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, helping users discover relevant products. Integrating multimodal data can mitigate challenges like data sparsity and cold-start, but also introduces risks such as information adjustment and inherent noise, posing robustness challenges. In this paper, we analyze multimodal recommenders from the perspective of flat local minima and leverage the denoising capability of BLIP, a Vision Language Model, to mitigate the inherent noise risk in multimodal inputs. We propose a concise yet effective recommendation training strategy that can implicitly enhance model robustness during optimization, addressing instability risks. Extensive theoretical and empirical analyses demonstrate the superiority of our approach across multimodal recommendation models and benchmarks. The proposed method: Sharpness-Aware Mirror Gradient with BLIP-Based Denoising (SGBD) complements existing robust training techniques and can be easily extended to advanced recommendation models, making it a promising paradigm for training robust multimodal recommender systems.

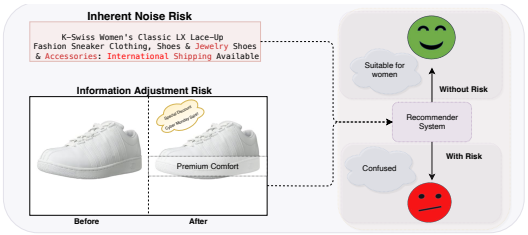
1 Introduction

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.

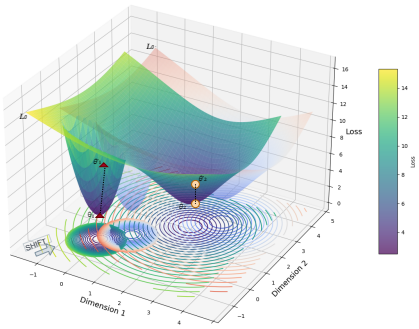
The **information adjustment risk** arises from the frequent modifications made to multimodal data, such as merchants updating keywords or images of items to keep up with trends and promotions. The **inherent noise risk** is present in the training phase, where the multimodal information, like subpar 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

*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

denoising capabilities of BLIP (Bootstrapped Language-Image Pre-training) (10) 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.

We address practical deployment challenges by proposing SGBD: Sharpness Aware Mirror Gradient with BLIP based denoising for building robust solutions for handling noisy inputs and information adjustments in production environments, with empirical validation across multiple recommendation models and datasets. Through strong theoretical analysis and extensive empirical experiments, we demonstrate the superiority of our proposed approach across various multimodal recommendation models and benchmarks. Additionally, we show that the integration of SAM with MG complements existing robust training methods and that the denoising capabilities of BLIP further enhance model performance. This makes our method a versatile and fundamental paradigm for training robust multimodal recommender systems, establishing a new benchmark for reliability and accuracy in the field.

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:

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

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.

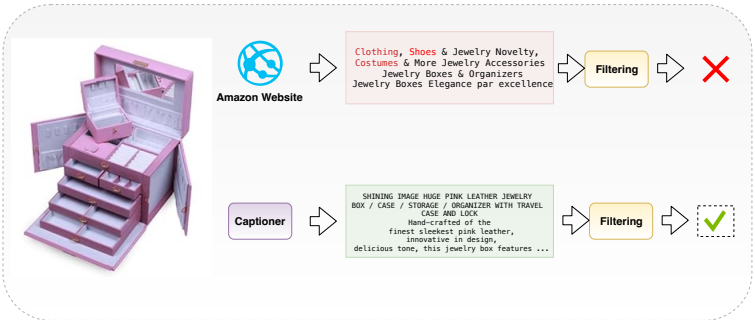


Figure 2: The captioning and filtering framework for BLIP. The captioner generates synthetic descriptions for image-text pairs (e.g., describing a jewelry box), while the filter removes noisy or irrelevant data. This process ensures that the resulting training dataset is cleaner and more representative,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

3 Methodology

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 (10).
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 (10) (9).

