PROTOTYPE-GUIDED RECALIBRATION AND INTER-MODAL SIGNAL RESHAPING FOR MULTI-MODAL SEQUENTIAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

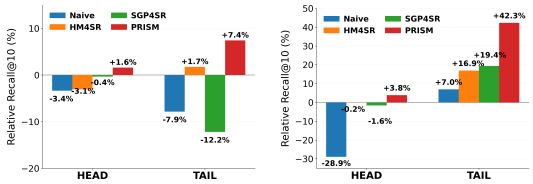
ABSTRACT

Multi-modal sequential recommendation incorporates textual and visual features to alleviate interaction sparsity. However, while uncalibrated multi-modal integration may improve tail-item recommendations, it often degrades head-item performance by distorting their robust collaborative representations. To address this limitation, we propose **P**rototype-Guided **R**ecalibration and **I**nter-Modal **S**ignal Reshaping for **M**ulti-Modal Sequential Recommendation (**PRISM**), which recalibrates multi-modal features before ID-based sequence modeling. Cross-Modal Semantic Reshaping separates shared semantics from modality-specific cues, aligning the former while using prototypes to preserve recurring patterns in the latter. Frequency-Gated Sequence Recalibration jointly analyzes textual and visual amplitude spectra to adaptively regulate each modality's sequence variations without uniformly smoothing informative dynamics. Cross-attention then integrates the recalibrated sequences with the ID-based sequence for prediction. Experiments on three datasets show that PRISM outperforms representative baselines and improves performance on both head and tail items.

1 INTRODUCTION

Sequential recommendation predicts a user's next interaction from historical behavior Boka et al. (2024). To alleviate interaction sparsity, multi-modal sequential recommendation incorporates textual and visual item features He & McAuley (2016); Li et al. (2023) into sequence modeling Liu et al. (2021); Xie et al. (2022b).

Existing multi-modal models have demonstrated promising performance in this field. Early studies integrate multi-modal features into sequence modeling to enrich item representations Qiu et al. (2021), while recent methods adopt adaptive interaction mechanisms Fu et al. (2024); Hu et al. (2023) or leverage explicit cross-modal alignment Liu et al. (2024b) to better capture user preferences. More recently, HM4SR Zhang et al. (2025) jointly models collaborative and multi-modal representations through a time-aware MoE architecture, whereas SGP4SR Li et al. (2026a) converts modality-specific information into item relations and interest representations to guide ID-based preference learning. Overall, existing methods predominantly focus on designing sophisticated mechanisms to couple pre-extracted multi-modal features with collaborative signals. However, these methods typically apply uncalibrated coupling across all items, neglecting that raw multi-modal features can inadvertently compromise the reliable collaborative representations of popular items.



(a) *Dogs*. (b) *Game*.

Figure 1: Relative Recall@10 compared to the ID-only baseline across item interaction frequencies. Here, ID-only (**w/o MM**) relies solely on collaborative signals, while Naive Fusion (**w/o CAL**) directly adds raw multimodal features to the collaborative representation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To investigate this issue, we conduct an empirical study on *Dogs* and *Game*, two real-world datasets from the Pixel benchmark Cheng et al. (2023). Specifically, we group items by training frequency into head (top 20%) and tail (bottom 50%). Figure 1 displays the relative Recall@10 changes for a direct-fusion variant, HM4SR, SGP4SR, and PRISM against their ID-only backbones. On *Dogs* and *Game*, existing baselines consistently suffer performance degradation on head items (frequently compromising head accuracy to boost tail performance, or even degrading both), whereas PRISM improves both groups simultaneously. This performance compromise exposes a critical flaw: existing methods simply inject multi-modal information to compensate for limited item interactions, failing to protect the well-learned collaborative representations of head items.

Overcoming this compromise poses two critical challenges: **(i) Semantic heterogeneity.** Raw multi-modal features inherently entwine modality-specific noise with intrinsic item semantics. Directly integrating these unrefined representations introduces semantic discrepancies that pollute the robust collaborative space of head items, requiring a selective distillation that filters noise without discarding the complementary cues vital for tail items. **(ii) Sequential volatility.** Multi-modal interaction histories often exhibit abrupt temporal fluctuations. Unrestricted integration disrupts the stable item transition patterns crucial for head items, whereas uniform smoothing erases the fine-grained dynamics essential for tail items. This demands adaptively regulating sequence variations while preserving informative preference dynamics.

To address these challenges, we propose **P**rototype-Guided **R**ecalibration and **I**nter-Modal **S**ignal Reshaping for **M**ulti-Modal Sequential Recommendation (**PRISM**). PRISM recalibrates multi-modal features at both item and sequence levels prior to ID-based sequence modeling. Specifically, to tackle the first challenge, we develop a *Cross-Modal Semantic Reshaping* module that disentangles shared semantics from modality-specific cues. This module aligns shared distributions, coordinates recurring modality-specific patterns through prototype-level optimal transport, and integrates the resulting components into item-level multi-modal representations.

To address the second challenge, we introduce a *Frequency-Gated Sequence Recalibration* module in the spectral domain. Guided by joint cross-modality amplitude spectra, it adaptively amplifies informative spectral components for each modality, dynamically emphasizing core preference patterns without suppressing original signal representations. Finally, the recalibrated sequences are fused with the core ID sequence via cross-attention for recommendation.

Experiments on three datasets show that PRISM outperforms representative baselines and improves both head and tail items. Ablation studies further validate its key components. Our main contributions are summarized as follows:

- We uncover a critical limitation in existing multi-modal sequential recommendation methods: directly injecting uncalibrated multi-modal features pollutes reliable collaborative representations, inadvertently degrading head-item performance or even undermining overall accuracy.
- We propose PRISM to address semantic heterogeneity and sequential volatility through Cross-Modal Semantic Reshaping and Frequency-Gated Sequence Recalibration before ID-based sequence modeling.
- Experiments on three datasets show that PRISM outperforms representative baselines and improves performance on both head and tail items. Ablation studies further validate its key components.

2 Related Work

Sequential recommendation models evolving user preferences from ordered interactions. Transformer-based models such as SASRec Kang & McAuley (2018) and BERT4Rec Sun et al. (2019) have become standard backbones, with subsequent studies extending sequential modeling through graph structures, contrastive learning, temporal dynamics, and frequency-domain representations. Despite these advances, ID-based models remain dependent on sufficient behavioral supervision and can struggle with sparsely interacted items. Multi-modal sequential recommendation alleviates interaction sparsity by incorporating textual and visual semantics, with recent methods exploring multi-modal representation learning, cross-modal alignment, and modality-aware preference modeling Yu et al. (2023); Liu et al. (2024b); Li et al. (2026a). However, directly coupling

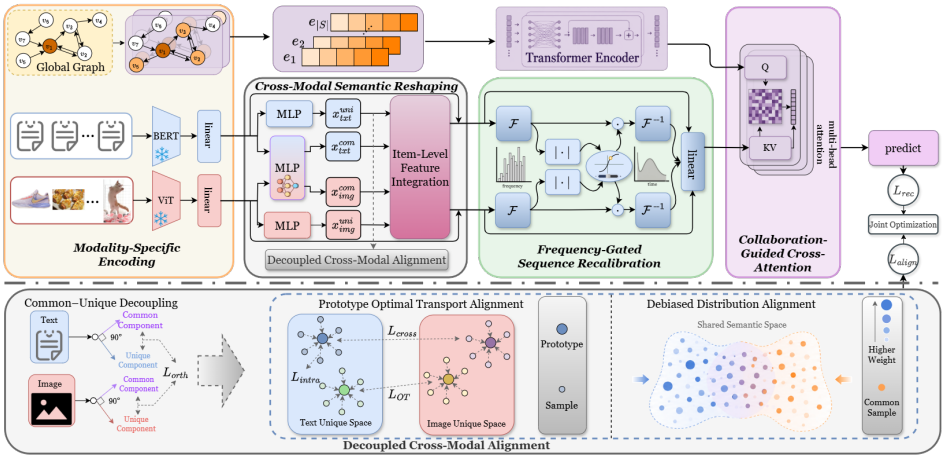


Figure 2: The overall architecture of the proposed PRISM

auxiliary semantics with collaborative representations can benefit tail items while perturbing reliable ID signals learned for frequently interacted items. PRISM therefore recalibrates multi-modal representations at both item and sequence levels before ID-guided fusion, allowing semantic evidence to complement sparse interactions while retaining collaborative signals as the prediction anchor. Extended related work and comparisons are provided in Appendix A.1.

3 Problem Definition

Multi-modal sequential recommendation predicts a user's next interaction from historical behavior and multi-modal item features. Let $\mathcal{U}$ and $\mathcal{V}$ denote the user and item sets, respectively. Each user is associated with a chronologically ordered sequence $\mathcal{S} = [v_1, v_2, \ldots, v_L]$, where $v_i \in \mathcal{V}$ and L is the sequence length. In addition to the collaborative signal derived from interactions, each item has textual and visual features. Given $\mathcal{S}$, the task is to estimate the conditional distribution of the next item and rank candidate items accordingly:

$$v^{\star}_{L+1} = \arg\max_{v \in \mathcal{V}} \; P(v_{L+1} = v \mid \mathcal{S}) \,. \tag{1}$$

We use B, L, and d for the batch size, sequence length, and hidden dimension, respectively. Let $\mathcal{M} = \{\mathrm{txt}, \mathrm{img}\}$ be the modality set and $M = |\mathcal{M}|$. Throughout, $m, m' \in \mathcal{M}$ and $m' \neq m$ denote two distinct modalities, while $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator.

4 Method

To enhance tail-item recommendation while preserving robust head-item ID representations, we propose **PRISM** (Figure 2). We first employ a **Modality-Specific Encoding** layer to project ID, textual, and visual features into a unified latent space. Within this space, a **Cross-Modal Semantic Reshaping** module disentangles and coordinates shared and modality-specific semantics. These item representations are then sequenced and processed via **Frequency-Gated Sequence Recalibration**, which adaptively regulates multi-modal variations using cross-modality spectral gates. Finally, a **Collaboration-Guided Cross-Attention and Prediction** module leverages the core ID sequence as a reference to selectively integrate the recalibrated multi-modal representations for next-item prediction.

4.1 Modality-Specific Encoding

4.1.1 Graph-Based Item Representation Learning.

We construct an undirected global item transition graph from all training sequences. Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E}^{\mathrm{tr}})$ denote the item graph, where $\mathcal{V}$ is the set of all unique items, and $\mathcal{E}^{\mathrm{tr}} = \{(v_i, v_{i+1}) \mid \exists u, v_i, v_{i+1} \in \mathcal{S}^{\mathrm{tr}}_u\}$ represents the transitions between consecutive items within any training sequence $\mathcal{S}^{\mathrm{tr}}_u$. LightGCN He et al. (2020) performs neighborhood propagation and layer-wise aggregation on $\mathcal{G}$ to learn an item embedding $\mathbf{e}_v \in \mathbb{R}^d$ for each $v \in \mathcal{V}$ from interaction data.

