arXiv:2603.17450v2 [cs.IR] 12 Aug 2026

VLM2Rec: Resolving Modality Collapse in Vision-Language Model Embedders for Multimodal Sequential Recommendation

Junyoung Kim[1], **Woojoo Kim**[1], **Wonbin Kweon**[2],
Jaehyung Lim[1], **Dongha Kim**[1], **Hwanjo Yu**[1,*],

[1]Pohang University of Science and Technology
[2]University of Illinois Urbana-Champaign

Abstract

Sequential Recommendation (SR) in multimodal settings typically relies on small frozen pretrained encoders, which limits semantic capacity and prevents Collaborative Filtering (CF) signals from being fully integrated into item representations. Inspired by the recent success of Large Language Models (LLMs) as high-capacity embedders, we investigate the use of Vision-Language Models (VLMs) as CF-aware multimodal embedders for SR. However, we find that standard contrastive Supervised Fine-Tuning (SFT), used to adapt VLMs for embedding generation and inject CF signals, can amplify inherent modality imbalance: optimization becomes dominated by one modality while the other degrades, ultimately undermining recommendation accuracy. To address this, we propose **VLM2Rec**, a VLM embedder-based framework for multimodal sequential recommendation designed to promote balanced modality utilization. Specifically, we introduce *Weak-modality Penalized Contrastive Learning* to mitigate gradient imbalance during optimization and *Cross-modal Relational Topology Regularization* to preserve geometric consistency between modalities. Experiments show that VLM2Rec consistently improves over strong baselines in both accuracy and robustness across diverse scenarios.

1 Introduction

Sequential Recommendation (SR) models dynamic user preferences from interaction histories [18, 13, 47, 68], yet conventional ID-based methods remain vulnerable to data sparsity and cold-start issues. Recent multimodal approaches alleviate this by incorporating text/image information, typically using small pretrained encoders [8, 41] whose outputs are kept frozen [62, 65, 16, 51] to maintain semantic integrity. While effective, this design yields static item embeddings that cannot fully absorb Collaborative Filtering (CF) signals, leaving CF injection entirely to the SR backbone and motivating the use of large modality encoders to produce CF-aware item representations with modality semantics.

In NLP, LLMs have been firmly established as powerful encoders for text context reasoning [26, 52, 23, 24, 48]. By transferring this capability to recommendation, recent studies [31, 56, 12] have begun to model item transition patterns and inject sequence-level CF signals into the embedding space with rich textual semantics. This motivates extending the paradigm to multimodal SR [65, 4] with Vision-Language Models (VLMs), and recent VLM-based SR studies [63, 38] have successfully integrated visual and textual signals, leading to improved recommendation performance. However, they remain confined to item-level modeling, missing crucial multimodal sequential transitions essential to SR modeling. To inherently encode these dynamic transitions into item representations,

*Corresponding author.

Preprint.

we leverage modern open-source VLMs [3, 30, 70], which offer not only strong *textual reasoning* but also *multi-image reasoning* across an interaction sequence. However, we find that naively porting the LLM pipeline to this setting introduces a critical **modality collapse**.

Prior studies [45, 22, 43, 44, 66, 9] indicate that VLMs frequently suffer from *modality collapse*, including in recommendation [61, 69], disproportionately relying on a dominant modality, typically text, while underutilizing the weaker one. This arises from their language-centric backbones, which align visual features into a linguistic space and yield embeddings that systematically underrepresent visual semantics. After defining the VLM embedder-based SR setup in Sec. 2, we first revisit simple prompting-based modality fusion strategies (Appendix B). Commonly used interleaved prompting [61, 45, 69, 63] biases self-attention toward the dominant modality, while separate-modality prompting [60] avoids cross-modal interference but still under-represents the weak-modality input. Both fail to mitigate this imbalance, suggesting that modality collapse is inherent to the pretrained VLM itself rather than the prompting strategy, motivating objective-level interventions.

Moreover, we identify the **Paradox of SFT**: the conventional contrastive Supervised Fine-Tuning (SFT) [36, 32], essential for adapting VLMs as embedders and injecting CF signals, counterintuitively exacerbates modality collapse. As detailed in Sec. 3, the model engages in *shortcut learning* [2, 35], relying on the strong modality while starving the weak one of gradients, degrading its ability to push weak-modality negatives away from positives in the representation space. Ultimately, this optimization imbalance paradoxically widens the modality gap and harms recommendation performance.

Building on these insights, we introduce **VLM2Rec**, a VLM embedder-based framework for multimodal SR that encodes the full interaction history at the sequence level, enabling high-capacity VLMs to capture dynamic behavioral patterns in the multimodal embedding space. To address the SFT paradox, we introduce two objective-level interventions in Sec. 4. *Weak-modality Penalized Contrastive Learning* ($\mathcal{L}_{\text{WPCL}}$) identifies the user-adaptive weak modality and improves its ability to distinguish positives from negatives, while *Cross-modal Relational Topology Regularization* ($\mathcal{L}_{\text{CRTR}}$) further aligns the relative sequence–item topology across modalities to encourage cross-modal geometric consistency. These objectives enable balanced utilization of textual and visual dynamics, yielding a VLM embedder that absorbs CF signals into discriminative multimodal representations. Across diverse benchmarks, our experiments in Sec. 5 show that VLM2Rec mitigates modality collapse and consistently improves recommendation accuracy and robustness across various real-world scenarios.

Our contributions are as follows:

- We empirically identify the paradox of SFT: standard contrastive fine-tuning can amplify modality collapse in VLMs by under-optimizing the weaker modality on SR datasets.
- We introduce VLM2Rec, a VLM embedder-based framework with two objective-level interventions that improve weak-modality discrimination while preserving cross-modal topology.
- We evaluate VLM2Rec, showing consistent improvements across datasets, backbones, model families, and various real-world scenarios.

2 Preliminaries

Background. Sequential recommendation predicts the next item i^+ from a user's interaction history S_u by learning sequence representations [18, 13, 47, 68, 58, 39]. Recent NLP studies repurpose large language models as general-purpose embedders [26, 52, 23, 24, 48], and recommender systems adopt this idea for semantic item or sequence representations [56, 31, 12]. Multimodal SR further incorporates visual signals through fusion-based sequence models [7, 16], modality-aware architectures [62, 51, 4], pretrained content encoders [62], or VLM-based representations [38, 63]. Yet, directly constructing a multimodal sequence–item representation space remains underexplored. Using VLMs as sequence embedders is a natural direction, but vanilla VLMs are often already modality-imbalanced, typically text-dominant, due to biased multimodal optimization or dominant-modality embedding geometry [54, 17, 45, 28]. We therefore study VLM-based SR, where sequence-level CF supervision must be injected without amplifying this modality imbalance. Appendix A provides a detailed discussion of these related lines of work.

Problem Formulation. Given user sequence S_u, we encode each modality with a VLM $\Phi(\cdot)$ using modality-specific prompts P_T and P_V (Appendix B), and extract the last-token hidden state of the

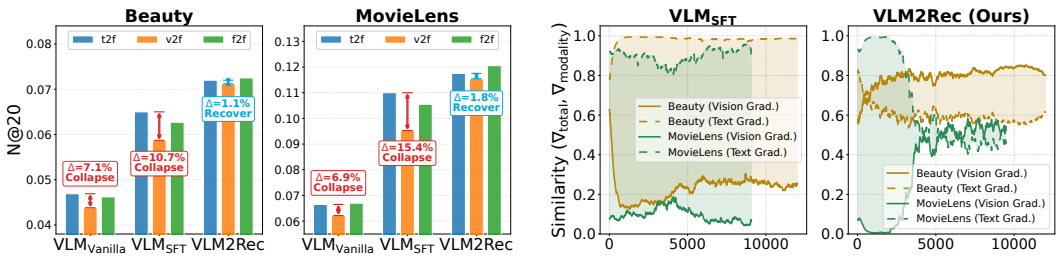

Figure 1: Analysis of modality collapse via **dropout test (Left, a)** and **gradient dynamics (Right, b)**. SFT amplifies the performance gap between text-only and vision-only inputs and steers the fused update toward the text gradient, while VLM2Rec reduces this gap by rebalancing modality gradients.

Table 1: VLM representation geometry (Uniformity U, Separability S) across three states.

Metric	Modality	Beauty			MovieLens		
		VLM$_{Vanilla}$	VLM$_{SFT}$	VLM2Rec (Ours)	VLM$_{Vanilla}$	VLM$_{SFT}$	VLM2Rec (Ours)
U (↓)	Fused	-0.3984	-1.7969	-1.9921	-0.2275	-0.8203	-4.6562
	Vision	**-0.0237**	-0.0398	**-0.2031**	**-0.0236**	**-0.0078**	**-1.7187**
	Text	-0.3691	-1.7422	-1.6718	-0.2177	-0.7891	-1.1788
S (↑)	Fused	1.0340	1.1910	1.1870	1.0117	1.2471	1.2263
	Vision	**1.0071**	**1.0000**	**1.2134**	**1.0035**	**0.9998**	**1.1134**
	Text	1.0381	1.1887	1.1639	1.0118	1.0116	1.1135

final transformer layer as the sequence representation: $\mathbf{z}_u^T = \Phi(P_T(S_u^T))$, $\mathbf{z}_u^V = \Phi(P_V(S_u^V))$, each ℓ_2-normalized. Item embeddings $\mathbf{e}_i^T, \mathbf{e}_i^V$ are obtained in the same way. The final fused representations are formed by simple parameter-free fusion: $\mathbf{z}_u = \mathbf{z}_u^T + \mathbf{z}_u^V$ and $\mathbf{e}_i = \mathbf{e}_i^T + \mathbf{e}_i^V$.

Standard Supervised Fine-Tuning Objective. To adapt generative VLMs for embedders and inject sequence-level CF signals, we employ conventional Supervised Fine-Tuning (SFT) [36, 32]. This objective aligns the representation space by maximizing the similarity between the user sequence and the positive item while pushing negative items away:

$$\mathcal{L}_{\text{SFT}} = -\sum_{u \in \mathcal{B}} \log \frac{e^{\text{sim}(\mathbf{z}_u, \mathbf{e}_{i^+})/\tau}}{e^{\text{sim}(\mathbf{z}_u, \mathbf{e}_{i^+})/\tau} + \sum_{i^- \in \mathcal{N}_u} e^{\text{sim}(\mathbf{z}_u, \mathbf{e}_{i^-})/\tau}}, \tag{1}$$

where $\mathcal{B}$ is the batch, $\text{sim}(\cdot,\cdot)$ is cosine similarity, τ is the temperature, and $\mathcal{N}_u$ is the negative set.

3 The Paradox of SFT: Analysis of Modality Collapse in VLM-based SR

We present evidence of the **Paradox of SFT**: when adapting VLMs as SR embedders, the standard SFT objective can worsen modality imbalance. On Amazon Beauty and MovieLens, which cover two different platforms, we first observe this effect in recommendation performance and then analyze its cause through optimization dynamics and representation geometry (Fig. 1, Tab. 1).

Performance Gap and Negative Transfer. We first examine how standard SFT (Eq. 1) affects modality-specific recommendation performance. In Fig. 1a, *f2f*, *t2f*, and *v2f* are evaluation settings that use fused ($\mathbf{e}_i$), text-only ($\mathbf{e}_i^T$), and vision-only ($\mathbf{e}_i^V$) item embeddings extracted from each model as input to the SR backbone (see Fig. 2, lower-left panel), while the target is the fused item embedding $\mathbf{e}_i$. The Vanilla model already shows a large gap of *t2f* over *v2f*, reflecting the inherent text-dominant prior of VLMs; yet the weaker visual modality is not uniformly harmful, as it can still provide synergy on some domains (*f2f*<*t2f* on Beauty but *f2f*>*t2f* on MovieLens). Critically, SFT not only fails to close this gap but further amplifies it, widening from Δ=7.1% to 10.7% on Beauty and from 6.9% to 15.4% on MovieLens. Moreover, *f2f* remains below *t2f* after SFT, indicating that visual information induces negative transfer rather than improving the fused representation.

Optimization Dynamics. To identify the cause of this degradation, we compute gradients of the SFT loss (Eq. 1) with respect to fused, text-only, and vision-only inputs, denoted by $\mathbf{g}_{\text{total}}$, $\mathbf{g}_T$, and $\mathbf{g}_V$, respectively, and then measure the cosine similarities $\cos(\mathbf{g}_{\text{total}}, \mathbf{g}_T)$ and $\cos(\mathbf{g}_{\text{total}}, \mathbf{g}_V)$ (Fig. 1b). This alignment quantifies each modality's contribution to the actual update direction. From the start,

$\mathbf{g}_{\text{total}}$ is strongly aligned with the text gradient $\mathbf{g}_T$, while alignment with the vision gradient $\mathbf{g}_V$ drops rapidly. As training proceeds, $\cos(\mathbf{g}_{\text{total}}, \mathbf{g}_T)$ rises toward 1, whereas $\cos(\mathbf{g}_{\text{total}}, \mathbf{g}_V)$ initially increases but then declines as modality bias accumulates. This suggests that the model tends to minimize the contrastive objective by relying on the easier text modality, while providing insufficient optimization for the visual modality to receive CF signals and distinguish positives from negatives.

