Beyond Uniform Alignment: Rethinking Representation Fusion in Personalized LLM-based Recommendation

ZICHEN YUAN*, Northeastern University, China
YICHEN WU*, National University of Singapore, Singapore
GUANYU ZHU, South China Agricultural University, China
JIACHENG LIU, South China Normal University, China
XINYU FANG, Hefei University of Technology, China
LIFAN SUN, University of California San Diego, USA
HANWEN DU, Soochow University, China
XINYUAN SONG, Emory University, USA
JOEMON M. JOSE, University of Glasgow, School of Computing Science, UK
JUNCHEN FU†, University of Glasgow, School of Computing Science, UK
YOUHUA LI†, City University of Hong Kong, China
YONGXIN NI, National University of Singapore, Singapore

Large Language Models (LLMs) are increasingly being jointly trained with domain-specific encoders as a powerful integration strategy to enhance domain performance. However, this approach faces critical challenges unique to recommender systems. However, empirical observations reveal that conventional uniform alignment strategies often lead to structural degradation in recommendation scenarios. Specifically, the inherent subjectivity of user preferences compromises representation quality during joint optimization, while interaction sparsity drives these models to overfit dominant active users. This inherent subjectivity, coupled with the domain's interaction-sparsity nature, drives data-driven LLMs to capture preference patterns that overfit to dominant active users and popular items. To address this, we propose PALER (Personalized Alignment for LLM-based Enhanced Recommendation), a unified framework combining modality-invariant alignment with an adaptive reweighting strategy. The alignment module employs contrastive learning to project textual and behavioral representations into a shared latent space and ground the LLM representations in personalized semantic spaces through joint optimization. Concurrently, the reweighting module dynamically tracks user-level loss to amplify the influence of long-tail users during gradient updates, effectively mitigating optimization imbalance skewed towards mainstream users. Experiments on real-world datasets demonstrate that PALER

*Both authors contributed equally to this research.
†Corresponding authors.

Authors' Contact Information: Zichen Yuan, yzc66633@gmail.com, Northeastern University, Shenyang, Liaoning, China; Yichen Wu, wuyichen@u.nus.edu, National University of Singapore, Singapore, Singapore; Guanyu Zhu, South China Agricultural University, Guangzhou, Guangdong, China, xiaoguanbx@gmail.com; Jiacheng Liu, South China Normal University, Guangzhou, Guangdong, China, helloliujiacheng@gmail.com; Xinyu Fang, Hefei University of Technology, Hefei, Anhui, China, 2021211921@mail.hfut.edu.cn; Lifan Sun, University of California San Diego, La Jolla, CA, USA, lis005@ucsd.edu; Hanwen Du, Soochow University, Suzhou, Jiangsu, China, hwdu@stu.suda.edu.cn; Xinyuan Song, Emory University, Atlanta, GA, USA, xsong69@emory.edu; Joemon M. Jose, University of Glasgow, School of Computing Science, Glasgow, Scotland, UK, Joemon.Jose@glasgow.ac.uk; Junchen Fu, University of Glasgow, School of Computing Science, Glasgow, UK, j.fu.3@research.gla.ac.uk; Youhua Li, City University of Hong Kong, Hong Kong, China, youhuali2-c@my.cityu.edu.hk; Yongxin Ni, National University of Singapore, Singapore, Singapore, niyongxin@u.nus.edu.

ACM 1557-735X/2026/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

outperforms strong baselines, especially for sparse-interaction users, with ablation studies confirming the effectiveness and complementarity of the proposed modules. The implemented code and further information are included in an anonymous code repository: https://anonymous.4open.science/r/Beyond-Uniform-Alignment-A-Personalized-Perspective-on-LLM-based-Recommendation-1DB4.

CCS Concepts: • **Information systems** → **Recommender systems**.

Additional Key Words and Phrases: Recommender Systems, Large Language Models, Personalized Alignment, Contrastive Learning

ACM Reference Format:
Zichen Yuan, Yichen Wu, Guanyu Zhu, Jiacheng Liu, Xinyu Fang, Lifan Sun, Hanwen Du, Xinyuan Song, Joemon M. Jose, Junchen Fu, Youhua Li, and Yongxin Ni. 2026. Beyond Uniform Alignment: Rethinking Representation Fusion in Personalized LLM-based Recommendation. *J. ACM* 37, 4, Article 111 (August 2026), 26 pages. https://doi.org/XXXXXXX.XXXXXXX

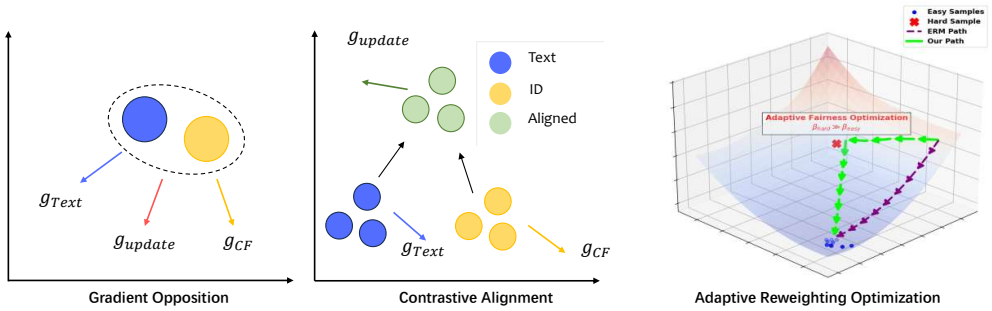

Fig. 1. Illustration of the optimization challenges and the proposed solutions. The left panel illustrates gradient opposition in naive fusion, where opposing textual (g_{Text}) and behavioral (g_{CF}) gradients yield a compromised joint update (g_{update}) that degrades personalized signals. In the middle, the modality-invariant module explicitly aligns g_{Text} and g_{CF}, generating a synergistic g_{update} that preserves inherently user-specific representations. On the right, the adaptive reweighting mechanism dynamically steers the optimization path away from easy samples toward hard samples, effectively mitigating the impact of interaction sparsity.

1 Introduction

The rapid evolution of large language models (LLMs) has facilitated the development of multimodal systems that synergize LLMs with task-specific encoders (e.g., sequence ID encoders or image encoders) to jointly process structured and unstructured inputs [2, 11, 21, 28]. The efficacy of these systems is predicated upon effective representation alignment across modalities, with prior work employing contrastive learning [8, 30, 39], instruction tuning [36], and preference alignment [13, 27] to mitigate semantic drift and improve task-grounded integration.

This paradigm, which typically follows a uniform alignment strategy, has recently been adopted in recommender systems, where LLMs are integrated with behavioral encoders such as collaborative filtering (CF) models to enhance user preference modeling [3, 59]. However, in contrast to CV and NLP datasets that typically provide objective ground-truth labels, recommendation targets are inherently subjective; for instance, an item favored by one user may be dismissed by another. Moreover, user–item interaction data are typically sparse and follow a long-tail distribution [9]. As LLMs are inherently data-driven, such uniform alignment strategies often converge toward

generalized preference patterns dominated by active users and popular items, thereby undermining the ID encoder's representation quality and diminishing personalization [1, 37].

To validate the hypothesis that such uniform alignment can introduce optimization imbalance in LLM-based recommendation, we designed a probe experiment to compare the standalone performance of recommendation encoders under two settings: (1) trained standalone, and (2) trained jointly with an LLM via naive fusion, which involves direct end-to-end joint optimization without explicit modality alignment or structural decoupling. As shown in Table 2, the evaluated encoders, when trained in the joint framework, consistently underperform their standalone counterparts, indicating that a naive fusion-based training strategy is detrimental to the representation learning of recommendation encoders. This degradation is attributable to the following two domain-specific challenges:

User subjectivity. Unlike CV and NLP tasks, recommendation targets are inherently subjective, as the same movie may perfectly match one user's interests while being entirely unappealing to another. Without explicit modality alignment, a uniform alignment strategy that optimizes representations toward globally consistent targets inevitably weakens the model's ability to capture personalized user preferences due to gradient opposition. This effect can be empirically validated by segmenting users based on interest consistency during evaluation, and by introducing a unified recommendation encoder as a baseline for comparison (Sec 4.4.2). When user interest consistency is low, models trained with uniform alignment strategies suffer greater performance degradation compared to a standalone encoder. This confirms that preference divergence undermines the efficacy of conventional joint training schemes due to target subjectivity.