4.1.2 TEXTUAL AND VISUAL ENCODING.

For each item, pretrained BERT and ViT models with frozen parameters extract features from its textual description $\mathcal{T}$ and image $\mathcal{I}$, respectively. Separate MLPs project these features, yielding $\mathbf{x}_{\text{txt}}, \mathbf{x}_{\text{img}} \in \mathbb{R}^d$. However, merely matching dimensions does not inherently guarantee cross-modal semantic alignment. We therefore reshape these multi-modal features prior to sequence modeling.

4.2 CROSS-MODAL SEMANTIC RESHAPING

Raw textual and visual features entangle shared concepts with modality-specific cues and incidental details. Directly incorporating such features may introduce semantic discrepancies that interfere with well-learned ID representations. To reduce this interference without discarding signals useful for tail items, Cross-Modal Semantic Reshaping separates the two components, aligns shared semantics across modalities, and organizes recurring modality-specific patterns before item-level integration.

Common–Unique Decoupling. For each modality $m \in \mathcal{M}$, a projection head shared across modalities extracts common semantics, while a modality-specific head retains unique semantic aspects, yielding $\mathbf{x}_m^{\text{com}}, \mathbf{x}_m^{\text{uni}} \in \mathbb{R}^d$: $\mathbf{x}_m^{\text{com}} = \text{MLP}_{\text{com}}(\mathbf{x}_m)$ and $\mathbf{x}_m^{\text{uni}} = \text{MLP}_{\text{uni}}^m(\mathbf{x}_m)$.

To prevent redundant encoding between the two components, we minimize their within-modality cosine similarity, $\mathcal{L}_{\text{orth}} = \sum_{m\in\mathcal{M}} \left(\frac{\langle \mathbf{x}_m^{\text{uni}}, \mathbf{x}_m^{\text{com}} \rangle}{\|\mathbf{x}_m^{\text{uni}}\|_2 \, \|\mathbf{x}_m^{\text{com}}\|_2} \right)^2$. The resulting components support separate objectives for aligning shared semantics and coordinating modality-specific patterns.

Decoupled Cross-Modal Alignment. Shared semantics and modality-specific features serve distinct roles and require tailored alignment strategies. Shared semantics should maintain cross-modal consistency, whereas modality-specific features should preserve recurring intra-modal patterns. During training, we collect item vectors from B sequences of length L and flatten them into $N = BL$ samples. The shared and modality-specific vectors are denoted by $\boldsymbol{\psi}_n^m, \boldsymbol{\phi}_n^m \in \mathbb{R}^d$, respectively, where $n \in \{1, \ldots, N\}$.

Prototype Optimal Transport Alignment. Instance-wise alignment on modality-specific features is counterproductive: they naturally differ across modalities, forcing item-level agreement erases modal-specific information and causes semantic conflict. Instead of instance-level matching, we organize modality-specific features into K Gaussian prototypes with learnable means $\boldsymbol{\mu}_k^m \in \mathbb{R}^d$ and diagonal covariances $\boldsymbol{\Sigma}_k^m \in \mathbb{R}^{d\times d}$ to capture recurring intra-modal patterns. The assignment probability of $\boldsymbol{\phi}_n^m$ to prototype k is

$$q_n^m(k) = \frac{\mathcal{N}(\boldsymbol{\phi}_n^m; \boldsymbol{\mu}_k^m, \boldsymbol{\Sigma}_k^m)}{\sum_{j=1}^K \mathcal{N}\big(\boldsymbol{\phi}_n^m; \boldsymbol{\mu}_j^m, \boldsymbol{\Sigma}_j^m\big)}. \tag{2}$$

Based on these probabilities, the intra-modal compactness loss $\mathcal{L}_{\text{intra}}$ minimizes the $q_n^m(k)$-weighted squared Mahalanobis distance between batch features and detached prototypes, pulling features toward cluster centers to capture recurring semantic patterns. Next, we measure cross-modal correspondence using the Bures–Wasserstein distance. For a multi-modal index $\mathbf{k} = (k_1, \ldots, k_M)$, the transport cost is

$$C^{m,m'}(k,k') = \|\boldsymbol{\mu}_k^m - \boldsymbol{\mu}_{k'}^{m'}\|_2^2 + \text{Tr}\big(\boldsymbol{\Sigma}_k^m + \boldsymbol{\Sigma}_{k'}^{m'} - 2(\boldsymbol{\Sigma}_k^m \boldsymbol{\Sigma}_{k'}^{m'})^{1/2}\big),\ C(\mathbf{k}) = \sum_{m<m'} C^{m,m'}(k_m, k_{m'}). \tag{3}$$

Let $p_k^m = N^{-1} \sum_n q_n^m(k)$ denote the empirical prototype distribution of modality m, and let $\Pi(p^1, \ldots, p^M)$ contain the transport plans with these marginals. We obtain cross-modal prototype correspondences by solving the entropy-regularized multi-marginal optimal transport problem (noting that in the two-modality setting, this formulation degenerates into standard two-marginal OT)

$$\mathcal{L}_{\text{OT}} = \min_{T\in\Pi(p^1,\ldots,p^M)} \sum_{\mathbf{k}} T(\mathbf{k}) C(\mathbf{k}) + \lambda \sum_{\mathbf{k}} T(\mathbf{k}) \log T(\mathbf{k}), \tag{4}$$

$$\mathcal{L}_{\text{cross}}^{(m,m')} = \frac{1}{N} \sum_{n,k,k'} T^\star_{m,m'}(k,k') q_n^m(k) \left\| \boldsymbol{\phi}_n^m - \text{sg}\left(\boldsymbol{\mu}_{k'}^{m'}\right) \right\|_2^2, \tag{5}$$

where $T^\star_{m,m'}$ is the pairwise marginal of the optimal plan $T^\star$ computed via the Sinkhorn algorithm, and $q_n^m(k)$ is defined in Eq. equation 2. The total cross loss is aggregated over distinct modalities

as $\mathcal{L}_{\text{cross}} = \sum_{m \neq m'} \mathcal{L}_{\text{cross}}^{(m,m')}$. A prototype-diversity regularizer $\mathcal{L}_{\text{div}}$ separates cluster centers by penalizing the mean exponentiated negative squared Euclidean distance between all distinct pairs of ℓ_2-normalized means. Simultaneously, a load-balancing regularizer $\mathcal{L}_{\text{uniform}}$ prevents assignment collapse by minimizing the KL divergence between the empirical prototype distribution p_k^m and a uniform prior. These regularizers prevent prototype and assignment collapse. The complete unique-component objective is $\mathcal{L}_{\text{uni}} = \lambda_{\text{dist}}\mathcal{L}_{\text{intra}} + \lambda_{\text{align}}\left(\mathcal{L}_{\text{OT}} + \mathcal{L}_{\text{cross}}\right) + \lambda_{\text{div}}\left(\mathcal{L}_{\text{div}} + \mathcal{L}_{\text{uniform}}\right)$.

Debiased Distribution Alignment. The common components represent shared semantics and are therefore aligned at the distribution level. We first normalize them to reduce scale and variance discrepancies. Because repeated items can dominate an in-batch distribution, each vector receives a non-differentiable inverse-frequency weight $w_n = c_n^{-1} / \sum_r c_r^{-1}$, where c_n is the occurrence count of its item in the current batch. We then apply weighted maximum mean discrepancy:

$$\mathcal{L}_{\text{MMD}}^{w} = \frac{2}{M(M-1)} \sum_{m<m'} \Big[\sum_{a,b} w_a w_b \kappa(\boldsymbol{\psi}_a^m, \boldsymbol{\psi}_b^m) + \sum_{r,s} w_r w_s \kappa(\boldsymbol{\psi}_r^{m'}, \boldsymbol{\psi}_s^{m'}) - 2 \sum_{a,r} w_a w_r \kappa(\boldsymbol{\psi}_a^m, \boldsymbol{\psi}_r^{m'}) \Big]. \tag{6}$$

Here, κ is a Gaussian kernel. We complement kernel matching with $\mathcal{L}_{\text{sem}}$, which aligns statistical moments by computing the sum of Euclidean distances between the batch-wise empirical mean, variance, and skewness across modalities. The common-component objective is $\mathcal{L}_{\text{com}} = \mathcal{L}_{\text{MMD}}^{w} + \mathcal{L}_{\text{sem}}$. These constraints align shared semantics while reducing bias toward items repeated within the batch.

Item-Level Feature Integration. To construct the item representation, we integrate the decoupled common and unique components. Specifically, an element-wise gate $\mathbf{g}_m$ is derived from the counterpart modality to balance cross-modal complementarity and intra-modal specificity:

$$\mathbf{g}_m = \sigma\left(\mathbf{x}_{m'}^{\text{uni}} \mathbf{W}_g^m + \mathbf{b}_g^m\right), \quad \widetilde{\mathbf{x}}_m^{\text{uni}} = \mathbf{g}_m \odot \mathbf{x}_m^{\text{uni}} + (1 - \mathbf{g}_m) \odot \mathbf{x}_{m'}^{\text{uni}}. \tag{7}$$

Next, we average the common components and concatenate them with the gated unique feature $\widetilde{\mathbf{x}}_m^{\text{uni}}$. A linear projection combined with a residual connection yields the final representation $\mathbf{x}_m^{\text{item}} \in \mathbb{R}^d$:

$$\overline{\mathbf{x}}^{\text{com}} = \tfrac{1}{2}\left(\mathbf{x}_{\text{txt}}^{\text{com}} + \mathbf{x}_{\text{img}}^{\text{com}}\right), \quad \mathbf{x}_m^{\text{item}} = \left[\overline{\mathbf{x}}^{\text{com}} \parallel \widetilde{\mathbf{x}}_m^{\text{uni}}\right] \mathbf{W}_{\text{fuse}}^m + \mathbf{b}_{\text{fuse}}^m + \text{LN}(\mathbf{x}_m), \tag{8}$$

where $\mathbf{W}_g^m \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_{\text{fuse}}^m \in \mathbb{R}^{2d \times d}$ are projection matrices. The resulting multimodal item representations are then fed into the subsequent sequence modeling layer. Due to space constraints, explicit formulations and architectural specifics are provided in Appendix A.2.1.

4.3 Frequency-Gated Sequence Recalibration

In sequential recommendation, incidental user interactions often introduce abrupt fluctuations in multi-modal feature sequences, deviating from a user's persistent preferences. Directly incorporating these volatile signals disrupts ID-based sequential modeling, whereas uniform smoothing risks discarding informative preference dynamics. To reconcile this, we propose Frequency-Gated Sequence Recalibration to learn dynamic, modality-specific weights across the frequency spectrum. By conditioning the gating mechanism on the joint textual and visual spectra, the model adaptively recalibrates spectral contributions while preserving essential sequential patterns.

