Adaptive Gradient Masking for Balancing ID and MLLM-based Representations in Recommendation

Yidong Wu *
Imperial College London
London, UK

Siyuan Chen *
University of Bristol
Bristol, UK

Binrui Wu
Fudan University
Shanghai, China

Fan Li
KuaiShou Technology
Beijing, China

Jiechao Gao †
Stanford University
Stanford, USA

Abstract

In large-scale recommendation systems, multimodal (MM) content is increasingly introduced to enhance the generalization of ID features. The rise of Multimodal Large Language Models (MLLMs) enables the construction of unified user and item representations. However, the semantic distribution gap between MM and ID representations leads to *convergence inconsistency* during joint training: the ID branch converges quickly, while the MM branch requires more epochs, thus limiting overall performance. To address this, we propose a two-stage framework including MM representation learning and joint training optimization. First, we fine-tune the MLLM to generate unified user and item representations, and introduce collaborative signals by post-aligning user ID representations to alleviate semantic differences. Then, we propose an Adaptive Gradient Masking (AGM) training strategy to dynamically regulate parameter updates between ID and MLLM branches. AGM estimates the contribution of each representation with mutual information, and applies non-uniform gradient masking at the sub-network level to balance optimization. We provide theoretical analysis of AGM's effectiveness and further introduce an unbiased variant, AGM*, to enhance training stability. Experiments on offline and online A/B tests validate the effectiveness of our approach in mitigating convergence inconsistency and improving performance.

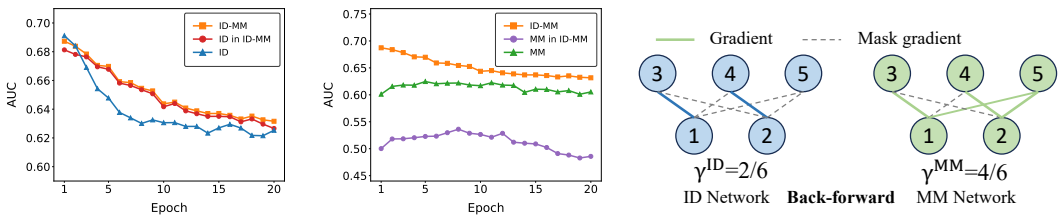
1 Introduction

Large-scale industrial recommendation systems have traditionally relied on ID-based features, such as cross ID features (e.g., FM [1], DCN [2]), list-wise ID features (e.g., DIN [3], TWIN [4]), to model user–item interactions. These features offer strong memorization capabilities and are effective in capturing co-occurrence patterns. However, they suffer from limited generalization, making them inadequate for addressing long-tail items, data sparsity, and cold-start scenarios. To mitigate these limitations, recent research [5, 6, 7] has incorporated multimodal (MM) content (e.g., images and textual descriptions) to enrich the semantic representations of users and items.

The emergence of Multimodal Large Language Models (MLLMs) facilitates the generation of unified, high-level semantic embeddings from diverse modalities, offering promising avenues for enhancing recommendation performance. As is demonstrated in many recent studies, integrating categorical features (e.g., ID and category) with MLLM-based representations (e.g., images and texts) can

*Equal Contribution
†Corresponding Author: jiechao@stanford.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) ID (b) MM (c) Gradient Masking.

Figure 1: (a) Validation-set AUC comparison between the ID-only model (ID) and the combined model (ID-MM), where "ID in ID-MM" denotes the contribution of the ID component within the combined model; (b) Validation-set AUC comparison between the multimodal-only model (MM) and the combined model (ID-MM), where "MM in ID-MM" denotes the contribution of the multimodal component within the combined model. (c) The illustration of Adaptive Gradient Masking.

effectively foster the performance of recommendation system [8, 9, 10, 11]. Works like FREEDOM [9] introduce auxiliary losses on multimodal data to refine ID embeddings through graph structures. Recent self-supervised learning models such as BM3 [10] adopt joint training of content-based and user-item objectives. AlignRec [12] proposes a unified alignment framework that aligns visual, textual, and categorical modalities using pretraining and contrastive learning.

Despite these advances, integrating ID and MLLM-based representations remains challenging due to convergence inconsistency during joint training. This stems from a semantic gap—ID embeddings encode co-occurrence patterns, while MLLM features capture high-level semantics—and an optimization imbalance, as ID embeddings are trainable whereas MLLM parameters are typically frozen. As a result, the ID branch quickly dominates training, leading to biased gradients that suppress the multimodal branch and ultimately degrade overall performance. To address this issue, although AlignRec [12] alleviates this convergence speed mismatch between multimodal and ID features by adopting a two-stage strategy: first, pre-training the alignment of content modes and then performing joint training, it does not fundamentally address the imbalance during the joint optimization phase. In contrast, we take a more fine-grained approach by dynamically balancing gradient updates between ID and multimodal branches.

In more detail, we propose a two-stage framework consisting of multimodal representation learning and joint training optimization. Firstly, we fine-tune a pretrained Multimodal Large Language Model (MLLM) to generate unified multimodal representations for items and users from visual and textual content. Specifically, Item representations are obtained from multimodal prompts through a designated output token, while user representations are derived by aggregating historical item representations. Besides, to bridge the semantic gap between ID and MM representations, we introduce collaborative alignment by post-aligning multimodal embeddings with their corresponding ID embeddings. In the second stage, we jointly train the recommendation model by combining ID and MLLM-based representations under the binary cross-entropy objective, and propose an Adaptive Gradient Masking (AGM) strategy to dynamically regulate their parameter updates during optimization. AGM estimates the informativeness of each representation through mutual information, and applies non-uniform gradient masking at the subnetwork level to encourage balanced convergence. This adaptive mechanism prevents the ID branch from dominating training and ensures consistent convergence across both branches. To demonstrate the effectiveness of our approach, we provide a theoretical analysis showing that AGM leads to more balanced gradient updates, thereby promoting consistent convergence of both branches. Furthermore, we propose an unbiased variant, AGM*, which improves training stability by correcting the bias introduced by binary masking.

Our contributions are summarized as follows:

- We analyze *convergence inconsistency* issue between ID and MLLM-based representations during joint training in recommendations, and propose a two-stage method to address it.
- We propose Adaptive Gradient Masking (AGM), a subnetwork-level optimization strategy that dynamically balances gradient updates between ID and multimodal branches. We also introduce AGM*, an unbiased variant that enhances stability.

- We fine-tune MLLMs to generate unified user and item representations, and introduce a collaborative alignment mechanism to bridge the semantic gap between ID and MM embeddings.
- We provide theoretical analysis on the convergence properties of our method, and validate their effectiveness through extensive offline experiments and online A/B testing in a large-scale industrial recommendation system.

2 Related works

Multi-modal recommendation Multi-modal recommendation extends the classic collaborative filtering paradigm by integrating diverse content modalities (e.g., images, text, videos) to capture richer contextual signals and thereby enhance recommendation accuracy. Early approaches, such as VBPR [13] and methods fusing visual features with ID embeddings [14, 15], showed that combining basic item side information can significantly improve user–item matching. Subsequently, attention-based architectures, including VECF [16] and MAML [17], explored finer-grained user preferences by leveraging mechanisms like image segmentation [18] and multi-modal feature interactions. With the surge of Graph Neural Networks (GNNs) in recommendation [19], models like MMGCN [20], GRCN [21], and DualGNN [22] pushed multi-modal recommendation further by injecting high-order neighbor relationships or user attentions across item modalities into node representations. To better reveal item–item semantic similarities, LATTICE [23] constructs separate item–item graphs for each modality and fuses them into a latent graph, while MVGAE [24] employs a modality-specific variational graph autoencoder to combine multi-modal embeddings. Later, MGCN [8] constructs separate graph views to fuse text, image, and user–item interactions more effectively. Recent work like GUME [25] focuses on leveraging semantic neighbors and refining user modality embeddings to strengthen long-tail item connectivity, while LGMRec [26] separates local user–item interactions from global attribute relationships via hypergraph modeling. FREEDOM [9] tackles noisy item–item structures by freezing precomputed graphs and pruning user–item edges. Self-supervised learning method such as BM3 [10] proposes a self-supervised learning framework that relies on latent embedding dropout to create view augmentations. AlignRec [12] addresses alignment challenges across different modalities by unifying multi-modal content and ID-based features through a multi-stage alignment process. In addition, the remarkable progress of foundation models in various modalities [27, 28, 29] has prompted researchers to adopt large-scale pretrained encoders for capturing more holistic multi-modal representations. Typical examples include VIP5 [30], which extends the text-based P5 [31] by incorporating a CLIP image encoder, and MMGRec [32], which reveals item IDs from both collaborative and multi-modal signals via a Graph RQ-VAE. Moreover, IISAN [33] proposes a lightweight Decoupled PEFT architecture that simultaneously tackles intra- and inter-modal adaptation in a plug-and-play manner.

