representations with advanced optimization. These findings establish new directions for noise-aware recommendation techniques. Comparative analysis of BLIP1 and BLIP2 performance will be addressed in future work with product dataset fine-tuning.

7. Conclusion

In this paper, we introduced SGBD (Sharpness-Aware Mirror Gradient with BLIP-Based Denoising), a novel framework designed to tackle the visual robustness challenges prevalent in multi-modal recommender systems within the retail sector. As the retail industry rapidly adopts AI-driven solutions across both online and physical environments, recommendation systems must contend with noisy, diverse, and frequently updated visual data—ranging from fine-grained recognition of visually similar products to entirely new item classes introduced daily. SGBD bridges core ideas from computer vision, optimization, and recommender systems, showcasing how vision-language models can be extended beyond traditional use cases to address key retail challenges.

The core innovation of our approach lies in its dual strategy for improving visual robustness. First, we utilize BLIP’s vision-language architecture to denoise product imagery, extracting semantically meaningful representations that emphasize product-relevant features while suppressing background clutter or irrelevant visual noise. Second, we enhance the training process using a sharpness-aware mirror gradient method that encourages convergence toward flat minima, resulting in models that are more resilient to shifts in visual distribution—a common issue in dynamic retail environments. Through both theoretical analysis and empirical evaluation on REC, PREC, MAP, and NDCG metrics, we demonstrate that SGBD consistently outperforms strong baselines in terms of stability and generalization. These results underline the potential of our approach as a scalable and practical solution for building robust multimodal recommendation systems in real-world retail applications, where visual variability and large-scale data are the norm. These findings highlight the robustness and scalability of our approach for real-world multi-modal applications.

References

- [1] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions, 2021.
- [2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.
- [3] Yali Du, Meng Fang, Jinfeng Yi, Chang Xu, Jun Cheng, and Dacheng Tao. Enhancing the robustness of neural collaborative filtering systems under malicious attacks. *IEEE Transactions on Multimedia*, 21(3):555–565, 2019.
- [4] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2021.
- [5] Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, page 1623–1625, New York, NY, USA, 2022. Association for Computing Machinery.
- [6] Bowei He, Xu He, Yingxue Zhang, Ruiming Tang, and Chen Ma. Dynamically expandable graph convolution for streaming recommendation. In *Proceedings of the ACM Web Conference 2023*, page 1457–1467. ACM, 2023.
- [7] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016.
- [8] Zhongzhan Huang, Senwei Liang, Mingfu Liang, and Haizhao Yang. Dianet: Dense-and-implicit attention network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4206–4214, 2020.
- [9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.
- [10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- [11] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton 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*, page 43–52, New York, NY, USA, 2015. Association for Computing Machinery.
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [13] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback, 2012.
- [14] Jiajie Su, Chaochao Chen, Weiming Liu, Fei Wu, Xiaolin Zheng, and Haoming Lyu. Enhancing hierarchy-aware graph networks with deep dual clustering for session-based recommendation. In *Proceedings of the ACM Web Conference 2023*, page 165–176, New York, NY, USA, 2023. Association for Computing Machinery.
- [15] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. Adversarial training towards robust multimedia recommender system, 2019.
- [16] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *Trans. Multi.*, 25: 5107–5116, 2022.
- [17] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xueming Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2023.
- [18] Xixi Wu, Yun Xiong, Yao Zhang, Yizhu Jiao, Jiawei Zhang, Yangyong Zhu, and Philip S. Yu. Consrec: Learning consensus behind interactions for group recommendation. In *Proceedings of the ACM Web Conference 2023*. ACM, 2023.
- [19] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 353–362, New York, NY, USA, 2016. Association for Computing Machinery.
- [20] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.
- [21] Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima, 2024.
- [22] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions, 2023.
- [23] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation, 2023.
- [24] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, page 845–854. ACM, 2023.