For a given modality $m \in \{\text{txt}, \text{img}\}$, we formulate the sequence of item-level vectors (incorporating positional embeddings and dropout) as a matrix $\mathbf{X}_m^{\text{item}} \in \mathbb{R}^{L \times d}$. We then perform a discrete Fourier transform along the sequence dimension, obtaining $\mathbf{Z}_m = \mathcal{F}\left(\mathbf{X}_m^{\text{item}}\right)$ and $\mathbf{A}_m = |\mathbf{Z}_m|$, where $\mathbf{Z}_m \in \mathbb{C}^{L \times d}$ represents the complex frequency coefficients and $\mathbf{A}_m \in \mathbb{R}^{L \times d}$ denotes their corresponding amplitude spectra. To adaptively determine the contribution of different frequency components, we design a gating network parameterized by $\mathbf{W}^m \in \mathbb{R}^{2d \times d}$ and $\mathbf{b}^m \in \mathbb{R}^d$, conditioned on the joint amplitude spectra of both modalities: $\mathbf{G}_m = \sigma\left(\left[\mathbf{A}_m \parallel \mathbf{A}_{m'}\right] \mathbf{W}^m + \mathbf{b}^m\right)$.

The real-valued gate matrix $\mathbf{G}_m \in \mathbb{R}^{L \times d}$ modulates the amplitude of each frequency coefficient while preserving its phase. Unlike static low-pass filters that attenuate high-frequency components, this data-driven gating performs frequency-specific residual amplification. Instead of explicitly suppressing spectral coefficients, it adaptively amplifies informative components while preserving the baseline signal, thereby reallocating relative spectral energy across the frequency domain. The modulated spectra are then transformed back to the time domain via the inverse Fourier transform alongside a residual connection:

$$\mathbf{X}_m^{\text{flt}} = \text{Re}\left(\mathcal{F}^{-1}(\mathbf{Z}_m \odot \mathbf{G}_m)\right) + \mathbf{X}_m^{\text{item}}. \tag{9}$$

By incorporating the residual connection, the gated branch provides an additive, frequency-dependent spectral boost without suppressing any original coefficients, thereby achieving true residual spectral recalibration. Further mathematical details and padding efficiency analyses are provided in Appendix A.3 and A.4.

The resulting frequency-recalibrated sequence $\mathbf{X}_m^{\text{flt}} \in \mathbb{R}^{L\times d}$ retains the core item-level semantics. Subsequently, the textual and visual recalibrated sequences are concatenated and mapped into a unified multi-modal representation $\mathbf{X}_{\text{fus}}^{\text{seq}} \in \mathbb{R}^{L\times d}$:

$$\mathbf{X}_{\text{fus}}^{\text{seq}} = \text{LayerNorm}\big(\big[\mathbf{X}_{\text{txt}}^{\text{flt}} \parallel \mathbf{X}_{\text{img}}^{\text{flt}}\big]\mathbf{W}_{\text{seq}} + \mathbf{b}_{\text{seq}}\big)\,. \tag{10}$$

While utilizing inter-modality cross-attention serves as an intuitive alternative for deriving $\mathbf{X}_m^{\text{flt}}$, empirical evaluations indicate that its performance is highly dataset-dependent and lacks stable generalization across diverse scenarios.

4.4 Collaboration-Guided Cross-Attention and Prediction

For an input sequence $(v_1, \ldots, v_L)$, we stack its item embeddings $\{\mathbf{e}_{v_l}\}_{l=1}^{L}$ into the ID sequence matrix $\mathbf{X}_{\text{id}}^{\text{item}} \in \mathbb{R}^{L\times d}$. The preceding module provides the multi-modal sequence matrix $\mathbf{X}_{\text{fus}}^{\text{seq}} \in \mathbb{R}^{L\times d}$. We encode the ID sequence with a Transformer and use it as the query to retrieve relevant multi-modal information:

$$\mathbf{X}_{\text{id}}^{\text{seq}} = \text{Enc}_{\text{id}}\big(\mathbf{X}_{\text{id}}^{\text{item}}\big)\,, \quad \mathbf{H} = \text{Enc}_{\text{mix}}(\mathbf{X}_{\text{id}}^{\text{seq}}, \mathbf{X}_{\text{fus}}^{\text{seq}})\,. \tag{11}$$

Here, $\mathbf{X}_{\text{id}}^{\text{seq}}, \mathbf{H} \in \mathbb{R}^{L\times d}$. Enc_{mix} is a Transformer cross-attention encoder in which $\mathbf{X}_{\text{id}}^{\text{seq}}$ provides the query and $\mathbf{X}_{\text{fus}}^{\text{seq}}$ provides the key and value. Thus, the sequence representation learned from item IDs and user interactions determines which recalibrated multi-modal features are incorporated at each position.

We take the final-position state of the sequence matrix $\mathbf{H} \in \mathbb{R}^{L\times d}$ as the user representation, denoted as $\mathbf{u}_{\text{fus}} \in \mathbb{R}^d$. This representation is then scored against the embedding of a candidate item v as $\mathbf{u}_{\text{fus}} = \mathbf{H}_{L,:}$ and $\widehat{y}_v = \langle \mathbf{u}_{\text{fus}}, \mathbf{e}_v \rangle$, where $\mathbf{H}_{L,:}$ denotes the L-th row vector of $\mathbf{H}$ (: denotes all dimensions). Candidate scoring therefore remains based on item embeddings learned from user interactions. The recommendation objective is jointly optimized with the semantic-reshaping regularizers:

$$\mathcal{L}_{\text{rec}} = -\log \frac{\exp(\langle \mathbf{u}_{\text{fus}}, \mathbf{e}_{v^+} \rangle)}{\sum_{v\in\mathcal{V}} \exp(\langle \mathbf{u}_{\text{fus}}, \mathbf{e}_v \rangle)}, \quad \mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{orth}} + \beta(\mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{com}})\,. \tag{12}$$

Here, α controls common–unique separation, while β balances recommendation with component-specific distribution alignment. The complete objective recalibrates multi-modal features while retaining the ID-based sequence as the reference for next-item prediction.

5 Experiment

We comprehensively evaluate PRISM below to validate our claims. Due to space constraints, additional studies (e.g., efficiency, hyper-parameter sensitivity, scalability, cross-modal visualizations, and case studies) are presented in Appendices A.5, A.6, A.8, A.9, and A.10.

5.1 Experimental Setup

Datasets. We evaluate our method on three representative category-specific subsets of the Pixel dataset Cheng et al. (2023): *Fitness*, *Game*, and *Dogs*. Dataset statistics after standard 5-core filtering Hou et al. (2022); Lin et al. (2024); Wang et al. (2024) are summarized in Table 1. Due to space limits, full dataset selection rationales and modality preprocessing protocols are deferred to Appendix A.11.

Table 1: Statistics of processed datasets.

Dataset	Users	Items	Interactions	Sparsity	Avg. Seq. Length
Fitness	8530	1316	59996	99.47%	7.03
Game	24787	7809	207114	99.89%	8.36
Dogs	27332	2078	225524	99.60%	8.25

Table 2: Overall performance comparison. Bold and underlined scores denote the best and second-best results. R@K and N@K denote Recall@K and NDCG@K. All multimodal models share identical image and text encoders. Improv. tracks PRISM's relative gain over the best baseline (statistically significant via paired t-test with p-value < 0.05).

Method	Fitness				Game				Dogs			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
SASRec	0.1377	0.2108	0.0736	0.0935	0.0581	0.0938	0.0272	0.0362	0.1736	0.2396	0.0968	0.1148
BERT4Rec	0.1196	0.1851	0.0672	0.0846	0.0378	0.0631	0.0189	0.0254	0.1689	0.2331	0.0972	0.1147
BSARec	0.1366	0.1993	0.0771	0.0945	0.0579	0.0964	0.0298	0.0395	0.1653	0.2342	0.0936	0.1126
SASRecF	0.1469	0.2138	0.0839	0.1023	0.0602	0.0946	0.0317	0.0405	0.1780	0.2422	0.1080	0.1249
NOVA	0.1193	0.1793	0.0639	0.0803	0.0391	0.0660	0.0197	0.0266	0.1677	0.2325	0.0967	0.1144
MISSRec	0.1520	0.2198	0.0856	0.1044	0.0291	0.0481	0.0140	0.0187	0.1685	0.2383	0.0945	0.1135
ATHWE	0.1508	0.2195	0.0875	0.1068	0.0618	0.0964	0.0327	0.0416	0.1797	0.2482	0.1075	0.1262
HM4SR	0.1498	0.2173	0.0841	0.1021	0.0582	0.0925	0.0307	0.0394	0.1798	0.2490	0.1062	0.1253
SGP4SR	0.1408	0.2053	0.0799	0.0981	0.0535	0.0877	0.0264	0.0351	0.1723	0.2375	0.1002	0.1179
PRISM (ours)	**0.1579**	**0.2255**	**0.0901**	**0.1091**	**0.0686**	**0.1077**	**0.0360**	**0.0459**	**0.1903**	**0.2577**	**0.1165**	**0.1349**
Improv.	+3.88%	+2.59%	+2.97%	+2.15%	+11.00%	+11.72%	+10.09%	+10.34%	+5.84%	+3.49%	+7.87%	+6.89%

Table 3: The effectiveness of different variants of PRISM.

Variants	Fitness		Game		Dogs	
	R@10	N@10	R@10	N@10	R@10	N@10
w/o MM	0.1330	0.0692	0.0571	0.0258	0.1833	0.1016
w/o CAL	0.1254	0.0760	0.0431	0.0219	0.1708	0.1036
w/o SEP	0.1477	0.0898	0.0628	0.0340	0.1776	0.1089
w/o ALN	0.1483	0.0802	0.0628	0.0341	0.1691	0.1024
w/o COM	0.1540	0.0883	0.0652	0.0346	0.1859	0.1124
w/o FG	0.1538	0.0880	0.0644	0.0340	0.1823	0.1096
PRISM	**0.1580**	**0.0901**	**0.0686**	**0.0360**	**0.1903**	**0.1165**

Compared Methods. To verify the effectiveness of PRISM, we select representative baselines from two categories: (1) ID-based sequential recommendation methods, including SASRec Kang & McAuley (2018), BERT4Rec Sun et al. (2019), and BSARec Shin et al. (2024); and (2) multi-modal sequential recommendation methods, including SASRecF Oh et al. (2024), NOVA Liu et al. (2021), MISSRec Wang et al. (2023), ATHWE Liu et al. (2026a), HM4SR Zhang et al. (2025), and SGP4SR Li et al. (2026a). For fairness, all multimodal baselines share the same BERT/ViT features and are tuned under identical experimental protocols.