Multi-modal Large Language Model Multi-modal Large Language Models (MLLMs) have recently achieved significant progress in integrating language with other modalities, driven by the surge in large-scale pretraining [29, 34, 35, 36]. Research efforts generally begin with multimodal understanding and text generation, with representative models such as BLIP-2 [37] and LLAVA [38]. Models like LLaMA-Adapter [39, 40] and mPLUG-Owl [41, 42] align text and image features via extensive image–text pairs, while InstructBLIP [43] reshapes multiple tasks into instruction-based formats. Despite such progress, enhancing the visual encoder resolution [44, 45, 46, 47, 48, 49] can result in prohibitive memory overhead, especially in multi-page scenarios. To address such a problem, TextMonkey [50] employs token resampling to reduce the visual token load. Similarly, more recent models such as Qwen2-VL [51] and GPT-4 [52] have exhibited outstanding proficiency in multimodal reasoning and generation.

3 Methodology

3.1 Problem formulation

We consider the Click-through Rate (CTR) task defined on a dataset $\mathcal{D} = \{(u_i, v_i, y_i)\}_{i=1}^{N}$, where each sample consists of a user $u = (\mathbf{e}_u^{\mathrm{id}}, \mathbf{e}_u^{\mathrm{mm}})$, an item $v = (\mathbf{e}_v^{\mathrm{id}}, \mathbf{e}_v^{\mathrm{mm}})$, and a binary label $y \in \{0, 1\}$ indicating whether the user engaged with the item. Here, $\mathbf{e}_u^{\mathrm{id}}$ and $\mathbf{e}_v^{\mathrm{id}}$ denote trainable ID embeddings

for user and item, respectively, while $\mathbf{e}_u^{\mathrm{mm}}$ and $\mathbf{e}_v^{\mathrm{mm}}$ represent multi-modal representations of user and item, extracted from MLLM. The goal is to learn a prediction function $f(u_i, v_i; \boldsymbol{\theta})$ that estimates the probability of user–item interaction. The model is trained to minimize binary cross-entropy loss.

3.2 Multimodal representation learning

Multimodal information, such as text, images, and other item-related metadata, provides substantial advantages by enhancing the representation of both items and users in recommendation systems. However, although pre-trained Multimodal Large Language Models (MLLMs) excel at understanding the data representation [52, 53, 54], their original evaluation metrics are not specifically designed to meet the unique demands of recommendation tasks. As a result, when faced with extensive user and item information, these pre-trained models often struggle to extract key, effective feature embeddings. To address this limitation and improve the extraction of multimodal features, we introduce a novel approach to efficiently generate task-relevant embeddings by harnessing multimodal features.

3.2.1 Item embedding

In this section, we fine-tune the MLLM using three novel alignment objectives aimed at enhancing cross-modal consistency. Furthermore, we append a special token, [Item_cls], to the end of each item description, which allows the model to condense lengthy multimodal token sequences into compact and informative embeddings.

For item i, we first combine its textual and visual attributes into a unified input description. This is accomplished using a specific prompt template designed to guide the model's multimodal understanding: "Integrate text and visual information into an embedding representation. Textual: `[Text]`, Visual: `[Image/video]`." Then the MLLM encodes the input and generates a corresponding token sequence including [item_cls], in the form of $\{t_1, t_2, ..., t_m, [item\_cls]\}$. Finally, the hidden state associated with the [Item_cls] token is extracted as the multimodal embedding for item i.

$$\mathbf{e}_v^{mm} = \text{MLLM}(\text{text}_i, \text{image}_i), \tag{1}$$

where $\mathbf{e}_v^{mm}$ donates the multimodal item embedding of item v.

In the fine-tuning phase, we introduce three specialized alignment tasks for multimodal recommendation, aimed at improving the MLLM's performance and suitability in recommendation scenarios.

Text-image alignment: To align visual and textual features, we adopt a method inspired by BERT [55]. For item i with image V_i and text T_i, we mask 20% of T_i's tokens with a special `[MASK]` token, obtaining $\hat{T}_i$. The model then takes $(V_i, \hat{T}_i)$ as input, with the corresponding original description T_i as the target output. This reconstruction task compels the model to leverage visual information to infer missing textual content, thereby learning the meaningful relationship between visual features and textual context for improved cross-modal understanding.

Meta-data processing: Recommendation systems leverage both structured metadata (e.g., title, price, tags) and unstructured descriptions. Since metadata directly reflects item characteristics, its effective processing enhances MLLMs' encoding performance. Thus, for item i, we propose predicting its detailed description T_i from its metadata, establishing a robust mapping between structured attributes and unstructured text.

User behavior understanding: The model explicitly captures interest evolution patterns by predicting users' future interactions based on their multimodal historical behavior sequences, enabling adaptive optimization of recommendation strategies. For this purpose, we create fine-tuning samples where a user's interaction history (containing both textual and visual item features) serves as input, while the next interacted item provides the supervision signal.

3.2.2 User embedding

Despite the basic user information, analyzing historical item sequences is also crucial for predicting user preferences. However, handling extensive user histories and aligning textual information with corresponding images presents a significant challenge in multimodal recommendation scenarios. To address the efficient aggregation of long multimodal sequences, we propose the User Embedding Generator (UEG). This module is designed to efficiently aggregate the sequence multimodal information

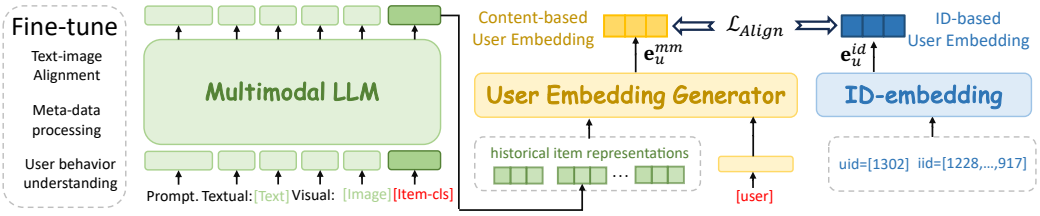


Figure 2: Multimodal representation learning

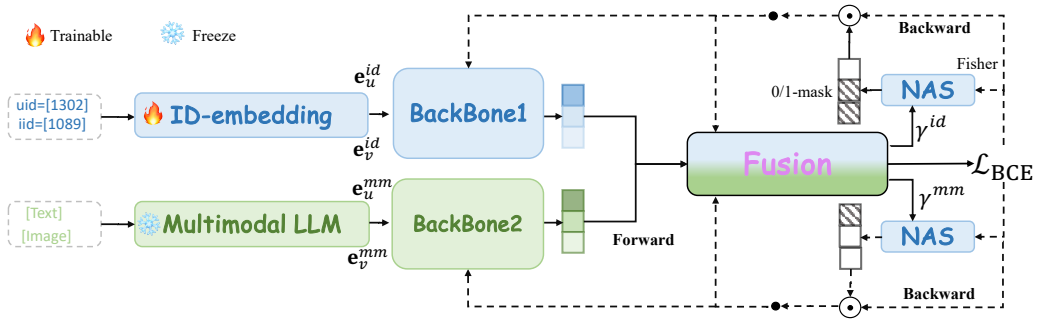


Figure 3: Adaptive Gradient Masking(AGM).

of user historical items into a unified user multimodal representation. For user u,

$$\mathbf{e}_u^{mm} = \text{UEG}(\mathbf{e}_{v_1}^{mm}, \mathbf{e}_{v_2}^{mm}, ..., \mathbf{e}_{v_s}^{mm}) \tag{2}$$

As illustrated in Fig. 2, the UEG is a learnable neural network module that takes both the multimodal representations of a user's historical interactions and their unique identifier as input, producing a comprehensive multimodal user embedding $\mathbf{e}_u^{mm}$. To stabilize the learning of multimodal user representations, we further incorporate a pre-trained ID-based embedding layer(with frozen parameters), which generates an ID-based user representation $\mathbf{e}_u^{id}$ from the user's identifier and historical interactions. The UEG module is optimized via an alignment loss:

$$\mathcal{L}_{align} = \|\mathbf{e}_u^{id} - \mathbf{e}_u^{mm}\|_2^2 \tag{3}$$

This objective ensures the learned multimodal representations maintain consistency with established ID-based embeddings while capturing rich multimodal patterns.