SGBD: Sharpness-Aware Mirror Gradient with BLIP-Based Denoising for Robust Multimodal Product Recommendation

Supplementary Material

Model	Baby							
	REC@10	REC@20	NDCG@10	NDCG@20	PREC@10	PREC@20	MAP@10	MAP@20
DualGNN								
Vanilla	0.0297	0.0460	0.0161	0.0204	0.0033	0.0026	0.0116	0.0127
Vanilla+MG	0.0375	0.0598	0.0199	0.0256	0.0041	0.0033	0.0141	0.0156
Vanilla+SGBD	0.0402	0.0626	0.0199	0.0256	0.0041	0.0033	0.0141	0.0156
BLIP	0.0418	0.0657	0.0222	0.0284	0.0047	0.0037	0.0158	0.0174
BLIP+MG	0.0432	0.0651	0.0235	0.0291	0.0047	0.0036	0.0169	0.0184
BLIP+SGBD	0.0452	0.0682	0.0262	0.0305	0.0049	0.0038	0.0177	0.0192
BLIP2	0.0509	0.0810	0.0276	0.0354	0.0057	0.0045	0.0198	0.0218
BLIP2+MG	0.0461	0.0697	0.0251	0.0312	0.0051	0.0038	0.0181	0.0197
BLIP2+SGBD	0.0482	0.0703	0.0262	0.0325	0.0056	0.0046	0.0196	0.0206
Improv.	71.38%	76.09%	124.84%	73.53%	72.73%	76.92%	70.69%	62.20%
Dragon								
Vanilla	0.0536	0.0847	0.0285	0.0364	0.0059	0.0047	0.0202	0.0223
Vanilla+MG	0.0544	0.0837	0.0291	0.0365	0.0057	0.0044	0.0211	0.0231
Vanilla+SGBD	0.0544	0.0837	0.0291	0.0365	0.0057	0.0044	0.0211	0.0231
BLIP	0.0638	0.0991	0.0344	0.0435	0.0070	0.0055	0.0246	0.0271
BLIP+MG	0.0625	0.0947	0.0335	0.0419	0.0069	0.0053	0.0239	0.0261
BLIP+SGBD	0.0625	0.0947	0.0335	0.0419	0.0069	0.0053	0.0239	0.0261
BLIP2	0.0644	0.0971	0.0346	0.0430	0.0071	0.0054	0.0247	0.0269
BLIP2+MG	0.0643	0.0978	0.0348	0.0434	0.0071	0.0054	0.0249	0.0272
BLIP2+SGBD	0.0671	0.1021	0.0364	0.0453	0.0075	0.0057	0.0261	0.0285
Improv.	25.19%	15.47%	27.72%	24.45%	27.12%	21.28%	29.21%	27.80%
SLMRec								
Vanilla	0.0508	0.0716	0.0282	0.0336	0.0056	0.0040	0.0206	0.0220
Vanilla+MG	0.0509	0.0728	0.0284	0.0340	0.0056	0.0040	0.0207	0.0222
Vanilla+SGBD	0.0530	0.0772	0.0301	0.0360	0.0059	0.0042	0.0219	0.0235
BLIP	0.0506	0.0741	0.0282	0.0343	0.0056	0.0041	0.0207	0.0223
BLIP+MG	0.0504	0.0758	0.0280	0.0346	0.0056	0.0041	0.0207	0.0223
BLIP+SGBD	0.0542	0.0813	0.0301	0.0373	0.0060	0.0045	0.0220	0.0239
BLIP2	0.0493	0.0738	0.0272	0.0335	0.0055	0.0041	0.0196	0.0213
BLIP2+MG	0.0506	0.0745	0.0276	0.0338	0.0056	0.0041	0.0199	0.0215
BLIP2+SGBD	0.0557	0.0819	0.0304	0.0372	0.0062	0.0045	0.0219	0.0237
Improv.	9.65%	14.39%	7.8%	11.01%	10.71%	12.50%	6.80%	8.64%
Avg. Improv.	35.41%	35.32%	53.45%	36.33%	36.85%	36.90%	35.57%	32.89%

Table 5. Recommendation performance on Amazon dataset Baby. Metrics in color represent best performance for the particular evaluation metric.