Interaction sparsity. As LLMs are inherently data-driven, traditional joint optimization schemes tend to fit interaction patterns dominated by high-frequency users and popular items [10, 44]. This creates a severe sample imbalance during training, which is further amplified as sparsity increases. In the sparsity experiment (Sec 4.4.3), we group users by interaction sparsity and include a standalone recommendation encoder as a baseline. Under identical conditions, models trained with a uniform alignment strategy exhibit more pronounced performance degradation than a standalone encoder as interaction data becomes increasingly sparse. This confirms that the prevailing LLM–encoder integration paradigm demonstrates diminished efficacy in modeling sparse or long-tail interactions.

To address these issues, we propose **PALER** (Personalized Alignment for LLM-based Enhanced Recommendation), a unified framework designed to reshape the optimization trajectories. As illustrated in Figure 1, PALER reshapes the optimization trajectories to resolve the aforementioned gradient opposition and sample imbalance through two core components:

- **Modality-Invariant Module.** To resolve the gradient opposition (Figure 1, left), this module aligns the behavioral representations from the CF encoder with the textual semantic space of the LLM via a contrastive objective (Figure 1, middle). Specifically, it projects CF outputs into a shared latent space and minimizes an InfoNCE loss to contrast textual and behavioral embeddings of the same item against negative samples. By decoupling this explicit alignment from conventional end-to-end fusion, the module prevents the degradation of personalized behavioral signals. Consequently, it establishes a representation space that is both semantically coherent and inherently user-specific.
- **Adaptive Reweighting Module.** We introduce an optimization mechanism that dynamically adjusts user-level learning weights based on individual training feedback. During training, the system monitors each user's prediction loss and updates their corresponding weight using an exponential moving average (EMA), where users with greater difficulty (e.g., sparse or atypical behaviors) are upweighted to receive more learning focus. This

mechanism mitigates the overfitting to hard samples (Figure 1, right), thereby enhancing the model's performance for long-tail users.

Our key contributions are summarized as follows:

- We identify the overlooked problem of personalized alignment in LLM-based recommendation through empirical studies, revealing that such structural degradation is attributed to user subjectivity and data sparsity.
- We propose the PALER framework, which integrates a modality-invariant alignment strategy with an adaptive reweighting mechanism to jointly optimize personalized representations that capture user subjectivity and enhance robustness against interaction sparsity, leading to improved recommendation performance.
- We conduct extensive experiments on real-world datasets, demonstrating that PALER consistently outperforms advanced baselines across diverse scenarios.

2 Related Work

The integration of Large Language Models (LLMs) has marked a significant paradigm shift in recommender systems, moving beyond traditional collaborative filtering techniques. The trajectory of this integration has followed a logical progression, starting with direct applications of pre-trained models and evolving toward more sophisticated hybrid architectures.

2.1 Zero-shot and Few-shot Approaches

Early explorations naturally focused on the most direct application: leveraging the vast world knowledge of pre-trained LLMs without any task-specific training. In this zero-shot or few-shot paradigm, user interaction histories and candidate items are formatted into natural language prompts, instructing the model to perform ranking or prediction tasks [19, 25, 47]. While this approach excels in cold-start scenarios, its performance is highly sensitive to prompt design and often fails to capture the nuanced, implicit collaborative patterns that are not explicitly present in the text.

2.2 Fine-tuning LLMs for Recommendation

To address the performance limitations of zero-shot approaches, subsequent research focuses on fine-tuning LLMs on domain-specific data to learn user patterns [4, 34, 56]. Common strategies include training models to directly generate or rerank items, often with explanations [4, 22, 25, 32, 60], and employing instruction tuning on a diverse set of recommendation-related tasks to improve generalization [34]. While fine-tuning significantly improves performance, these text-based models still struggle to fully internalize the crucial collaborative signals embedded in user-item interaction structures.

2.3 Hybrid Models: Fusing LLMs with Collaborative Signals

Recognizing that purely text-based models overlook the rich, latent signals in user-item interaction graphs, the current state-of-the-art has converged on hybrid architectures. This paradigm enriches LLMs by fusing non-textual collaborative information—typically user and item ID embeddings from traditional CF models—directly into the LLM's input space. By providing richer, more fine-grained signals, these hybrid models have demonstrated superior performance [15, 23, 24, 29, 40, 41, 46, 52, 57]. This fusion, however, introduces new challenges in aligning representations from different modalities, a problem that our work directly addresses. Furthermore, the application of these powerful hybrid LLMs is expanding into more dynamic scenarios, including conversational systems [18, 19, 50] and interactive agents [54, 55].

2.4 Representation Alignment in LLM-based Recommendation

The efficacy of LLM-based recommendation systems fundamentally hinges on the precise alignment of heterogeneous modality spaces, specifically bridging collaborative behavioral signals with textual semantics. This alignment paradigm has evolved from rudimentary linear projections toward deep architectural integration, exemplified by CoLLM [58], which pioneered the embedding of collaborative signals directly into the LLM's input space. Subsequent research has expanded this trajectory by systematically investigating internal representation learning mechanisms [41] and developing alignment strategies tailored for controllable recommendation tasks [35]. More recently, the state-of-the-art has shifted toward more robust, multi-stage, and iterative alignment processes. For instance, Semantic Convergence [27] utilizes a two-stage alignment framework with behavioral semantic tokenization to mitigate modality discrepancies, while OneRec [13] introduces iterative preference alignment to synchronize generative outputs with nuanced user intent. Despite these advancements, existing frameworks predominantly rely on a uniform alignment objective, which often fails to maintain representation integrity under high user subjectivity, a limitation addressed by the modality-invariant module in our proposed PALER framework.

2.5 Long-tail Recommendation and Sparsity Adaptation

Interaction sparsity and long-tail distributions remain persistent challenges in recommendation. Traditional approaches primarily rely on heuristic re-weighting or sampling to balance data distribution [1]. Recently, researchers have leveraged LLMs to mitigate these issues via semantic enhancement. For instance, LLMRec [49] employs LLM-driven graph augmentation to synthesize interactions for sparse items, while LLM-ESR [33] enhances long-tailed sequential recommendation through semantic distillation. Unlike these data-level augmentation strategies, PALER introduces an Adaptive Reweighting Module that dynamically prioritizes sparse users based on real-time training feedback via an exponential moving average of loss.

3 Methodology

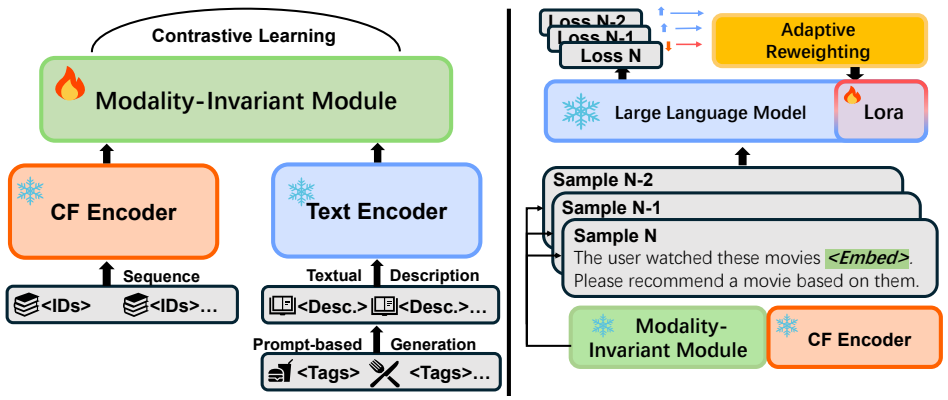

Fig. 2. The architecture of PALER. The left panel illustrates the contrastive alignment, where a CF encoder and a text encoder are aligned in a shared latent space. The right panel shows the prompt-based generation stage, where an LLM is fine-tuned using prompts embedded with the unified representations and adaptive reweighting optimization strategy.

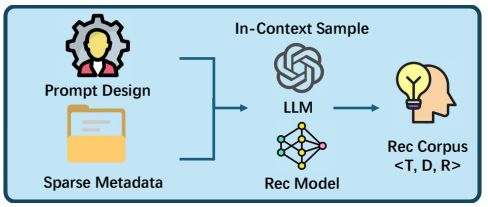

Fig. 3. The instruction construction pipeline. We employ a hybrid strategy to enrich the sparse metadata: an LLM utilizes prompt engineering (ICL & CoT) to generate detailed descriptions (D), while a traditional Rec Model (e.g., SASRec) acts as a teacher to provide reliable recommendation lists (R). These components are combined with original tags (T) to form the final Enriched Corpus $< T, D, R >$.