3.3 AGM

Forward propagation and convergence inconsistency Before computing interaction logits, we first obtain ID-based embeddings $(\mathbf{e}_u^{id}, \mathbf{e}_v^{id})$ through trainable embedding layers and multimodal embeddings $(\mathbf{e}_u^{mm}, \mathbf{e}_v^{mm})$ using the method described in Section 3.2. These are processed through separate backbones $g^{id}(\cdot,\cdot;\boldsymbol{\theta}^{id})$ and $g^{mm}(\cdot,\cdot;\boldsymbol{\theta}^{mm})$, with their outputs concatenated and fused through φ to produce the final logit:

$$f(u,v;\boldsymbol{\theta}) = \varphi([g^{id}(\mathbf{e}_u^{id}, \mathbf{e}_v^{id}; \boldsymbol{\theta}^{id}); g^{mm}(\mathbf{e}_u^{mm}, \mathbf{e}_v^{mm}; \boldsymbol{\theta}^{mm})]) \tag{4}$$

We train using binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \left(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right), \tag{5}$$

where $\hat{y}_i = \sigma(f(u,v;\boldsymbol{\theta}))$. By decomposing φ's weight matrix into W^{id} and W^{mm}, we can express Equation 4 as:

$$f(u,v;\boldsymbol{\theta}) = W^{id} * g^{id}(\mathbf{e}_u^{id}, \mathbf{e}_v^{id}; \boldsymbol{\theta}^{id}) + W^{mm} * g^{mm}(\mathbf{e}_u^{mm}, \mathbf{e}_v^{mm}; \boldsymbol{\theta}^{mm}) + b \tag{6}$$

As shown in Appendix C, the ID and MM branches update nearly independently. However, ID features converge faster with stronger signals, dominating predictions and gradient updates. This creates a feedback loop where MM components receive weakened optimization signals, remaining under-utilized at convergence. Consequently, the ID pathway determines the final logit while the MM pathway stays inadequately optimized, causing the observed convergence inconsistency.

Mask ratio via modal significance As discussed, ID and MM representations carry signals of varying strengths during training. To quantify the influence of different representations within the training objective, we introduce a contribution score s. For a given sample, the contribution score is formulated as:

$$s = y * p + (1 - y) * (1 - p), \quad \text{where} \quad p = \sigma(W * g(\mathbf{e}_u, \mathbf{e}_v; \boldsymbol{\theta}) + b/2) \tag{7}$$

The relative contribution ratios ρ^{id} and ρ^{mm} are calculated per mini-batch $\mathcal{B}$:

$$\rho^{id} = \frac{\sum_{x_i \in \mathcal{B}} s_i^{id}}{\sum_{x_i \in \mathcal{B}} s_i^{mm}}, \quad \rho^{mm} = 1/\rho^{id} \tag{8}$$

To prevent abrupt fluctuations, we employ exponential moving average (EMA) smoothing with momentum λ when updating these ratios across iterations.

$$\rho_t = \lambda \rho_t + (1 - \lambda)\rho_{t-1} \tag{9}$$

In order to mitigate optimization imbalance, we need to provide adequate optimization opportunities to the non-dominant modality while suppressing the parameter updates of the dominant modality. Therefore, inspired by softmax normalization [56], we define the update ratio γ^{id} and γ^{mm} as follows:

$$\gamma^{id} = \frac{\exp(\rho^{mm})}{\exp(\rho^{id}) + \exp(\rho^{mm})}, \quad \gamma^{mm} = 1 - \gamma^{id} \tag{10}$$

A higher value of γ indicates fewer parameters are frozen and more parameters are updated in the corresponding branch.

Adaptive Gradient Masking To implement modality-specific gradient updates, we utilize the Fisher Information Matrix (FIM) [57] , donated as $\mathbf{F}(\boldsymbol{\theta})$, which allows us to effectively measure the relative importance of model parameters across modalities. Specifically, the Fisher Information Matrix is defined as:

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \log p(\hat{y} \mid x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log p(\hat{y} \mid x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\top}\right] \tag{11}$$

Following [58], given a batch of data, we estimate the importance of parameters using the diagonal elements of $\mathbf{F}(\boldsymbol{\theta})$. Formally, the Fisher information for the j-th parameter is calculated as follows:

$$\mathbf{F}_j(\boldsymbol{\theta}) = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\frac{\partial \log p(\hat{y}_i \mid x_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right)^2 \tag{12}$$

Subsequently, we normalize these diagonal elements and denote π_j as the importance score of the j-th parameter. Aggregating the importance scores of all parameters, we obtain a probability distribution $\boldsymbol{\pi}$ over the parameter space:

$$\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_{|\boldsymbol{\theta}|}\}, \quad \text{where} \quad \pi_j = \frac{F_j(\boldsymbol{\theta})}{\sum_{j=1}^{|\boldsymbol{\theta}|} F_j(\boldsymbol{\theta})} \tag{13}$$

Given the update ratio γ of each modality and the parameter-wise probability distribution $\boldsymbol{\pi}$, we employ the non-uniform adaptive sampling [59] to generate the gradient mask $\mathbf{m}(t) \in \{0, 1\}^{|\boldsymbol{\theta}|}$, where 1 indicates the parameter will be updated during backpropagation and 0 means it remains frozen. This sampling method primarily directs our focus toward parameters carrying richer information. Concurrently, probabilistic sampling extends coverage to a broader range of parameters, promoting more thorough exploration and enhancing the model's generalization capability across the entire parameter space.

Finally, the parameter update rule with gradient masking thus becomes:

$$\boldsymbol{\theta}(t + 1) = \boldsymbol{\theta}(t) - \eta * \nabla \mathcal{L}(\boldsymbol{\theta}(t)) * \mathbf{m}(t) \tag{14}$$

3.4 Theoretical analysis and AGM*

In this section, we present our theoretical analysis of the Asymmetric Gradient Masking (AGM) approach and its improved variant AGM. We begin by establishing the convergence properties of the original AGM method, then introduce an importance weighting scheme that leads to better convergence guarantees in AGM. The convergence of AGM is complicated by the bias introduced through gradient masking. Theorem 1 formalizes this behavior:

Theorem 1 (Convergence of AGM)**.** *Suppose the loss function $\mathcal{L}(\cdot)$ is L-smooth and $\nabla l(\boldsymbol{\theta}(t))$ is unbiased, i.e. $\mathbb{E}(\nabla l(\boldsymbol{\theta}(t))) = \nabla \mathcal{L}(\boldsymbol{\theta}(t))$, which is commonly used in non-convex optimization. However, 0/1 mask makes $\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)$ biased, i.e., $\mathbb{E}[\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)] \neq \nabla \mathcal{L}(\boldsymbol{\theta}(t))$, since $\nabla l(\boldsymbol{\theta}(t))$ and $\mathbf{m}(t)$ are not independent. Under the Mask-Incurred Error assumption, we have the following convergence result for AGM over T steps:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2] \leq \mathcal{O}\left(\frac{1+(1+\nu)^2}{(1-\delta^2)(1+\nu)\sqrt{T}}\right), \tag{15}$$

where $\delta \in (0,1)$ and $\nu \geq 0$ are two constants.

The key limitation here is the bias in gradient estimates caused by the interaction between the mask $\mathbf{m}(t)$ and the stochastic gradients. This bias manifests in the $(1-\delta^2)$ term in the denominator, which slows down convergence. To address this issue, we propose AGM* which incorporates importance weighting through a modified mask $\hat{\mathbf{m}}(t)$. The weights are defined as:

$$\hat{\mathbf{m}}_j(t) = \begin{cases} \dfrac{1}{\pi_j + c}, & \text{if } \mathbf{m}_j(t) = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

where π_j represents the probability of the j-th parameter being unmasked and c is a small constant for numerical stability. This weighting scheme helps compensate for the bias introduced by the original masking operation. The update rule for AGM* becomes:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla \mathcal{L}(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t). \tag{17}$$

The importance weighting in AGM* leads to better theoretical guarantees, as shown in Theorem 2:

Theorem 2 (Convergence of AGM*)**.** *Under some assumptions for $\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)$, we have:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2] \leq \mathcal{O}\left(\frac{1+(1+\nu)^2}{(1+\nu)\sqrt{T}}\right), \tag{18}$$

Comparing Theorems 1 and 2, we see that AGM* removes the problematic $(1-\delta^2)$ term from the denominator, leading to faster convergence. This improvement comes from the fact that the importance weights in $\hat{\mathbf{m}}(t)$ help maintain the unbiasedness of the gradient estimates despite the masking operation. The complete proofs and additional technical details can be found in the Appendix.