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

The denoising capabilities of BLIP extend beyond the training phase, providing robust handling of noisy input data during inference. Robust representations learned during pre-training enable BLIP to generalize effectively to unseen noisy data, ensuring consistent

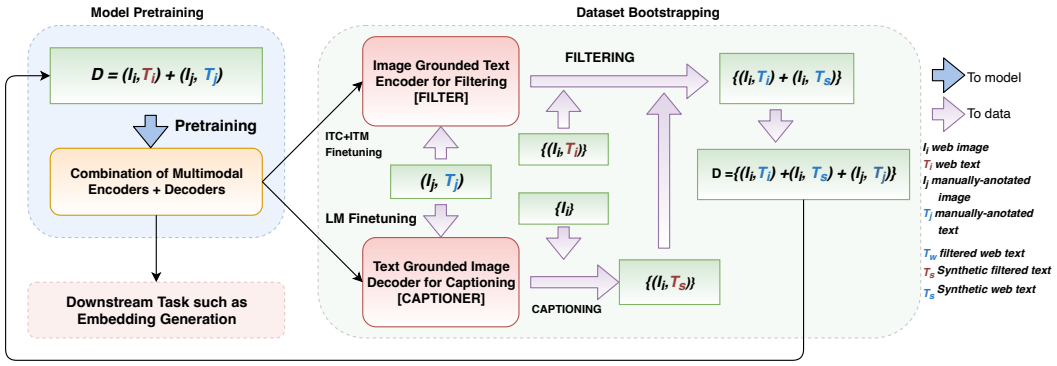


Figure 3: The learning framework of BLIP highlights its bootstrapping-based denoising capability. A captioner generates synthetic captions for web images, while a filter discards noisy image-text pairs. Both the captioner and filter are initialized from the same pre-trained model and fine-tuned on a small-scale human-annotated dataset to ensure quality. The resulting bootstrapped dataset, free from noise, is used for pre-training a new model, enhancing its robustness and generalization. During inference, the denoising mechanism is retained, allowing the model to filter noisy inputs dynamically and maintain high-quality performance in real-world scenarios.

recommendation performance across diverse input conditions. The key advantage of BLIP is its robustness to modality gaps and domain shifts during inference. By employing contrastive learning objectives during training, the model learns to associate semantically similar image-text pairs while distinguishing dissimilar ones. The learned representation not only encapsulates high-level semantics from individual modalities but also integrates cross-modal context. For example, in an image captioning task, visual regions are dynamically weighted by their relevance to specific textual tokens, enabling nuanced and contextually aware caption generation (10). Similarly, in retrieval tasks, BLIP computes similarity scores in the fused latent space, facilitating precise matching of visually and textually aligned inputs.

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:

$$\mathcal{L}_{\text{SAM}}(\theta) = \mathcal{L}\left(\theta + \epsilon \frac{\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\| + \delta}\right), \quad \tilde{g} = \nabla_\theta \mathcal{L}_{SAM} \tag{2}$$

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, which are known to generalize better (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 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 2
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. To address the challenge of inherent noise in multimodal data during training, BLIP models employ a combination of robust architectural and optimization strategies. Central to this approach is contrastive learning, which aligns semantically meaningful image-text pairs while separating noisy or irrelevant pairs, as established in prior work (10) (9). This ensures that the model focuses on high-quality relationships during training.

To further enhance robustness, BLIP incorporates a cross-modal bootstrapping mechanism (10), where high-quality signals from one modality (e.g., visual) guide the refinement of noisy embedding in the other (e.g., textual). This interplay ensures balanced learning across modalities, leveraging alignment to reduce noise impact effectively.

The use of frozen pre-trained language models (e.g., FLAN-T5 or OPT)(20)(9)(2) provides robust semantic grounding. These models map noisy textual inputs to consistent embedding, as demonstrated in large-scale pretraining studies. Pretraining on carefully curated multimodal datasets further enhances foundational robustness by minimizing empirical risk on clean distributions (12; 10). This allows the model to adapt effectively during fine-tuning on noisier downstream datasets.

Together, these strategies enable BLIP to construct robust multimodal representations, significantly mitigating noise-related risks in large-scale datasets.

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: step 5 and step 6 described in algorithm 1 are equivalent to introducing an additional regularization term $\nabla^2_\theta \mathcal{L}(\theta) \nabla_\theta \mathcal{L}(\theta) / (\|\nabla_\theta \mathcal{L}(\theta)\| + \delta)$ *and an additional multiplicative factor of* $[\alpha_1 / (\|\nabla_\theta \mathcal{L}(\theta)\| + \delta) - \alpha_2]$ *to the original loss objective* $\mathcal{L}_\theta$

Proof: Substituting $\tilde{\theta}$ from Step 5 into Step 6 in alg. 1 and applying a Taylor expansion for $\mathcal{L}_{SAM}(\theta)$ as defined in Eq. 2, we can write the final update θ' as:

$$\theta' = \theta - \nabla_\theta \left(\mathcal{L}_\theta(\theta) \left(\frac{\alpha_1}{||\nabla_\theta \mathcal{L}_\theta(\theta)|| + \delta} - \alpha_2 \right) + \alpha_1 \alpha_2 \nabla_\theta^2 \left(\frac{\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\| + \delta} \right) \right) \tag{3}$$

Thus, the effective loss objective becomes:

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

This formulation introduces two key adjustments:

1. A multiplicative factor scaling the original loss, controlling gradient updates based on their magnitude relative to α_1/α_2.
2. A regularization term proportional to the curvature (Hessian) of the loss landscape. This term dynamically adjusts the optimization trajectory, penalizing sharp minima with high curvature while incentivizing flat minima.