Geometric Analysis: Representation Collapse. To characterize how optimization bias translates into embedding geometry, we measure uniformity U^m and alignment A^m, widely used in representation learning [53, 39], for each modality $m \in \{F, T, V\}$ (Tab. 1):

$$\begin{aligned} U^m &\triangleq \log \mathbb{E}_{(u,i)\sim\mathcal{P}_{\text{uniform}}}\left[e^{-2\|\mathbf{z}_u^m - \mathbf{e}_i^m\|_2^2}\right], \\ A^m_{\text{pos}} &\triangleq \mathbb{E}_{(u,i^+)\sim\mathcal{D}}\left[\|\mathbf{z}_u^m - \mathbf{e}_{i^+}^m\|_2^2\right], \\ A^m_{\text{neg}} &\triangleq \mathbb{E}_{(u,i^+)\sim\mathcal{D}}\mathbb{E}_{i^-\sim\mathcal{P}_n(\cdot|u)}\left[\|\mathbf{z}_u^m - \mathbf{e}_{i^-}^m\|_2^2\right], \end{aligned} \tag{2}$$

where $\mathcal{D}$ is the evaluation set, $\mathcal{P}_n(\cdot \mid u)$ is the negative sampler, and $\mathcal{P}_{\text{uniform}}$ uniformly samples sequence–item pairs. Since ranking tasks depend on relative distance between positives and negatives rather than absolute alignment, we derive separability S^m from the alignment terms:

$$S^m \triangleq A^m_{\text{neg}}/(A^m_{\text{pos}} + \epsilon), \tag{3}$$

where ϵ is a small constant. $S^m > 1$ indicates that negatives are properly farther than positives, else the representation space has lost its ability to distinguish positives from negatives. As shown in Tab. 1, the text and fused spaces show similar U^m and S^m values, indicating that the fused representation is dominated by text. This is because contrastive SFT gradients flow predominantly through the easier text modality, improving its uniformity and separability, which in turn improves the fused space. However, the visual modality receives insufficient contrastive signal, leaving S^V at approximately 1 or below and U^V degraded, confirming that modality collapse manifests in the VLM-generated representation space.

Conclusion Overall, our results suggest that the recommendation drop arises from SFT biasing VLM updates toward the easier text modality. This leaves the weaker visual modality insufficiently discriminative, so visual features cause negative transfer when fused. The same pattern appears on Toys and the visually informative Clothing domains (Appendix E.1) and across diverse VLM families and parameter scales (Appendix E.2).

4 VLM2Rec

As shown in Fig. 2, **VLM2Rec** is designed to mitigate the modality collapse identified in Sec. 3 at the objective level. It follows the VLM embedder-based formulation in Sec. 2 and deliberately avoids sophisticated fusion modules or architecture-specific modifications. This minimalist design isolates multimodal representation learning while improving compatibility with off-the-shelf VLMs.

4.1 Weak-modality Penalized Contrastive Learning

In Sec. 3, we show that the standard contrastive objective (Eq. 1) can induce an optimization imbalance in VLM-based embedders, encouraging shortcut learning on the strong modality and leaving the weak modality with poor positive-negative item discrimination. To overcome this limitation, we propose *Weak-modality Penalized Contrastive Learning* ($\mathcal{L}_{\text{WPCL}}$).

User-adaptive Modality Gap Estimation First, to quantify how clearly each modality discriminates the ground-truth item from negatives, we define the *Discriminative Margin* $\mathcal{M}$. Recognizing that users may differ in how much they rely on textual cues or visual cues in their purchasing behavior, we calculate this margin on a per-user basis.

For a user u and modality $m \in \{T, V\}$, the margin $\mathcal{M}_{u,m}$ measures the confidence with which the model distinguishes the positive target item embedding $\mathbf{e}^m_{i^+}$ from a set of negative item embeddings $\mathbf{e}^m_{i^-}$ ($i^- \in \mathcal{N}_u$):

$$\mathcal{M}_{u,m} = \text{sim}(\mathbf{z}^m_u, \mathbf{e}^m_{i^+}) - \frac{1}{|\mathcal{N}_u|}\sum_{i^-\in\mathcal{N}_u} \text{sim}(\mathbf{z}^m_u, \mathbf{e}^m_{i^-}), \tag{4}$$

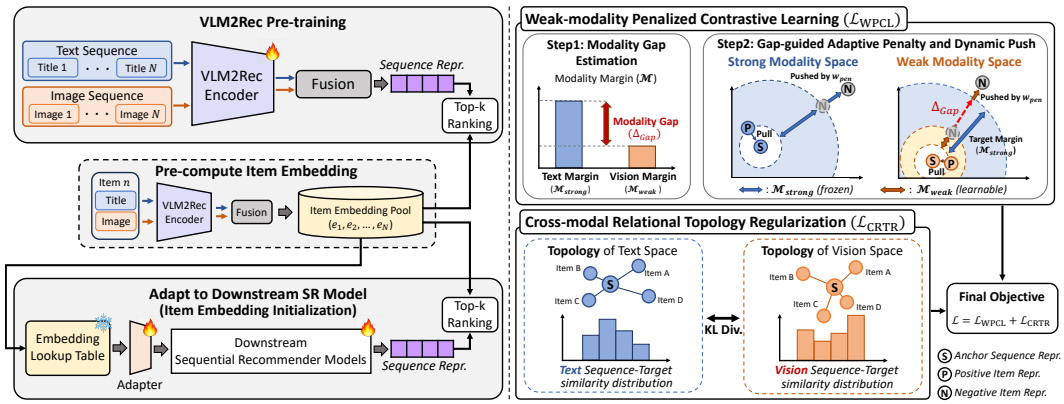

Figure 2: Overview of VLM2Rec. **Left**: VLM2Rec pretraining and the initialization of SR backbone item embeddings. **Right**: Proposed training objectives. $\mathcal{L}_{\text{WPCL}}$ adaptively penalizes the weak modality to improve positive-negative discrimination, while $\mathcal{L}_{\text{CRTR}}$ aligns cross-modal relational topology to promote modality-consistent neighborhoods.

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity and $\mathcal{N}_u$ denotes the negative samples per user. A larger $\mathcal{M}_{u,m}$ implies stronger discriminative capability for that modality. Based on these margins, we dynamically identify the user-adaptive strong and weak modalities at each training step:

$$\mathcal{M}_{u,\text{strong}} = \max_m(\mathcal{M}_{u,m}), \quad \mathcal{M}_{u,\text{weak}} = \min_m(\mathcal{M}_{u,m}). \tag{5}$$

This user-adaptive assignment avoids assuming a globally weak modality, allowing the correction to target either text or vision according to each user's current discriminative margins. We then define the *Modality Gap* $\Delta_{u,\text{gap}}$ as the disparity between these margins:

$$\Delta_{u,\text{gap}} = \text{sg}[\mathcal{M}_{u,\text{strong}}] - \mathcal{M}_{u,\text{weak}}, \tag{6}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator. By applying $\text{sg}[\cdot]$ to $\mathcal{M}_{u,\text{strong}}$, it fixes the discriminative level of the strong modality as a *target lower bound*, encouraging the model to minimize $\Delta_{u,\text{gap}}$ by enhancing $\mathcal{M}_{u,\text{weak}}$ rather than degrading $\mathcal{M}_{u,\text{strong}}$.

Dynamic Penalty from Modality Gap To explicitly reduce this gap, we convert $\Delta_{u,\text{gap}}$ into a penalty weight $w_{u,\text{pen}}$ that grows monotonically with the modality gap:

$$w_{u,\text{pen}} = 1 + \beta \cdot \text{Softplus}(\alpha \cdot \Delta_{u,\text{gap}}), \tag{7}$$

where $\alpha, \beta > 0$ are learnable scalars. Since Softplus is positive, $w_{u,\text{pen}} > 1$ and increases monotonically with $\Delta_{u,\text{gap}}$. We then formulate $\mathcal{L}_{\text{WPCL}}$ by integrating $w_{u,\text{pen}}$ into the contrastive objective (Eq. 1):

$$\mathcal{L}_{\text{WPCL}} = -\sum_{u\in\mathcal{B}} \log \frac{e^{\text{sim}(\mathbf{z}_u,\mathbf{e}_{i^+})/\tau_{\text{WPCL}}}}{e^{\text{sim}(\mathbf{z}_u,\mathbf{e}_{i^+})/\tau_{\text{WPCL}}} + w_{u,\text{pen}} \sum_{i^-\in\mathcal{N}_u} e^{\text{sim}(\mathbf{z}_u,\mathbf{e}_{i^-})/\tau_{\text{WPCL}}}}, \tag{8}$$

where τ_{WPCL} is the temperature. Because $w_{u,\text{pen}}$ grows with $\Delta_{u,\text{gap}}$, users with larger modality gaps receive a stronger negative-sample pushing penalty. Appendix C.1 shows that $w_{u,\text{pen}}$ imposes an adaptive margin $\tau_{\text{WPCL}} \log w_{u,\text{pen}}$ on the contrastive decision boundary, and Appendix C.2 decomposes the full $\mathcal{L}_{\text{WPCL}}$ gradient into a standard contrastive term and a $w_{u,\text{pen}}$-mediated term proportional to $\nabla_\theta \mathcal{M}_{u,\text{weak}}$, injecting additional gradient signal for weak-modality discrimination.

4.2 Cross-modal Relational Topology Regularization

While $\mathcal{L}_{\text{WPCL}}$ improves weak-modality discrimination, multimodal recommendation also requires text/vision spaces to preserve compatible sequence–item relations. We therefore introduce *Cross-modal Relational Topology Regularization* ($\mathcal{L}_{\text{CRTR}}$). Rather than enforcing rigid point-wise matching that erases modality-specific characteristics, $\mathcal{L}_{\text{CRTR}}$ aligns the relative similarity structures across modalities, promoting geometric consistency while preserving distinct multimodal semantics (Fig. 2).

Formally, we define a candidate item set $\mathcal{C}^m$ for each modality. We then compute a probability distribution $\mathbf{P}_u^m \in \mathbb{R}^{|\mathcal{C}^m|}$, where each entry $\mathbf{P}_{u,j}^m$ is the softmax-normalized similarity between the sequence representation $\mathbf{z}_u^m$ of user u and the candidate item embedding $\mathbf{e}_j^m$ of item j over $\mathcal{C}^m$:

$$\mathbf{P}_{u,j}^m = \frac{e^{\text{sim}(\mathbf{z}_u^m, \mathbf{e}_j^m)/\tau_{\text{CRTR}}}}{\sum_{k=1}^{|\mathcal{C}^m|} e^{\text{sim}(\mathbf{z}_u^m, \mathbf{e}_k^m)/\tau_{\text{CRTR}}}}, \tag{9}$$

where τ_{CRTR} is a temperature parameter. $\mathbf{P}_u^m$ captures how user u ranks candidate items in the modality-specific representation space. We then align these relational topologies across modalities by minimizing the bidirectional Kullback-Leibler (KL) divergence between $\mathbf{P}_u^T$ and $\mathbf{P}_u^V$ of each user u:

$$\mathcal{L}_{\text{CRTR}} = \frac{1}{2|\mathcal{B}|} \sum_{u \in \mathcal{B}} \left(\text{KL}(\mathbf{P}_u^T \| \mathbf{P}_u^V) + \text{KL}(\mathbf{P}_u^V \| \mathbf{P}_u^T) \right). \tag{10}$$

The bidirectional form symmetrically co-regularizes text and vision to share a consistent neighborhood structure, rather than relying on the dominant modality as a fixed alignment target. Appendix C.3 shows that $\mathcal{L}_{\text{CRTR}}$ preserves modality-specific embedding geometry, unlike point-wise alignment, and bounds cross-modal neighborhood discrepancy.

4.3 Final Objective

Finally, we combine the proposed loss functions to optimize the model jointly. The total objective function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{WPCL}} + \lambda \cdot \mathcal{L}_{\text{CRTR}}, \tag{11}$$

where λ is a hyperparameter that balances discrimination and structural consistency. Together, $\mathcal{L}_{\text{WPCL}}$ improves weak-modality discrimination and $\mathcal{L}_{\text{CRTR}}$ provides structural cross-modal alignment, yielding a balanced multimodal representation space.

Once trained, VLM2Rec can be utilized in two ways (Fig. 2, left): providing item embeddings as a frozen initialization for the item table of a downstream SR backbone (Sec. 5.2), or directly extracting a sequence representation and ranking items by similarity in the VLM embedding space (Sec. 5.5).