4 Experiments

4.1 Setup

Dataset We conduct offline experiments on four open-source datasets from diverse recommendation domains. First, we choose the Microlens dataset [60], which features user-item interactions, video introductions, and video cover images. In addition, we adopt three categories from the Amazon dataset–Baby, Sports, and Electronics [61, 62]–which contain user-item interactions, product descriptions, and images. All raw datasets are preprocessed with a 5-core setting on both items and users, as described in [12, 10]. Detailed statistics of the datasets are provided in Appendix A.1.

Baselines and Evaluation In our experiments, we conduct two parts of evaluation. We first compare AGM with several recent advanced multimodal recommendation models, including VBPR [13], BM3 [10], FREEDOM [9], AlignRec [12], MGCN [8], LGMRec [26], GUME [25], and MM-Rec [63], to demonstrate its effectiveness. Next, to examine the generalization capability of AGM, we test our framework with diverse backbones, including MLP [64], DCN [65], and Fibinet [66]. To evaluate the performance of all models, we adopt two widely-used classification metrics: AUC (Area Under the ROC Curve) [67] and LogLoss (Logarithmic Loss) [68].

Table 1: Performance comparison of AGM and ID, MM, ID+MM models across different backbone architectures, measured by AUC.

Model	MLP				DCN				Fibinet			
	Baby	Elec.	Sports	Micro.	Baby	Elec.	Sports	Micro.	Baby	Elec.	Sports	Micro.
ID	0.6741	0.7215	0.7012	0.6883	0.6696	0.7192	0.6854	0.6825	0.6792	0.7196	0.7043	0.6901
MM	0.6237	0.6577	0.6295	0.6279	0.6202	0.6527	0.6251	0.6227	0.6154	0.6548	0.6223	0.6245
ID+MM	0.6719	0.7218	0.7072	0.6904	0.6685	0.7123	0.6827	0.6857	0.6641	0.7269	0.7115	0.6915
AGM*	**0.6864**	**0.7308**	**0.7145**	**0.6992**	**0.6827**	**0.7256**	**0.7062**	**0.6965**	**0.6832**	**0.7310**	**0.7129**	**0.6973**

Table 2: Comparison of AUC (higher is better) and LogLoss (lower is better) between AGM and other multimodal recommendation methods. Boldface indicates the best performance, and underlined values indicate the second-best.

Model	Baby		Electronics		Sports		MicroLens	
	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss
VBPR	0.6729	0.6739	0.7158	0.6032	0.6985	0.6533	0.6758	0.6054
FREEDOM	0.6802	0.6708	0.7221	0.5973	0.7023	0.6472	0.6772	0.5957
BM3	0.6715	0.6712	0.7119	0.6084	0.6932	0.6515	0.6705	0.6021
AlignRec	0.6832	0.6681	0.7274	0.5988	0.7101	0.6438	0.6869	0.5906
MGCN	0.6810	0.6695	0.7239	0.5994	0.7085	0.6417	0.6851	0.5881
LGMRec	0.6823	0.6687	0.7247	0.6012	0.7009	0.6480	0.6778	0.5935
GUME	0.6834	<u>0.6679</u>	0.7270	0.5991	0.7119	<u>0.6399</u>	0.6968	0.5876
MM-Rec	0.6643	0.6691	0.7136	0.5987	0.6703	0.6415	0.6735	0.5885
AGM	<u>0.6852</u>	0.6683	<u>0.7285</u>	<u>0.5974</u>	<u>0.7126</u>	0.6405	<u>0.6974</u>	<u>0.5856</u>
AGM*	**0.6864**	**0.6656**	**0.7310**	**0.5969**	**0.7145**	**0.6391**	**0.6992**	**0.5841**
Δ_{AGM^*-AGM}	+0.0012	-0.0027	+0.0025	-0.0005	+0.0019	-0.0014	+0.0018	-0.0015

4.2 Performance Comparison

Compared to different backbones Table 1 illustrates the AUC performance of four methods (AGM* and other three traditional model frameworks) evaluated on different backbone architectures. More specifically, to analyze the individual and combined effects of different training features and compare their performance to AGM*, we conducted experiments on three traditional model frameworks: (i) ID: A baseline model that utilizes only ID features (e.g., user ID, item ID). (ii) MM: A variant that relies solely on multimodal features (e.g., image, text). (iii) ID+MM: A straightforward combination of ID and multimodal features, without specialized fusion or alignment mechanisms.

The results reveal the following key insights: (i) AGM* consistently achieves the highest AUC across all backbone architectures, demonstrating that AGM* not only achieves superior performance but also maintains robustness and generalizability across various backbones. (ii) Models that rely solely on multimodal features (MM) consistently exhibit the lowest AUC scores across all settings. This suggests that multimodal signals alone are insufficient to capture user preferences, likely due to noise and sparse semantics in text or image modalities.

Compared to different baselines Table 2 presents the AUC and LogLoss results of our proposed AGM* and AGM framework in comparison with several state-of-the-art multimodal recommendation baselines across four benchmark datasets. From the experimental results, we derive the following observations: (i) AGM* consistently achieves the best performance on all datasets, although it is based on relatively simple neural network architectures. These gains can be attributed to AGM*'s ability to adaptively modulate feature contributions during training. (ii) Compared with other MLLM-based methods such as AlignRec [12] and GUME [25], AGM* achieves consistently better performance across all datasets. This superiority can be partially attributed to our fine-tuning strategy, which enhances the semantic alignment of multimodal features and ensures better adaptation of the pretrained MLLM to the recommendation domain. (iii) Experimental results prove the effectiveness of our proposed unbiased version AGM*, as it outperforms the biased AGM.

4.3 Ablation Study

In this part, we conduct ablation studies to evaluate the contribution of each core component in AGM and AGM*. Specifically, we compare the full models with the following variants: For AGM: (i)

Table 3: Ablation results (AUC) of different modules in AGM.

Dataset	ID+MM	Random	w/o BM	w/o FM	AGM	w/o BM*	w/o FM*	AGM*
Baby	0.6719	0.6792	0.6773	0.6815	**0.6852**	0.6789	0.6837	**0.6864**
Elec.	0.7218	0.7267	0.7252	0.7271	**0.7285**	0.7254	0.7281	**0.7308**
Sports	0.7072	0.7103	0.7095	0.7112	**0.7126**	0.7107	0.7121	**0.7145**
Micro.	0.6904	0.6949	0.6938	0.6954	**0.6974**	0.6946	0.6967	**0.6992**

Table 4: The performance of Online A/B Testing at the platform.

Main	watch-time	app usage	long view	short view
+MM	+0.022%	+0.008%	+0.132%	-0.160%
AGM*	+0.175%	+0.124%	+0.678%	-0.235%

Random: When generating the gradient mask $\mathbf{m}(t)$, this variant adopts purely random sampling that ignores parameter importance distributions, instead of the non-uniform adaptive sampling [59]. (ii) **w/o Backbone Masking (-BM)**: This variant disables gradient modulation on the backbone1,2 in Fig. 3. In other words, no additional gradient masking is applied to the backbone network layers; the gradients flow through these two layers in their original form. (iii) **w/o Fusion Masking (-FM)**: This variant omits gradient modulation on the fusion block in Fig. 3, so the gradients are propagated through this block in their original form without adaptive gradient masking. For AGM*, we apply the same ablations: (iv) **w/o Backbone Masking (-BM*)**: Gradient modulation on the backbone1,2 is removed in AGM*. (v) **w/o Fusion Masking (-FM*)**: Gradient modulation on the fusion block is removed in AGM*.