When $\|\nabla_\theta \mathcal{L}(\theta)\| + \delta \geq \alpha_1/\alpha_2$, the gradient's sign reverses, effectively pushing the weights away from sharp minima. The curvature-dependent regularization term further discourages convergence to regions with high curvature, promoting exploration of flatter solutions. Notably, when escaping sharp minima, the regularization term may become negative, amplifying the gradient to accelerate movement toward smoother regions. As the model approaches a saddle point, the curvature term diminishes in magnitude, allowing stable convergence.

By balancing gradient scaling and curvature-dependent regularization, the proposed framework ensures robustness against noisy gradients and sharp minima, enabling the model to consistently converge to flat, generalizable solutions.

A detailed discussion on how BLIP based denoising complements the proposed Sharpness Aware Mirror Gradient is stated in the Appendix.

4 Experiments

To establish the competitiveness of the proposed method, we conduct thorough experiments where we train a variety of multimodal recommendation models including Graph Neural Network based models: DualGNN (17) and DRAGON (23), and self supervised learning based model: SLMRec (16) on 4 Amazon Product Recommendation datasets (11). **Dataset** The experiments are conducted on four multimodal Amazon datasets: Baby, Sports, Electronics and Clothing where each sample is a pair of product images and their text description. The data processing follows same steps as outlined by Zhou et al (22). The exact statistics of the dataset is mentioned in table 1. **Metric** We compare the top-k precision (PREC), recall (REC), mean average precision (MAP) and normalized discounted cumulative gain (NDCG) as these top-k metrics help us identify the most important products for recommendation (6)(14)(18)(22). These four evaluation metrics capture complementary aspects of system performance: REC assesses user interest coverage, PREC measures recommendation accuracy, MAP evaluates average ranking accuracy, and NDCG highlights ranking quality. Together, they provide a holistic evaluation of the recommender system. **Baselines** We use the DualGNN, DRAGON and SLMRec as the baseline along with their variants trained using Mirror Gradient from (21). The baselines are trained using the original multimodal features where description based feature has been computed as the sentence embedding from all-MiniLM-L6-v2 while the visual features are generated using deep CNN from (11). **Implementation Details** For product image and decription 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 (9). 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.

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

	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 3: 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 4: 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

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 incremental individual benefit from both BLIP based denoising in the product representation and Sharpness Aware Mirror Gradient can be observed in table 2 and 3. We demonstrate improvement for higher top-k values in appendix. In table 4, we observe that the proposed method delivers more robust flat minima generalized solution that doesn't change much with respect to injection of Gaussian noise in input feature as compared to that in the existing baseline.

5 Discussion

This work presents a significant advancement in training recommender systems by integrating BLIP's noise-robust image-text fused representations with the enhanced sharpness aware optimization framework. The use of BLIP ensures high-quality multimodal embeddings by effectively mitigating inherent noise through cross-modal bootstrapping and pretraining on curated datasets. This enables the model to capture rich, noise-tolerant representations critical for improving recommendation accuracy in complex multimodal settings. We make use of cross attention based fused embedding against individual image-text embedding for product representation due to superior performance of fused embedding on downstream task (will be shared in appendix).

The proposed SGBD framework further enhances the training process by dynamically adjusting gradients to guide optimization toward flat local minima, which are associated with improved generalization and robustness. By penalizing sharp minima and amplifying escape from suboptimal solutions, SGBD ensures a stable and effective optimization trajectory even

in the presence of noisy gradients and challenging loss landscapes. As established in (21) the Mirror Gradient technique outperforms Sharpness Aware Minimization (4) in a one-to-one setting. We demonstrate in this work that when combined together, the two techniques complement each other to deliver an even more robust solution.