3.1 Overall Architecture

As illustrated in Figure 2, PALER adopts a two-stage architecture that is bifurcated into a contrastive alignment stage (left panel) and a prompt-based generation stage (right panel). The initial stage pre-trains a modality-invariant module via a dual-stream framework that aligns collaborative filtering (CF) sequences and semantic textual descriptions into a shared latent space through contrastive learning. In the subsequent stage, the output of the CF encoder is mapped through the pre-trained MIM to generate a unified embedding. This embedding is then integrated into the textual prompt as a soft token, which is denoted as $\langle Embed \rangle$. The LLM is then fine-tuned via LoRA using an adaptive reweighting strategy. This strategy recalibrates the loss contribution of individual samples during training to prioritize optimization on more challenging sequences characterized by low interest consistency.

3.2 Corpus preparation

To address the issue of sparse metadata, we first enrich item information by employing a prompt-based LLM method that utilizes In-Context Learning (ICL) [7] and Chain-of-Thought (CoT) [48] to transform minimal tags into comprehensive textual descriptions. As illustrated in Figure 3, We identify pipeline for corpus enrichment. Specifically, we developed three key aspects for building these instructions: Tags, Descriptions, and Recommendations. By structuring our training data around these aspects, we guide the model to understand the relationships between item attributes and user preferences more effectively. Building upon these principles, we now present the detailed construction process of the corpus.

3.2.1 Prompt-Based Description Generation. Our methodology begins with addressing the common problem of data sparsity in recommendation datasets, where item information is often limited to a set of tags. To create a richer understanding of items and better model user preferences, we developed a prompt-based description generation technique. This technique leverages the In-Context Learning (ICL) and Chain-of-Thought (CoT) reasoning abilities of LLMs to convert sparse tags into detailed natural language descriptions.

Specifically, for an item i with an incomplete set of tags, represented as $\tau_i = \{t_1, t_2, ..., t_n\}$, we use a predefined prompt template. This template combines the raw tags with contextual examples to steer the LLM towards generating a comprehensive description, T_i. The generation function is formally defined as:

$$T_i = P(\tau_i, \Phi). \tag{1}$$

Here, $P(\cdot)$ is the LLM's generation function guided by the prompt Φ. The prompt is designed to first instruct the model to summarize the item's core attributes (e.g., name, category) and then

use contextual examples to infer supplementary features like background or storyline, thereby achieving information completion.

3.2.2 Key Aspects in Instructions. We designed a flexible instruction format based on a multi-level data generation strategy. This involves an item-level analysis to generate rich representations, which are then used to construct multi-faceted user profiles. We identified three fundamental building blocks for these user-centric instructions:

- **Tag (T):** This represents the original, discrete metadata for an item, such as genre or keywords, expressed as a set $T_i = \{t_1, t_2, ..., t_n\}$. While essential, tags alone often fail to capture the full characteristics of an item.
- **Description (D):** This refers to the expanded, natural language narrative, D_i, generated from the sparse tags using our ICL and CoT prompting method. These descriptions provide a much deeper content analysis, including details on themes, style, and a summary, revealing more nuanced attributes of the item.
- **Recommendation (R):** This is the final recommendation list, R_i, which is the target output for a given user. To ensure the quality and reliability of our training targets, we use a label augmentation strategy where a traditional recommender model (SASRec) acts as a "teacher" to produce high-quality, behavior-based recommendation lists.

3.3 Modality Alignment for LLM–ID Fusion

Due to the subjectivity of user preferences, we have demonstrated that the current paradigm -which directly fuses LLM-based text representations with ID embeddings - performs particularly poorly in scenarios where user interests are inconsistent (see detailed results in Section 4). Since users' historical interactions already reflect their preference patterns, which should be effectively captured by the ID encoder, we attribute this poor performance to the degradation of the ID encoder caused by naive alignment. This naive combination of two modalities often leads to gradient opposition [51] and geometric decay [53] of the representation space. This section first empirically establishes the existence of these issues and then proposes a modality-invariant module based on a contrastive alignment framework, which is explicitly designed to address these two issues and restore the ID encoder's ability to represent user preferences under high subjectivity.

We instantiate this modality-invariant module with an asymmetric, decoupled alignment strategy that fundamentally differs from conventional joint training. Specifically, the CF encoder is retained as a frozen behavioral backbone that captures user-dependent preference patterns. Instead of updating the CF encoder alongside the LLM, we introduce a lightweight modality-invariant module (MIM) that is trained independently to project CF representations into the semantic space established by the text encoder. This decoupled formulation prevents recommendation gradients from distorting the behavioral representation space and enables MIM to specialize in learning a semantics-consistent mapping, which is crucial for preserving user subjectivity under high-variance preference patterns.

3.3.1 Geometric Decay: The Inherent Flaw of Naive Fusion.
Directly fusing a CF encoder with a text encoder incentivizes shortcut learning [20]. The model learns to rely on superficial features (e.g., item popularity) instead of shared semantics. This drifts the shared representation geometry, making embeddings overly concentrated and less separable across items and users—an instance of representation degradation akin to "geometric decay" discussed in representation learning [53]. As this geometry deteriorates, optimization signals from the two modalities begin to diverge.

This issue ultimately manifests as a conflict at the gradient level. For a multi-modal model with a CF encoder f_{CF} and a text encoder f_{txt}, a downstream task loss L_{Task} produces separate gradient

components, i.e., g_{CF} and g_{txt} on their shared parameters. We found that these gradients often conflict, which can be attributed to:

Assumption 1 (Gradient Opposition). *For a given sample, if the gradient components g_{CF} and g_{txt} satisfy*

$$\cos(g_{\text{CF}}, g_{\text{txt}}) < \epsilon, \tag{2}$$

where ϵ is a small positive constant, then the modalities are in a state of gradient opposition.

This persistent conflict directly degrades the model's representational power, a consequence formalized in the following theorem. Here, we use $I(\cdot;\cdot)$ to denote the mutual information [42] between a representation and the target variable Y, which represents the ground-truth next item.

Theorem 1 (Degradation of Representation Power). *Let $Z_{\text{CF}}^{\text{fused}}$ and $Z_{\text{txt}}^{\text{fused}}$ be the representations from a naively fused framework, and $Z_{\text{CF}}^{\text{indep}}$ and $Z_{\text{txt}}^{\text{indep}}$ be from independently trained encoders. Then*

$$I(Z_{\text{CF}}^{\text{fused}}; Y) < I(Z_{\text{CF}}^{\text{indep}}; Y), I(Z_{\text{txt}}^{\text{fused}}; Y) < I(Z_{\text{txt}}^{\text{indep}}; Y). \tag{3}$$

This result reveals that naive fusion can lead to a measurable degradation in the mutual information between each modality's representation and the target, indicating a loss of discriminative power for both CF and textual features. By contrast, our method is explicitly designed to alleviate such degradation through altering the training objectives. It preserves modality-specific information so as to enhance downstream performance.

Now we want to provide the rigid mathematical proof for Theorem 1.

