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

Multimodal recommender systems leverage diverse information to model user preferences and item features. While integrating multimodal data mitigates sparsity and cold-start challenges, it introduces information adjustment and noise risks. We analyze these systems through flat local minima and use BLIP's denoising capability to address inherent noise. Our proposed training strategy enhances model robustness during optimization. Through theoretical and empirical analyses, we demonstrate our approach's effectiveness across multiple recommendation models. The proposed Sharpness-Aware Mirror Gradient with BLIP-Based Denoising (SGBD) complements existing techniques and extends to advanced models, establishing a robust paradigm for multimodal recommendation.

1. Introduction

Multimodal recommender systems leverage texts, images, and videos to model user preferences and item features, helping users discover relevant items. While this integration mitigates challenges like data sparsity and cold-start issues [1] [5] [8] [19], it introduces two key risks that challenge system robustness.

The **information adjustment risk** stems from frequent updates to multimodal data, such as merchants modifying item keywords or images. The **inherent noise risk** occurs during training through subpar image quality, noisy text, or irrelevant features. These risks impair accurate user targeting, leading to suboptimal recommendations [15] [19].

To address these challenges [3], we propose a novel optimization strategy combining Sharpness-Aware Minimization (SAM) [4] with Mirror Gradient (MG) [21] to promote flat minima solutions. We leverage BLIP [9] for denoising multimodal inputs, significantly improving representation quality and recommendation accuracy.

*Equal contribution.

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

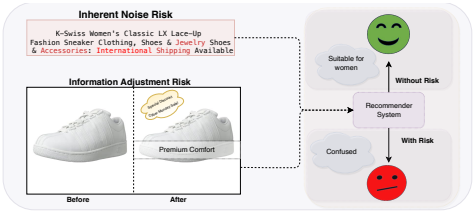
2. Preliminaries

Multimodal Product Recommender Let the set of customers be $\mathcal{U} = \{u_0, u_1, \ldots, u_n\}$ and the set of products be $I = \{i_0, i_1, \ldots, i_m\}$. Each customer $u \in \mathcal{U}$ has given an explicit positive feedback about product $\mathcal{I}_u \in \mathcal{I}$. For each product $\mathcal{I}_u \in \mathcal{I}$ the multimodal information is constituted by the visual features as $v_i \in \mathcal{V}$, textual features as $t_i \in \mathcal{T}$ and the multimodal recommendation model is represented by $\mathcal{R}$. The multimodal product preference score $y_{u,i}$ is computed as: $y_{u,i} = \mathcal{R}(u, i, v_i, t_i, \mathcal{I}_u|\theta)$ where θ represents the parameters of $\mathcal{R}$ and $y_{u,i}$ is the preference score that a customer u has for the product i. A high $y_{u,i}$ implies a high probability of customer u buying product i, hence the products with high $y_{u,i}$ form the recommendation set for a customer u. **Loss Function for Recommender Sytem** 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 $\mathcal{L}(.)$ to represent the overall loss function.

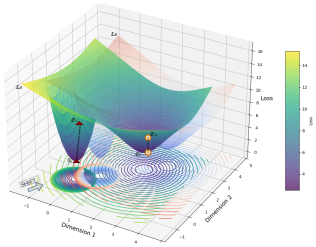
3. Methodologies

3.1. Overcoming Noise in Images and Text

Multimodal product data faces significant noise challenges, including low-resolution images, compression artifacts, and



(a) Multimodal risks illustration: Unrelated information in product descriptions and images can mislead feature generation, resulting in incorrect customer preference modeling and suboptimal recommendations.



(b) Impact of information adjustment on loss landscape: Sharp minima (ΔL_1) show greater loss increase than flat minima (ΔL_2), demonstrating the advantage of flat minima optimization for robustness.

Figure 1. Multimodal risks in production recommendation systems

inconsistent text descriptions. These issues compromise feature quality and recommendation accuracy.

BLIP (Bootstrapped Language-Image Pre-training) addresses these challenges through noise-robust contrastive learning and masked modeling. Key mechanisms include:

1. **Enhancement of Image Representations:** BLIP refines visual features through large-scale pre-training, filtering irrelevant artifacts while preserving salient item attributes [9].
2. **Refinement of Textual Descriptions:** Using masked language modeling and caption generation, BLIP denoises product descriptions and metadata, ensuring text captures relevant image aspects [9] [10].

The model's cross-attention mechanism aligns image features with text tokens, enabling effective semantic correlation extraction.