Table 3 presents the experimental results, demonstrating two key observations: (i) The removal of any module leads to a noticeable drop in AUC performance, from which we can conclude that all components make contributions to AGM and AGM*. (ii) Among all the ablation variants, removing the dynamic gradient masking on the backbone1,2 (-BM/-BM*) results in the most significant performance drop. This may be because without proper gradient regulation at this early level, imbalanced learning signals can lead to biased feature extraction from each modality. Consequently, these biases are carried forward and accumulated through the subsequent layers which substantially undermines the downstream fusion process, leading to a more pronounced overall performance loss compared to removing other modules.

4.4 Industrial Application

To further assess the real-world effectiveness of our model, we integrate AGM into the industrial recommendation system of a large-scale short video platform that serves hundreds of millions of users. The model is deployed in a 14-day online A/B test to evaluate its performance in a production environment.

We adopt widely-used industry metrics, such as app usage time and watch time, to measure performance. As shown in Table 4, our model achieves substantial improvements over the baseline, further confirming AGM's effectiveness. Notably, the model has now been fully deployed across the platform, actively serving hundreds of millions of users every day.

5 Conclusion

In this paper, we tackle the convergence inconsistency problem in joint training of ID-based and MLLM-based representations within large-scale recommendation systems. We propose a two-stage framework that first learns semantically aligned multimodal representations through MLLM fine-tuning and post-alignment with ID features, and then introduces a novel Adaptive Gradient Masking (AGM) strategy to balance optimization across modalities. Our theoretical analysis and extensive empirical results—across both offline benchmarks and real-world A/B testing—demonstrate that the proposed framework effectively mitigates the convergence gap, stabilizes training, and significantly boosts recommendation performance. These findings highlight the importance of

coordinated optimization in multimodal recommendation and pave the way for more robust integration of pretrained models into industrial systems.

Acknowledgement

This work was partially supported by the Yonghua Foundation.

References

[1] S. Rendle, "Factorization machines," in *2010 IEEE International conference on data mining*. IEEE, 2010, pp. 995–1000.

[2] R. Wang, R. Shivanna, D. Cheng, S. Jain, D. Lin, L. Hong, and E. Chi, "Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems," in *Proceedings of the web conference 2021*, 2021, pp. 1785–1797.

[3] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1059–1068.

[4] J. Chang, C. Zhang, Z. Fu, X. Zang, L. Guan, J. Lu, Y. Hui, D. Leng, Y. Niu, Y. Song *et al.*, "Twin: Two-stage interest network for lifelong user behavior modeling in ctr prediction at kuaishou," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3785–3794.

[5] X.-R. Sheng, F. Yang, L. Gong, B. Wang, Z. Chan, Y. Zhang, Y. Cheng, Y.-N. Zhu, T. Ge, H. Zhu *et al.*, "Enhancing taobao display advertising with multimodal representations: Challenges, approaches and insights," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 4858–4865.

[6] Q. Liu, J. Zhu, Y. Yang, Q. Dai, Z. Du, X.-M. Wu, Z. Zhao, R. Zhang, and Z. Dong, "Multimodal pretraining, adaptation, and generation for recommendation: A survey," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6566–6576.

[7] Y. Ye, Z. Zheng, Y. Shen, T. Wang, H. Zhang, P. Zhu, R. Yu, K. Zhang, and H. Xiong, "Harnessing multimodal large language models for multimodal sequential recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 13 069–13 077.

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

[9] X. Zhou and Z. Shen, "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 935–943.

[10] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap latent representations for multi-modal recommendation," in *Proceedings of the ACM web conference 2023*, 2023, pp. 845–854.

[11] K. Zhang, Y. Qin, J. Jin, Y. Liu, R. Su, W. Zhang, and Y. Yu, "Dream: A dual representation learning model for multimodal recommendation," *arXiv preprint arXiv:2404.11119*, 2024.

[12] Y. Liu, K. Zhang, X. Ren, Y. Huang, J. Jin, Y. Qin, R. Su, R. Xu, Y. Yu, and W. Zhang, "Alignrec: Aligning and training in multimodal recommendations," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 1503–1512.

[13] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[14] Q. Liu, S. Wu, and L. Wang, "Deepstyle: Learning user preferences for visual recommendation," in *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 2017, pp. 841–844.

[15] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," *arXiv preprint arXiv:1205.2618*, 2012.

[16] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 765–774.

[17] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, and M. Kankanhalli, "User diverse preference modeling by multimodal attentive metric learning," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1526–1534.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: a survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

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

[21] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3541–3549.

[22] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, "Dualgnn: Dual graph neural network for multimedia recommendation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1074–1084, 2021.

[23] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3872–3880.

[24] J. Yi and Z. Chen, "Multi-modal variational graph auto-encoder for recommendation systems," *IEEE Transactions on Multimedia*, vol. 24, pp. 1067–1079, 2021.

[25] G. Lin, M. Zhen, D. Wang, Q. Long, Y. Zhou, and M. Xiao, "Gume: Graphs and user modalities enhancement for long-tail multimodal recommendation," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 1400–1409.

[26] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "Lgmrec: local and global graph learning for multimodal recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8454–8462.

[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

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

[30] S. Geng, J. Tan, S. Liu, Z. Fu, and Y. Zhang, "Vip5: Towards multimodal foundation models for recommendation," *arXiv preprint arXiv:2305.14302*, 2023.

[31] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)," in *Proceedings of the 16th ACM conference on recommender systems*, 2022, pp. 299–315.

[32] H. Liu, Y. Wei, X. Song, W. Guan, Y.-F. Li, and L. Nie, "Mmgrec: Multimodal generative recommendation with transformer model," *arXiv preprint arXiv:2404.16555*, 2024.

[33] J. Fu, X. Ge, X. Xin, A. Karatzoglou, I. Arapakis, J. Wang, and J. M. Jose, "Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 687–697.

[34] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[35] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[36] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[37] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[38] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[39] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[40] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[41] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[42] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, G. Xu, C. Li, J. Tian, Q. Qian, J. Zhang *et al.*, "Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model," *arXiv preprint arXiv:2310.05126*, 2023.

[43] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2305.06500

[44] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023. [Online]. Available: https://arxiv.org/abs/2308.12966

[45] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[46] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llavanext: Improved reasoning, ocr, and world knowledge," 2024.

[47] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao *et al.*, "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[48] Z. Guo, R. Xu, Y. Yao, J. Cui, Z. Ni, C. Ge, T.-S. Chua, Z. Liu, and G. Huang, "Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images," in *European Conference on Computer Vision*. Springer, 2024, pp. 390–406.

[49] G. Luo, Y. Zhou, Y. Zhang, X. Zheng, X. Sun, and R. Ji, "Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models," *arXiv preprint arXiv:2403.03003*, 2024.

[50] Y. Liu, B. Yang, Q. Liu, Z. Li, Z. Ma, S. Zhang, and X. Bai, "Textmonkey: An ocr-free large multimodal model for understanding document," *arXiv preprint arXiv:2403.04473*, 2024.

[51] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[52] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[53] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[54] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[56] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, Architectures and Applications*. Springer, 1990, pp. 227–236.

[57] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368, 1922.

[58] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[59] S. Gopal, "Adaptive sampling for sgd by exploiting side information," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 364–372.

[60] Y. Ni, Y. Cheng, X. Liu, J. Fu, Y. Li, X. He, Y. Zhang, and F. Yuan, "A content-driven micro-video recommendation dataset at scale," *arXiv preprint arXiv:2309.15379*, 2023.

[61] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 43–52.

[62] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.

[63] C. Wu, F. Wu, T. Qi, C. Zhang, Y. Huang, and T. Xu, "Mm-rec: Visiolinguistic model empowered multimodal news recommendation," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 2560–2564.

[64] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[65] R. Wang, B. Fu, G. Fu, and M. Wang, "Deep & cross network for ad click predictions," in *Proceedings of the ADKDD'17*, 2017, pp. 1–7.

[66] T. Huang, Z. Zhang, and J. Zhang, "Fibinet: combining feature importance and bilinear feature interaction for click-through rate prediction," in *Proceedings of the 13th ACM conference on recommender systems*, 2019, pp. 169–177.