Proof. Decompose the mutual information of the fused representation using the chain rule:

$$I(Z^{\text{fused}}; Y) = I(Z_{CF}^{\text{fused}}; Y) + I(Z_{\text{txt}}^{\text{fused}}; Y \mid Z_{CF}^{\text{fused}}). \tag{4}$$

This separates the contribution of the CF component and the text component (conditioned on CF) toward explaining the target Y. Under the gradient opposition Assumption 1, the update for the shared parameters θ becomes:

$$\Delta\theta = -\eta\big(g_{CF} + g_{\text{txt}}\big), \tag{5}$$

with the squared norm bounded by

$$\|g_{CF} + g_{\text{txt}}\|_2^2 \le \big(\|g_{CF}\|_2 + \|g_{\text{txt}}\|_2\big)^2 - 2(1-\epsilon)\|g_{CF}\|_2\|g_{\text{txt}}\|_2. \tag{6}$$

This inequality quantifies how the conflicting directions (measured by the small cosine similarity) strictly bound the squared magnitude of the effective update, hence distorting the learned representation. Let $p^{\text{indep}}(z \mid y)$ denote the representation distribution obtained from independently trained (unperturbed) models. Under fusion, we have a perturbed version:

$$p^{\text{fused}}(z \mid y) = p^{\text{indep}}(z \mid y) + \Delta p(z \mid y), \tag{7}$$

where $\Delta p(z \mid y)$ quantifies the distortion induced by gradient conflict. Note that the magnitude of $\Delta p(z \mid y)$ is linked to the conflict intensity:

$$\|\Delta p(z \mid y)\|_1 \propto 1 - \cos(g_{CF}, g_{\text{txt}}). \tag{8}$$

By a standard property of KL divergence (a version of Pinsker's inequality [**?**]), we have

$$D_{KL}\Big(p^{\text{indep}}(z \mid y)\Big\|p^{\text{fused}}(z \mid y)\Big) \ge \frac{1}{2}\|\Delta p(z \mid y)\|_1^2. \tag{9}$$

Taking an expectation over Y we obtain:

$$\mathbb{E}_Y\left[D_{KL}\Big(p^{\text{indep}}(z \mid y)\Big\|p^{\text{fused}}(z \mid y)\Big)\right] \ge \frac{1}{2}\mathbb{E}_Y\left[\|\Delta p(z \mid y)\|_1^2\right]. \tag{10}$$

By the continuity of mutual information, the information loss caused by the perturbation can be lower-bounded. We can express the mutual information of the fused representation as being strictly bounded above by the independent one minus a penalty proportional to this perturbation:

$$I(Z^{\text{fused}}; Y) \leq I(Z^{\text{indep}}; Y) - c \cdot \mathbb{E}_Y \left[D_{KL} \left(p^{\text{indep}}(z \mid y) \middle\| p^{\text{fused}}(z \mid y) \right) \right], \tag{11}$$

where $c > 0$ is a constant dependent on the representation space geometry. Substitute the inequality (10) into equation (11):

$$I(Z^{\text{fused}}; Y) \leq I(Z^{\text{indep}}; Y) - \frac{c}{2} \mathbb{E}_Y \left[\|\Delta p(z \mid y)\|_1^2 \right]. \tag{12}$$

Since the gradient conflict implies a nonzero perturbation, when the conflict is severe ($\epsilon \to 0$), we have

$$\|\Delta p(z \mid y)\|_1 > 0, \tag{13}$$

it follows that:

$$I(Z^{\text{fused}}; Y) < I(Z^{\text{indep}}; Y). \tag{14}$$

This proof shows that gradient conflict inevitably introduces a perturbation to the conditional representation distribution, which translates into a quantifiable loss of mutual information with the target. Consequently, naive fusion under severe conflict leads to degraded representation quality. Our method, by mitigating gradient conflict, effectively suppresses such perturbations and preserves the mutual information, thereby yielding stronger downstream performance.

3.3.2 *Contrastive Alignment as a Geometric Regularizer.*

To counteract this geometric decay, we propose a modality-invariant representation learning approach that uses contrastive learning as a powerful geometric regularizer [6]. The core idea is to disincentivize shortcut learning by enforcing a strict cross-modality alignment objective. By forcing the model to pull representations of the same item from different modalities together, it must learn their shared semantic essence rather than relying on superficial, modality-specific features.

This process actively promotes a more isotropic (uniform) representation space, which in turn mitigates geometric decay and preserves the information capacity required for true alignment [17]. Our architecture employs a dual-stream framework to implement this strategy. A CF encoder f_{CF} and a text encoder f_{txt} separately process the inputs. Their outputs are then projected into a shared latent space by a modality-invariant module, which is optimized via the InfoNCE loss[38]. For a batch of B items, the loss is defined as:

$$L_{\text{item}} = -\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(z_i^{\text{ID}}, z_i^{\text{txt}})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(z_i^{\text{ID}}, z_j^{\text{txt}})/\tau\right)}, \tag{15}$$

where z_i^{ID} and z_i^{txt} are the representations for item i; $\text{sim}(\cdot, \cdot)$ denotes a similarity metric (e.g., cosine similarity); and τ is a temperature hyperparameter. By minimizing L_{item}, the model reshapes the embedding space to be more geometrically sound, pulling positive pairs together and pushing apart negative pairs.

3.3.3 *Theoretical Analysis: From Geometric Decay to Gradient Opposition.*

As established earlier, naive fusion between the CF encoder and the text encoder distorts the shared representation space, leading to geometric decay that undermines the model's ability to capture fine-grained, cross-modal semantics. This degradation ultimately manifests as gradient opposition between modalities during optimization, where updates beneficial to one modality may conflict with those needed by the other.

In this section, we theoretically show that our proposed contrastive alignment framework addresses geometric decay (Lemma 1) and, as a direct consequence, alleviates gradient opposition in downstream tasks (Theorem 2).

First, we formalize how the proposed contrastive objective counteracts geometric decay by explicitly reducing the distance between representations of the same item across modalities. This enforces a consistent shared geometry, preventing the drift and concentration effects and lays the foundation for mitigating gradient opposition.

Lemma 1 (Representation Alignment Property of the Contrastive Objective). *Let f_{MIM} be the mapping learned by the modality-invariant module, and let $d(\cdot,\cdot)$ be a distance metric in the latent space. Optimizing L_{item} is equivalent to solving*

$$\min_{\theta_{\text{MIM}}} \sum_{i\in\text{Batch}} d\big(f_{\text{MIM}}(z_i^{\text{ID}}), f_{\text{MIM}}(z_i^{\text{txt}})\big). \tag{16}$$

Building on this result, we next examine its implications for gradient behavior and prove that alleviating geometric decay leads to reduced gradient opposition.

Theorem 2 (Reducing Gradient Opposition through Contrastive Learning). *Let M_{con} be a model pre-trained with the contrastive objective described in Lemma 1. Assume that g_{CF} and g_{txt} satisfy Assumption 1. When M_{con} is fine-tuned on a downstream task, the corresponding task gradients g_{CF} and g_{txt} satisfy*

$$\mathbb{E}_{(u,i)\sim\mathcal{D}}\big[\cos(g_{\text{CF}}, g_{\text{txt}})\big] \geq \gamma, \tag{17}$$

where γ is a constant close to 1, and $\mathcal{D}$ denotes the data distribution.

This result demonstrates that contrastive pre-training enforces a high degree of gradient alignment between modalities during downstream fine-tuning, ensuring that optimization steps benefit both representations simultaneously. Such alignment mitigates gradient conflict and facilitates more effective multi-modal fusion, enabling the CF encoder to better capture user subjectivity and ultimately leading to improved task performance. Therefore, this theorem provides a theoretical justification for the empirical superiority of our method.

We are given the InfoNCE loss for a positive pair $(z_i^{\text{ID}}, z_i^{\text{txt}})$ as

$$\mathcal{L}_{\text{item}} = -\mathbb{E}\log\frac{e^{s(z_i^{\text{ID}}, z_i^{\text{txt}})/\tau}}{\sum_j e^{s(z_i^{\text{ID}}, z_j^{\text{txt}})/\tau}}, \tag{18}$$

where $s(\cdot,\cdot)$ is a similarity function (often an inner product) and τ is a temperature parameter. Minimizing $\mathcal{L}_{\text{item}}$ forces the representations to be aligned in the shared space, driving the Euclidean distance between them toward zero:

$$f_{\text{MIM}}(z_i^{\text{ID}}) \approx f_{\text{MIM}}(z_i^{\text{txt}}). \tag{19}$$

To formalize gradient alignment, we must evaluate the gradients of the downstream task loss $\mathcal{L}_{\text{task}}$ with respect to these shared representations, rather than the disparate network parameters, to ensure dimensional compatibility. Let the aligned representations be denoted as $\tilde{z}_{CF} = f_{\text{MIM}}(z_i^{\text{ID}})$ and $\tilde{z}_{txt} = f_{\text{MIM}}(z_i^{\text{txt}})$. We define the representation-level gradients as:

$$g_{CF} \triangleq \nabla_{\tilde{z}_{CF}}\mathcal{L}_{\text{task}}, \tag{20a}$$

$$g_{txt} \triangleq \nabla_{\tilde{z}_{txt}}\mathcal{L}_{\text{task}}. \tag{20b}$$