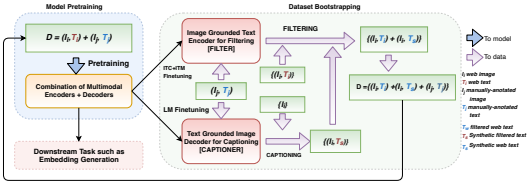


Figure 2. BLIP's bootstrapping-based denoising framework: A captioner generates synthetic captions while a filter removes noisy pairs. Both components, initialized from a pre-trained model and fine-tuned on human-annotated data, create clean training data. The framework maintains denoising capabilities during inference, ensuring robust real-world performance.

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 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 it's gradient is given by:

$$\mathcal{L}_{\text{SAM}}(\theta) = \mathcal{L}\left(\theta + \epsilon \frac{\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\| + \delta}\right), \quad \tilde{g} = \nabla_\theta \mathcal{L}_{SAM} \tag{1}$$

where $\mathcal{L}(.)$ is the loss function, θ represents the model parameters, ϵ is an adversarial perturbation, and ρ is a predefined 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 1
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 pretrained 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 $\nabla^2_\theta \mathcal{L}(\theta) \nabla_\theta \mathcal{L}(\theta) / (\|\nabla_\theta \mathcal{L}(\theta)\| + \delta)$ and multiplicative factor $[\alpha_1 / (\|\nabla_\theta \mathcal{L}(\theta)\| + \delta) - \alpha_2]$ to $\mathcal{L}_\theta$.

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

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

The effective loss objective becomes:

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

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

We evaluate our method using Graph Neural Network models (DualGNN [17], DRAGON [23]) and self-supervised learning (SLMRec [16]) on Amazon Product Recommendation datasets [11]. **Dataset:** Experiments use four multimodal Amazon datasets (Baby, Sports, Electronics, Clothing), following Zhou et al.'s [22] processing steps. Dataset statistics are in Table 1. **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:** We use BLIP encoder and BLIP2 Q-former for feature generation, with Adam optimizer ($\beta = 3$) on NVIDIA Tesla T4 GPU. Training runs for 1000 epochs with 20-step early stopping.

4.1. Results

From Table 2 and 3, we compare the proposed method's performance against the baselines and observe that proposed method delivers consistently higher performance across different models by an average of 24.5%. The additional benefit from both representation and modified SAM

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.

	Baby				Sports			
Model	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.

	Clothing				Electronics			
Model	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.

can be observed in Table 2 and 3. We demonstrate improvement for higher top-k values in appendix. In Table 4 in the supplementary material, 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. 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.

6. Conclusion

The proposed method SGBD addresses inherent noise and information adjustment risks in multimodal learning through BLIP-based noise-robust product representations and a modified SAM framework with Mirror Gradient, driving optimization toward flat local minima. Theoretical analysis and experiments on REC, PREC, MAP, and NDCG metrics demonstrate that our method outperforms baselines, effectively mitigating noise and enhancing generalization. These findings highlight the robustness and scalability of our approach for real-world multimodal 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. 1

[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

[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. 1

[4] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2021. 1, 2

[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. 1

[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. 3

[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. 1

[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. 1

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1, 2, 3

[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. 2, 3

[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. 3

[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. 3

[13] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback, 2012. 1

[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. 3

[15] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. Adversarial training towards robust multimedia recommender system, 2019. 1

[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. 1, 3

[17] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2023. 3

[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. 3

[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. 1

[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. 3

[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. 1, 3

[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. 3

[23] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation, 2023. 1, 3

[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. 1

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

Supplementary Material

Theoretical Connection Between BLIP and Sharpness Aware Mirror Gradient

The complementary nature of BLIP's denoising and Sharpness Aware Mirror Gradient optimization can be formally established through their distinct but synergistic effects on the loss landscape. Let $\mathcal{L}(\theta, x, y)$ be the loss function for parameters θ and input-output pairs (x, y).

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

$$\mathcal{R}_{total} = \mathcal{R}_{inherent} + \mathcal{R}_{adjustment} \tag{3}$$

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:

$$\mathcal{R}_{inherent}(f_{BLIP}(x)) \leq \mathcal{R}_{inherent}(x) \tag{4}$$

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

$$\mathbb{E}_{x\sim\mathcal{D}}[\|f_{BLIP}(x) - x^*\|] \leq \mathbb{E}_{x\sim\mathcal{D}}[\|x - x^*\|] \tag{5}$$

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:

$$\min_{\theta} \max_{\|\epsilon\|\leq\rho} \mathcal{L}(\theta + \epsilon, f_{BLIP}(x), y) \tag{6}$$

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

$$\mathcal{R}_{total}(\theta_{SAM_{MG}}, f_{BLIP}(x)) \leq \tag{7}$$

$$\min(\mathcal{R}_{total}(\theta, x), \mathcal{R}_{total}(\theta_{SAM_{MG}}, x)) \tag{8}$$

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:

$$\|\nabla_\theta \mathcal{L}(\theta, f_{BLIP}(x+\eta), y) - \nabla_\theta \mathcal{L}(\theta, f_{BLIP}(x), y)\| \leq K\rho \tag{9}$$

		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 $\epsilon \sim \mathcal{N}(0, 10^{-6})$.