5 Experiments

5.1 Experimental Setup

Datasets and Baselines We evaluate VLM2Rec on four Amazon domains[2] [34] (Toys, Beauty, Clothing, Sports) and MovieLens[3] [11] to test generalization across domains and platforms, applying 5-core filtering (keeping users/items with $\geq$5 interactions) and the leave-one-out protocol (the last/second-to-last items for test/validation). We adopt Qwen2.5-VL-3B [3] as the VLM backbone and Qwen2.5-3B [40] as the LLM backbone, used across all baselines, both fine-tuned with LoRA [15]. As our goal is not to propose a novel negative sampling strategy, both $\mathcal{N}_u$ and $\mathcal{C}^m$ are constructed from in-batch users/target items for simplicity and efficiency; see Appendix E.4 for experiments with hard negative sampling strategies for $\mathcal{N}_u$. The baselines cover 1) ID-based SR models, 2) small text/image encoders, 3) recommendation-specific foundation model embedders, and 4) our additionally implemented LLM/VLM-based embedder variants with or without SFT. Tab. 2 positions baselines along five properties: foundation-model (FM), sequence-level CF (Seq.CF), multimodal inputs (MM), embedder tuning (Emb.tune), and modality balancing (Bal.). VLM2Rec uniquely addresses all five properties. Detailed baseline implementations are in Appendix D.

Table 2: Positioning against baselines. ✓: yes, △: partial/limited.

Baselines	FM	Seq.CF	MM	Emb.tune	Bal.
ID-based [13, 18, 68, 27]		✓		✓	
BERT [8]				✓	
CLIP [41]			✓	✓	
SLIM [56]		✓			
SLIM$^+$ [56]	✓	✓			
LLMEmb [31]	✓	△		✓	
LLM2Rec [12]	✓	✓		✓	
$\text{VLM}_{\text{Prompt}}$ [38]	✓		✓		
NoteLLM-2 [63]	✓		✓	✓	△
$\text{LLM}_{\text{Vanilla}}$	✓				
LLM_{SFT}	✓	✓		✓	
$\text{VLM}_{\text{Vanilla}}$	✓		✓		
VLM_{SFT}	✓	✓	✓	✓	
VLM2Rec	✓	✓	✓	✓	✓

[2] https://jmcauley.ucsd.edu/data/amazon/

[3] https://grouplens.org/datasets/movielens/1m/

Table 3: Performance on downstream SR backbone. The best and second-best results are highlighted by **boldface** and underline. * indicates statistical significance ($p < 0.05$, paired t-test over 5 seeds).

Data	Metric	ID	Text-based Models							Multimodal-based Models					Ours	
		Base	BERT	SLIM	SLIM^{+}	LLMEmb	LLM2Rec	LLM_{Van}	LLM_{SFT}	CLIP	VLM_{Prm}	Note-2	VLM_{Van}	VLM_{SFT}	**VLM2Rec**	*Imp.(%)*
Backbone: SASRec																
Toys	H@10	0.0904	0.0702	0.0733	0.0957	0.0982	0.1005	0.0808	0.1012	0.0728	0.0796	0.1049	0.0848	0.1063	**0.1183***	11.29
	H@20	0.1226	0.1039	0.1082	0.1340	0.1374	0.1403	0.1144	0.1412	0.1054	0.1130	0.1436	0.1182	0.1459	**0.1589***	8.91
	N@10	0.0524	0.0391	0.0411	0.0559	0.0574	0.0594	0.0460	0.0619	0.0408	0.0450	0.0598	0.0493	0.0603	**0.0683***	10.34
	N@20	0.0604	0.0473	0.0495	0.0657	0.0671	0.0687	0.0537	0.0710	0.0483	0.0529	0.0687	0.0574	0.0692	**0.0785***	10.56
Beauty	H@10	0.0712	0.0600	0.0606	0.0806	0.0844	0.0891	0.0643	0.0949	0.0662	0.0681	0.0903	0.0691	0.0915	**0.1083***	14.12
	H@20	0.1056	0.0953	0.0959	0.1213	0.1251	0.1307	0.0985	0.1376	0.1000	0.1014	0.1325	0.1029	0.1337	**0.1511***	9.81
	N@10	0.0397	0.0308	0.0312	0.0445	0.0465	0.0493	0.0339	0.0530	0.0358	0.0369	0.0504	0.0377	0.0515	**0.0617***	16.42
	N@20	0.0482	0.0390	0.0395	0.0540	0.0563	0.0588	0.0421	0.0604	0.0437	0.0453	0.0621	0.0462	0.0627	**0.0725***	15.63
Clothing	H@10	0.0403	0.0303	0.0308	0.0434	0.0454	0.0469	0.0337	0.0475	0.0347	0.0359	0.0490	0.0366	0.0498	**0.0537***	7.83
	H@20	0.0651	0.0520	0.0528	0.0680	0.0698	0.0718	0.0566	0.0724	0.0586	0.0596	0.0743	0.0604	0.0754	**0.0820***	8.75
	N@10	0.0199	0.0144	0.0147	0.0220	0.0234	0.0244	0.0166	0.0248	0.0176	0.0180	0.0257	0.0184	0.0261	**0.0276***	5.75
	N@20	0.0259	0.0198	0.0201	0.0281	0.0294	0.0305	0.0217	0.0309	0.0229	0.0234	0.0320	0.0239	0.0325	**0.0347***	6.77
Sports	H@10	0.0433	0.0317	0.0322	0.0427	0.0444	0.0464	0.0367	0.0491	0.0354	0.0375	0.0469	0.0381	0.0474	**0.0567***	15.48
	H@20	0.0651	0.0513	0.0520	0.0653	0.0673	0.0694	0.0579	0.0719	0.0564	0.0590	0.0723	0.0598	0.0735	**0.0846***	15.1
	N@10	0.0240	0.0157	0.0161	0.0228	0.0238	0.0251	0.0200	0.0269	0.0192	0.0205	0.0258	0.0210	0.0260	**0.0307***	14.13
	N@20	0.0295	0.0208	0.0212	0.0286	0.0300	0.0313	0.0251	0.0323	0.0241	0.0254	0.0325	0.0259	0.0330	**0.0377***	14.24
MovieLens	H@10	0.1259	0.0889	0.0899	0.1539	0.1583	0.1597	0.1094	0.1607	0.1026	0.1119	0.1627	0.1129	0.1642	**0.1791***	9.07
	H@20	0.1848	0.1329	0.1338	0.2355	0.2365	0.2384	0.1662	0.2391	0.1563	0.1690	0.2418	0.1705	0.2437	**0.2684***	10.14
	N@10	0.0631	0.0394	0.0399	0.0780	0.0802	0.0830	0.0543	0.0878	0.0508	0.0553	0.0904	0.0561	0.0918	**0.0980***	6.75
	N@20	0.0728	0.0550	0.0555	0.0907	0.0975	0.0992	0.0655	0.1012	0.0625	0.0663	0.1044	0.0670	0.1055	**0.1205***	14.22
Backbone: DiffuRec																
Toys	H@10	0.0890	0.0600	0.0610	0.0905	0.0928	0.0950	0.0825	0.0964	0.0750	0.0840	0.0972	0.0855	0.0979	**0.1148***	17.26
	H@20	0.1240	0.0870	0.0880	0.1275	0.1305	0.1338	0.1180	0.1371	0.1080	0.1200	0.1352	0.1215	0.1356	**0.1573***	14.73
	N@10	0.0500	0.0310	0.0315	0.0515	0.0530	0.0544	0.0465	0.0549	0.0410	0.0485	0.0554	0.0495	0.0557	**0.0654***	17.41
	N@20	0.0590	0.0390	0.0395	0.0605	0.0620	0.0638	0.0550	0.0646	0.0490	0.0572	0.0652	0.0585	0.0656	**0.0761***	16.01
Beauty	H@10	0.0851	0.0691	0.0694	0.0940	0.0964	0.0987	0.0812	0.1003	0.0825	0.0831	0.0984	0.0841	0.0989	**0.1105***	10.17
	H@20	0.1225	0.1013	0.1025	0.1345	0.1383	0.1411	0.1193	0.1430	0.1209	0.1216	0.1412	0.1221	0.1417	**0.1543***	7.9
	N@10	0.0471	0.0350	0.0354	0.0532	0.0545	0.0564	0.0439	0.0575	0.0440	0.0452	0.0567	0.0465	0.0568	**0.0629***	9.39
	N@20	0.0568	0.0433	0.0439	0.0627	0.0647	0.0657	0.0535	0.0668	0.0527	0.0552	0.0672	0.0562	0.0683	**0.0739***	8.2
Clothing	H@10	0.0488	0.0331	0.0336	0.0457	0.0473	0.0491	0.0404	0.0500	0.0420	0.0431	0.0506	0.0436	0.0512	**0.0596***	16.41
	H@20	0.0730	0.0546	0.0554	0.0698	0.0714	0.0737	0.0641	0.0753	0.0651	0.0670	0.0758	0.0677	0.0767	**0.0892***	16.3
	N@10	0.0252	0.0158	0.0162	0.0233	0.0244	0.0254	0.0202	0.0260	0.0216	0.0223	0.0263	0.0226	0.0268	**0.0311***	16.04
	N@20	0.0315	0.0212	0.0218	0.0296	0.0307	0.0317	0.0263	0.0324	0.0278	0.0288	0.0328	0.0292	0.0333	**0.0385***	15.62
Sports	H@10	0.0473	0.0353	0.0359	0.0456	0.0468	0.0478	0.0424	0.0485	0.0413	0.0435	0.0496	0.0440	0.0508	**0.0598***	17.72
	H@20	0.0696	0.0560	0.0565	0.0685	0.0704	0.0726	0.0647	0.0736	0.0630	0.0660	0.0740	0.0669	0.0751	**0.0889***	18.38
	N@10	0.0250	0.0174	0.0178	0.0239	0.0247	0.0256	0.0223	0.0264	0.0215	0.0226	0.0255	0.0231	0.0258	**0.0326***	23.48
	N@20	0.0309	0.0229	0.0233	0.0299	0.0308	0.0318	0.0277	0.0334	0.0269	0.0283	0.0317	0.0287	0.0320	**0.0399***	19.46
MovieLens	H@10	0.1296	0.0960	0.0970	0.1488	0.1536	0.1584	0.1200	0.1620	0.1164	0.1248	0.1637	0.1272	0.1651	**0.1800***	9.02
	H@20	0.1901	0.1392	0.1406	0.2208	0.2285	0.2342	0.1776	0.2375	0.1687	0.1838	0.2395	0.1862	0.2410	**0.2614***	8.46
	N@10	0.0643	0.0480	0.0488	0.0749	0.0806	0.0864	0.0595	0.0896	0.0586	0.0624	0.0904	0.0634	0.0912	**0.0981***	7.57
	N@20	0.0787	0.0605	0.0614	0.0902	0.0970	0.1037	0.0730	0.1070	0.0737	0.0758	0.1087	0.0768	0.1094	**0.1186***	8.41

Evaluation Protocol All modality-embedding methods, including VLM2Rec, produce frozen item representations and are mapped to the backbone hidden size (d=128) by a single linear adapter (Fig. 2, lower-left panel). We use each baseline's adapter when specified in their own papers, and ours otherwise. Downstream training updates only the adapter and SR backbone, making inference cost identical across modality-embedding methods. We adopt full-ranking evaluation [21] and report two ranking metrics, Hit Ratio (H@K) and Normalized Discounted Cumulative Gain (N@K), at $K \in \{10, 20\}$ averaged over five random seeds; all reported VLM2Rec gains pass a paired t-test at $p < 0.05$ against the strongest baseline.

5.2 Main Results

Based on the main results (Tab. 3), we highlight five observations. 1) **Multimodal signals are useful but insufficient without CF injection:** Generally, CLIP > BERT and $\text{VLM}_{\text{Vanilla}} > \text{LLM}_{\text{Vanilla}}$, showing the benefit of multimodal semantics. However, both remain weaker than ID-based SR models, indicating that foundation-model semantics alone cannot replace sequence-level CF supervision. 2) **Sequence-level CF injection improves initialization quality:** LLM_{SFT} and VLM_{SFT} generally outperform their vanilla counterparts, suggesting that sequence-level SFT transfers interaction structure into the embedding space. This also explains why methods without explicit sequence-level CF, such as BERT, CLIP, and $\text{VLM}_{\text{Prompt}}$, are limited as downstream initializers. 3) **Existing FM baselines address only partial aspects of SR initialization:** LLMEmb distills CF-aware ID embeddings and SLIM$^{(+)}$ and LLM2Rec inject sequential signals through next-item generation; all incorporate CF indirectly rather than through an explicit sequence-level contrastive objective. NoteLLM-2 strengthens multimodal item representations via modality-aware visual enhancement, but remains at item-level encoding without sequence-transition modeling. This helps explain their weaker transfer compared with VLM2Rec, which directly shapes sequence–item alignment in a balanced multi-

modal embedding space. 4) **Naive multimodal SFT suffers from a modality paradox:** although $\text{VLM}_{\text{Vanilla}}$ often outperforms $\text{LLM}_{\text{Vanilla}}$, VLM_{SFT} can underperform LLM_{SFT} after sequence-level tuning. SFT-induced modality collapse leaves the visual modality under-optimized, causing negative transfer when fused. VLM2Rec avoids this through explicit weak-modality balancing. 5) **VLM2Rec provides backbone-agnostic robustness:** VLM2Rec consistently achieves the best initialization quality across Transformer-based (SASRec), Diffusion-based (DiffuRec), and additionally MLP-based (FMLP) and RNN-based (GRU4Rec) backbones (Appendix E.3). Since VLM2Rec produces balanced, CF-aware item embeddings offline, the resulting representations serve as a plug-and-play initializer that transfers across diverse SR backbones.