[67] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[68] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and scope of the paper are included in the abstract and Section 1. Please refer to the first and last paragraph of Section 1 for scope and contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of our work in the Appendix F

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide required assumptions and some proofs in Section 3 and the detailed proofs are shown in Appendix C, D, and E.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Model architectures can be seen in Section 3, and datasets as well as experimental hyperparameters can be seen in Section 4 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A public code repository is included in Appendix A.3

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Datesets and training details can be seen in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The statistical metrics we used are all officially defined and well-established. The experimental results are the average of five experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information on the computer resources can be seen in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive impacts that our model will bring in Appendix G.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release high risk models and the datasets used in the paper are open-sourced.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original paper for all the existing assets used in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We don not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: MLLM is utilized as part of our model for multimodal embeddings.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Supplemental Material: Adaptive Gradient Masking for Balancing ID and MLLM-based Representations in Recommendation

A Experimental details

A.1 Dataset

The detailed descriptions of the datasets used in the main text are as follows:

- **Amazon Baby**: This dataset consists of user-generated reviews on baby-related products sold on Amazon, such as bottles, diapers, and infant toys. It typically includes product names, review texts, and star ratings, making it a valuable resource for research in recommendation systems.
- **Amazon Sports**: This dataset includes Amazon user reviews pertaining to sports and outdoor products, including equipment, camping gear, and fitness devices. Alongside textual reviews and ratings, the dataset provides insights into consumer preferences and opinions, supporting various applications in recommendation modeling.
- **Amazon Electronics**: This dataset is derived from Amazon reviews on electronic products, covering items such as mobile phones, cameras, laptops, and home appliances. In addition to review content and ratings, it contains product metadata, facilitating extensive research in product recommendation.
- **Microlens**: This dataset contains user–item interaction records, video introductions, and video cover images. Each video within MicroLens contains multiple modalities, including text descriptions, images, audio, and raw video information. This extensive coverage enables robust benchmarking of both classical and state-of-the-art recommendation systems.

Table 5: Statistical Information of datasets.

Dataset	Baby	Sports	Electronics	Microlens
#User	19,445	35,598	192,403	25,411
#Item	7,050	18,357	63,001	20,276
#Interaction	160,792	296,337	1,689,188	223,263
Sparsity	99.88%	99.95%	99.99%	99.96%

A.2 Baselines

We summarize the key characteristics of the baseline methods used in our comparative evaluation:

- **VBPR** [13]: This model incorporates visual features into matrix factorization by treating them as auxiliary information for user preference, learning with BPR loss.
- **BM3** [10]: This model simplifies the self-supervised multimodal recommendation framework by using a latent representation dropout mechanism instead of graph augmentation to generate contrastive views, enhancing representation learning.
- **FREEDOM** [9]: This model refines ID-based item representations through auxiliary contrastive losses on multimodal data, and leverages graph structures to align different modalities in a unified embedding space.
- **AlignRec** [12]: This model addresses modality misalignment and optimization imbalance by proposing a two-stage training strategy that first pre-trains inter-content alignment, then jointly optimizes with the recommendation task.

- **MGCN** [8]: This model purifies modality features using item behavior data to reduce noise, and models user-modal preference through multi-view graph convolution networks.
- **LGMRec** [26]: This model enhances multimodal recommendation by learning both local and global semantic relations in item-user graphs, effectively capturing fine-grained user preferences.
- **GUME** [25]: This model improves long-tail multimodal recommendation by incorporating user-specific modality preferences and behavior graphs to enhance personalized modeling.
- **MM-Rec** [63]: This model enhances multimodal news recommendation by jointly encoding news text and image ROIs using a visiolinguistic model, and introduces a candidate-aware attention mechanism to identify relevant historical news.

A.3 Training details

The code and model are available at: **AGM**.

In the fine-tuning phase, we adopted Qwen2vl-2b[51] as the backbone model. The model was fine-tuned for 5 epochs across four distinct datasets, utilizing a batch size of 128 on 4 A100 GPUs.

For AGM, offline evaluations were conducted using TensorFlow 2.15.0 on a single RTX 4090 GPU, selecting Adam as the optimizer. Hyperparameters, including batch size and learning rate, were systematically tuned across candidate sets of $\{256, 512, 1024, 2048\}$ and $\{1e-3, 1e-4, 1e-5\}$, respectively. The best model was selected based on the minimum validation loss, and early stopping was applied with a patience of 5 to prevent over-fitting.

B In-depth analysis

B.1 Evaluation of multimodal representations

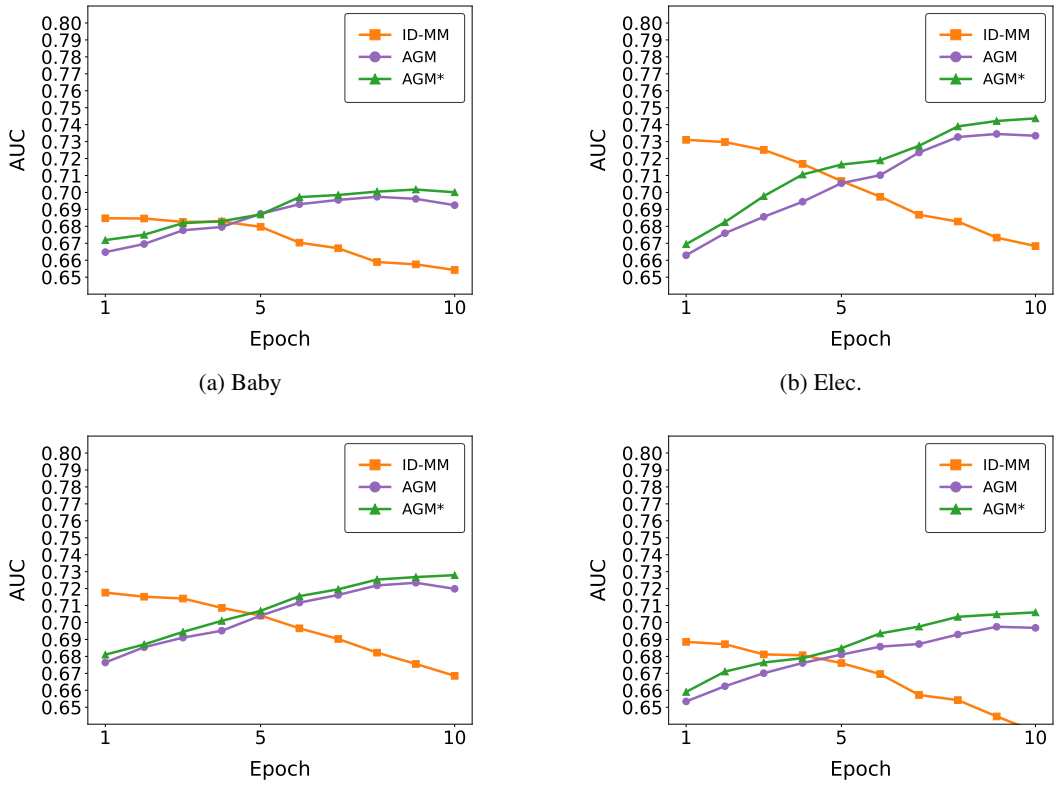
To assess the quality of the multimodal representations generated by AGM, we conduct zero-shot recommendation experiments on the Amazon dataset following the protocol of AlignRec [12]. For each user, the last interacted item is regarded as the target item, and the rest form the historical sequence. We average the multimodal features of historical items to construct the user representation, then compute its similarity with the candidate items' features to evaluate if the target item ranks within top-K. We compare the following methods: (i) Amazon, which uses separately trained visual (CNN) and textual (Transformer) encoders; (ii) MLLM, which directly uses the frozen MLLM outputs; (iii) w/o $\mathcal{L}_{align}$, a variant of AGM that disables the feature alignment loss; and (iv) Ours (AGM), which includes all proposed components. We report Recall@20 and Recall@50 in Table 6. The results reveal three key observations: (i) AGM significantly outperforms traditional feature extractors and raw MLLM features, indicating the benefit of task-specific representation refinement. (ii) Removing the alignment loss leads to noticeable performance drops, highlighting its importance in guiding effective feature selection and fusion. (iii) AGM achieves consistent improvements across all categories, demonstrating its robust capability in modeling multimodal user-item relationships in a zero-shot scenario.