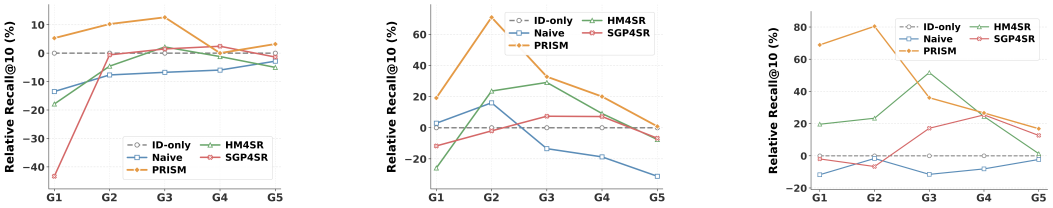
Evaluation Protocol. We chronologically partition interactions into training (80%), validation (10%), and test (10%) sets by global timestamps. Across all splits, user interaction sequences are uniformly expanded into prefix sub-sequences for sequential next-item prediction. Under the full-set ranking protocol without negative sampling, we evaluate model performance using Recall@K and NDCG@K ($K \in \{10, 20\}$).

Implementation Details. We implement PRISM in PyTorch with a batch size of 256 and hidden dimension $d = 128$. Behavior sequence length is capped at 300; during training, we adopt dynamic batch-level padding aligned to the maximum length per batch to minimize padding noise on the Frequency-Gated Sequence Recalibration module. Both Transformer encoders (Enc_{id} and Enc_{mix}) consist of 1 layer with 2 heads. Hyperparameters are determined via grid search: LightGCN propagation layers $L_{gcn} \in \{1, 2, 3, 4\}$, Gaussian prototypes $K \in \{5, 10, 20, 30\}$ per modality, main loss weights $\alpha \in \{0.001, 0.01, 0.1, 0.5, 1.0\}$ and $\beta \in \{0.01, 0.1, 0.5, 1.0\}$, and auxiliary weights $\lambda_{dist}, \lambda_{align}, \lambda_{div} \in \{0.1, 0.5, 1.0\}$. Models are optimized using Adam with a learning rate of 0.001 on an NVIDIA GeForce RTX 4070 Ti GPU. All experiments are averaged over 5 random seeds.

5.2 Overall Comparison

Table 2 presents overall performance across three datasets. PRISM consistently outperforms all baselines with a 6.58% average gain (+10.79% on average for Game), confirming its efficacy in item retrieval and top-ranked recommendation.

Notably, ID-only baselines (e.g., SASRec, BSARec) remain highly competitive on PIXEL datasets. Rich interaction signals in social media allow them to learn reliable sequential preferences without



(a) *Dogs*. (b) *Game*. (c) *Fitness*.

Figure 3: Relative Recall@10 gains over ID-only backbone across frequency groups (G_1–G_5).

modal features, explaining why naive multimodal fusion fails to consistently outperform ID-only methods.

Among multimodal approaches, while SASRecF and ATHWE perform competitively, methods like NOVA and MISSRec lag behind ID-only baselines due to semantic discrepancies and cross-modal alignment challenges in social-media content. In contrast, PRISM consistently outperforms all multimodal baselines, indicating that prototype-guided semantic reshaping and inter-modal spectral recalibration better exploit complementary modal signals while reducing their interference with collaborative sequential modeling.

5.3 Ablation Studies

To evaluate the contribution of each module, we construct six PRISM variants: (1) **w/o MM** removes multimodal branches, retaining only collaborative GCN and ID encoders; (2) **w/o CAL** discards cross-modal semantic reshaping and frequency-gated recalibration, directly summing raw textual and visual features; (3) **w/o SEP** removes the orthogonality constraint for common–unique feature separation; (4) **w/o ALN** omits both prototype OT and debiased alignment losses, whereas (5) **w/o COM** removes only the latter; (6) **w/o FG** replaces frequency-gated sequence recalibration with an MLP-based transformation for sequence fusion. As reported in Table 3, we observe the following:

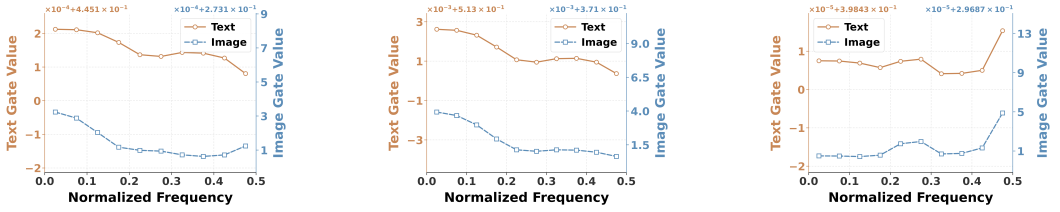
First, **w/o MM** and **w/o CAL** consistently underperform PRISM, confirming the value of multimodal cues and the necessity of cross-modal semantic reshaping before integrating them into collaborative sequential modeling. Second, the degradation of **w/o SEP** verifies that common–unique representation disentanglement effectively reduces feature redundancy. Both **w/o ALN** and **w/o COM** yield consistent performance drops, demonstrating the contribution of distribution alignment, while the larger degradation of **w/o ALN** further shows the complementary roles of prototype-level and debiased alignment in preserving top-K ranking quality. Finally, the degradation of **w/o FG** indicates that frequency-aware sequence recalibration is more effective than a generic MLP transformation for sequence fusion, providing adaptive spectral modulation while preserving the original sequence representation through residual enhancement.

5.4 Performance across Varying Interaction Frequencies

To evaluate robustness across interaction densities, we partition items into quintiles (G_1–G_5, from sparsest to densest) based on training interaction frequencies. Figure 3 reports the relative Recall@10 gains of PRISM against Naive Fusion and state-of-the-art baselines (HM4SR, SGP4SR) over the ID-only backbone. The results highlight critical disparities in multi-modal feature integration. Direct feature injection (Naive Fusion) consistently degrades performance below the ID-only baseline, confirming that uncalibrated multi-modal noise distorts collaborative signals. Existing baselines (HM4SR, SGP4SR) display volatile performance across frequency regimes, often failing at extreme densities. In contrast, PRISM delivers substantial gains in low-frequency regimes (G_1–G_2) while sustaining competitive or superior accuracy across denser items (G_3–G_5). This demonstrates that PRISM's gains stem fundamentally from pre-fusion feature recalibration rather than simple auxiliary feature addition.

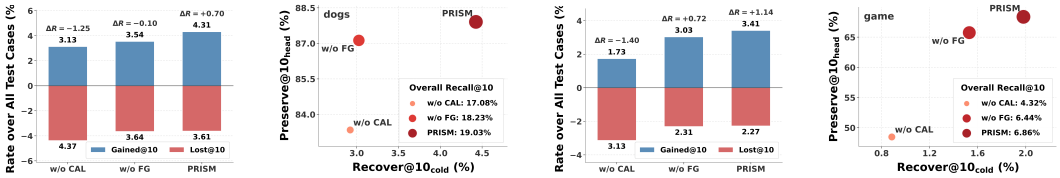
5.5 Modality-Specific Spectral Gate Analysis

To examine how Frequency-Gated Sequence Recalibration adaptively adjusts multi-modal dynamics, we compute the dataset-level average gate $\bar{g}_m(k) = \frac{1}{Nd}\sum_{i=1}^{N}\sum_{j=1}^{d}\mathbf{G}_m[i,k,j]$ across ten equal-width frequency bins (N: sequence count, d: hidden dimension).



(a) *Dogs*. (b) *Fitness*. (c) *Game*.

Figure 4: Frequency-gate curves across spectral bins. Left/right y-axes: textual/visual gate values.



(a) Gained/lost hits. (b) Recovery & pres. (c) Gained/lost hits. (d) Recovery & pres.

Figure 5: Hit preservation and recovery on *Dogs* (a,b) and *Game* (c,d). (a,c) Gained vs. lost Recall@10; (b,d) low-frequency recovery vs. head-item preservation (bubble size: overall Recall@10).

As shown in Figure 4, the learned gates exhibit clear modality asymmetry and domain adaptivity across all three datasets. Textual gates consistently dominate visual ones, confirming modality-specific recalibration. Furthermore, the gating networks yield subtle yet non-uniform band variations tailored to domain semantics: on *Fitness* and *Dogs*, low-frequency components are slightly emphasized to capture slower-varying sequential dynamics, whereas on *Game*, high-frequency gates display a modest rebound to accommodate faster position-level transitions. These distinct trends demonstrate that the proposed module achieves flexible, data-driven spectral recalibration rather than static or uniform frequency scaling.

5.6 Hit Preservation and Recovery Analysis

To examine whether multimodal features rectify ID backbone errors or corrupt correct predictions, we decompose net Recall change ($\Delta R@K$) into *gained hits* (ID-missed targets recovered) and *lost hits* (ID-correct targets corrupted by noise). We further evaluate item-group dynamics via *low-frequency item recovery rate* ($\mathrm{Recover}_{\mathrm{cold}}$, proportion of rescued ID-missed low-frequency targets) and *head-item preservation rate* ($\mathrm{Preserve}_{\mathrm{head}}$, proportion of retained correct ID head predictions).

As shown in Figures 5(a, c), raw multimodal features (**w/o CAL**) cause severe ranking disruption (lost hits $>$ gained hits, $\Delta R@10 < 0$). Semantic calibration (**w/o FG**) substantially mitigates this disruption, achieving net positive gains on *Game* while remaining slightly below the ID baseline on *Dogs*, whereas PRISM consistently yields maximum gained hits and minimum lost hits across both datasets. Figures 5(b, d) further illustrate the trade-off between low-frequency item recovery and head preservation. Occupying the top-right quadrant, PRISM achieves superior low-frequency recovery alongside high head-item hit retention. This confirms that semantic calibration resolves cross-modal alignment while frequency-gated recalibration adaptively amplifies informative spectral dynamics, successfully balancing tail-item recovery with strong head-signal retention. Additional results on the *Fitness* dataset appear in Appendix A.7.

6 Conclusion

In this paper, we proposed PRISM, a multi-modal sequential recommendation framework that recalibrates pre-trained textual and visual features before collaborative integration. Cross-Modal Semantic Reshaping decouples shared semantics from modality-specific cues through distribution matching and prototype-guided optimal transport, while Frequency-Gated Sequence Recalibration adaptively adjusts multi-modal variations across interaction sequences. The refined representations are selectively incorporated through Collaboration-Guided Cross-Attention. Experiments on three benchmark datasets show consistent improvements over representative baselines, with robust gains for both head and tail items across different interaction sparsity levels.

ACKNOWLEDGMENTS

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation and editing of research code, as well as with language editing and improving the readability of the manuscript. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was manually reviewed and revised to ensure technical correctness and consistency with the authors' intended claims. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work. The main paper describes the proposed model architecture, learning objectives, experimental protocol, datasets, evaluation metrics, and implementation settings. Additional implementation details, hyperparameter settings, analyses, and supplementary experimental results are provided in the appendix. Together, these materials are intended to provide the information necessary to reproduce the reported experiments and results.