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 $\alpha < 1$
- Sharpness Aware Mirror Gradient's flat minima ensure: $\|\nabla^2_\theta \mathcal{L}(\theta, x, y)\| \leq \beta$ for some bounded β
- The composition of these properties yields the final bound

This 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 improving the efficacy and reliability of systems deployed in production.

Baby								
Model	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	0.0362	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	0.0362	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.

Clothing								
Model	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	0.0267	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.

Sports								
Model	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	0.0250	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.

Electronics								
Model	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	0.0019	0.0015	0.0064	0.0070
BLIP+MG	0.0199	0.0303	0.0108	0.0135	0.0023	0.0017	0.0077	0.0084
BLIP+SGBD	0.0211	0.0316	0.0115	0.0139	0.0032	0.0022	0.0079	0.0089
BLIP2	0.0166	0.0260	0.0090	0.0114	0.0019	0.0015	0.0063	0.0070
BLIP2+MG	0.0208	0.0322	0.0113	0.0142	0.0023	0.0018	0.0081	0.0089
BLIP2+SGBD	0.0209	0.0331	0.0122	0.0156	0.0036	0.0019	0.0095	0.0096
Improv.	8.30%	8.88%	17.31%	17.29%	63.64%	29.41%	28.38%	18.52%
Dragon								
Vanilla	0.0317	0.0482	0.0175	0.0217	0.0036	0.0027	0.0126	0.0138
Vanilla+MG	0.0324	0.0492	0.0177	0.0220	0.0036	0.0028	0.0127	0.0138
Vanilla+SGBD	0.0324	0.0492	0.0177	0.0220	0.0036	0.0028	0.0127	0.0138
BLIP	0.0324	0.0485	0.0178	0.0220	0.0036	0.0027	0.0129	0.0140
BLIP+MG	0.0317	0.0483	0.0172	0.0215	0.0036	0.0027	0.0123	0.0134
BLIP+SGBD	0.0323	0.0485	0.0173	0.0215	0.0038	0.0028	0.0135	0.0144
BLIP2	0.0336	0.0512	0.0185	0.0230	0.0038	0.0029	0.0134	0.0146
BLIP2+MG	0.0325	0.0494	0.0179	0.0222	0.0037	0.0028	0.0129	0.0141
BLIP2+SGBD	0.0331	0.0496	0.0181	0.0235	0.0039	0.0036	0.0131	0.0153
Improv.	5.68%	6.22%	5.71%	8.29%	8.33%	33.33%	3.97%	10.87%
SLMRec								
Vanilla	0.0432	0.0641	0.0243	0.0297	0.0049	0.0037	0.0178	0.0193
Vanilla+MG	0.0434	0.0649	0.0246	0.0301	0.0049	0.0037	0.0181	0.0195
Vanilla+SGBD	0.0435	0.0651	0.0256	0.0323	0.0052	0.0039	0.0193	0.0198
BLIP	0.0448	0.0654	0.0254	0.0307	0.0051	0.0037	0.0187	0.0202
BLIP+MG	0.0448	0.0657	0.0254	0.0308	0.0051	0.0037	0.0187	0.0202
BLIP+SGBD	0.0457	0.0669	0.0256	0.0312	0.0058	0.0051	0.0193	0.0217
BLIP2	0.0448	0.0657	0.0254	0.0312	0.0051	0.0037	0.0187	0.0202
BLIP2+MG	0.0449	0.0657	0.0254	0.0307	0.0051	0.0038	0.0187	0.0201
BLIP2+SGBD	0.0483	0.0709	0.0270	0.0327	0.0054	0.0038	0.0202	0.0218
Improv.	11.81%	10.61%	11.11%	10.10%	18.37%	37.84%	13.48%	12.95%
Avg. Improv.	8.60%	19.79%	8.03%	11.89%	30.11%	33.53%	15.28%	14.11%

Table 8. Recommendation performance on Amazon dataset Electronics. Metrics in color represent best performance for the particular evaluation metric.