Table 6: Evaluation of multimodal representations

Generation Methods	Baby		Elec.		Sports	
	R@20	R@50	R@20	R@50	R@20	R@50
Amazon	0.0052	0.0150	0.0093	0.0135	0.0040	0.0072
MLLM	0.0140	0.0345	0.0202	0.0364	0.0053	0.0089
w/o $\mathcal{L}_{align}$	0.0225	0.0426	0.0251	0.0417	0.0120	0.0159
Ours	**0.0276**	**0.0509**	**0.0293**	**0.0478**	**0.0175**	**0.0206**

B.2 Convergence analysis

To better illustrate the convergence performance of different methods, we plot the AUC values across training epochs for AGM, AGM*, and the combined model (ID-MM) on Amazon Baby, Electronics, Sports and Microlens. As shown in Fig.4, both AGM and AGM* demonstrate increasing



(a) Baby (b) Elec.

(c) Sports (d) Micro.

Figure 4: Validation-set AUC comparison of the combined model (ID-MM), AGM, and AGM* across four datasets

AUC throughout training epochs, indicating effective joint learning. In contrast, ID-MM exhibits a downward trend in AUC as training progresses. This degradation can be attributed to the ID features dominating the training process of ID-MM, especially in the absence of any modulation mechanism on gradients during the backpropagation phase. Consequently, as the model overfits the ID space over time, the performance of ID-MM degrades. This dominance suppresses the contribution of multimodal features, resulting in suboptimal representations and overall performance decline. In contrast, our AGM and AGM* introduce Gradient Modulation, which dynamically balances the gradient flow between the ID and multimodal branches, preventing ID features from overwhelming MM features, allowing both to contribute meaningfully to the final prediction.

In addition, AGM* shows better overall performance compared to AGM, especially in larger datasets (e.g., Amazon Electronics). Specifically, by compensating for the gradient masking bias introduced by the original masking operation in the AGM, AGM* improves the gradient update process, promoting faster convergence, and ultimately improving the final AUC of the model.

C Convergence inconsistency analysis of ID-MM combine model

In the joint learning of ID and multimodal representations for recommendation, we observe an optimization imbalance phenomenon, where one representation dominates the learning process, causing the other to be under-optimized. We introduce the analysis of the optimization imbalance phenomenon for the model with concatenation as fusion method. In our recommendation model, the logits output is formulated as:

$$\varphi(x_i) = W[f(\mathbf{e}^{id}; \boldsymbol{\theta}^{id}); g(\mathbf{e}^{mm}; \boldsymbol{\theta}^{mm})] + b \tag{19}$$

To observe the optimization process of each component individually, W can be represented as the combination of two matrix: W^{id} and W^{mm}. The Equation 19 can be rewritten as:

$$\varphi(x_i) = W^{id} * f(\mathbf{e}^{id}; \boldsymbol{\theta}^{id}) + W^{mm} * g(\mathbf{e}^{mm}; \boldsymbol{\theta}^{mm}) + b \tag{20}$$

The model is trained using binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \left(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\right), \tag{21}$$

where the predicted probability is $\hat{y} = \sigma(\varphi(x_i))$. The gradient of the loss with respect to the logit is:

$$\frac{\partial L}{\partial \varphi(x_i)} = \sigma(\varphi(x_i)) - y_i. \tag{22}$$

Applying the chain rule, the gradients for the ID weights and backbone parameters of the ID representation are as follows:

$$\frac{\partial L}{\partial W^{id}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial L}{\partial \varphi(x_i)} * f(\mathbf{e}^{id}; \boldsymbol{\theta}^{id}) \tag{23}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}^{id}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial L}{\partial \varphi(x_i)} W^{id} \frac{\partial f(\mathbf{e}^{id}; \boldsymbol{\theta}^{id})}{\partial \boldsymbol{\theta}^{id}} \tag{24}$$

According to Equations 23 and 24, the optimization of W^{id} and $f(\cdot)$ is nearly independent of the optimization of multimodal parameters W^{mm} and $g(\cdot)$, except for the term associated with the training loss. As a result, the backbone have limited ability to make adjustments based on feedback from one another. The analysis of the feed-forward and back-propagation stages reveals that both the model predictions and gradients are governed by the sum of ID and multimodal (MM) components. Since ID features converge faster and contain stronger discriminative information, they dominate the model prediction $\varphi(x_i)$ and gradient $\frac{\partial L}{\partial \varphi(x_i)}$ through $W^{id} \cdot f(\mathbf{e}^{id}; \boldsymbol{\theta}^{id})$. Even if the MM representations remain under-optimized and produce erroneous outputs during training, the more informative ID components can still "correct" these errors through summation, thereby influencing both the feed-forward and back-propagation processes. Consequently, according to Eq. (9) and Eq. (11), the MM, which has relatively lower confidence in the correct category, receives limited optimization, leading to its under-utilization. Based on this analysis, ID features play a dominant role in the optimization process. As the model approaches convergence, MM components may still require further training to compensate for their under-optimized features.

D Convergence analysis of AGM

In this section, we give a detailed proof of Theorem 1. Recall update step of stochastic gradient descent (SGD) for AGM is:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \cdot \nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t), \tag{25}$$

where $\nabla \ell(\boldsymbol{\theta}(t))$ is the stochastic gradient of $\nabla \mathcal{L}(\boldsymbol{\theta}(t))$ and $\eta > 0$ is the learning rate, and $\mathbf{m}(t)$ is a binary mask vector. To analyze the convergence of AGM, we have the three following common assumptions for $\mathcal{L}(\cdot)$.

Assumption 1 (Smoothness)**.** *The loss function $\mathcal{L}$ is L-smooth, which is common for non-convex optimization. That is, for any $\boldsymbol{\theta}, \boldsymbol{\theta}'$, we have:*

$$\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}') \leq \langle \nabla \mathcal{L}(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle + \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{26}$$

Assumption 2 (Bounded Variance)**.** *We assume that the stochastic gradient $\nabla \ell(\boldsymbol{\theta}) \odot \mathbf{m}(t)$ is biased and its variance is bounded. That is, for any $\boldsymbol{\theta}(t)$ and $\mathbf{m}(t)$, we have*

$$\mathbb{E}\left[\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)\right] = \nabla \mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t)), \tag{27}$$

and

$$\mathbb{E}\left[\|\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t) - \mathbb{E}\left[\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)\right]\|\right]^2 \leq \nu \left\|\nabla \mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\right\|^2 + \sigma^2, \tag{28}$$

where $\sigma^2 \geq 0$ and $\nu \geq 0$ are two constants.

Assumption 3 (Mask-Incurred Error)**.** *For any $\boldsymbol{\theta}(t)$ and $\mathbf{m}(t)$, we have*

$$\left\|\mathbb{E}\left[\nabla \ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t) - \nabla \ell(\boldsymbol{\theta}(t))\right]\right\| \leq \delta \left\|\mathbb{E}\left[\nabla \ell(\boldsymbol{\theta}(t))\right]\right\|, \tag{29}$$

where the constant $\delta \in [0, 1]$.

Proof of Theorem 1. *As Theorem 1 claims:*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 \leq \mathcal{O}\left(\frac{1+(1+\nu)^2}{(1-\delta^2)(1+\nu)\sqrt{T}}\right), \tag{30}$$

By Assumption 1, and let $h(t) := \nabla\ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)$, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}(t+1)) - \mathcal{L}(\boldsymbol{\theta}(t)) &\leq \langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \boldsymbol{\theta}(t+1)-\boldsymbol{\theta}(t)\rangle + \frac{L}{2}\|\boldsymbol{\theta}(t+1)-\boldsymbol{\theta}(t)\|^2 \\ &= -\eta\langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), h(t)\rangle + \frac{\eta^2 L}{2}\|h(t)\|^2, \quad \text{(using eq. 25)}\end{aligned} \tag{31}$$

Taking expectation over both sides of 31 and by using Assumption 2:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(t+1)) - \mathcal{L}(\boldsymbol{\theta}(t))] &\leq -\eta\langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\rangle + \frac{\eta^2 L}{2}\mathbb{E}\left[\|h(t)\|^2\right] \\ &= -\eta\langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\rangle + \frac{\eta^2 L}{2}\left(\mathbb{E}[\|h(t)-\mathbb{E}[h(t)]\|^2] + \mathbb{E}[\|\mathbb{E}[h(t)]\|^2]\right) \\ &\leq -\eta\langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\rangle + \frac{\eta^2 L}{2}\left((1+\nu)\|\nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\|^2 + \sigma^2\right) \\ &\leq -\eta\langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\rangle + \frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2},\end{aligned} \tag{32}$$

where the last inequality is due to $\eta \leq \frac{1}{(1+\nu)L}$.