REFERENCES

Tesfaye Fenta Boka, Zhendong Niu, and Rama Bastola Neupane. A survey of sequential recommendation systems: Techniques, evaluation, and future directions. *Information Systems*, 125:102427, 2024.

Yu Cheng, Yunzhu Pan, Jiaqi Zhang, Yongxin Ni, Aixin Sun, and Fajie Yuan. An image dataset for benchmarking recommender systems with raw pixels. *arXiv preprint arXiv:2309.06789*, 2023.

Haoyan Fu, Zhida Qin, Shixiao Yang, Haoyao Zhang, Bin Lu, Shuang Li, Tianyu Huang, and John CS Lui. Time matters: Enhancing sequential recommendations with time-guided graph neural odes. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 637–648, 2025.

Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M Jose. Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 687–697, 2024.

Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

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

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*, pp. 585–593, 2022.

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*, pp. 843–853, 2023.

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

Changhong Li, Zhiqiang Guo, Guohui Li, Zhong Yang, and Chuhang Hong. Sgp4sr: Seperated-modality guided user perference learning for multimodal sequential reconmmendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15054–15062, 2026a.

Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1258–1267, 2023.

Zilong Li, Jia Zhu, Chenglei Huang, Zhangze Chen, Hanghui Guo, Guoqing Ma, and Jianxia Ling. Capturing dynamic user interests under modality imbalance for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15198–15206, 2026b.

Xiaolin Lin, Jinwei Luo, Junwei Pan, Weike Pan, Zhong Ming, Xun Liu, Shudong Huang, and Jie Jiang. Multi-sequence attentive user representation learning for side-information integrated sequential recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 414–423, 2024.

Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4249–4256, 2021.

Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. *Advances in Neural Information Processing Systems*, 37:26701–26727, 2024a.

Shiyu Liu, Jinzhao Su, Chang-Dong Wang, Shengli Sun, Chen Lin, and Zhenhua Huang. Adaptive temporal expert routing with hierarchical wavelet enhancement for multi-modal sequential recommendation. *ACM Trans. Inf. Syst.*, 44(4), May 2026a. ISSN 1046-8188. doi: 10.1145/3801160. URL https://doi.org/10.1145/3801160.

Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. Alignrec: Aligning and training in multimodal recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1503–1512, 2024b.

Zemu Liu, Zhida Qin, Pengzhan Zhou, Tianyu Huang, and Gangyi Ding. Bridging time and domains: A time-aware framework for cross-domain sequential recommendation. In *Proceedings of the ACM Web Conference 2026*, pp. 6091–6102, 2026b.

Hyungtaik Oh, Wonkeun Jo, and Dongil Kim. Attention-based sequential recommendation system using multimodal data. *arXiv preprint arXiv:2405.17959*, 2024.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

Zhida Qin, Zemu Liu, Haoyan Fu, Chong Zhang, Tianyu Huang, Yidong Li, and Gangyi Ding. Bridging behavior and semantics for time-aware cross-domain sequential recommendation. *arXiv preprint arXiv:2605.02369*, 2026.

Ruihong Qiu, Zi Huang, and Hongzhi Yin. Memory augmented multi-instance contrastive predictive coding for sequential recommendation. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 519–528. IEEE, 2021.

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*, pp. 813–823, 2022.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 811–820, 2010.

Yehjin Shin, Jeongwhan Choi, Hyowon Wi, and Noseong Park. An attentive inductive bias for sequential recommendation beyond the self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8984–8992, 2024.

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*, pp. 1441–1450, 2019.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 565–573, 2018.

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*, pp. 6548–6557, 2023.

Maolin Wang, Yutian Xiao, Binhao Wang, Sheng Zhang, Shanshan Ye, Wanyu Wang, Hongzhi Yin, Ruocheng Guo, and Zenglin Xu. Findrec: Stein-guided entropic flow for multi-modal sequential recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3008–3018, 2025.

Shuhan Wang, Bin Shen, Xu Min, Yong He, Xiaolu Zhang, Liang Zhang, Jun Zhou, and Linjian Mo. Aligned side information fusion method for sequential recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 112–120, 2024.

Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM web conference 2023*, pp. 790–800, 2023.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 346–353, 2019.

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)*, pp. 1259–1273. IEEE, 2022a.

Yueqi Xie, Peilin Zhou, and Sunghun Kim. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1611–1621, 2022b.

Huayang Xu, Huanhuan Yuan, Guanfeng Liu, Junhua Fang, Lei Zhao, and Pengpeng Zhao. Wavelet enhanced adaptive frequency filter for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16058–16065, 2026.

Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 6576–6585, 2023.

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*, pp. 3672–3682, 2025.

Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 935–943, 2023.

A Appendix

A.1 Extended Related Work

Sequential recommendation predicts future interactions from chronologically ordered behavioral histories. Early approaches model sequential dependencies through Markov chains Rendle et al. (2010), recurrent networks Hidasi et al. (2015), and convolutional architectures Tang & Wang

(2018). Transformer-based models such as SASRec Kang & McAuley (2018) and BERT4Rec Sun et al. (2019) subsequently established self-attention as a standard mechanism for long-range preference modeling. Later studies extend sequential encoders with graph propagation Wu et al. (2019), contrastive objectives Xie et al. (2022a); Qiu et al. (2022), temporal dynamics Fu et al. (2025), and cross-domain transfer Liu et al. (2026b); Qin et al. (2026). Frequency-domain modeling provides another perspective on sequential dependencies. BSARec Shin et al. (2024) incorporates frequency-aware representations into sequential recommendation, while recent work further explores adaptive spectral regulation through wavelet representations Xu et al. (2026). PRISM uses spectral modeling specifically to recalibrate multi-modal sequences before their integration with collaborative representations, with each modality conditioned on spectral information from the other modality.

Multi-modal recommendation introduces semantic information from text, images, and other content to complement interaction-derived representations. Early studies incorporate visual features into collaborative recommendation He & McAuley (2016), followed by generative multi-modal sequential modeling Qiu et al. (2021), graph-based representation learning Yu et al. (2023); Zhou & Shen (2023), and self-supervised multi-modal learning Wei et al. (2023). Recent approaches explicitly coordinate heterogeneous modalities through cross-modal alignment Liu et al. (2024b); Wang et al. (2025), time-aware mixture architectures Zhang et al. (2025), modality-guided sequential modeling Li et al. (2026a), and dynamic preference modeling under modality imbalance Li et al. (2026b). Related work on long-tailed sequential recommendation also demonstrates the value of semantic knowledge when behavioral supervision is sparse Liu et al. (2024a). These studies primarily improve the extraction, alignment, or integration of auxiliary semantics. PRISM instead focuses on regulating multi-modal information before it enters collaborative sequential modeling, motivated by the different effects that auxiliary semantics can exert on sparsely and frequently interacted items.

DecAlign Qian et al. (2026) provides a closely related perspective on decoupled cross-modal representation learning by separating common and modality-specific semantics and organizing them through distribution alignment and prototype-level optimal transport. PRISM adopts this principle as an item-level component and tailors it to multi-modal sequential recommendation, introducing frequency-debiased distribution matching and moment alignment for shared semantics, together with an OT-guided cross-modal loss that explicitly reinforces prototype correspondences for modality-specific semantics. More importantly, semantic alignment alone does not resolve the central issue studied in PRISM: auxiliary semantics may benefit sparsely interacted items while interfering with reliable collaborative signals. PRISM therefore extends item-level reshaping with sequence-level recalibration and collaboration-guided fusion, jointly regulating how multi-modal information is organized, evolves along interaction sequences, and enters ID-based prediction. Thus, the two methods share part of their representation-level motivation, while PRISM targets the distinct challenge of coordinating multi-modal semantics with collaborative signals throughout sequential recommendation.

A.2 Methodology and Implementation Details

To guarantee full reproducibility and provide complete algorithmic transparency, this section presents the precise mathematical formulations, numerical stabilization techniques, and detailed hyperparameter configurations of our proposed framework. These concrete specifications serve to supplement the primary methodology presented in the main text with self-contained rigor.

A.2.1 Methodological Details

In this section, we provide the detailed mathematical formulations, numerical stabilization techniques, and concrete loss functions that constitute the core modules of our proposed method.

Details of Auxiliary Objectives This section provides the exact mathematical formulations and numerical stabilization techniques for the auxiliary objectives introduced in the Cross-Modal Semantic Reshaping module.

Intra-Modal Compactness Loss ($\mathcal{L}_{\text{intra}}$). To compute the Mahalanobis distance while guaranteeing a positive definite diagonal covariance matrix and maintaining numerical stability, we parameterize the covariance in the log-space. Specifically, the log-variance is clamped to the range $[-10, 10]$ before exponentiation, i.e., $\mathbf{\Sigma}_k^m = \exp(\text{clamp}(\log \mathbf{\Sigma}_k^m, -10, 10))$. To prevent trivial solutions, the

gradients of the prototypes and variances are detached during the distance computation. The loss is formalized as:

$$\mathcal{L}_{\text{intra}} = \sum_{m\in\mathcal{M}} \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} q_n^m(k)\sum_{j=1}^{d}\frac{(\phi_{n,j}^m - \bar{\mu}_{k,j}^m)^2}{\bar{\Sigma}_{k,j}^m}, \tag{13}$$

where $\bar{\boldsymbol{\mu}}_k^m$ and $\bar{\boldsymbol{\Sigma}}_k^m$ denote the detached mean and covariance of prototype k, respectively.

Prototype Diversity Loss ($\mathcal{L}_{\text{div}}$). The diversity regularizer aims to separate the prototypes within the same modality. It computes the mean exponentiated negative squared Euclidean distance exclusively over all distinct prototype pairs ($k \neq k'$), explicitly masking out the self-distance diagonal:

$$\mathcal{L}_{\text{div}} = \sum_{m\in\mathcal{M}} \frac{1}{K(K-1)}\sum_{k=1}^{K}\sum_{k'\neq k}\exp\Big(-\|\boldsymbol{\mu}_k^m - \boldsymbol{\mu}_{k'}^m\|_2^2\Big). \tag{14}$$

Load-Balancing Regularizer ($\mathcal{L}_{\text{uniform}}$). This regularizer minimizes the Kullback-Leibler (KL) divergence from the empirical assignment distribution to a uniform prior. To avoid numerical underflow, the average assignment probability $\bar{p}_k^m = \frac{1}{N}\sum_n q_n^m(k)$ is smoothed with a lower bound $\epsilon = 10^{-8}$. The objective is simplified as:

$$\mathcal{L}_{\text{uniform}} = \sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\hat{p}_k^m \log(K\hat{p}_k^m), \tag{15}$$

where $\hat{p}_k^m = \max(\bar{p}_k^m, \epsilon)$.