Model	Sports							
	REC@10	REC@20	NDCG@10	NDCG@20	PREC@10	PREC@20	MAP@10	MAP@20
DualGNN								
Vanilla	0.0443	0.0693	0.0241	0.0305	0.0049	0.0039	0.0173	0.0190
Vanilla+MG	0.0437	0.0668	0.0241	0.0301	0.0049	0.0038	0.0175	0.0191
Vanilla+SGBD	0.0477	0.0707	0.0265	0.0325	0.0053	0.0039	0.0194	0.0210
BLIP	0.0442	0.0669	0.0240	0.0299	0.0049	0.0037	0.0172	0.0188
BLIP+MG	0.0416	0.0623	0.0227	0.0281	0.0046	0.0035	0.0164	0.0178
BLIP+SGBD	0.0509	0.0771	0.0286	0.0354	0.0057	0.0043	0.0210	0.0228
BLIP2	0.0457	0.0694	0.0259	0.0312	0.0051	0.0039	0.0181	0.0197
BLIP2+MG	0.0450	0.0695	0.0248	0.0311	0.0050	0.0039	0.0180	0.0197
BLIP2+SGBD	0.0492	0.0754	0.0275	0.0342	0.0055	0.0042	0.0201	0.0219
Improv.	14.90%	11.26%	18.67%	16.67%	16.33%	10.26%	21.39%	20.00%
Dragon								
Vanilla	0.0633	0.0944	0.0339	0.0420	0.0070	0.0052	0.0242	0.0264
Vanilla+MG	0.0623	0.0931	0.0340	0.0419	0.0068	0.0051	0.0246	0.0268
Vanilla+SGBD	0.0636	0.0975	0.0344	0.0431	0.0071	0.0054	0.0246	0.0270
BLIP	0.0638	0.0940	0.0341	0.0419	0.0070	0.0052	0.0243	0.0264
BLIP+MG	0.0602	0.0902	0.0326	0.0403	0.0066	0.0050	0.0234	0.0255
BLIP+SGBD	0.0622	0.0916	0.0356	0.0423	0.0088	0.0067	0.0258	0.0287
BLIP2	0.0638	0.0962	0.0347	0.0430	0.0070	0.0053	0.0250	0.0273
BLIP2+MG	0.0626	0.0937	0.0343	0.0423	0.0069	0.0052	0.0249	0.0270
BLIP2+SGBD	0.0652	0.0978	0.0358	0.0443	0.0072	0.0054	0.0260	0.0283
Improv.	3.00%	3.60%	5.6%	5.50%	25.71%	18.85%	7.40%	8.70%
SLMRec								
Vanilla	0.0668	0.0985	0.0373	0.0455	0.0074	0.0055	0.0274	0.0296
Vanilla+MG	0.0673	0.0989	0.0373	0.0455	0.0074	0.0055	0.0272	0.0294
Vanilla+SGBD	0.0702	0.1030	0.0389	0.0474	0.0077	0.0057	0.0285	0.0308
BLIP	0.0658	0.0964	0.0367	0.0446	0.0073	0.0054	0.0269	0.0290
BLIP+MG	0.0652	0.0968	0.0366	0.0448	0.0073	0.0054	0.0269	0.0291
BLIP+SGBD	0.0685	0.1016	0.0384	0.0471	0.0077	0.0057	0.0283	0.0306
BLIP2	0.0649	0.0974	0.0364	0.0448	0.0072	0.0055	0.0268	0.0290
BLIP2+MG	0.0664	0.0977	0.0370	0.0451	0.0074	0.0055	0.0271	0.0293
BLIP2+SGBD	0.0724	0.1075	0.0410	0.0497	0.0082	0.0060	0.0299	0.0323
Improv.	8.38%	0.41%	9.92%	9.23%	10.81%	9.09%	9.12%	9.12%
Avg. Improv.	8.76%	5.09%	11.40%	4.31%	17.62%	12.73%	12.64%	12.61%

Table 6. Recommendation performance on Amazon dataset Sports. Metrics in color represent best performance for the particular evaluation metric.