Empirical results demonstrate the efficacy of the proposed method, achieving 24.5% average improvement across key metrics (REC, PREC, MAP, and NDCG) across top 5 recommendations under the Bayesian Personalized Ranking (BPR) loss framework. This significant performance gain underscores the synergy between robust multimodal representations and advanced optimization strategies in building state-of-the-art recommender systems. These findings open avenues for further exploration of noise-aware training and optimization techniques in recommendation tasks. The difference in performance of BLIP1 and BLIP2 will be discussed in a future work where we finetune these models on product dataset.

6 Conclusion

The proposed method SGBD: Sharpness Aware Mirror Gradient with BLIP based denoising 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 adaptivity of our approach for real-world multimodal applications.

References

[1] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. Bias and debias in recommender system: A survey and future directions, 2021.

[2] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022.

[3] Du, Y., Fang, M., Yi, J., Xu, C., Cheng, J., and Tao, D. Enhancing the robustness of neural collaborative filtering systems under malicious attacks. *IEEE Transactions on Multimedia 21*, 3 (2019), 555–565.

[4] Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization, 2021.

[5] Gao, C., Wang, X., He, X., and Li, Y. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2022), WSDM '22, Association for Computing Machinery, p. 1623–1625.

[6] He, B., He, X., Zhang, Y., Tang, R., and Ma, C. Dynamically expandable graph convolution for streaming recommendation. In *Proceedings of the ACM Web Conference 2023* (Apr. 2023), WWW '23, ACM, p. 1457–1467.

[7] He, R., and McAuley, J. Vbpr: Visual bayesian personalized ranking from implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence 30*, 1 (Feb. 2016).

[8] Huang, Z., Liang, S., Liang, M., and Yang, H. Dianet: Dense-and-implicit attention network. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 04 (Apr. 2020), 4206–4214.

[9] Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[10] Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[11] McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2015), SIGIR '15, Association for Computing Machinery, p. 43–52.

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763.

[13] Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. Bpr: Bayesian personalized ranking from implicit feedback, 2012.

[14] Su, J., Chen, C., Liu, W., Wu, F., Zheng, X., and Lyu, H. Enhancing hierarchy-aware graph networks with deep dual clustering for session-based recommendation. In *Proceedings of the ACM Web Conference 2023* (New York, NY, USA, 2023), WWW '23, Association for Computing Machinery, p. 165–176.

[15] Tang, J., Du, X., He, X., Yuan, F., Tian, Q., and Chua, T.-S. Adversarial training towards robust multimedia recommender system, 2019.

[16] Tao, Z., Liu, X., Xia, Y., Wang, X., Yang, L., Huang, X., and Chua, T.-S. Self-supervised learning for multimedia recommendation. *Trans. Multi. 25* (June 2022), 5107–5116.

[17] Wang, Q., Wei, Y., Yin, J., Wu, J., Song, X., and Nie, L. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia 25* (2023), 1074–1084.

[18] Wu, X., Xiong, Y., Zhang, Y., Jiao, Y., Zhang, J., Zhu, Y., and Yu, P. S. Consrec: Learning consensus behind interactions for group recommendation. In *Proceedings of the ACM Web Conference 2023* (Apr. 2023), WWW '23, ACM.

[19] Zhang, F., Yuan, N. J., Lian, D., Xie, X., and Ma, W.-Y. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, Association for Computing Machinery, p. 353–362.

[20] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022.

[21] Zhong, S., Huang, Z., Li, D., Wen, W., Qin, J., and Lin, L. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima, 2024.

[22] Zhou, H., Zhou, X., Zeng, Z., Zhang, L., and Shen, Z. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions, 2023.

[23] Zhou, H., Zhou, X., Zhang, L., and Shen, Z. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation, 2023.

[24] Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., and Jiang, F. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023* (Apr. 2023), WWW '23, ACM, p. 845–854.

Appendix

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:

$$\mathcal{R}_{total} = \mathcal{R}_{inherent} + \mathcal{R}_{adjustment} \tag{4}$$

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:

$$\mathcal{R}_{inherent}(f_{BLIP}(x)) \leq \mathcal{R}_{inherent}(x) \tag{5}$$

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

$$\mathbb{E}_{x\sim\mathcal{D}}[\|f_{BLIP}(x) - x^*\|] \leq \mathbb{E}_{x\sim\mathcal{D}}[\|x - x^*\|] \tag{6}$$

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:

$$\min_{\theta} \max_{\|\epsilon\|\leq\rho} \mathcal{L}(\theta + \epsilon, f_{BLIP}(x), y) \tag{7}$$

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

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

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:

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

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.

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	0.0362	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.

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	0.0250	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.

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	0.0267	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.

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	0.0019	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.