Statistical Alignment Loss ($\mathcal{L}_{\text{sem}}$). To enforce cross-modal consistency, statistical alignment matches the batch-wise empirical moments of the shared semantics using the inverse-frequency weights w_n. For each modality $m \in \{\text{txt}, \text{img}\}$, we calculate the weighted mean ($\boldsymbol{\mu}_m$), standard deviation ($\boldsymbol{\sigma}_m$), and skewness ($\mathbf{s}_m$) over $\{\boldsymbol{\psi}_n^m\}_{n=1}^N$:

$$\boldsymbol{\mu}_m = \sum_{n=1}^{N} w_n \boldsymbol{\psi}_n^m, \quad \boldsymbol{\sigma}_m = \sqrt{\sum_{n=1}^{N} w_n(\boldsymbol{\psi}_n^m - \boldsymbol{\mu}_m)^2}, \quad \mathbf{s}_m = \frac{\sum_{n=1}^{N} w_n(\boldsymbol{\psi}_n^m - \boldsymbol{\mu}_m)^3}{\boldsymbol{\sigma}_m^3 + \epsilon}, \tag{16}$$

where all vector powers and divisions are element-wise, and $\epsilon = 10^{-6}$ avoids division by zero. The statistical alignment loss is then defined as:

$$\mathcal{L}_{\text{sem}} = \|\boldsymbol{\mu}_{\text{txt}} - \boldsymbol{\mu}_{\text{img}}\|_2^2 + \|\boldsymbol{\sigma}_{\text{txt}} - \boldsymbol{\sigma}_{\text{img}}\|_2^2 + \|\mathbf{s}_{\text{txt}} - \mathbf{s}_{\text{img}}\|_2^2. \tag{17}$$

Item-Transition Graph Construction and Propagation To capture global transition dynamics while strictly preventing temporal data leakage, we construct the item-transition graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}^{\text{tr}})$ exclusively from historical training sequences $\mathcal{S}^{\text{tr}}$ (strictly isolated from validation and test sets). To filter noise and improve efficiency, we apply Top-K pruning by retaining only the K most frequent transitions for each item.

We treat the pruned graph $\mathcal{G}$ as undirected and adopt LightGCN for neighborhood aggregation. The layer-wise message passing for item $v \in \mathcal{V}$ is defined as:

$$\mathbf{e}_v^{(l+1)} = \sum_{u\in\mathcal{N}(v)} \frac{1}{\sqrt{|\mathcal{N}(v)|\,|\mathcal{N}(u)|}}\mathbf{e}_u^{(l)}, \tag{18}$$

where $\mathcal{N}(v)$ denotes the neighbors of item v, and $\mathbf{e}_v^{(0)}$ is its initial learnable embedding. After L propagation layers, we average the representations across all layers to capture multi-hop collaborative signals:

$$\mathbf{e}_v = \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{e}_v^{(l)}. \tag{19}$$

The resulting structural embedding $\mathbf{e}_v \in \mathbb{R}^d$ is subsequently fed into the cross-attention module.

A.2.2 Training Details and Hyperparameters

In this section, we detail the specific hyperparameter configurations for the auxiliary objectives and the overall training strategy.

Hyperparameters for Auxiliary Losses. For the prototype optimal transport alignment, the entropy-regularized multi-marginal optimal transport problem is solved via the Sinkhorn algorithm. We set the entropy regularization coefficient (λ) to 0.1 and the maximum number of Sinkhorn iterations to 50. For the debiased distribution alignment, the bandwidth of the Gaussian kernel utilized in the weighted Maximum Mean Discrepancy (MMD) calculation is set to 1.0.

Warm-up Training Strategy. The overall loss function integrates the primary sequential recommendation objective ($\mathcal{L}_{\text{rec}}$) with the auxiliary semantic reshaping constraints ($\mathcal{L}_{\text{uni}}$ and $\mathcal{L}_{\text{com}}$). To prevent the cross-modal alignment losses from interfering with the initial learning phase of user preferences and item representations, we employ a warm-up training strategy. Specifically, we strictly set the alignment loss weight $\beta = 0$ for the first 3 epochs. This warm-up period ensures that the core recommendation modules—particularly the ID-based item embeddings and the sequence modeling layer—establish a robust and task-oriented initialization guided solely by $\mathcal{L}_{\text{rec}}$. Following the warm-up phase, β is restored to its predefined value to jointly optimize the multi-modal semantic reshaping and the next-item prediction task.

A.3 Residual Spectral Interpretation of Frequency-Gated Sequence Recalibration

This section characterizes the spectral behavior of the proposed Frequency-Gated Sequence Recalibration (FGSR). We show that, owing to its residual formulation and bounded gating mechanism, FGSR acts as a bounded, frequency-dependent differential gain operator over the original spectrum, rather than an attenuating low-pass filter or hard frequency mask.

Notation and Conjugate Symmetry. Let $\mathbf{X}_m^{\text{item}} \in \mathbb{R}^{L\times d}$ denote the real-valued sequence embedding shared by the Fourier branch and the identity shortcut. Let $\mathbf{Z}_m = \mathcal{F}\left(\mathbf{X}_m^{\text{item}}\right) \in \mathbb{C}^{L\times d}$ denote its Discrete Fourier Transform (DFT). Since $\mathbf{X}_m^{\text{item}}$ is strictly real, its spectrum exhibits conjugate symmetry, i.e., $\mathbf{Z}_{m,k,j} = \mathbf{Z}^*_{m,L-k,j}$, and its amplitude spectrum satisfies $\mathbf{A}_{m,k,j} = |\mathbf{Z}_{m,k,j}| = \mathbf{A}_{m,L-k,j}$.

Because the gating network applies a weight-shared transformation across sequence positions, the derived real-valued gate inherits this symmetry ($\mathbf{G}_{m,k,j} = \mathbf{G}_{m,L-k,j}$). Consequently, the modulated spectrum $\mathbf{Z}_m \odot \mathbf{G}_m$ preserves conjugate symmetry, guaranteeing that its inverse transform $\mathcal{F}^{-1}\left(\mathbf{Z}_m \odot \mathbf{G}_m\right)$ is strictly real-valued.

Proposition 1 (Residual Spectral Equivalence)**.** *Let $\mathbf{G}_m = \sigma(\cdot) \in (0,1)^{L\times d}$ denote the learned frequency gating matrix. The recalibrated sequence embedding*

$$\mathbf{X}_m^{\text{flt}} = \mathcal{F}^{-1}\left(\mathbf{Z}_m \odot \mathbf{G}_m\right) + \mathbf{X}_m^{\text{item}} \tag{20}$$

is algebraically equivalent to modulating $\mathbf{Z}_m$ via a single effective gain matrix $\mathbf{1} + \mathbf{G}_m$ prior to the inverse transform:

$$\mathbf{X}_m^{\text{flt}} = \mathcal{F}^{-1}\left(\mathbf{Z}_m \odot (\mathbf{1} + \mathbf{G}_m)\right). \tag{21}$$

Proof. By the linearity of the inverse transform $\mathcal{F}^{-1}$, and substituting $\mathbf{X}_m^{\text{item}} = \mathcal{F}^{-1}\left(\mathbf{Z}_m\right)$:

$$\begin{aligned}\mathbf{X}_m^{\text{flt}} &= \mathcal{F}^{-1}\left(\mathbf{Z}_m \odot \mathbf{G}_m\right) + \mathcal{F}^{-1}\left(\mathbf{Z}_m\right) \\ &= \mathcal{F}^{-1}\left(\mathbf{Z}_m \odot \mathbf{G}_m + \mathbf{Z}_m\right) \\ &= \mathcal{F}^{-1}\left(\mathbf{Z}_m \odot (\mathbf{1} + \mathbf{G}_m)\right).\end{aligned} \tag{22}$$

□

Proposition 2 (Bounded Differential Gain)**.** *For every frequency index $k \in \{1,\dots,L\}$ and channel $j \in \{1,\dots,d\}$, the effective gain satisfies $1 + G_{m,k,j} \in (1,2)$, yielding*

$$\left|(1 + G_{m,k,j})\, Z_{m,k,j}\right| > \left|Z_{m,k,j}\right|, \qquad \forall\, k, j. \tag{23}$$

Proof. Since $\mathbf{G}_m = \sigma(\cdot)$, each entry strictly satisfies $G_{m,k,j} \in (0,1)$, which bounds $1 + G_{m,k,j} \in (1,2)$. As the gain strictly exceeds unity, multiplying $Z_{m,k,j}$ by this factor strictly increases its magnitude. □

Table 4: Performance and training efficiency comparison between standard batched FFT (**PRISM**) and dynamic sequence-length FFT (**PRISM (Dynamic)**). Bold scores denote superior performance (higher recommendation accuracy or lower training time). *Improv.* represents the relative accuracy gain or training time overhead introduced by dynamic FFT.

Dataset	Method	R@10	R@20	N@10	N@20	Time (s)
Fitness	**PRISM**	0.1579	0.2255	0.0901	0.1091	**14.2**
	PRISM (Dynamic)	**0.1594**	**0.2288**	**0.0903**	**0.1104**	43.8
	Improv.	+0.95%	+1.46%	+0.22%	+1.19%	+208.45%
Game	**PRISM**	0.0686	0.1077	0.0360	0.0459	**44**
	PRISM (Dynamic)	**0.0702**	**0.1085**	**0.0374**	**0.0471**	130
	Improv.	+2.33%	+0.74%	+3.89%	+2.61%	+195.45%
Dogs	**PRISM**	0.1903	0.2577	0.1165	0.1349	**45**
	PRISM (Dynamic)	**0.1905**	**0.2586**	**0.1166**	**0.1353**	139
	Improv.	+0.11%	+0.35%	+0.09%	+0.30%	+208.89%

Proposition 2 formalizes that FGSR does not suppress or eliminate any frequency component below its original magnitude. Instead of zeroing out specific bands, $\mathbf{G}_m$ reweights the relative magnitude profile across frequencies through differential amplification. For any two spectral entries (k_1, j_1) and (k_2, j_2):

$$\frac{|(1+G_{m,k_1,j_1})Z_{m,k_1,j_1}|}{|(1+G_{m,k_2,j_2})Z_{m,k_2,j_2}|} = \frac{1+G_{m,k_1,j_1}}{1+G_{m,k_2,j_2}} \cdot \frac{|Z_{m,k_1,j_1}|}{|Z_{m,k_2,j_2}|}. \tag{24}$$

Components assigned higher gate values receive greater relative amplification while retaining the full baseline magnitude of lower-gated components, enabling adaptive enhancement of informative frequency patterns without explicit frequency suppression.

Remark 1 (Implications for Sequential Recommendation). This residual formulation provides two functional advantages. First, retaining the original identity shortcut ensures that baseline item representations remain intact, reducing the risk of catastrophic degradation under noisy spectral gate estimates. Second, constraining the effective gain within $(1, 2)$ prevents explosive magnitude updates, promoting stable optimization. Because $\mathbf{G}_m$ is derived from joint cross-modal amplitude spectra, the differential amplification of each modality is contextually regularized by signals from the other.