Since $-\langle a, b\rangle + \frac{\|b\|^2}{2} = \frac{\|a-b\|^2}{2} - \frac{\|a\|^2}{2}$, then 32 implies that

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(t+1))] - \mathcal{L}(\boldsymbol{\theta}(t)) &\leq -\eta\langle\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \nabla\mathcal{L}(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\rangle + \frac{\eta^2 L}{2}\mathbb{E}\left[\|h(t)\|^2\right] \\ &\leq \frac{\eta}{2}\|b(\boldsymbol{\theta}(t))\|^2 - \frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2}.\end{aligned} \tag{33}$$

Next, by 27 in Assumption 2 and Assumption 3, we know

$$\|b(\boldsymbol{\theta}(t))\| = \|\mathbb{E}[\nabla\ell(\boldsymbol{\theta}(t)) \odot \mathbf{m}(t)] - \nabla\ell(\boldsymbol{\theta}(t))\| \leq \delta\|\mathbb{E}[\nabla\ell(\boldsymbol{\theta}(t))]\| = \delta\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|. \tag{34}$$

Therefore, by (33) and (34) we have

$$\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(t+1)) - \mathcal{L}(\boldsymbol{\theta}(t))] \leq -\frac{\eta(1-\delta^2)}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L\sigma^2}{2}. \tag{35}$$

which implies

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 \leq \frac{2\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}(\boldsymbol{\theta}(t+1))]}{\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2}. \tag{36}$$

By summing up for $t = 1, \ldots, T$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 &\leq \frac{2\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(1)) - \mathcal{L}(\boldsymbol{\theta}(T+1))]}{T\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2} \\ &\leq \frac{2\mathcal{L}(\boldsymbol{\theta}(1))}{T\eta(1-\delta^2)} + \frac{\eta L\sigma^2}{1-\delta^2}.\end{aligned} \tag{37}$$

By setting $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 \leq \frac{2(1+\nu)L\mathcal{L}(\boldsymbol{\theta}(1))}{\sqrt{T}(1-\delta^2)} + \frac{\sigma^2}{(1+\nu)\sqrt{T}(1-\delta^2)} = \mathcal{O}\left(\frac{1+(1+\nu)^2}{(1-\delta^2)(1+\nu)\sqrt{T}}\right). \tag{38}$$

E Convergence analysis of AGM*

For unbiased stochastic gradient, recall that the update step of SGD is

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t), \tag{39}$$

where $\nabla \ell(\boldsymbol{\theta}(t))$ is the stochastic version of the gradient of loss function $\nabla \mathcal{L}(\boldsymbol{\theta}(t))$ at $\boldsymbol{\theta}(t)$, and $\eta > 0$ is the learning rate.

The element of $\hat{\mathbf{m}}(t)$ is given by

$$\hat{\mathbf{m}}_j(t) = \begin{cases} \dfrac{1}{\pi_j(t)}, & \text{if } \mathbf{m}_j(t) = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

Suppose that stochastic gradient $\nabla \ell(\boldsymbol{\theta}(t))$ is unbiased, i.e., $\mathbb{E}[\nabla \ell(\boldsymbol{\theta}(t))] = \nabla \mathcal{L}(\boldsymbol{\theta}(t))$. Then we have

$$\begin{aligned} \mathbb{E}[\nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t)] &= \mathbb{E}[\nabla \ell(\boldsymbol{\theta}(t)) \odot \boldsymbol{\pi}(t)^{-1} \odot \mathbf{m}(t)] \\ &= \mathbb{E}_{\mathbf{m}(t)} \left[\nabla \ell(\boldsymbol{\theta}(t)) \odot \boldsymbol{\pi}(t)^{-1} \odot \mathbf{m}(t) \mid \nabla \ell(\boldsymbol{\theta}(t)) \right] \\ &= \mathbb{E}[\nabla \ell(\boldsymbol{\theta}(t)) \odot \boldsymbol{\pi}(t)^{-1} \odot \mathbb{E}_{\mathbf{m}(t)}[\mathbf{m}(t) \mid \nabla \ell(\boldsymbol{\theta}(t))]] \\ &= \mathbb{E}[\nabla \ell(\boldsymbol{\theta}(t))] \\ &= \nabla \mathcal{L}(\boldsymbol{\theta}(t)), \end{aligned} \tag{41}$$

indicating that $\nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t)$ is also unbiased.

Assumption 4 (Bounded Variance)**.** *We assume that the stochastic gradient $\nabla \ell(\boldsymbol{\theta}) \odot \hat{\mathbf{m}}(t)$ is unbiased and its variance is bounded. That is, for any $\boldsymbol{\theta}(t)$ and $\hat{\mathbf{m}}(t)$, we have*

$$\mathbb{E}\left[\nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t)\right] = \nabla \mathcal{L}(\boldsymbol{\theta}(t)), \tag{42}$$

and

$$\mathbb{E}\left[\|\nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t) - \mathbb{E}\left[\nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t)\right]\|\right]^2 \leq \nu \left\|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\right\|^2 + \sigma^2, \tag{43}$$

where $\sigma^2 \geq 0$ and $\nu \geq 0$ are two constants.

Proof of Theorem 2. As Theorem 2 claims: under assumptions 1, 4, we have:

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 \leq \mathcal{O}\left(\frac{1 + (1+\nu)^2}{(1+\nu)\sqrt{T}} \right), \tag{44}$$

where the learning rate is set as $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$.

By Assumption 1, and we set $\hat{h}(t) := \nabla \ell(\boldsymbol{\theta}(t)) \odot \hat{\mathbf{m}}(t)$, then we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}(t+1)) - \mathcal{L}(\boldsymbol{\theta}(t)) &\leq \langle \nabla \mathcal{L}(\boldsymbol{\theta}(t)), \boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t) \rangle + \frac{L}{2} \|\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t)\|^2 \\ &= -\eta \left\langle \nabla \mathcal{L}(\boldsymbol{\theta}(t)), \hat{h}(t) \right\rangle + \frac{\eta^2 L}{2} \left\| \hat{h}(t) \right\|^2, \end{aligned} \tag{45}$$

Taking expectation over both sides of (45) and by using Assumption 4, we have

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(t+1))] - \mathcal{L}(\boldsymbol{\theta}(t)) &\leq -\eta \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L}{2} \mathbb{E}[\|\hat{h}(t)\|^2] \\ &= -\eta \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L}{2} \left(\mathbb{E}\left[\|\hat{h}(t) - \nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 \right] + \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 \right) \\ &\leq -\eta \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L}{2} \left((1+\nu) \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \sigma^2 \right) \\ &\leq -\eta \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta}{2} \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 + \frac{\eta^2 L \sigma^2}{2}, \end{aligned} \tag{46}$$

where the last inequality is due to $\eta \leq \frac{1}{(1+\nu)L}$. Therefore, we have

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 \leq \frac{2\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}(\boldsymbol{\theta}(t+1))]}{\eta} + \eta L \sigma^2. \tag{47}$$

By summing up for $t = 1, \ldots, T$, we have

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 &\leq \frac{2\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}(1)) - \mathcal{L}(\boldsymbol{\theta}(T+1))]}{T\eta} + \eta L \sigma^2 \\ &\leq \frac{2\mathcal{L}(\boldsymbol{\theta}(1))}{T\eta} + \eta L \sigma^2. \end{aligned} \tag{48}$$

Since $\eta = \frac{1}{(1+\nu)L\sqrt{T}}$, we get

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|^2 \leq \frac{2(1+\nu)L\mathcal{L}(\boldsymbol{\theta}(1))}{\sqrt{T}} + \frac{\sigma^2}{(1+\nu)\sqrt{T}}. \tag{49}$$

F Limitation

While our proposed AGM framework demonstrates promising results, several limitations remain. First, the framework's performance may depend on the quality and diversity of the multimodal data available during training. Second, the current experiments focus on recommendation tasks, and extending the approach to other multimodal applications may require further adaptation.

G Broader Impacts

Our work on Adaptive Gradient Masking for recommendation systems presents several important societal implications. The improved ability to handle multimodal content could significantly enhance recommendation quality, particularly for niche and cold-start items, potentially benefiting both users through more relevant suggestions and content creators through better exposure. The framework's ability to balance different feature types may also lead to more diverse recommendations, mitigating some common filter bubble effects.