5.3 Ablation Study

Fig. 3 evaluates the contribution of each component. Without $\mathcal{L}_{\text{WPCL}}$, the model underperforms VLM_{SFT}, as it neither injects sequence-level CF signals nor improves weak-modality discrimination. Within $\mathcal{L}_{\text{WPCL}}$, excluding $w_{u,\text{pen}}$ lowers performance, showing the importance of the user-adaptive weak-modality penalty. Without the stop-gradient in $\mathcal{L}_{\text{WPCL}}$, optimization can close the gap by degrading the strong-modality margin rather than improving the weak one, which also harms performance. *w/o* $\mathcal{L}_{\text{CRTR}}$ also underperforms VLM2Rec, confirming the contribution of cross-modal topology alignment.

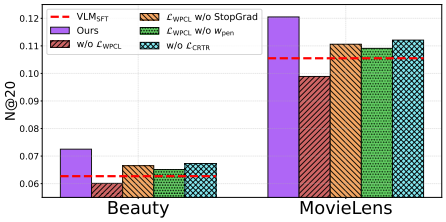

Figure 3: Ablation study of VLM2Rec.

5.4 Analysis of Mitigating Modality Collapse

Fig. 1a shows that VLM2Rec recovers the gap of *t2f* over *v2f* that had been amplified by SFT. Once the gap is closed, the two modalities act synergistically: the fused setting (*f2f*) surpasses either single-modality input and yields the best overall performance. Fig. 1b further shows that VLM2Rec mitigates the weak-modality gradient drop under SFT, yielding more balanced optimization between text and vision. Tab. 1 gives consistent geometric evidence: VLM2Rec restores vision-space separability S^V from the collapsed SFT state ($S^V \approx 1$ or $S^V < 1$) to $S^V > 1$ and improves vision uniformity U^V. Together, these results show that VLM2Rec mitigates SFT-amplified modality collapse and turns visual signals into useful recommendation evidence; these findings are consistently observed across additional domains, model families, and scales (Appendix E.1 and E.2).

5.5 Analysis of Intrinsic CF Encoding

To assess whether the learned embedding space intrinsically encodes CF signals, we use a retrieval setting that extracts both sequence and item representations from the same embedder and ranks items by their similarities, without training an SR backbone. As shown in Fig. 4, Vanilla model embeddings underperform SASRec, indicating that semantic representations alone do not capture sufficient CF signal. VLM2Rec surpasses SASRec even in this backbone-free setting, showing that its embeddings already contain sufficient CF signals in addition to rich multimodal semantics. Appendix E.5 confirms that these signals remain complementary when combined with ID embeddings in state-of-the-art hybrid multimodal SR backbones.

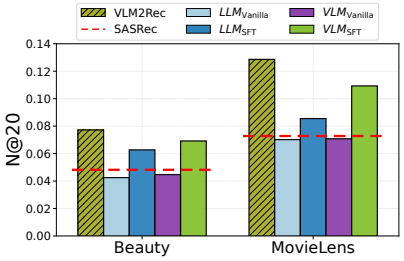

Figure 4: Intrinsic CF signal analysis.

5.6 Analysis of User-level Modality-Gap and Rank Recovery

We further ask whether VLM2Rec reduces modality collapse across modality-balanced and modality-imbalanced users, and whether this recovery explains ranking gains. Users are stratified into Low/High (bottom/top 20% by Vanilla modality gap), corresponding to initially modality-balanced/-imbalanced cases; *Collapse* isolates Low users most affected by SFT, defined as the top 50% within Low by the gap increase $\Delta^{\text{SFT}}_{u,\text{gap}} - \Delta^{\text{Vanilla}}_{u,\text{gap}}$. Tab. 4 shows that VLM2Rec substantially reduces the SFT-induced gap across all groups; User Red. denotes the fraction of users whose gap is reduced, confirming that

Table 4: User-level modality-gap stratification. User Red. denotes the fraction of users whose gap is reduced by VLM2Rec over SFT.

Data	Group	$\Delta^{\text{SFT}}_{u,\text{gap}}$	$\Delta^{\text{Ours}}_{u,\text{gap}}$	User Red.
MovieLens	Low	0.4048	0.0585	0.9461
	Collapse	0.6322	0.0691	1.0000
	High	0.3642	0.0679	0.9245
Clothing	Low	0.1711	0.0473	0.8410
	Collapse	0.2800	0.0561	0.9928
	High	0.1696	0.0566	0.8039

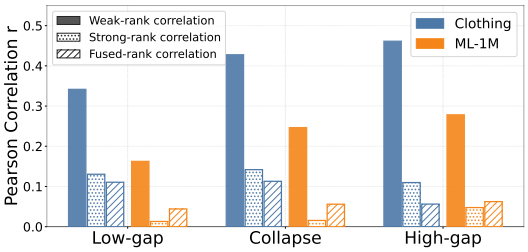

Figure 5: Correlation between $\Delta_{u,\text{gap}}$ and rank.

the effect is consistent. To link this recovery to recommendation quality, we correlate per-user gap recovery, $\Delta^{\text{SFT}}_{u,\text{gap}} - \Delta^{\text{Ours}}_{u,\text{gap}}$, with modality-specific sequence-target rank recovery. Fig. 5 shows the strongest correlation with weak-modality rank recovery, indicating that weak-modality improvement most directly drives the performance gain. This supports our central claim that resolving modality collapse is crucial for VLM-based SR, beyond merely refining the fused representation. Appendix E.6 further shows that the weak modality varies across users even within the same dataset, motivating user-adaptive measurement via $\Delta_{u,\text{gap}}$, and Appendix E.7 shows a real-world collapse-user case whose weak modality is recovered by VLM2Rec.

5.7 Analysis of Generalization under Scarce CF Signals

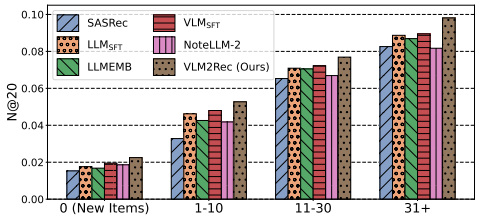

Figure 6: Performance in the cold-start scenario on the Beauty dataset.

Table 5: Performance comparison on cross-domain sequential recommendation (N@20).

Model	Clothing → Beauty	Sports → Clothing
LLMEmb	0.0426	0.0204
LLM$_{\text{SFT}}$	0.0489	0.0227
NoteLLM-2	0.0506	0.0236
VLM$_{\text{SFT}}$	0.0531	0.0254
VLM2Rec	**0.0588**	**0.0288**

We evaluate generalization under scarce CF signals in two settings: *cold-start items* (Fig. 6) and *cross-domain transfer* (Tab. 5). In the cold-start analysis, items are grouped by training frequency, where lower-frequency groups require greater reliance on semantics over CF memorization. In cross-domain transfer, the embedder trained on a source domain generates item embeddings for the target domain, which initialize the SR backbone trained and evaluated on that target; domain shift makes memorized CF signals unreliable. VLM2Rec achieves the best results in both settings, confirming that balanced multimodal representations transfer effectively when CF signals are scarce. Appendix E.8 further shows that this robustness extends across VLM families and parameter scales.

5.8 Analysis of Computational Cost and Few-shot Efficiency

Tab. 6 reports training time (mins/epoch) and item embedding generation time (seconds) across LLM/VLM methods. VLM$_{\text{SFT}}$ and VLM2Rec are more expensive than other methods because they process image sequence tokens, but their similar runtimes show that this cost arises from the VLM encoding stage rather than our proposed loss terms. Moreover, as noted in Sec. 5.1, downstream inference cost is identical across modality-embedding methods. To mitigate this cost, we report a few-shot variant trained on K_u random users: at K_u=128, it outperforms some full-training baselines, and at K_u=1024 (about 5–6% of training data), it performs comparably to the strongest baseline with shorter training time. The complexity and hyperparameter analyses in Appendix E.9 and Appendix E.10 confirm that our losses preserve SFT's asymptotic complexity and are stable across hyperparameters.

Table 6: Few-shot performance and efficiency analysis.

Model	Input	Beauty			Toys		
		N@20	Train	Emb	N@20	Train	Emb
LLMEmb	Item	0.0563	22	58	0.0671	15	38
LLM2Rec	Seq./Item	0.0588	33	58	0.0687	26	39
LLM$_{\text{SFT}}$	Seq.	0.0604	35	56	0.0710	26	39
NoteLLM-2	Item	0.0621	26	132	0.0687	20	91
VLM$_{\text{SFT}}$	Seq.	0.0627	117	113	0.0692	92	79
VLM2Rec (K_u=128)	Seq.	0.0618	1	113	0.0675	1	79
VLM2Rec (K_u=256)		0.0635	2		0.0689	2	
VLM2Rec (K_u=512)		0.0651	4		0.0707	4	
VLM2Rec (K_u=1024)		0.0662	9		0.0721	7	
VLM2Rec (Full)		**0.0725**	118		**0.0785**	95	

6 Conclusion

We presented VLM2Rec, a framework for adapting VLMs into CF-aware multimodal embedders for sequential recommendation. The key motivation is that multimodal capacity alone does not guarantee better recommendations: under standard contrastive SFT, VLMs can amplify modality imbalance and suppress the weak modality, turning visual information into negative transfer. VLM2Rec targets this failure mode with weak-modality penalization and cross-modal relational topology regularization, improving weak-modality discrimination while preserving multimodal consistency. Experiments show that this simple objective-level correction yields stronger downstream initialization across datasets, model families, and real-world scenarios. We acknowledge and discuss limitations in Appendix F, and hope this simple, general framework motivates further research on VLM-based recommender systems.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. Multi-modal mixture of experts representation learning for sequential recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 110–119, 2023.

[5] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.

[6] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2022*, pages 2172–2182, 2022.

[7] Qiang Cui, Shu Wu, Qiang Liu, Wen Zhong, and Liang Wang. Mv-rnn: A multi-view recurrent neural network for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 32(2):317–331, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, 2019.

[9] Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22, 2023.

[10] Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Honggang Chen, Donglin Wang, Qingsen Yan, and Siteng Huang. Filter, correlate, compress: Training-free token reduction for mllm acceleration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4601–4609, 2026.

[11] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[12] Yingzhi He, Xiaohao Liu, An Zhang, Yunshan Ma, and Tat-Seng Chua. Llm2rec: Large language models are powerful embedding models for sequential recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 896–907, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737029. URL https://doi.org/10.1145/3711896.3737029.

[13] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

[14] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 585–593, 2022.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[16] Hengchang Hu, Wei Guo, Yong Liu, and Min-Yen Kan. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 843–853, 2023.

[17] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pages 9226–9259. PMLR, 2022.

[18] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE, 2018.

[19] WooJoo Kim, JunYoung Kim, JaeHyung Lim, SeongJin Choi, SeongKu Kang, and HwanJo Yu. Flame: Condensing ensemble diversity into a single network for efficient sequential recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, page 823–833, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809560. URL https://doi.org/10.1145/3805712.3809560.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1748–1757, 2020.

[22] JuneHyoung Kwon, MiHyeon Kim, Eunju Lee, Juhwan Choi, and YoungBin Kim. See-saw modality balance: See gradient, and sew impaired vision-language balance to mitigate dominant modality bias. *arXiv preprint arXiv:2503.13834*, 2025.

[23] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

[24] Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*, 2024.

[25] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023.

[26] Shiyu Li, Yang Tang, Ruijie Liu, Shi-Zhe Chen, and Xi Chen. Conan-embedding-v2: Training an llm from scratch for text embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15011–15027, 2025.

[27] Zihao Li, Aixin Sun, and Chenliang Li. Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, 42(3):1–28, 2023.

[28] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[29] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342, 2025.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[31] Qidong Liu, Xian Wu, Wanyu Wang, Yejing Wang, Yuanshao Zhu, Xiangyu Zhao, Feng Tian, and Yefeng Zheng. Llmemb: large language model can be a good embedding generator for sequential recommendation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i11.33327. URL `https://doi.org/10.1609/aaai.v39i11.33327`.

[32] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[35] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[37] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[38] Claudio Pomo, Matteo Attimonelli, Danilo Danese, Fedelucio Narducci, and Tommaso Di Noia. Do recommender systems really leverage multimodal content? a comprehensive analysis on multimodal representations for recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 2377–2387, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720406. doi: 10.1145/3746252.3761398. URL `https://doi.org/10.1145/3746252.3761398`.

[39] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 813–823, 2022.

[40] A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint*, 2024.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[42] Ruiyang Ren, Zhaoyang Liu, Yaliang Li, Wayne Xin Zhao, Hui Wang, Bolin Ding, and Ji-Rong Wen. Sequential recommendation with self-attentive multi-adversarial network. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 89–98, 2020.

[43] Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. *arXiv preprint arXiv:2404.07983*, 2024.

[44] Peiyang Shi, Michael C Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in clip. In *ICLR 2023 workshop on multimodal representation learning: perks and pitfalls*, 2023.