We note that the spectral recalibration described above establishes the relative magnitude ratios prior to multimodal concatenation and layer normalization, which subsequently apply shared affine rescaling across modalities.

A.4 Engineering FGSR: Sequence Padding vs. Parallel Efficiency

This section discusses the engineering considerations underlying the batched implementation of Frequency-Gated Sequence Recalibration (FGSR), with a particular focus on sequence padding and computational efficiency.

In standard batch processing for sequential recommendation, interaction sequences of varying lengths are zero-padded to a common length L to enable efficient parallel computation. The FFT length determines the discrete frequency sampling grid, and padding can therefore change the sampled spectral representation and introduce boundary-related spectral leakage. To evaluate the practical impact of this implementation choice, we compare the standard batched FFT with a dynamic sequence-length variant (**PRISM (Dynamic)**), which iterates through the batch and computes exact-length FFTs for individual sequences without padding. As shown in Table 4, the dynamic variant yields only modest performance gains, including $+0.11\%$ to $+2.33\%$ on Recall@10, while increasing training time by approximately 200%.

Several properties of the implementation help limit the downstream effect of padded spectral computation:

- FGSR employs a learnable gating network parameterized by $\mathbf{W}^m$ and $\mathbf{b}^m$ rather than a fixed spectral filter. This provides the model with flexibility to adapt to padding-related

spectral patterns observed during training instead of imposing the same spectral response across all sequences.

- For a fixed finite sequence, appending zeros leaves the observed time-domain values unchanged while sampling its discrete-time Fourier transform on a different frequency grid. The resulting spectral representation is therefore related to the same underlying finite sequence, although the sampled frequency coefficients depend on the FFT length.
- Spectral modulation followed by the inverse FFT can produce non-zero responses at padded temporal positions. We apply a deterministic boolean mask based on the valid sequence length L_{valid} after the IFFT and modality fusion operations, explicitly removing activations at invalid positions before subsequent sequential modeling. This masking prevents padded positions from directly contributing downstream, while not assuming that valid-position representations are invariant to the FFT length.

Based on this empirical comparison, we adopt the zero-padded batched FFT as the default implementation of PRISM. The results in Table 4 indicate that this choice provides substantially higher parallel efficiency with only a limited impact on final recommendation performance.

A.5 Efficiency Analysis

In this section, we analyze the time complexity of the proposed PRISM framework. During the training phase, the computational overhead primarily originates from graph propagation, cross-modal semantic reshaping, and sequence modeling. For a global transition graph with $|\mathcal{E}^{\text{tr}}|$ edges, the LightGCN encoding requires $\mathcal{O}(l|\mathcal{E}^{\text{tr}}|d)$ operations per epoch, where l is the number of propagation layers and d is the latent dimension. In the cross-modal semantic reshaping module, the computation of Prototype Optimal Transport and weighted Maximum Mean Discrepancy (MMD) operates within each mini-batch, incurring a complexity of $\mathcal{O}(BLKd + B^2L^2d)$, where B is the batch size, L is the maximum sequence length, and K denotes the number of Gaussian prototypes. The frequency-gated sequence recalibration leverages the Fast Fourier Transform (FFT) along the sequence dimension, which is strictly bounded by $\mathcal{O}(BdL \log L)$. Finally, the collaboration-guided cross-attention operates at $\mathcal{O}(BdL^2)$, which aligns with the standard complexity for Transformer-based architectures. Overall, the training complexity scales linearly with the graph size and quadratically with the sequence length, remaining highly comparable to mainstream multi-modal recommendation paradigms.

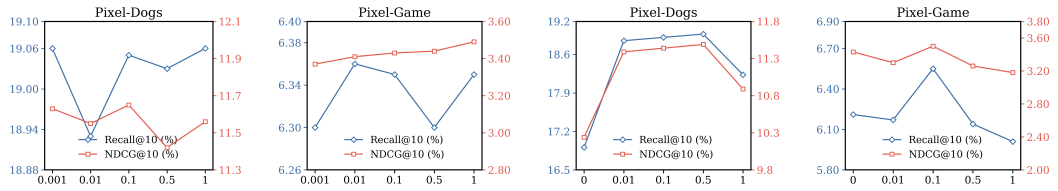
During the inference phase, the computational overhead is substantially reduced for real-time serving. Specifically, the graph-based item representations are pre-computed offline to eliminate redundant graph convolutions across the candidate pool during scoring. To avoid excessive GPU memory consumption from caching full multi-modal embeddings across all items, textual and visual projections are dynamically processed only for items within the active sequence. Consequently, online serving merely evaluates the active user's historical sequence through the FFT-based recalibration and cross-attention modules, yielding an online complexity of $\mathcal{O}(dL \log L + dL^2)$. This design enables PRISM to satisfy strict low-latency requirements in real-time sequential recommendation while effectively exploiting rich multi-modal semantics.

A.6 Hyper-Parameter Study

A.6.1 Hyperparameter Analysis.

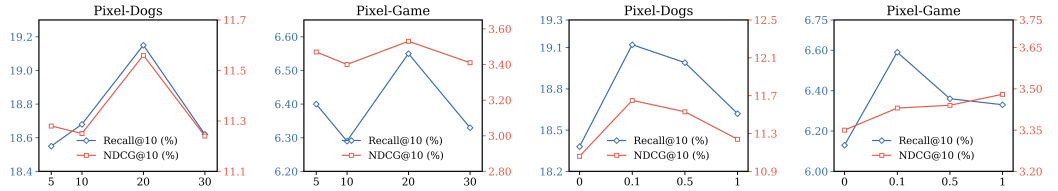
We investigate four key hyperparameters of PRISM, including the orthogonality weight α, the semantic recalibration weight β, the number of Gaussian prototypes K, and the prototype alignment coefficient λ_{align}. Recall@10 and NDCG@10 are adopted to evaluate their effects on retrieval and ranking performance.

As illustrated in Figure 6a and 6b, α controls the orthogonality constraint between the common and unique representations. PRISM remains relatively stable across the tested range, demonstrating limited sensitivity to this parameter. On *Dogs*, $\alpha = 0.1$ achieves the best overall performance with an NDCG@10 of 0.1165. A weak constraint may leave redundant information between the two paths, whereas an excessively strong constraint may over-separate shared and modality-specific semantics.



(a) Effect of α on *Dogs*. (b) Effect of α on *Game*. (c) Effect of β on *Dogs*. (d) Effect of β on *Game*.

Figure 6: Parameter sensitivity analysis. (a)-(b) Performance under different weights of the common–unique orthogonality constraint (α). (c)-(d) Performance under different weights of cross-modal semantic recalibration (β).



(a) Effect of K on *Dogs*. (b) Effect of K on *Game*. (c) Effect of λ_{align} on *Dogs*. (d) Effect of λ_{align} on *Game*.

Figure 7: Parameter sensitivity analysis. (a)-(b) Performance under different numbers of Gaussian prototypes (K). (c)-(d) Performance under different strengths of prototype-level cross-modal alignment (λ_{align}).

Figure 6c and 6d reports the effect of β, which balances the complete cross-modal semantic recalibration objective against the recommendation loss. The model exhibits a clear non-monotonic trend, achieving the best results at $\beta = 0.5$ on *Dogs* and $\beta = 0.1$ on *Game*. In particular, $\beta = 0.1$ yields a Recall@10 of 0.0655 and an NDCG@10 of 0.0350 on *Game*. A small β provides insufficient semantic correction, while an overly large value causes the auxiliary alignment objectives to dominate recommendation supervision.

The influence of the prototype number K is shown in Figure 7a and 7b. The two datasets exhibit a consistent trend, with $K = 20$ achieving the best performance. On *Dogs*, it obtains a Recall@10 of 0.1915 and an NDCG@10 of 0.1156. Too few prototypes cannot adequately characterize heterogeneous modality-specific semantics, whereas too many prototypes may fragment coherent semantic structures and reduce the reliability of prototype estimation.

Figure 7c and 7d analyzes λ_{align}, which controls the strength of prototype-level optimal transport and cross-modal sample guidance for unique features. A moderate value of 0.1 provides the most balanced performance, achieving the highest Recall@10 on both datasets and an NDCG@10 of 0.1165 on *Dogs*. Without this alignment, cross-modal prototype correspondence cannot be explicitly established, while excessive alignment may suppress the complementary semantics that the unique path is designed to preserve.

Overall, these results validate the necessity of separately controlling common–unique decoupling, semantic recalibration, prototype granularity, and prototype-level correspondence. Moderate regularization and alignment strengths generally provide the best balance between cross-modal consistency and the preservation of modality-specific information.

A.7 Extended Results for Hit Preservation and Recovery Analysis

Figure 8 illustrates the extended hit preservation and recovery analysis on the *Fitness* dataset. Prominently positioned in the top-right quadrant, PRISM consistently achieves a superior Pareto-optimal trade-off by maximizing low-frequency item recovery while sustaining high head-item hit retention. These supplementary observations fully corroborate the main-text findings, confirming that PRISM effectively balances tail-item discovery with robust head-signal preservation across diverse domain distributions.



(a) Gained / lost hits. (b) Recovery & preservation.

Figure 8: Hit preservation and recovery analysis on *Fitness*. (a) Net Recall@10 decomposition into gained vs. lost hits. (b) low-frequency item recovery vs. head-item preservation trade-off (bubble size indicates overall Recall@10).

Table 5: Statistics of the large-scale evaluation datasets.

Dataset	Users	Items	Interactions	Sparsity	Avg. Seq. Length
Song	161,720	7,954	1,353,687	99.89%	8.37
Mix	39,882	42,200	709,190	99.96%	17.78

A.8 Scalability Evaluation on Large-Scale Datasets

To evaluate PRISM's scalability and generalization, we extend our experiments to two large-scale Pixel benchmarks: *Song* and *Mix*. Unlike standard benchmarks constrained by limited interaction scales and short sequences, *Song* contains over 1.35M single-domain interactions, serving to test model efficiency and stability under dense collaborative signals. Meanwhile, *Mix* synthesizes multi-category sequences (42,200 items, average length 17.78) with high semantic heterogeneity and temporal fluctuations, testing PRISM's ability to recalibrate multi-modal sequences and capture long-term preference dynamics at scale. Detailed statistics are summarized in Table 5.

A.8.1 Scalability Evaluation

Table 6 presents the evaluation results on two large-scale datasets. PRISM consistently outperforms all baselines across all metrics, proving its scalability and generalization under complex preference dynamics. Key observations include:

On the single-domain *Song* dataset, while graph-based SGP4SR Li et al. (2026a) yields a strong baseline through structural sequence modeling, PRISM achieves substantial further gains. This confirms that prototype-guided recalibration effectively stabilizes representation learning and prevents representation collapse under dense collaborative feedback. Furthermore, on the multi-category *Mix* dataset—characterized by longer sequences and higher semantic variance—PRISM exhibits even more pronounced relative improvements over competitive methods like HM4SR Zhang et al. (2025). This demonstrates that frequency-gated spectral recalibration effectively regulates abrupt sequence variations while preserving informative preference dynamics at scale.