Assuming the downstream loss function $\mathcal{L}_{\text{task}}$ is continuously differentiable and locally smooth, evaluating its gradient at two points that are geometrically proximate yields highly similar gradient

vectors. Since contrastive pre-training ensures $\|\tilde{z}_{CF} - \tilde{z}_{txt}\|_2 \to 0$, it naturally follows that:

$$\nabla_{\tilde{z}_{CF}} \mathcal{L}_{\text{task}} \approx \nabla_{\tilde{z}_{txt}} \mathcal{L}_{\text{task}}, \tag{21}$$

which simplifies to $g_{CF} \approx g_{txt}$. The cosine similarity between these representation-level gradient vectors is defined as:

$$\cos(g_{CF}, g_{txt}) = \frac{\langle g_{CF}, g_{txt} \rangle}{\|g_{CF}\|_2 \|g_{txt}\|_2}. \tag{22}$$

Because the vectors are approximately equal in both magnitude and direction within the shared latent space, their inner product approaches the product of their respective magnitudes:

$$\frac{\langle g_{CF}, g_{txt} \rangle}{\|g_{CF}\|_2 \|g_{txt}\|_2} \to 1. \tag{23}$$

Thus, there exists some $\gamma \sim 1$ such that:

$$\mathbb{E}\left[\cos(g_{CF}, g_{txt})\right] \geq \gamma. \tag{24}$$

This proof establishes that contrastive pre-training via the InfoNCE loss drives the two modalities' representations to be closely aligned in the shared space. As a result, the downstream task gradients with respect to these shared representations become highly correlated, yielding a cosine similarity close to one. This strong gradient alignment mitigates optimization conflict and supports more effective multi-modal learning, thereby explaining the empirical advantages of our method.

3.4 Adaptive Reweighting for Personalized Optimization

Beyond the representation-level challenges caused by user subjectivity, a core issue in LLM-based recommendation stems from interaction sparsity. The standard Empirical Risk Minimization (ERM) framework, by optimizing for average performance, naturally leads models to overfit the interaction patterns of high-frequency users while neglecting the long tail. To address this tendency in sparse-interaction scenarios, we introduce our adaptive reweighting strategy. This approach is designed to achieve a more personalized optimization by ensuring the model learns effectively from the entire user distribution.

3.4.1 Limitations of Uniform Optimization.

Real-world user interaction data is characterized by a long-tail distribution [12]. A significant proportion of users have sparse or noisy interaction histories, which can be characterized as having a low signal-to-noise ratio (low-SNR). The standard Empirical Risk Minimization (ERM) framework, by optimizing for the average loss, is naturally skewed by this structure. Consequently, the model tends to overfit the dominant patterns of high-frequency users while underfitting the nuanced preferences of these low-SNR users. This leads to a parameter solution θ^*_{ERM} that is suboptimal for this large user subset, failing to deliver effective personalization, which motivates a more adaptive optimization strategy.

3.4.2 The Adaptive Reweighting Strategy.

To counteract the limitations of the ERM framework, we introduce the adaptive reweighting strategy. The core idea is to enhance personalization during training by dynamically adjusting each user's contribution to the total loss. This compels the model to allocate sufficient learning resources to users within the long tail, thereby mitigating the underfitting issue caused by interaction sparsity. The implementation proceeds as follows:

(1) **Compute Individual Losses:** For each user in a training batch, compute the loss values $L(u_1), L(u_2), \ldots, L(u_n)$ produced by the LLM recommendation module.
(2) **Solve for Optimal Weights:** Given these losses, solve the regularized optimization defined in Lemma 2 to obtain the optimal weight coefficients β^*_i for each user in the batch.

(3) **Weight Smoothing:** To prevent drastic fluctuations between batches, apply an Exponential Moving Average (EMA) to smooth the weights over iterations. At iteration t, update:

$$\beta_i^{(t)} = (1 - \eta_t)\, \beta_i^{(t-1)} + \eta_t\, \beta_i^*, \tag{25}$$

where the decay rate η_t decreases as training progresses.

(4) **Update with a Weighted Loss:** Construct the final objective as a weighted loss:

$$L_{\text{weighted}} = \sum_{i=1}^{n} \beta_i^{(t)}\, L(u_i) \tag{26}$$

Importantly, to maintain training efficiency and prevent catastrophic forgetting, we freeze the LLM backbone parameters and exclusively update the LoRA parameters, denoted as Θ_{LoRA}. The gradient update step explicitly incorporates the adaptive weights:

$$\Theta_{\text{LoRA}}^{(t+1)} \leftarrow \Theta_{\text{LoRA}}^{(t)} - \gamma \cdot \nabla_{\Theta_{\text{LoRA}}} \left(\sum_{i=1}^{n} \beta_i^{(t)} L(u_i) \right), \tag{27}$$

where γ is the learning rate. This weighted gradient flow ensures that the optimization trajectory is proactively corrected by high-difficulty samples (i.e., sparse-interaction users), rather than being dominated by the average patterns of mainstream users.

Algorithm 1 Adaptive Reweighting Optimization Strategy

Require: Training data $\mathcal{D}$, model parameters θ, batch size B, smoothing schedule $\{\eta_t\}$
Ensure: Updated model parameters θ

for iteration $t = 1, 2, \ldots$ **do**
  Sample a batch of users $\{u_1, \ldots, u_B\} \subset \mathcal{D}$
  for each user u_i in batch **do** ▷ *Individual loss for each user*
    Compute loss $L_i \leftarrow L(u_i; \theta)$
  end for
  $\beta^* \leftarrow \underset{\beta_i \geq 0, \sum_i \beta_i = K}{\arg\max} \left(\max_i \beta_i L_i - \alpha \sum_i \beta_i^2 \right)$ ▷ *Solve for optimal weights via Lemma 3*
  for each user u_i in batch **do** ▷ *Smooth with EMA*
    $\beta_i \leftarrow (1 - \eta_t)\, \beta_i + \eta_t\, \beta_i^*$
  end for
  $L_{\text{fair}} \leftarrow \sum_{i=1}^{B} \beta_i\, L_i$ ▷ *Compute weighted loss*
  $\theta \leftarrow \theta - \lambda\, \nabla_\theta L_{\text{fair}}$ ▷ *Update model*
end for

3.4.3 *Theoretical Analysis of the Reweighting Strategy.*

The effectiveness of our proposed adaptive reweighting strategy is not merely empirical but is also supported by a solid theoretical foundation. The core principle and resulting performance improvements are rigorously described by the following lemma and theorem.

First, we establish that the designed weight-solving method has a key monotonic property linking each weight to its loss. In the optimization problem below, we introduce two key hyperparameters: a regularization coefficient $\alpha > 0$ that controls the smoothness of the weight distribution, and a constant $K > 0$ that fixes the total magnitude of the weights, which is typically set to the batch size n.

Lemma 2 (Loss Monotonicity of Adaptive Weights). *Consider the regularized optimization problem to solve for weights β:*

$$\begin{aligned} \beta^* = \underset{\beta}{\operatorname{argmax}} \left(\max_{i=1,\ldots,n} \beta_i L_i - \alpha \sum_{i=1}^{n} \beta_i^2 \right) \\ s.t. \quad \sum_{i=1}^{n} \beta_i = K, \quad \beta_i \geq 0. \end{aligned} \tag{28}$$

Its optimal solution β_i^* is a monotonically increasing function of the corresponding loss L_i.

Building on this key property, we now demonstrate that our reweighting method leads to a provable performance improvement for the low-SNR user group.

Theorem 3 (Loss Improvement on Low-SNR Users). *Let $\theta^*_{\text{reweight}}$ be the model parameters obtained by optimizing our weighted loss L_{weighted}, and let θ^*_{ERM} be those from standard ERM. Then*

$$\sum_{u \in U_{\text{hard}}} L(\theta^*_{\text{reweight}}) < \sum_{u \in U_{\text{hard}}} L(\theta^*_{\text{ERM}}). \tag{29}$$

This theoretical confirms that our adaptive reweighting strategy offers a principled, provable method to address ERM's suboptimality via more personalized learning.

Next we want to provide the proof sketch for Theorem 3.