[45] Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, 2025.

[46] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7614–7623, 2025.

[47] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019.

[48] Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. Llms are also effective embedding models: An in-depth overview. *arXiv preprint arXiv:2412.12591*, 2024.

[49] Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. Flashsloth: Lightning multimodal large language models via embedded visual compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14570–14581, 2025.

[50] Haicheng Wang, Zhemeng Yu, Gabriele Spadaro, Chen Ju, Victor Quétu, Shuai Xiao, and Enzo Tartaglione. Folder: Accelerating multi-modal large language models with enhanced performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23614–23625, 2025.

[51] Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6548–6557, 2023.

[52] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, 2024.

[53] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[54] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[55] Yimu Wang, Mozhgan Nasr Azadani, Sean Sedwards, and Krzysztof Czarnecki. Leo-mini: An efficient multimodal large language model using conditional token reduction and mixture of multi-modal experts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7257–7272, 2025.

[56] Yuling Wang, Changxin Tian, Binbin Hu, Yanhua Yu, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, Liang Pang, and Xiao Wang. Can small language models be good reasoners for sequential recommendation? In *Proceedings of the ACM Web Conference 2024*, pages 3876–3887, 2024.

[57] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

[58] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 1259–1273. IEEE, 2022.

[59] Lanling Xu, Zhen Tian, Bingqian Li, Junjie Zhang, Daoyuan Wang, Hongyu Wang, Jinpeng Wang, Sheng Chen, and Wayne Xin Zhao. Sequence-level semantic representation fusion for recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5015–5022, 2024.

[60] Yu Ye, Junchen Fu, Yu Song, Kaiwen Zheng, and Joemon M Jose. Are multimodal embeddings truly beneficial for recommendation? a deep dive into whole vs. individual modalities. In *European Conference on Information Retrieval*, pages 66–81. Springer, 2026.

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

[62] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2639–2649, 2023.

[63] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 2815–2826, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709440. URL `https://doi.org/10.1145/3690624.3709440`.

[64] Junjie Zhang, Ruobing Xie, Hongyu Lu, Wenqi Sun, Wayne Xin Zhao, Yu Chen, and Zhanhui Kang. Frequency-augmented mixture-of-heterogeneous-experts framework for sequential recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 2596–2605, 2025.

[65] Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. Hierarchical time-aware mixture of experts for multi-modal sequential recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 3672–3682, 2025.

[66] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023.

[67] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1893–1902, 2020.

[68] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. Filter-enhanced mlp is all you need for sequential recommendation. In *Proceedings of the ACM web conference 2022*, pages 2388–2399, 2022.

[69] Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, et al. When large vision language models meet multimodal sequential recommendation: An empirical study. In *Proceedings of the ACM on Web Conference 2025*, pages 275–292, 2025.

[70] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[71] Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. *arXiv preprint arXiv:2510.02745*, 2025.

A Related Work

Multimodal Sequential Recommendation Sequential Recommendation (SR) models dynamic preferences from interaction histories [18, 13, 47, 68, 58, 39, 6, 64, 19]. While fusion strategies [62, 7, 16] and modern architectures [51, 4, 14] incorporate side information, they typically rely on *frozen* encoders. Unlike ID-based models [18, 13], where learnable tables absorb collaborative signals, frozen encoders shift the learning burden to the backbone. This necessitates a shift toward CF-aware modality embeddings that enable direct ranking without ID dependence.

Large Model Embedders for Multimodal Sequential Recommendation Recent NLP work repurposes LLMs as high-capacity encoders for representation learning [26, 52, 23, 24, 48]. These methods provide important foundations for CF-aware foundation-model embedders. SLIM [56] distills sequential recommendation knowledge from ChatGPT [1] into compact representations, and LLMEmb [31] distills CF-aware ID embeddings into language-model representations. LLM2Rec [12] further combines generative next-item supervision with an item-level contrastive stage to inject sequential recommendation signals. In multimodal recommendation, NoteLLM-2 [63] introduces modality-aware visual enhancement for stronger multimodal item embeddings. Nevertheless, the use of VLMs as sequence-level multimodal embedders remains relatively underexplored, especially when contrastive SFT is expected to inject sequential transition patterns while maintaining balanced text-vision utilization.

Modality Imbalance and Balancing in Multimodal Learning Multimodal models often over-rely on an easier or dominant modality while under-utilizing the others. Prior studies analyze how dataset and model biases induce imbalanced multimodal optimization [54, 17, 57, 9, 45], and related work further shows that vision-language embedding spaces can exhibit systematic modality gaps or dominant-modality geometry [44, 66, 28]. In general multimodal learning, several methods such as adaptive gradient modulation [37, 25] have been proposed to address this imbalance at the optimization level. In recommendation, however, recent studies [69] have only begun to document this phenomenon, without offering objective-level solutions for the VLM-as-embedder paradigm. We find that contrastive SFT, commonly used to inject CF signals, further exacerbates this imbalance, and address it with explicit weak-modality balancing.

B Prompt Design and Representation Extraction

The exact templates for both the interleaved baseline and our proposed separate prompts are detailed below:

Interleaved Prompt
Describe this item sequence in one word,
0) <item1_title|> <item1_image|>, ..., L) <item$L$_title|> <item$L$_image|>

Proposed Separate-Modality Prompts
[P_T]: Describe this item text sequence in one word,
0) <item1_title|>, ..., L) <item$L$_title|>
[P_V]: Describe this item image sequence in one word,
0) <item1_image|>, ..., L) <item$L$_image|>

We deliberately adopt minimalist instruction templates to reduce prompt sensitivity. The phrase "*Describe this ... in one word,*" encourages the autoregressive model to compress the preceding sequence into its final hidden state, from which we extract a single embedding vector. For candidate items, we reuse the same template with L=0, ensuring structural alignment between sequence and item representations. All extracted modality embeddings ($\mathbf{z}_u^T, \mathbf{z}_u^V, \mathbf{e}_i^T, \mathbf{e}_i^V$) are ℓ_2-normalized prior to additive fusion, so that the fused representations $\mathbf{z}_u$ and $\mathbf{e}_i$ lie on a common scale and downstream cosine-similarity-based objectives are not dominated by the modality with the larger norm.

C Additional Methodological Analysis

C.1 Theoretical Analysis of WPCL as Adaptive Margin Learning

We next show that the negative-sample pushing penalty in $\mathcal{L}_{\text{WPCL}}$ is equivalent to imposing a contrastive margin that grows with the modality gap $\Delta_{u,\text{gap}}$. For a user u, define

$$A_u = e^{\text{sim}(\mathbf{z}_u, \mathbf{e}_{i^+})/\tau_{\text{WPCL}}}, \quad B_u = \sum_{i^- \in \mathcal{N}_u} e^{\text{sim}(\mathbf{z}_u, \mathbf{e}_{i^-})/\tau_{\text{WPCL}}}.$$

The per-user WPCL term can be written as

$$\ell_u = -\log \frac{A_u}{A_u + w_{u,\text{pen}} B_u} = \log\left(1 + w_{u,\text{pen}} \frac{B_u}{A_u}\right). \tag{12}$$

For the single-negative case with positive similarity s^+ and negative similarity s^-, this reduces to

$$\ell_u = \log\left(1 + e^{\frac{s^- - s^+}{\tau_{\text{WPCL}}} + \log w_{u,\text{pen}}}\right). \tag{13}$$

Viewed as a logistic loss $\log(1 + e^x)$ with $x = (s^- - s^+)/\tau_{\text{WPCL}} + \log w_{u,\text{pen}}$, the inflection (loss-symmetric) point—the effective contrastive margin—shifts from $s^+ = s^-$ to

$$s^+ - s^- = \tau_{\text{WPCL}} \log w_{u,\text{pen}}. \tag{14}$$

Achieving low loss therefore requires $s^+ - s^- > \tau_{\text{WPCL}} \log w_{u,\text{pen}}$, and since $w_{u,\text{pen}}$ increases monotonically with $\Delta_{u,\text{gap}}$, users with larger modality gaps are assigned a larger required separation margin. For multiple negatives, B_u is the log-sum-exp aggregate over negative similarities; hence, the same interpretation holds with respect to the aggregate negative evidence. WPCL is therefore not merely a constant rescaling of the negative-similarity term, but an adaptive-margin objective whose margin is controlled by the user-level modality gap.

C.2 Theoretical Analysis of WPCL Gradient Pathway

The full gradient of the per-user WPCL loss decomposes as

$$\nabla_\theta \ell_u = \underbrace{\nabla_\theta \ell_u \big|_{w_{u,\text{pen}}=\text{const}}}_{\text{standard contrastive}} + \underbrace{\nabla_\theta \ell_u^{\text{gap}}}_{w_{u,\text{pen}}\text{-mediated}}. \tag{15}$$

The first term is the standard contrastive gradient with $w_{u,\text{pen}}$-amplified negatives. Since it operates on the fused representation $\mathbf{z}_u = \mathbf{z}_u^T + \mathbf{z}_u^V$, it updates both modalities and empirically inherits the text-dominant optimization dynamics observed in Section 3. We now analyze the second term and show that it selectively targets $\mathcal{M}_{u,\text{weak}}$ under a fixed strong/weak assignment.

From the per-user loss above, we obtain

$$\frac{\partial \ell_u}{\partial w_{u,\text{pen}}} = \frac{B_u}{A_u + w_{u,\text{pen}} B_u} > 0. \tag{16}$$

Thus, increasing $w_{u,\text{pen}}$ monotonically increases the loss contribution of the negative-similarity term. Because Eq. 7 uses the original softplus form, $w_{u,\text{pen}}$ is larger than one even when $\Delta_{u,\text{gap}} = 0$; this constant offset preserves the baseline positive-negative discrimination, while the derivative below determines how the penalty changes as the modality gap grows. From the definition of $w_{u,\text{pen}}$, we have

$$\frac{\partial w_{u,\text{pen}}}{\partial \Delta_{u,\text{gap}}} = \alpha\beta\, \sigma(\alpha \Delta_{u,\text{gap}}) > 0, \tag{17}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and the positivity is guaranteed since α and β are implemented to be strictly positive in Eq. 7. Therefore,

$$\frac{\partial \ell_u}{\partial \Delta_{u,\text{gap}}} = \frac{\partial \ell_u}{\partial w_{u,\text{pen}}} \cdot \frac{\partial w_{u,\text{pen}}}{\partial \Delta_{u,\text{gap}}} > 0. \tag{18}$$

Let θ denote the trainable parameters. The gradient of this branch through $w_{u,\mathrm{pen}}$ can be decomposed as

$$\nabla_\theta \ell_u^{\mathrm{gap}} = \frac{\partial \ell_u}{\partial w_{u,\mathrm{pen}}} \frac{\partial w_{u,\mathrm{pen}}}{\partial \Delta_{u,\mathrm{gap}}} \nabla_\theta \Delta_{u,\mathrm{gap}}. \tag{19}$$

Since Eq. 6 applies stop-gradient to $\mathcal{M}_{u,\mathrm{strong}}$,

$$\nabla_\theta \Delta_{u,\mathrm{gap}} = -\nabla_\theta \mathcal{M}_{u,\mathrm{weak}}. \tag{20}$$

Hence,

$$\nabla_\theta \ell_u^{\mathrm{gap}} = -\gamma_u \nabla_\theta \mathcal{M}_{u,\mathrm{weak}}, \qquad \gamma_u = \frac{\partial \ell_u}{\partial w_{u,\mathrm{pen}}} \frac{\partial w_{u,\mathrm{pen}}}{\partial \Delta_{u,\mathrm{gap}}} > 0. \tag{21}$$

Under gradient descent, the full update combines the standard contrastive gradient, which updates both modalities with a text-dominant bias (Section 3), with the $w_{u,\mathrm{pen}}$-mediated term whose update direction is proportional to $+\nabla_\theta \mathcal{M}_{u,\mathrm{weak}}$. This second term injects additional gradient signal for weak-modality discrimination, either by raising the positive similarity or by lowering the average negative similarity of the weak modality. At ties where the max/min assignment changes, the objective is treated with the standard subgradient convention used for piecewise differentiable neural losses.

C.3 Theoretical Analysis of CRTR as Topology Discrepancy Control

We formalize the role of $\mathcal{L}_{\mathrm{CRTR}}$ from three perspectives: why distributional alignment is preferred over point-wise alignment, how it controls neighborhood discrepancy, and why the trivial solution is excluded by the joint objective.