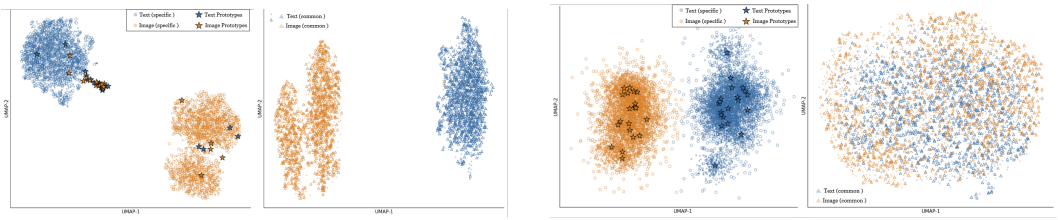
A.9 Visual Analysis of Cross-Modal Decoupling

To qualitatively inspect the behavioral characteristics of the proposed Cross-Modal Semantic Reshaping module, we visualize the decoupled common and modality-specific embeddings derived from Eq. (2) using 2D UMAP, as shown in Figure 9. Prior to alignment, the uncalibrated representations exhibit a pronounced modality gap, where textual and visual features occupy distinct, disjoint clusters. Following optimization guided by $\mathcal{L}_{\text{com}}$, the common embeddings from both modalities display substantial distributional overlap. This demonstrates that the shared subspace effectively aligns multi-modal feature distributions, yielding structural organization consistent with our alignment and decoupling objectives.

Conversely, the modality-specific embeddings optimized via $\mathcal{L}_{\text{uni}}$ behave exactly as formulated: rather than collapsing into a uniform cross-modal distribution—a suboptimal naive fusion that inevitably erases valuable modality-specific cues and induces semantic conflict—they maintain distinct modal boundaries. Furthermore, guided by the Prototype Optimal Transport Alignment, these

Table 6: performance comparison on large-scale datasets. Bold and underlined scores denote the best and second-best results. R@K and N@K represent Recall@K and NDCG@K. All multimodal models share identical image and text encoders. Improv. tracks PRISM's relative gain over the best baseline (statistically significant via paired t-test with p-value < 0.05).

Method	Song				Mix			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
SASRec	0.2093	0.2799	0.1152	0.1331	0.0137	0.0220	0.0062	0.0083
ATHWE	0.2079	0.2802	0.1205	0.1390	0.0159	0.0258	0.0078	0.0103
HM4SR	0.2057	0.2769	0.1212	0.1394	<u>0.0164</u>	<u>0.0278</u>	<u>0.0081</u>	<u>0.0110</u>
SGP4SR	<u>0.2197</u>	<u>0.2894</u>	<u>0.1299</u>	<u>0.1475</u>	0.0122	0.0214	0.0058	0.0081
PRISM (ours)	**0.2325**	**0.3019**	**0.1448**	**0.1624**	**0.0198**	**0.0318**	**0.0100**	**0.0130**
Improv.	+5.83%	+4.32%	+11.47%	+10.10%	+20.73%	+14.39%	+23.46%	+18.18%



(a) Before training. (b) After training.

Figure 9: UMAP visualizations of decoupled common and modality-specific embeddings before and after cross-modal semantic reshaping.

unique features transition from an unorganized state into well-structured, compact clusters anchored by the learned prototypes. This structural evolution confirms that our decoupling mechanism strictly prevents representation degeneration, successfully preserving recurring intra-modal patterns while cleanly isolating them from the shared cross-modal semantics.

A.10 Case Study

We present a representative head-item case to examine how PRISM handles a locally inconsistent interaction. In this case, direct multi-modal integration loses an ID-only hit, semantic reshaping alone does not fully recover the target, whereas the complete PRISM model restores the hit and ranks the target above the ID-only model.

Figure 10(a) shows the recent interaction sequence and the target item. The highlighted interaction is identified as the largest local discontinuity under the original multi-modal features, indicating a departure from the surrounding preference pattern. Figure 10(b) reports the textual and visual frequency gates learned by PRISM. The modality-specific offset notation exposes their small but non-uniform variations across frequency bands. This result indicates fine-grained sequence-level recalibration rather than frequency suppression. Since each Fourier coefficient is computed from the complete sequence, an individual frequency band should not be attributed to a particular historical position.

To examine how each model uses individual interactions, Figure 10(c) reports the interaction removal effect. For each model (q) and position (ℓ), we delete the complete interaction and compute

$$I^{\mathrm{del}}_{q,\ell} = \log p_q(y \mid \mathcal{S}) - \log p_q(y \mid \mathcal{S}_{-\ell}) \tag{25}$$

while keeping the target and candidate set unchanged. A positive value indicates that the interaction supports the target prediction, whereas a negative value indicates that removing it increases the target probability. Because the intervention removes the complete interaction, it measures dependence on the historical item rather than the isolated contribution of its multi-modal content.

Direct multi-modal integration exhibits strong dependence on the highlighted interaction, although this dependence does not lead to an effective target ranking. Cross-modal semantic reshaping reduces this sensitivity, while the complete PRISM model makes the highlighted interaction nearly neutral and preserves stronger evidence from a more informative recent interaction. Consequently,

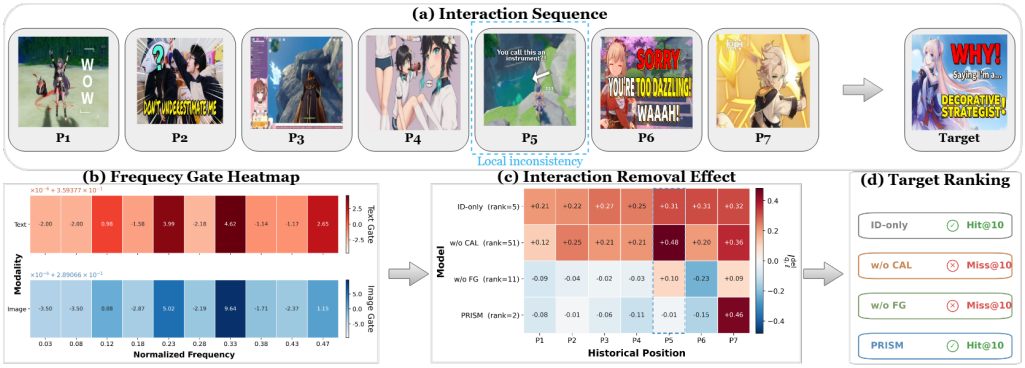


Figure 10: Case study of a head-item prediction with a locally inconsistent interaction. (a) The recent interaction sequence and target item, where the detected local inconsistency is highlighted. (b) The textual and visual frequency gates learned by PRISM, displayed with modality-specific offsets to reveal small variations across normalized frequency bands. (c) The change in target log-probability after removing each complete historical interaction. Positive and negative values indicate supportive and adverse interaction effects, respectively. (d) The target-ranking outcomes of ID-only, w/o CAL, w/o FG, and PRISM. Only visual thumbnails are shown in (a); the textual features used by the model are omitted for readability.

Figure 10(d) shows that PRISM recovers the target hit and ranks it above the ID-only model. The case illustrates how semantic reshaping and frequency-gated recalibration jointly reorganize the use of historical evidence under locally inconsistent multi-modal signals.

A.11 Dataset Specifications and Preprocessing

In this section, we provide comprehensive details on the benchmark datasets evaluated in our study, including the underlying rationale for dataset selection, protocols for data filtering, and the explicit pipeline for multimodal feature construction.

A.11.1 Rationale for Dataset Selection

We evaluate our framework on the Pixel benchmark Cheng et al. (2023), which encompasses three distinct content ecosystems (*Fitness*, *Game*, and *Dogs*). As a representative user-generated content (UGC) social benchmark, Pixel presents a realistic yet challenging testbed for multimodal sequential recommendation. Specifically, its unconstrained cover images and casual metadata introduce substantial semantic heterogeneity and cross-modal noise, while social-driven interactions exhibit high sequential volatility due to impulsive user engagement. Evaluating across these three heterogeneous domains allows for a rigorous assessment of model recalibration and cross-domain generalizability under complex multimodal noise.

A.11.2 Preprocessing Protocol and Feature Construction

Following standard protocols in multimodal sequential recommendation Hou et al. (2022); Lin et al. (2024); Wang et al. (2024), we iteratively apply 5-core filtering on both user and item sides to eliminate sparse entities with fewer than five interactions. For multimodal feature extraction, we concatenate the three textual metadata fields—*title*, *tag*, and *description*—using space separators into a unified text sequence per item, and adopt the primary user-uploaded cover image to extract visual representations. Finally, user interaction logs are arranged chronologically by timestamps to construct historical behavior sequences for sequential preference modeling.

A.12 Limitations and Future Directions

While PRISM achieves superior performance in balancing multi-modal semantics and collaborative signals, we acknowledge certain potential extensions and trade-offs that open avenues for future research:

Non-Uniform Temporal Dynamics in Spectral Recalibration. The Frequency-Gated Sequence Recalibration module applies the Discrete Fourier Transform (DFT) under the standard assumption that user interactions occur at uniform discrete sequence steps. In real-world recommendation systems, however, user behavior often exhibits temporal non-uniformity and irregular interaction burstiness Fu et al. (2025). Directly incorporating continuous temporal embeddings into spectral filtering tends to disturb spectral phase coherence, whereas integrating Non-Uniform Fast Fourier Transforms (NUFFT) introduces non-trivial computational overhead during training. To prioritize training stability and spectral efficiency, PRISM focuses on ordinal sequence-based frequency recalibration. Extending frequency-gated recalibration to intrinsically model continuous time intervals without incurring substantial computational costs represents an intriguing direction for temporal-spectral sequence modeling.

Inference Complexity under Cross-Attention Integration. Although the proposed frequency recalibration operates with sub-quadratic $\mathcal{O}(L \log L)$ complexity per modality, PRISM adopts a Transformer cross-attention encoder ($\mathrm{Enc}_{\mathrm{mix}}$) to ensure high-capacity interaction between collaborative ID sequence representations and multi-modal features. Consequently, the overall sequence fusion retains an $\mathcal{O}(L^2)$ time complexity with respect to the sequence length L. This design intentionally prioritizes cross-modal expressive power over pure inference speed, which is optimal for standard sequence lengths. Future work may explore replacing cross-attention with pure spectral cross-fusion or state-space paradigms to achieve end-to-end sub-quadratic inference, thereby fully unlocking the asymptotic efficiency advantages of frequency-domain processing for ultra-long sequence recommendation.