Proof. The standard ERM objective is

$$L_{\text{ERM}}(\theta) = \sum_{u \in U} L_u(\theta), \tag{30}$$

and the fairness weighted loss is

$$L_{\text{fairness}}(\theta) = \sum_{u \in U} w(u)\, L_u(\theta), \tag{31}$$

where the weight function $w(u)$ assigns higher weight to low-SNR (or "hard") users, i.e., $w(u) > 1 \quad \text{for} \quad u \in U_{\text{hard}}$. Let

$$\theta^*_{\text{ERM}} = \arg\min_{\theta} L_{\text{ERM}}(\theta) \tag{32}$$

and

$$\theta^*_{\text{fairness}} = \arg\min_{\theta} L_{\text{fairness}}(\theta). \tag{33}$$

Since the fairness objective puts more emphasis on users in U_{hard}, the optimization drives the loss for these users lower. Moreover, because $\theta^*_{\text{fairness}}$ minimizes the weighted loss, we have

$$L_{\text{fairness}}(\theta^*_{\text{fairness}}) \leq L_{\text{fairness}}(\theta^*_{\text{ERM}}). \tag{34}$$

Expanding the weighted losses, we write

$$\sum_{u \in U} w(u) L_u(\theta^*_{\text{fairness}}) \leq \sum_{u \in U} w(u) L_u(\theta^*_{\text{ERM}}). \tag{35}$$

Splitting Equation (35) into hard and easy user sets, we get

$$\begin{aligned} &\sum_{u \in U_{\text{hard}}} w(u) L_u(\theta^*_{\text{fairness}}) + \sum_{u \in U \setminus U_{\text{hard}}} w(u) L_u(\theta^*_{\text{fairness}}) \\ &\leq \sum_{u \in U_{\text{hard}}} w(u) L_u(\theta^*_{\text{ERM}}) + \sum_{u \in U \setminus U_{\text{hard}}} w(u) L_u(\theta^*_{\text{ERM}}). \end{aligned} \tag{36}$$

Since $w(u) > 1$ only for $u \in U_{\text{hard}}$, even if the losses for the easy users change slightly, the heavy weighting on the hard users forces the inequality

$$\sum_{u \in U_{\text{hard}}} L_u(\theta^*_{\text{fairness}}) < \sum_{u \in U_{\text{hard}}} L_u(\theta^*_{\text{ERM}}). \tag{37}$$

This proof sketch shows that by assigning higher weights to hard users in the optimization objective, the fairness-aware training procedure explicitly prioritizes reducing their losses. As a result, the solution $\theta^*_{\text{fairness}}$ achieves strictly lower loss on the hard-user subset compared to the standard ERM solution, thereby validating the effectiveness of our method in improving performance for disadvantaged groups. □

4 Experiments

To systematically evaluate the effectiveness of our proposed PALER framework, we design a comprehensive experimental study addressing the following four research questions:

- **Representation Quality and Robustness:** To what extent does naive multimodal fusion damage personalized representations, and how does PALER mitigate the representation degradation? (RQ1)
- **Evaluation of Overall Performance:** Does PALER outperform state-of-the-art recommendation baselines, including both traditional ID-based and LLM-enhanced frameworks? (RQ2)
- **Ablation Study:** What are the individual contributions of the modality-invariant module and adaptive reweighting module to the overall performance of PALER? (RQ3)

4.1 Datasets

We conduct experiments on three real-world datasets:

(1) **Yelp:** A dataset originates from a well-known business review website, providing rich record for user behaviors and business attributes for the restaurant.
(2) **Movies&TV:** A large-scale dataset of product reviews from Amazon, including user ratings and detailed textual feedback for movies and television shows.
(3) **Steam:** A dataset comprised of user reviews for video games on the Steam platform, which includes review texts, user activity, and game metadata.

Detailed dataset statistics are provided in Table 1.

Table 1. Statistics of the datasets employed in our experiments, including user and item scales, total interactions, and overall sparsity.

Dataset	# Users	# Items	# Actions	Sparsity
Yelp	10,684	11,544	1.05M	99.15%
Movies&TV	152,776	65,631	2.94M	99.97%
Steam	69,993	27,560	0.69M	99.96%

4.2 Evaluation Metrics

Following established practices in sequential recommendation, we adopt a leave-one-out evaluation strategy. We use two widely adopted metrics: Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) for accuracy measurement.

(1) **HR@K:** Measures the fraction of cases that the ground-truth next item appears among the top K recommendations, over the given m candidates, where $K \leq m$.

(2) **NDCG@K:** Evaluates ranking quality by assigning higher importance to correct items that appear among the top K recommendations.

4.3 Baselines

We compare our model against representative baselines from two main categories.

- *Traditional ID-based Recommenders:* These models capture user behavior sequences from item IDs. Our selection includes SASRec [26], BERT4Rec [43], S^3-Rec [61], and Caser [45].
- *LLM-based Recommenders:* This category covers methods that leverage large language models for recommendation tasks. We compare against TALLRec [5], LLMRank [25], RecLM [35], LlamaRec [52], LLaRA [31], and CoLLM [58].

To enable a fair and transparent comparison, we summarize the key characteristics of each baseline as follows.

(1) **SASRec:** [26] A causal sequential model using a unidirectional transformer to predict the next item in a sequence of IDs.
(2) **BERT4Rec:** [43] A sequential method using a bidirectional transformer to learn user behavior sequences for recommendations.
(3) **S^3-Rec:** [61] A self-supervised learning approach, primarily leveraging Mutual Information Maximization (MIM) during its pre-training phase to learn from the correlations between different elements.
(4) **Caser:** [45] A sequential method treats recent item sequence embeddings as an 'image,' using horizontal convolutional filters to capture union-level patterns.
(5) **TALLRec:** [5] A framework that fine-tunes LLMs for recommendation tasks, aligning pre-trained models with recommendation.
(6) **LLMRank:** [25] An LLM-based recommendation system that pairs with sequential models with careful prompting.
(7) **RecLM:** [35] A framework combining supervised and reinforcement learning to improve LLMs' instruction-following abilities and generalize across various recommendation tasks.
(8) **LlamaRec:** [52] A two-stage recommendation framework based on LLMs. It first retrieves candidate items using an efficient sequential recommender and then ranks them using an LLM.
(9) **Llara** [31] A LLM-based sequential recommendation by combining behavioral patterns from traditional models and a hybrid text prompt representation.
(10) **CoLLM** [58] A method enhances LLMs for recommendation by converting collaborative embeddings into text-like sequence prompts.

4.4 Representation Quality and Robustness (RQ1)

This section addresses RQ1 by investigating the representation degradation caused by naive fusion. We conducted interconnected experiments examining core challenges in LLM-based recommendation systems, demonstrating the necessity of PALER's design.

4.4.1 The Probe Experiment.

We designed a 'probe experiment' to validate our core hypothesis: that naive fusion degrades the performance of the recommendation encoder. The experiment quantifies this degradation by isolating and assessing the quality of representations generated by an encoder trained within the fused model. Our probe experiment reveals that naive fusion between LLMs and domain-specific encoders introduces destructive interference, degrading the quality of task-specific representations. By comparing standalone CF encoders (e.g., SASRec and Caser) with their representations in a naively fused multimodal model, as summarized in Table 2, we observe significant performance

drops: In the Yelp dataset, SASRec's HR@10 decreases by 24.2% and Caser's by 28.5% compared to their independent baselines. This degradation occurs regardless of the CF architecture (transformer-based or CNN-based), indicating the issue stems from the fusion strategy itself rather than model-specific limitations. The results confirm that cross-modal feature interactions during naive fusion disrupt the CF encoder's ability to capture user-specific behavioral patterns, leading to geometric decay in the shared latent space. These findings underscore the critical need for PALER's modality-invariant alignment and adaptive reweighting mechanisms to preserve representation quality and mitigate the negative impact of sparsity on recommendation performance To support the findings discussed above, we provide the detailed probe experiment protocol below.

Objective. The experiment is designed to quantitatively measure the degradation in a Collaborative Filtering (CF) encoder's representation quality when it is trained via naive end-to-end fusion with a text modality, compared to being trained independently.

Architectures Under Test. To ensure the generality of our findings, we test two sequential recommendation models with distinct underlying architectures:

- **SASRec**: A model based on the Transformer architecture, representing self-attention mechanisms.
- **Caser**: A model based on Convolutional Neural Networks (CNNs), representing convolutional filter-based methods.