Modality-specific geometry preservation. A natural alternative to $\mathcal{L}_{\mathrm{CRTR}}$ is point-wise alignment, e.g., minimizing $\|\mathbf{z}_u^T - \mathbf{z}_u^V\|_2^2$. However, $\mathcal{L}_{\mathrm{CRTR}}$ is strictly weaker: it permits modality-specific embedding geometries as long as their ranking structures agree. Concretely, for any orthogonal matrix $R \in \mathbb{R}^{d\times d}$ and scalar $\rho > 0$, replacing all vision embeddings $\{\mathbf{z}_u^V, \mathbf{e}_j^V\}$ with $\{\rho R\,\mathbf{z}_u^V, \rho R\,\mathbf{e}_j^V\}$ preserves $\mathbf{P}_u^V$ and hence $\mathcal{L}_{\mathrm{CRTR}}$, because cosine similarity is invariant to joint positive scaling and rotation: $\mathrm{sim}(\rho R\,\mathbf{a},\ \rho R\,\mathbf{b}) = \frac{\rho^2 \mathbf{a}^\top R^\top R\,\mathbf{b}}{\rho^2 \|\mathbf{a}\|\|\mathbf{b}\|} = \mathrm{sim}(\mathbf{a},\mathbf{b})$. Since $\mathbf{P}_{u,j}^V$ depends only on $\{\mathrm{sim}(\mathbf{z}_u^V, \mathbf{e}_j^V)\}_{j=1}^{|\mathcal{C}^V|}$, the distribution is unchanged. By symmetry, an analogous invariance holds for text embeddings with an independent (ρ', R'). In contrast, point-wise alignment $\|\mathbf{z}_u^T - \rho R\,\mathbf{z}_u^V\|_2^2$ is generally non-zero under such transformations, meaning it would collapse modality-specific structure that $\mathcal{L}_{\mathrm{CRTR}}$ preserves.

Neighborhood discrepancy bound. For each user u, $\mathbf{P}_u^T$ and $\mathbf{P}_u^V$ are probability distributions over the same candidate item set. The CRTR term minimizes the bidirectional KL divergence

$$D_{\mathrm{SKL}}(\mathbf{P}_u^T, \mathbf{P}_u^V) = \mathrm{KL}(\mathbf{P}_u^T \| \mathbf{P}_u^V) + \mathrm{KL}(\mathbf{P}_u^V \| \mathbf{P}_u^T). \tag{22}$$

By Pinsker's inequality,

$$\|\mathbf{P}_u^T - \mathbf{P}_u^V\|_1 \le \sqrt{2\mathrm{KL}(\mathbf{P}_u^T \| \mathbf{P}_u^V)} \le \sqrt{2 D_{\mathrm{SKL}}(\mathbf{P}_u^T, \mathbf{P}_u^V)}. \tag{23}$$

Therefore, reducing the CRTR objective also reduces an upper bound on the total variation distance between the text-induced and vision-induced neighborhood distributions. Moreover, for any bounded neighborhood statistic h satisfying $\|h\|_\infty \le M$,

$$\left|\mathbb{E}_{j\sim \mathbf{P}_u^T}[h(j)] - \mathbb{E}_{j\sim \mathbf{P}_u^V}[h(j)]\right| \le M \|\mathbf{P}_u^T - \mathbf{P}_u^V\|_1. \tag{24}$$

Thus, minimizing $\mathcal{L}_{\mathrm{CRTR}}$ controls the discrepancy between modality-specific candidate-neighborhood statistics.

Trivial solution exclusion under the joint objective. A potential concern is that $\mathcal{L}_{\mathrm{CRTR}}$ alone admits a trivial solution: both distributions collapse to uniform $\mathbf{P}_u^T = \mathbf{P}_u^V = (1/|\mathcal{C}^m|, \ldots, 1/|\mathcal{C}^m|)$, achieving $D_{\mathrm{SKL}} = 0$ with zero discrimination. However, the joint objective $\mathcal{L} = \mathcal{L}_{\mathrm{WPCL}} + \lambda\,\mathcal{L}_{\mathrm{CRTR}}$ excludes this solution. Under uniform $\mathbf{P}_u^m$, per-modality cosines $\mathrm{sim}(\mathbf{z}_u^m, \mathbf{e}_j^m)$ are constant in j, so $\mathcal{M}_{u,T} = \mathcal{M}_{u,V} = 0$ and $\nabla_\theta \mathcal{L}_{\mathrm{CRTR}} = \mathbf{0}$. Yet $\nabla_\theta \mathcal{L}_{\mathrm{WPCL}} \neq \mathbf{0}$ generically, because

$$\nabla_\theta \ell_u^{\mathrm{WPCL}} = \frac{1}{\tau_{\mathrm{WPCL}}} \left(\sum_j p_j\, \nabla_\theta \mathrm{sim}(\mathbf{z}_u, \mathbf{e}_j) - \nabla_\theta \mathrm{sim}(\mathbf{z}_u, \mathbf{e}_{i^+}) \right), \tag{25}$$

where p_j are the WPCL softmax weights and the sum ranges over all in-batch candidates. Even when per-modality cosines are constant in j, the fused-space cosines $\mathrm{sim}(\mathbf{z}_u, \mathbf{e}_j)$ depend on cross-modal inner products $(\mathbf{z}_u^T)^\top \mathbf{e}_j^V$ and $(\mathbf{z}_u^V)^\top \mathbf{e}_j^T$, which vary with j on a generic configuration. The configurations where this does not hold form a measure-zero subset, so the trivial solution is generically not a stationary point of $\mathcal{L}$.

In summary, WPCL and CRTR play complementary roles: WPCL maintains discrimination and prevents collapse, while CRTR aligns modality-specific ranking structures without constraining the modality-specific embedding geometry.

D Experimental Details

D.1 Datasets and Preprocessing

Table 7: Statistics of datasets used in the experiments.

Dataset	#Users	#Items	#Interactions	Avg. Length	Sparsity (%)
Toys	15,921	8,383	108,336	6.8	99.92
Beauty	19,757	9,311	137,300	6.9	99.93
Clothing	30,757	17,087	196,614	6.4	99.96
Sports	32,127	14,820	222,591	6.9	99.95
MovieLens	6,033	2,177	58,234	9.65	99.56

Dataset statistics are summarized in Tab. 7. To mitigate data sparsity, we apply 5-core filtering [18, 42, 67, 14], retaining only users and items with at least five interactions; items lacking titles or images are also removed. Following [12], we truncate each interaction sequence to a maximum length of 10 and adopt the leave-one-out evaluation protocol [18, 42, 67]: for each user, the last item is held out for testing, the second-to-last for validation, and the remaining items for training.

D.2 Implementation Details

VLM2Rec pretraining. For efficient adaptation, we apply LoRA [15] with rank 16, alpha 32, and dropout 0.2. Hyperparameters are selected by grid search with $\tau_{\text{WPCL}} \in \{0.05, 0.1, 0.5, 1.0\}$, $\tau_{\text{CRTR}} \in \{0.001, 0.01, 0.1, 1.0\}$, and $\lambda \in \{0.1, 0.5, 1.0\}$. Training is conducted for 3 epochs using AdamW [33] with learning rate 1e-5 and batch size 8. Tab. 11 compares stronger negative-sampling strategies. Gradient checkpointing is enabled, and most experiments are run on a single RTX 3090. Experiments requiring full-parameter tuning or larger models are conducted on a single A100 80GB GPU. For large-model baselines, we use the same Qwen2.5 backbone family whenever applicable to isolate method differences from backbone effects.

SR backbone training. For modality-embedding methods, the offline item representations are kept as a frozen item embedding table; downstream training updates only the one-layer adapter and SR backbone parameters. The SR backbone models are trained with the Adam optimizer [20], using a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, a batch size of 256, and a dropout rate of 0.5. For downstream SR training, we use a patience of 10 epochs for early stopping on validation N@20. All SR backbone models are trained on a single RTX 3090 GPU. The hyperparameters for all baseline models are tuned according to the search spaces provided in their original papers. All reported main results are averaged over five random seeds, and significance markers compare VLM2Rec against the strongest baseline under a paired t-test over matched seeds.

D.3 Baselines

Following the grouping in Tab. 2, we organize the baselines into ID-based, small-encoder, recommendation-specific foundation-model, and our additionally implemented LLM/VLM-based embedder variants. In Tab. 2, ✓ indicates that a property is explicitly part of the method, △ that it is indirect or partial, and an empty cell that the property is absent.

ID-based baselines. We include GRU4Rec [13], FMLP [68], SASRec [18], and DiffuRec [27] as representative RNN-, MLP-, Transformer-, and Diffusion-based SR backbones, respectively. They receive ✓ for sequence-level CF and Embedding tuning because they train directly on interaction sequences and optimize to generate embeddings, but lack foundation-model semantics, multimodality, and modality balancing since they do not use external content embedders.

Small-encoder baselines. We use BERT[4] [8] for text-only item encoding and CLIP[5] [41] for multimodal item encoding. Both provide pretrained content features and are adapted through the common frozen item-table plus trainable adapter protocol, but their item embeddings are produced from item content without sequence-level recommendation supervision or weak-modality correction, so neither receives ✓ for sequence-level CF or modality balancing. BERT is text-only, while CLIP receives ✓ for multimodality through its joint text and image encoders.

Recommendation-specific foundation-model baselines. We include SLIM [56], SLIM$^+$, LLMEmb [31], LLM2Rec [12], VLM$_{\text{Prompt}}$ [38], and NoteLLM-2 [63], which adapt foundation-model embedders specifically for recommendation. Following the original paper, SLIM extracts final representations using BERT rather than the distilled student LLM from the ChatGPT API[6], whereas our additional SLIM$^+$ baseline directly extracts representations from the student foundation model; therefore, only SLIM$^+$ receives ✓ for FM in Tab. 2. LLMEmb is marked △ for sequence-level CF because it transfers CF indirectly from ID embeddings, while LLM2Rec and SLIM$^+$ inject sequence-level CF signals through a next-item generation objective. Since both SLIM and SLIM$^+$ are optimized for next-item generation rather than directly for embedding quality, neither receives ✓ for Embedding tuning. Among multimodal methods, VLM$_{\text{Prompt}}$ relies on zero-shot prompting without sequence-level supervision, and NoteLLM-2 is marked △ for modality balancing because it fuses the visual modality through a simple gated sum without a dedicated modality-aware design; NoteLLM-2 also lacks sequence-level CF as it operates at the item level.

LLM/VLM-based embedder variants. We additionally implement LLM$_{\text{Vanilla}}$, LLM$_{\text{SFT}}$, VLM$_{\text{Vanilla}}$, and VLM$_{\text{SFT}}$ to isolate the effects of (i) sequence-level SFT and (ii) the multimodal capacity of the backbone. The Vanilla variants directly extract sequence and item embeddings from the foundation model without further training, while the SFT variants apply sequence-level contrastive fine-tuning. LLM$_{\text{Vanilla}}$, LLM$_{\text{SFT}}$, and VLM$_{\text{Vanilla}}$ therefore lack at least one of sequence-level CF, multimodality, or embedder tuning, and none of these variants explicitly balance weak and strong modalities.

E Additional Empirical, Robustness, and Efficiency Analyses

E.1 Additional Modality-collapse Analysis

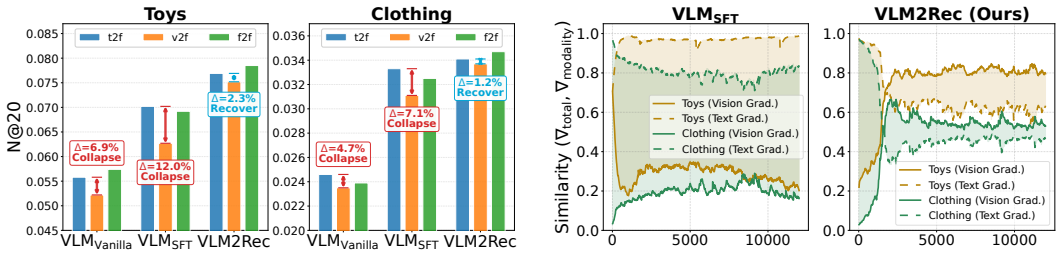

Figure 7: Additional analysis of modality collapse on Toys and Clothing via **dropout test (a)** and **gradient dynamics (b)**. Clothing is included as a domain where visual information is known to be highly informative in multimodal recommendation.

We extend the modality-collapse analysis to Toys and Clothing, which provide complementary cases: Toys is a sparse domain where image signals can be noisy, while Clothing is commonly

[4] google-bert/bert-large-uncased
[5] openai/clip-vit-large-patch14
[6] gpt-4.1-mini

Table 8: Comparison of VLM representation geometry across three states. Metrics include uniformity (U), and separability ($S = A_{\text{neg}}/A_{\text{pos}}$).

Metric	Modality	Toys			Clothing		
		$\text{VLM}_{\text{Vanilla}}$	VLM_{SFT}	**VLM2Rec (Ours)**	$\text{VLM}_{\text{Vanilla}}$	VLM_{SFT}	**VLM2Rec (Ours)**
$\boldsymbol{U}$ (↓)	Fused	-0.4707	-2.2500	-1.7969	-0.4336	-1.8359	-1.9951
	Vision	-0.0237	**-0.0039**	**-0.5703**	**-0.2368**	**-0.0237**	**-1.2136**
	Text	-0.4453	-2.2188	-0.9335	-0.4102	-1.7187	-1.3759
$\boldsymbol{S}$ (↑)	Fused	1.0609	1.2418	1.1609	1.0444	1.0474	1.1274
	Vision	**1.0039**	**0.9911**	**1.1551**	**1.0110**	**0.9997**	**1.0706**
	Text	1.0673	1.2480	1.1218	1.0551	1.0562	1.0610

considered a visually informative domain. As shown in Fig. 7 and Tab. 8, standard SFT again biases the fused update toward the stronger modality and leaves vision separability S^V near or below 1 with degraded uniformity, consistent with the findings in Section 3. Notably, even on Clothing where visual information is expected to be useful, standard SFT still pushes the vision space close to collapse, confirming that the pattern is not limited to visually noisy domains. VLM2Rec restores S^V well above 1, substantially improves U^V, and yields a more balanced gradient pattern. As a result, the fused setting (*f2f*) surpasses either single-modality input and yields the best overall performance, consistent with the mitigation results in Section 5.4.