Model	Clothing							
	REC@10	REC@20	NDCG@10	NDCG@20	PREC@10	PREC@20	MAP@10	MAP@20
DualGNN								
Vanilla	0.0301	0.0458	0.0158	0.0198	0.0031	0.0024	0.0113	0.0124
Vanilla+MG	0.0302	0.0457	0.0158	0.0198	0.0032	0.0024	0.0113	0.0124
Vanilla+SGBD	0.0320	0.0485	0.0166	0.0210	0.0034	0.0025	0.0119	0.0130
BLIP	0.0459	0.0670	0.0243	0.0297	0.0047	0.0035	0.0175	0.0190
BLIP+MG	0.0351	0.0503	0.0185	0.0224	0.0037	0.0026	0.0133	0.0144
BLIP+SGBD	0.0380	0.0543	0.0199	0.0242	0.0040	0.0028	0.0143	0.0155
BLIP2	0.0472	0.0695	0.0258	0.0314	0.0049	0.0036	0.0190	0.0205
BLIP2+MG	0.0362	0.0546	0.0192	0.0238	0.0038	0.0029	0.0138	0.0151
BLIP2+SGBD	0.0387	0.0563	0.0216	0.0251	0.0041	0.0036	0.0156	0.0158
Improv.	56.81%	51.75%	63.29%	58.59%	58.06%	50.00%	68.14%	65.32%
Dragon								
Vanilla	0.0512	0.0760	0.0273	0.0336	0.0053	0.0039	0.0198	0.0215
Vanilla+MG	0.0512	0.0766	0.0274	0.0339	0.0053	0.0040	0.0199	0.0217
Vanilla+SGBD	0.0553	0.0824	0.0298	0.0364	0.0057	0.0043	0.0215	0.0235
BLIP	0.0667	0.0983	0.0362	0.0443	0.0069	0.0051	0.0257	0.0289
BLIP+MG	0.0535	0.0795	0.0295	0.0361	0.0056	0.0041	0.0220	0.0237
BLIP+SGBD	0.0559	0.0827	0.0308	0.0374	0.0059	0.0043	0.0229	0.0245
BLIP2	0.0690	0.1012	0.0378	0.0460	0.0072	0.0053	0.0280	0.0302
BLIP2+MG	0.0651	0.0935	0.0358	0.0430	0.0068	0.0049	0.0266	0.0286
BLIP2+SGBD	0.0695	0.1003	0.0383	0.0461	0.0073	0.0052	0.0287	0.0309
Improv.	35.74%	33.16%	40.29%	37.20%	37.74%	35.90%	44.95%	43.72%
SLMRec								
Vanilla	0.0447	0.0662	0.0245	0.0300	0.0047	0.0035	0.0181	0.0196
Vanilla+MG	0.0449	0.0667	0.0245	0.0301	0.0047	0.0035	0.0181	0.0196
Vanilla+SGBD	0.0477	0.0714	0.0262	0.0321	0.0050	0.0038	0.0195	0.0210
BLIP	0.0438	0.0650	0.0239	0.0293	0.0046	0.0034	0.0176	0.0191
BLIP+MG	0.0447	0.0671	0.0245	0.0302	0.0047	0.0035	0.0181	0.0196
BLIP+SGBD	0.0468	0.0702	0.0257	0.0317	0.0050	0.0037	0.0192	0.0208
BLIP2	0.0464	0.0682	0.0251	0.0306	0.0049	0.0036	0.0184	0.0199
BLIP2+MG	0.0459	0.0689	0.0250	0.0308	0.0048	0.0036	0.0184	0.0200
BLIP2+SGBD	0.0483	0.0726	0.0263	0.0324	0.0051	0.0038	0.0195	0.0211
Improv.	8.05%	9.67%	7.35%	8.00%	8.50%	8.57%	7.73%	7.65%
Avg. Improv.	35.53%	31.53%	36.98%	34.60%	34.76%	31.49%	40.27%	38.90%

Table 7. Recommendation performance on Amazon dataset Clothing. Metrics in color represent best performance for the particular evaluation metric.

Model	Electronics							
	REC@10	REC@20	NDCG@10	NDCG@20	PREC@10	PREC@20	MAP@10	MAP@20
DualGNN								
Vanilla	0.0193	0.0304	0.0104	0.0133	0.0022	0.0017	0.0074	0.0081
Vanilla+MG	0.0195	0.0307	0.0102	0.0132	0.0022	0.0018	0.0071	0.0079
Vanilla+SGBD	0.0203	0.0312	0.0116	0.0138	0.0027	0.0021	0.0081	0.0089
BLIP	0.0166	0.0255	0.0090	0.0113	0.0			