Experimental Steps. The experiment follows a three-step protocol for each CF architecture:

Step 1: Establishing the Baseline. We first train a standalone version of the CF model (e.g., SASRec) on the Yelp dataset. The model is fully optimized for the recommendation task. We then evaluate its performance (NDCG@10) on the test set. This result serves as the "gold standard" or "undisturbed" performance benchmark for that architecture.
Step 2: Training the Naive Fusion Model. We construct a multimodal model where the CF encoder (e.g., SASRec) and a text encoder are combined. Their outputs are fused and fed to a prediction head, and the entire model is trained end-to-end on the same dataset. This represents the "naive fusion" condition described in Section 3.4.
Step 3: Probing the Internal Representations. After the naive fusion model has been trained, we freeze all of its parameters. We then extract the intermediate representation vectors produced by the internal CF encoder module. A new, lightweight classification head is attached to these frozen representations and is trained on the original recommendation task. The final performance of this head on the test set measures the inherent quality of the "disturbed" CF representations. The performance drop is calculated by comparing this result against the baseline established in Step 1.

4.4.2 *User Subjectivity Analysis.*

To evaluate the impact of user subjectivity on recommendation performance, we partition users by interest consistency and compare results under different alignment strategies. Interest consistency is quantified using a category-diversity metric $D(u)$, which represents the number of unique category labels in a user's interaction history. To ensure a distinct contrast between focused and divergent preferences, we employ an extreme group design based on the quartiles of $D(u)$. Specifically, we isolate the High-Consistency group (those below the first quartile, $D(u) \leq Q_1$) and the Low-Consistency group (those above the third quartile, $D(u) \geq Q_3$), while excluding the middle 50% of users from this analysis.

Table 2. Extended comparison of representation quality, showing performance recovery by our proposed alignment method. Best performance is marked by Standalone. Our model's results are underlined for emphasis.

CF Encoder Architecture	Dataset	Method	HR@10		NDCG@10	
			Value	Change(%)	Value	Change(%)
SASRec	Yelp	Standalone	0.7650	-	0.4323	-
		Probed (Naive Fusion)	0.5801	-24.2	0.3373	-22.0
		Probed (Ours)	0.7243	+78.0 Rec.	0.4152	+82.0 Rec.
	Steam	Standalone	0.8793	-	0.6340	-
		Probed (Naive Fusion)	0.6267	-28.7	0.5027	-20.7
		Probed (Ours)	0.8190	+76.1 Rec.	0.6121	+83.3 Rec.
	Movies	Standalone	0.7262	-	0.5108	-
		Probed (Naive Fusion)	0.5238	-27.9	0.3936	-23.0
		Probed (Ours)	0.6791	+76.7 Rec.	0.4907	+82.9 Rec.
Caser	Yelp	Standalone	0.6780	-	0.4246	-
		Probed (Naive Fusion)	0.4847	-28.5	0.3271	-23.0
		Probed (Ours)	0.6350	+77.8 Rec.	0.4065	+81.4 Rec.
	Steam	Standalone	0.5339	-	0.3283	-
		Probed (Naive Fusion)	0.3806	-28.7	0.2494	-24.0
		Probed (Ours)	0.5011	+78.6 Rec.	0.3138	+81.6 Rec.
	Movies	Standalone	0.5162	-	0.3388	-
		Probed (Naive Fusion)	0.3803	-26.3	0.2628	-22.4
		Probed (Ours)	0.4852	+77.2 Rec.	0.3243	+81.0 Rec.

The results in Table 3 demonstrate that low-consistency users consistently yield lower HR@10 scores than high-consistency users across all datasets. This performance deficit is markedly exacerbated by naive fusion strategies. For the Yelp dataset, the performance gap increases from 19.0% in the standalone model to 38.7% under naive fusion. Such trends indicate that uniform alignment fails to capture individual variability within heterogeneous interaction histories, as averaged representations cannot adequately accommodate divergent preferences.

PALER effectively mitigates this fusion-induced degradation through contrastive alignment. Our method achieves recovery ratios of 80.2%, 84.9%, and 66.3% for the Yelp, Steam, and Movies datasets, respectively. These figures quantify the proportion of the performance penalty induced by naive fusion that is successfully recovered by PALER. The substantial recovery ratios across all datasets validate that our modality-invariant module effectively aligns textual and behavioral representations while preserving personalized semantics critical for recommendation.

4.4.3 *Interaction Sparsity Impact.*

To investigate the impact of data sparsity on recommendation robustness, we partition users based on their interaction density and compare performance across different alignment strategies. We isolate two extreme segments for a clear contrast: the Active group, representing the top 25% of users with the most interactions, and the Sparse group, representing the bottom 25% of users with the fewest interactions.

Table 3. Performance comparison across consistency groups. *Gap* is defined as $(P_{High} - P_{Low})/P_{High}$. *Recovery Ratio* is calculated as $(\Delta G_{Naive} - \Delta G_{Ours})/\Delta G_{Naive}$, where $\Delta G = Gap_{Method} - Gap_{Standalone}$ represents the additional performance deficit induced by fusion baseline.

Dataset	Method	HR@10 Performance			Recovery
		High-Consistency	Low-Consistency	Gap (Δ%)	Ratio
Yelp	Standalone	0.7845	0.6354	19.0%	-
	Fused (Naive)	0.7210	0.4420	38.7%	-
	PALER (Ours)	**0.7890**	**0.6080**	22.9%	**80.2%**
Steam	Standalone	0.8120	0.6341	21.9%	-
	Fused (Naive)	0.7550	0.5096	32.5%	-
	PALER (Ours)	**0.8180**	**0.6260**	23.5%	**84.9%**
Movies	Standalone	0.6550	0.5790	11.6%	-
	Fused (Naive)	0.5980	0.4754	20.5%	-
	PALER (Ours)	**0.6590**	**0.5630**	14.6%	**66.3%**

As reported in Table 4, sparse users consistently exhibit lower HR@10 performance compared to active users. This performance drop is significantly exacerbated by naive fusion. For instance, in the Steam dataset, the performance gap between active and sparse users widens from 30.1% in the standalone model to 48.9% under naive fusion. This suggests that simple modality concatenation introduces substantial noise for data-scarce users, further degrading the model's ability to provide accurate recommendations.

PALER effectively reverses this trend, achieving recovery ratios of 104.8%, 101.6%, and 114.1% for the Yelp, Steam, and Movies datasets, respectively. Notably, these ratios exceed 100%, which indicates that PALER does more than merely mitigate the degradation introduced by naive fusion. It actually reduces the performance gap between active and sparse users to a level lower than that of the standalone model. For example, in the Movies dataset, PALER's performance drop is only 20.8%, surpassing the 22.1% achieved by the standalone baseline. These results validate that our modality-invariant alignment not only restores lost performance but also enhances the model's inherent robustness by effectively leveraging cross-modal semantic information to compensate for interaction sparsity.

4.5 Overall Performance Comparison (RQ2)

This section addresses RQ2 by comparing PALER against the established baselines. Our comprehensive evaluation in Table 5 demonstrates the substantial performance gains offered by PALER over existing baselines. The model achieves consistent state-of-the-art performance across three real-world datasets (Yelp, Steam, and Movies&TV) under both HR and NDCG metrics. Notably, PALER outperforms traditional ID-based recommenders like SASRec by significant margins while surpassing contemporary LLM-based competitors including LLaRA and CoLLM, securing top performance across all datasets.

This performance dominance arises from PALER's innovative architecture that bridges the gap between CF and language modeling. Unlike conventional LLM-based approaches that adopt generalized pre-training strategies, our model employs a domain-adaptive instruction tuning framework. This specialized design creates a synergistic integration of two critical components: (1) the fine-grained sequential patterns captured by CF modules, and (2) the deep semantic reasoning capabilities of large language models. The result is a system that not only preserves the precision

Table 4. Performance comparison across interaction density groups. *Drop* is defined as $(P_{Active} - P_{Sparse})/P_{Active}$. *Recovery Ratio* is calculated as $(\Delta D_{Naive} - \Delta D_{Ours})/\Delta D_{Naive}$, where $\Delta D = Drop_{Method} - Drop_{Standalone}$ represents the additional performance degradation induced by the fusion baseline.