E.2 Modality-collapse Analysis across VLM Backbones

Table 9: Modality dropout test across VLM backbones (Beauty, SASRec, N@20).

VLM Backbone	VLM_{SFT}				**VLM2Rec**			
	t2f	*v2f*	*f2f*	**Gap (%)**	*t2f*	*v2f*	*f2f*	**Gap (%)**
InternVL3-2B	0.0614	0.0558	0.0608	10.04	0.0703	0.0686	0.0702	**2.48**
Qwen2.5-VL-7B	0.0636	0.0588	0.0632	8.16	0.0746	0.0733	0.0754	**1.77**
LLaVA-1.5-7B	0.0625	0.0572	0.0620	9.27	0.0733	0.0725	0.0742	**1.10**
InternVL3-8B	0.0645	0.0597	0.0636	8.04	0.0744	0.0727	0.0750	**2.34**
Qwen2.5-VL-32B	0.0653	0.0612	0.0647	6.70	0.0761	0.0744	0.0776	**2.28**

We further check whether the modality-collapse pattern observed in Section 3 is specific to a particular VLM backbone. We rerun the modality dropout test on Beauty across five backbones spanning three model families (InternVL3 [70], Qwen2.5-VL [3], LLaVA-1.5 [30]) and four parameter scales (2B, 7B, 8B, 32B), and report *t2f*, *v2f*, and *f2f* together with the relative *t2f*–*v2f* gap in Tab. 9. For every backbone, VLM_{SFT} exhibits a sizable *t2f*–*v2f* gap (6.70–10.04%), indicating that the vision-only representation is systematically weaker than the text-only one regardless of the model family or parameter scale. VLM2Rec narrows this gap to 1.10–2.48% and simultaneously improves *f2f* performance, confirming that the SFT-induced modality collapse and its mitigation by VLM2Rec are consistent properties rather than artifacts of a specific backbone choice.

E.3 Additional Backbone Results

To verify the backbone robustness of VLM2Rec, we additionally evaluate our embeddings on two distinct sequential recommenders in Tab. 10: the MLP-based FMLP [68] and the RNN-based GRU4Rec [13]. VLM2Rec consistently delivers the best performance, indicating that the item embeddings learned by our framework are effective across diverse backbone architectures.

E.4 Effect of Negative Sampling for Modality Gap Estimation

We analyze the effect of negative sampling strategies for estimating $\Delta_{u,\text{gap}}$ in Eq. 6. In the main experiments, we use in-batch negatives for simplicity and efficiency. As shown in Tab. 11, stronger negatives improve N@20 over the in-batch setting. SASRec Top-k selects the top-k items predicted by SASRec for each user, while Text Similar and Image Similar sample target-similar

Table 11: Effect of negative sampling strategies for modality gap estimation (SASRec, N@20).

Dataset	In-batch	Random	SASRec Top-k	Text Sim.	Image Sim.
Clothing	0.0347	0.0342	0.0352	0.0357	0.0360
MovieLens	0.1205	0.1193	0.1218	0.1222	0.1229

Table 10: Performance on additional SR backbones. The best and second-best results are highlighted by **boldface** and underline. * indicates statistical significance ($p < 0.05$, paired t-test over 5 seeds).

Data	Metric	ID	Text-based Models							Multimodal-based Models					Ours	
		Base	BERT	SLIM	SLIM$^+$	LLMEmb	LLM2Rec	LLM_{Van}	LLM_{SFT}	CLIP	VLM_{Prm}	Note-2	VLM_{Van}	VLM_{SFT}	**VLM2Rec**	*Imp(%)*
Backbone: FMLP																
Toys	H@10	0.0840	0.0653	0.0680	0.0889	0.0912	0.0933	0.0738	0.0945	0.0676	0.0752	0.0974	0.0788	0.0988	**0.1097***	11.03
	H@20	0.1138	0.0965	0.1005	0.1245	0.1276	0.1303	0.1049	0.1310	0.0979	0.1063	0.1335	0.1098	0.1356	**0.1513***	11.58
	N@10	0.0487	0.0363	0.0381	0.0520	0.0533	0.0551	0.0418	0.0575	0.0378	0.0427	0.0558	0.0458	0.0564	**0.0626***	8.87
	N@20	0.0561	0.0440	0.0460	0.0610	0.0623	0.0638	0.0491	0.0660	0.0448	0.0499	0.0638	0.0533	0.0643	**0.0731***	10.76
Beauty	H@10	0.0665	0.0562	0.0566	0.0753	0.0788	0.0833	0.0600	0.0888	0.0618	0.0636	0.0852	0.0647	0.0863	**0.1025***	15.43
	H@20	0.0987	0.0891	0.0897	0.1133	0.1169	0.1222	0.0921	0.1286	0.0935	0.0948	0.1241	0.0961	0.1250	**0.1478***	14.93
	N@10	0.0369	0.0288	0.0293	0.0417	0.0434	0.0460	0.0318	0.0494	0.0334	0.0345	0.0473	0.0353	0.0480	**0.0581***	17.61
	N@20	0.0451	0.0365	0.0368	0.0504	0.0526	0.0550	0.0394	0.0565	0.0409	0.0424	0.0576	0.0431	0.0586	**0.0693***	18.26
Clothing	H@10	0.0349	0.0255	0.0267	0.0377	0.0394	0.0407	0.0292	0.0410	0.0289	0.0306	0.0425	0.0317	0.0432	**0.0508***	17.59
	H@20	0.0564	0.0440	0.0458	0.0589	0.0605	0.0623	0.0491	0.0628	0.0488	0.0509	0.0644	0.0524	0.0654	**0.0796***	21.71
	N@10	0.0173	0.0122	0.0127	0.0191	0.0203	0.0212	0.0144	0.0216	0.0146	0.0154	0.0223	0.0159	0.0226	**0.0267***	18.14
	N@20	0.0224	0.0168	0.0175	0.0243	0.0255	0.0264	0.0188	0.0268	0.0191	0.0200	0.0278	0.0207	0.0282	**0.0340***	20.57
Sports	H@10	0.0375	0.0275	0.0280	0.0377	0.0390	0.0409	0.0318	0.0426	0.0307	0.0325	0.0406	0.0330	0.0410	**0.0532***	24.88
	H@20	0.0564	0.0445	0.0451	0.0572	0.0585	0.0618	0.0502	0.0622	0.0489	0.0512	0.0627	0.0519	0.0637	**0.0784***	23.08
	N@10	0.0208	0.0136	0.0139	0.0199	0.0212	0.0226	0.0174	0.0233	0.0167	0.0178	0.0219	0.0182	0.0222	**0.0292***	25.32
	N@20	0.0255	0.0181	0.0184	0.0249	0.0258	0.0273	0.0218	0.0277	0.0209	0.0220	0.0282	0.0224	0.0286	**0.0354***	23.78
MovieLens	H@10	0.1091	0.0771	0.0779	0.1305	0.1340	0.1360	0.0949	0.1385	0.0889	0.0970	0.1410	0.0979	0.1423	**0.1600***	12.44
	H@20	0.1602	0.1152	0.1160	0.1995	0.2020	0.2035	0.1441	0.2055	0.1355	0.1465	0.2097	0.1478	0.2113	**0.2335***	10.51
	N@10	0.0547	0.0341	0.0346	0.0650	0.0675	0.0705	0.0470	0.0735	0.0440	0.0479	0.0783	0.0487	0.0796	**0.0874***	9.8
	N@20	0.0631	0.0476	0.0481	0.0765	0.0820	0.0845	0.0568	0.0865	0.0542	0.0574	0.0905	0.0581	0.0915	**0.1059***	15.74
Backbone: GRU4Rec																
Toys	H@10	0.0662	0.0547	0.0553	0.0637	0.0651	0.0665	0.0591	0.0689	0.0579	0.0612	0.0674	0.0627	0.0678	**0.0864***	25.4
	H@20	0.0963	0.0837	0.0846	0.0958	0.0977	0.0995	0.0915	0.1031	0.0860	0.0926	0.1008	0.0945	0.1013	**0.1229***	19.2
	N@10	0.0362	0.0284	0.0288	0.0344	0.0355	0.0363	0.0310	0.0379	0.0285	0.0330	0.0368	0.0343	0.0371	**0.0484***	27.7
	N@20	0.0436	0.0358	0.0365	0.0423	0.0435	0.0446	0.0390	0.0453	0.0349	0.0410	0.0462	0.0424	0.0466	**0.0577***	23.82
Beauty	H@10	0.0638	0.0590	0.0597	0.0689	0.0702	0.0715	0.0614	0.0733	0.0614	0.0629	0.0717	0.0635	0.0720	**0.0828***	12.96
	H@20	0.0955	0.0913	0.0919	0.1017	0.1032	0.1047	0.0928	0.1053	0.0925	0.0943	0.1065	0.0950	0.1072	**0.1206***	12.5
	N@10	0.0350	0.0310	0.0314	0.0378	0.0385	0.0390	0.0327	0.0392	0.0330	0.0338	0.0396	0.0345	0.0398	**0.0451***	13.32
	N@20	0.0428	0.0387	0.0391	0.0460	0.0465	0.0471	0.0402	0.0475	0.0404	0.0415	0.0481	0.0423	0.0484	**0.0547***	13.02
Clothing	H@10	0.0312	0.0282	0.0296	0.0326	0.0337	0.0346	0.0298	0.0350	0.0293	0.0301	0.0358	0.0310	0.0362	**0.0398***	9.94
	H@20	0.0519	0.0483	0.0502	0.0544	0.0558	0.0572	0.0507	0.0577	0.0497	0.0506	0.0588	0.0518	0.0592	**0.0627***	5.91
	N@10	0.0153	0.0132	0.0138	0.0159	0.0167	0.0174	0.0145	0.0177	0.0144	0.0149	0.0181	0.0155	0.0183	**0.0206***	12.57
	N@20	0.0203	0.0183	0.0191	0.0211	0.0221	0.0230	0.0194	0.0233	0.0191	0.0197	0.0238	0.0205	0.0241	**0.0263***	9.13
Sports	H@10	0.0326	0.0296	0.0309	0.0377	0.0391	0.0405	0.0314	0.0412	0.0309	0.0320	0.0417	0.0328	0.0420	**0.0469***	11.67
	H@20	0.0504	0.0476	0.0500	0.0586	0.0600	0.0618	0.0491	0.0652	0.0483	0.0495	0.0633	0.0508	0.0637	**0.0715***	9.66
	N@10	0.0171	0.0146	0.0153	0.0193	0.0203	0.0212	0.0163	0.0215	0.0161	0.0167	0.0219	0.0174	0.0221	**0.0244***	10.41
	N@20	0.0214	0.0193	0.0202	0.0246	0.0256	0.0266	0.0205	0.0280	0.0201	0.0209	0.0269	0.0217	0.0272	**0.0304***	8.57
MovieLens	H@10	0.0962	0.0841	0.0851	0.1190	0.1242	0.1269	0.0911	0.1305	0.0860	0.0928	0.1344	0.0944	0.1352	**0.1437***	6.29
	H@20	0.1495	0.1257	0.1269	0.1832	0.1907	0.1962	0.1442	0.2020	0.1308	0.1455	0.2076	0.1474	0.2085	**0.2225***	6.71
	N@10	0.0471	0.0375	0.0379	0.0586	0.0632	0.0660	0.0437	0.0685	0.0442	0.0451	0.0708	0.0460	0.0712	**0.0765***	7.44
	N@20	0.0593	0.0523	0.0528	0.0739	0.0791	0.0828	0.0567	0.0868	0.0560	0.0577	0.0898	0.0586	0.0903	**0.0964***	6.76

items based on BERT and CLIP embedding similarity, respectively. Notably, image-similar negatives achieve the best performance, suggesting that exposing the weak visual modality to fine-grained distinctions against hard negatives enables sharper modality-gap estimation. We adopt in-batch negatives to avoid additional retrieval or preprocessing overhead; nevertheless, advanced modality-aware negative sampling can further enhance VLM2Rec.

E.5 Additional Results on Hybrid SR Backbones with ID Embeddings

We test whether VLM2Rec remains useful when the downstream recommender already includes trainable ID embeddings. We replace the modality embedding component of UniSRec [14], TedRec [59], M3SRec [4], and HM4SR [65] with BERT, CLIP, LLM_{SFT}, VLM_{SFT}, or VLM2Rec embeddings. For UniSRec and TedRec, which accept text-side representations, VLM_{SFT} (Text) and Ours (Text) use only the text representation from the VLM embedder.

Table 12: Modality embedding replacement experiments on state-of-the-art models.

Embedding	UniSRec		TedRec	
	Clothing	MovieLens	Clothing	MovieLens
BERT	0.0308	0.1086	0.0332	0.1167
LLM_{SFT}	0.0325	0.1119	0.0348	0.1180
VLM_{SFT} (Text)	0.0328	0.1124	0.0355	0.1189
Ours (Text)	**0.0336**	**0.1137**	**0.0362**	**0.1208**
Embedding	**M3SRec**		**HM4SR**	
	Clothing	MovieLens	Clothing	MovieLens
CLIP	0.0322	0.1134	0.0277	0.1219
VLM_{SFT}	0.0338	0.1158	0.0328	0.1261
Ours	**0.0352**	**0.1176**	**0.0396**	**0.1315**

As shown in Tab. 12, VLM2Rec achieves the best performance across all hybrid settings. Notably, both VLM_{SFT} (Text) and Ours (Text) outperform LLM_{SFT}, indicating that the visual modality contributes during representation learning even when only the text representation is used downstream. Ours (Text) further improves over VLM_{SFT} (Text), confirming that our weak-modality balancing better exploits this visual signal. These results indicate that VLM2Rec's learned representations are complementary to ID embeddings and benefit both multimodal and text-oriented hybrid recommenders.

E.6 Anchor Weak-modality Analysis

Table 13: Anchor weak-modality split across the *All* and *Collapse* groups for each dataset. *Ratio* reports the fraction of users assigned to each anchor weak modality.

Data	Group	Weak	Ratio	$\Delta^{\text{SFT}}_{u,\text{gap}}$	$\Delta^{\text{Ours}}_{u,\text{gap}}$	User Red.
MovieLens	All	All	1.0000	0.3658	0.0522	0.9362
		Text	0.2728	0.2798	0.0228	0.9064
		Vision	0.7272	0.3981	0.0632	0.9473
	Collapse	All	1.0000	0.6255	0.0684	1.0000
		Text	0.2090	0.5497	0.0243	1.0000
		Vision	0.7910	0.6456	0.0801	1.0000
Clothing	All	All	1.0000	0.1366	0.0166	0.8290
		Text	0.4621	0.1301	0.0149	0.8235
		Vision	0.5379	0.1422	0.0180	0.8337
	Collapse	All	1.0000	0.2755	0.0510	0.9938
		Text	0.4891	0.2752	0.0477	0.9934
		Vision	0.5109	0.2758	0.0540	0.9943

We split each diagnostic user group by the anchor weak modality, defined as the weaker modality under the vanilla VLM. *All* and *Collapse* follow the same definitions as in Sec. 5.6. Tab. 13 shows two findings. First, vision is assigned as the weak modality for the majority of users (72.7% on MovieLens, 53.8% on Clothing), consistent with the text-dominance of VLMs. The ratio drops on the visually rich Clothing dataset, aligning with the expectation that visual information is more informative in this domain, but vision still remains the weaker modality overall. Second, VLM2Rec reduces $\Delta^{\text{SFT}}_{u,\text{gap}}$ by at least $5\times$ across every splits, with *User Red.* ≥ 0.82 everywhere, confirming broad recovery in both text-weak and vision-weak cases.

E.7 Case Study

Fig. 8 presents a representative user from the collapse-sensitive group defined in Section 5.6. For this user, standard sequence-level SFT produces recommendations that are dominated by the stronger modality, illustrating how the weak-modality signal can be suppressed even after CF supervision is injected. By contrast, VLM2Rec recovers a more balanced multimodal representation and ranks candidates that better reflect the target preference. This qualitative example complements the user-level stratification and gap-rank correlation analyses by showing how weak-modality recovery translates into an individual recommendation case.

E.8 Model Scalability and Robustness

To verify generalizability, we evaluate various VLM families (e.g., Qwen2.5-VL [3], InternVL3 [70], LLaVA-1.5 [30]) with parameter sizes ranging from 2B to 32B on the SASRec backbone (N@20), as shown in Fig. 9. Models within similar parameter groups (e.g., 2–3B, 7–8B) show comparable performance, suggesting robustness across different architectures. Overall, performance improves with capacity by leveraging richer prior knowledge, suggesting that parameter size can be chosen according to deployment objectives.

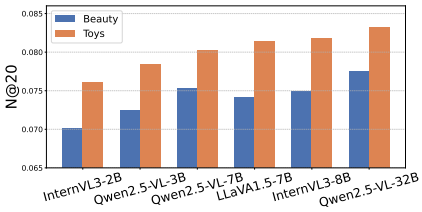

Figure 9: Model scalability across VLM families and parameter sizes.

E.9 Computational Complexity Analysis

We compare the computational overhead of $\mathcal{L}_{\text{SFT}}$, $\mathcal{L}_{\text{WPCL}}$, and $\mathcal{L}_{\text{CRTR}}$, focusing on the cost introduced by the loss computation itself after backbone encoding. Let $B = |\mathcal{B}|$ denote the batch size, $N = |\mathcal{N}_u|$ is the number of negatives per user, $C = |\mathcal{C}|$ the candidate set size used in $\mathcal{L}_{\text{CRTR}}$, and d the embedding dimension. As shown in Tab. 14, $\mathcal{L}_{\text{WPCL}}$ adds two modality-wise similarity passes over $\mathcal{L}_{\text{SFT}}$ for estimating discriminative margins, increasing time by a constant factor while keeping the same space complexity. $\mathcal{L}_{\text{CRTR}}$ constructs modality-wise sequence-to-candidate similarity matrices, yielding $O(4BCd)$ time and $O(BC)$ space. Under the standard in-batch setting, where $N = B - 1$ and the

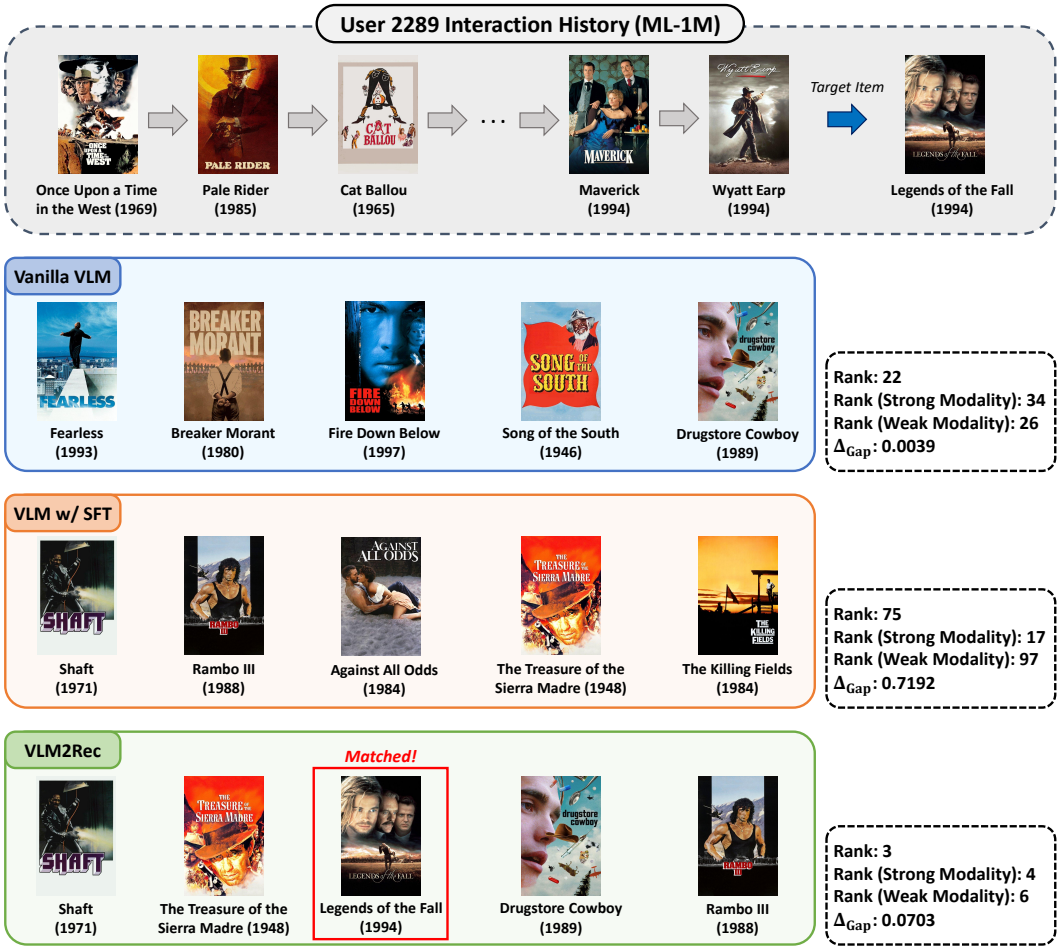

Figure 8: Top-5 prediction lists of each model for the case study.

Table 14: Time and space complexity comparison of the loss objectives. Here, space complexity denotes additional scratch memory required by the loss computation, excluding backbone activations.

Objective	Time Complexity	Space Complexity	In-batch Time	In-batch Space
$\mathcal{L}_{\text{SFT}}$	$O(2B(N+1)d)$	$O(B(N+1))$	$O(2B^2d)$	$O(B^2)$
$\mathcal{L}_{\text{WPCL}}$	$O(6B(N+1)d)$	$O(B(N+1))$	$O(6B^2d)$	$O(B^2)$
$\mathcal{L}_{\text{CRTR}}$	$O(4BCd)$	$O(BC)$	$O(4B^2d)$	$O(B^2)$

candidate set coincides with the batch ($|\mathcal{C}| = B$), all three losses reduce to $O(B^2d)$ time and $O(B^2)$ space, differing only in constant factors.

E.10 Hyperparameter Study

We conduct a sensitivity analysis of VLM2Rec by varying hyperparameters across the search space. In Fig. 10, VLM2Rec consistently maintains superior performance over baselines with minimal variance across varying hyperparameter values, demonstrating its robust stability. This reduces the additional tuning cost often required by multi-objective training, making VLM2Rec a stable solution across diverse domains.

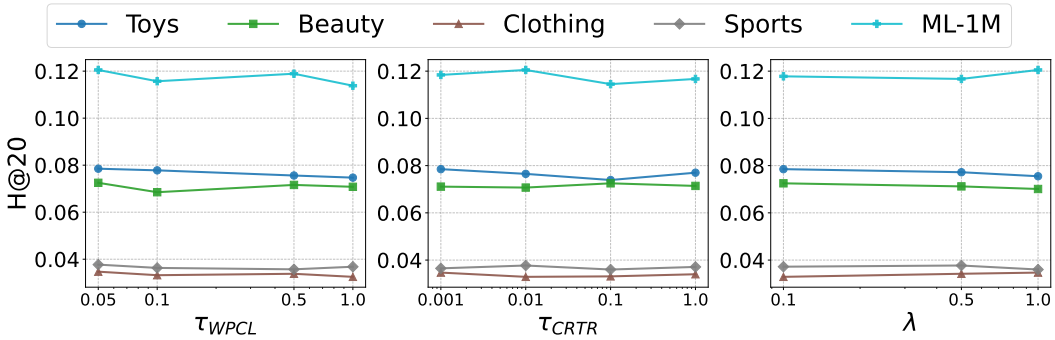

Figure 10: Impact of hyperparameters τ_{WPCL}, τ_{CRTR}, and λ.

F Limitations

We acknowledge two main limitations of the current framework. First, although we study multimodal recommendation, our setting is limited to vision–language modalities, i.e., images and texts. Extending the framework to additional modalities such as audio or video remains an important direction for future work. Second, directly encoding image sequences incurs high training cost due to the large number of visual tokens. Although our proposed losses preserve the same asymptotic complexity as standard SFT (Appendix E.9) and our few-shot variant matches the strongest baseline using only 5–6% of the training data (Section 5.8), the visual token cost from the VLM backbone itself remains a practical concern. Recent advances in token condensation or pruning [49, 50, 55, 71, 5, 10, 46, 29] suggest promising directions for reducing this cost, and adapting such techniques to recommendation-specific settings is a meaningful avenue for future research.

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: Please see Abstract and Section 1.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Please see Appendix F.

   Guidelines:

   - The answer [N/A] 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: Please see Appendix C.

Guidelines:

- The answer [N/A] 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: Please see Section 4 and Appendix D.2.

   Guidelines:

   - The answer [N/A] 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: We provide the code and data as a .zip file in the Supplementary Material. An anonymized code repository is also available in Appendix D.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Section 5.1 and Appendix D.

   Guidelines:

   - The answer [N/A] 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: Please see Table 3 and Table 10.

   Guidelines:

   - The answer [N/A] 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: Please see Table 6 and Appendix D.2.

   Guidelines:

   - The answer [N/A] 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: Yes, we confirm.

   Guidelines:

   - The answer [N/A] 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 conduct experiments on real-world recommendation scenarios that directly affect users' decision-making.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] 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: The paper cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: We include an anonymized zip file containing the source code with the submission. An anonymized code repository is also publicly available as noted in Appendix D.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Please see Section 2 and Section 3 and Section 5.1.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.