Dataset	Method	HR@10 Performance			Recovery
		Active (Top 25%)	Sparse (Bottom 25%)	Drop (Δ%)	Ratio
Yelp	Standalone	0.7798	0.6309	19.1%	-
	Fused (Naive)	0.7347	0.4732	35.6%	-
	PALER (Ours)	**0.7865**	**0.6426**	18.3%	**104.8%**
Steam	Standalone	0.8498	0.5940	30.1%	-
	Fused (Naive)	0.7895	0.4034	48.9%	-
	PALER (Ours)	**0.8520**	**0.5977**	29.8%	**101.6%**
Movies	Standalone	0.6795	0.5293	22.1%	-
	Fused (Naive)	0.6204	0.4262	31.3%	-
	PALER (Ours)	**0.6885**	**0.5455**	20.8%	**114.1%**

Table 5. Performance of PALER and baseline methods. The best and the second-best performance in each column is bolded and underlined, respectively. Statistically significant improvements are marked with †

		Yelp		Steam		Movies	
		NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5
Traditional	SASRec	0.241	0.325	0.251	0.340	0.240	0.322
	BERT4Rec	0.235	0.315	0.246	0.333	0.233	0.314
	S^3-Rec	0.245	0.330	0.255	0.345	0.243	0.328
	Caser	0.228	0.305	0.239	0.324	0.225	0.303
LLM-based	TALLRec	0.261	0.349	0.269	0.362	0.255	0.346
	LLMRank	0.267	0.355	0.273	0.367	0.259	0.351
	RecLM	0.281	0.379	0.290	0.391	0.276	0.369
	LlamaRec	0.283	0.382	0.288	0.388	0.274	0.372
	LLaRA	<u>0.308</u>	<u>0.415</u>	<u>0.318</u>	<u>0.428</u>	<u>0.301</u>	<u>0.399</u>
	CoLLM	0.295	0.398	0.302	0.405	0.288	0.381
Ours	PALER	**0.315**†	**0.421**†	**0.325**†	**0.435**†	**0.309**†	**0.408**†

of traditional recommendation signals but also enhances them with contextual understanding of user preferences. This dual-modality synergy, combined with our adaptive reweighting strategy, explains why PALER demonstrates both superior accuracy and exceptional resilience in challenging recommendation scenarios.

To ensure the statistical reliability, all experiments were conducted over five independent runs with distinct random seeds. Statistical significance was evaluated using a paired t-test between PALER and the strongest baseline, LLaRA. The results confirm that PALER significantly outperforms LLaRA across all datasets and metrics under the 95% confidence level ($p < 0.05$).

Table 6. Ablation study of PALER. We report results on three datasets. "w/o" indicates the removal of a specific component from our full model (PALER). 'Baseline" refers to a simple baseline with all our contributions removed.

Model	Yelp		Steam		Movies	
	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5
PALER	**0.315**	**0.421**	**0.325**	**0.435**	**0.309**	**0.408**
w/o Align	0.301	0.409	0.312	0.419	0.295	0.395
w/o Reweight	0.285	0.388	0.296	0.399	0.279	0.377
w/o Corpus	0.287	0.385	0.294	0.402	0.277	0.380
Baseline	0.259	0.352	0.271	0.370	0.251	0.349

4.6 Ablation Study (RQ3)

This section addresses RQ3 by isolating the effects of the proposed modules. To systematically evaluate the contribution of each component in PALER, we conduct an ablation study with five variants:

(1) **PALER**: The complete model incorporates our three core innovations - the domain-adaptive instruction corpus, contrastive modality alignment, and adaptive reweighting strategy.
(2) $\textbf{PALER}_{w/o\textbf{Align}}$: This variant removes the contrastive alignment mechanism (Sec 3.3.2), reverting to a naive fusion approach. This configuration isolates the impact of our modality-invariant representation learning.
(3) $\textbf{PALER}_{w/o\textbf{Reweight}}$: This variant eliminates the adaptive reweighting strategy during LLM fine-tuning (Sec 3.4.2), adopting equal weighting for all training samples (ERM).
(4) $\textbf{PALER}_{w/o\textbf{Corpus}}$: This variant removes our instruction corpus formulation strategy (Sec **??**), utilizing only raw, un-augmented data with basic prompting templates. This configuration tests the value of our domain-specific instruction engineering.
(5) $\textbf{PALER}_{\textbf{Baseline}}$: A baseline model that removes all three core contributions, representing a standard LLM fine-tuned on raw interaction data without any architectural enhancements.

As shown in Table 6, the ablation results demonstrate consistent performance degradation across all three datasets when any component is removed. The baseline $\textbf{PALER}_{\textbf{Baseline}}$ model achieves the lowest performance metrics, establishing a clear benchmark for comparison. Notably, the most pronounced performance degradation is observed in the variants lacking either the adaptive reweighting or contrastive alignment components.

On the Yelp dataset, the removal of the reweighting strategy leads to a 9.5% decrease in NDCG@5, while eliminating the alignment mechanism results in a 4.4% drop in the same metric. These patterns are consistently replicated across the Steam and Movies&TV datasets, which again highlights the merits of the proposed framework: (1) resolving modality representation conflicts via contrastive alignment, and (2) capturing personalized user preferences through adaptive sample weighting.

5 Efficiency and Complexity Analysis

5.1 Complexity Analysis

To rigorously evaluate the scalability of PALER, we analyze the time complexity per training step. Let $\mathcal{B}$, L, and d denote the batch size, input sequence length, and hidden dimension size, respectively.

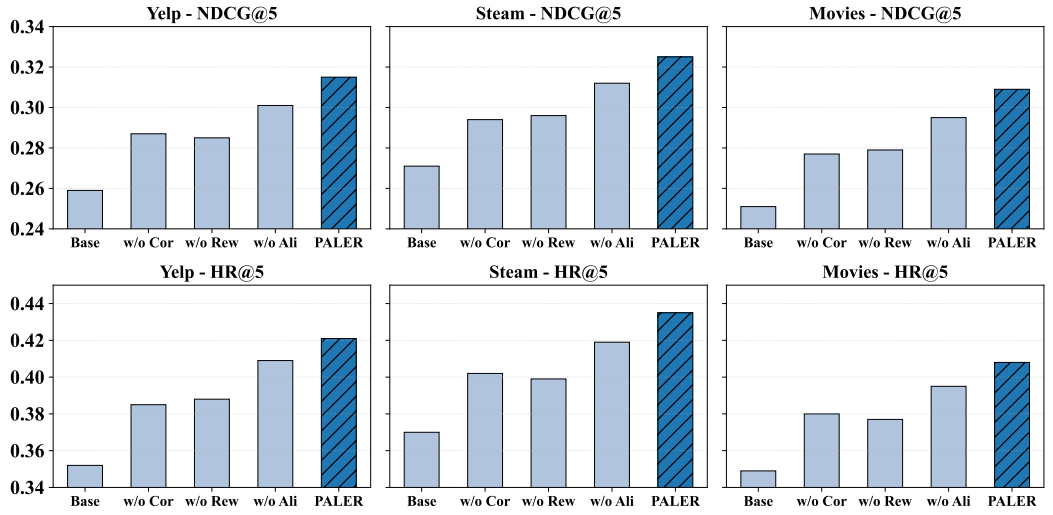

Fig. 4. Performance comparison of PALER and its variants on all datasets. Removing any component consistently degrades NDCG@5 and HR@5, confirming the necessity of all modules.

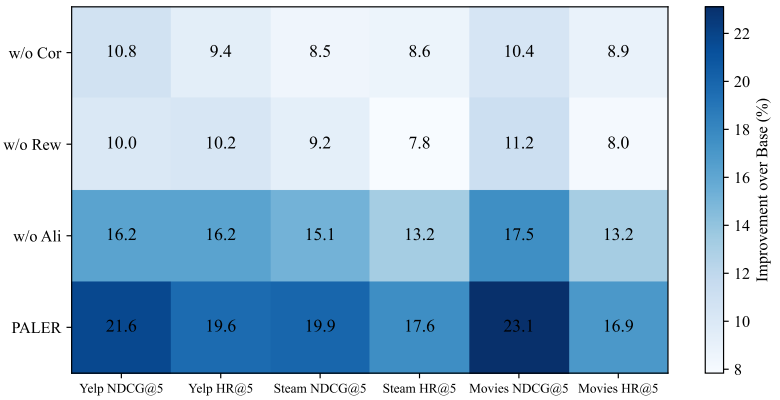

Fig. 5. Relative HR@5 and NDCG@5 improvements of each ablated model over the baseline across all datasets. Each cell reports the percentage gain computed as (variant − baseline)/baseline × 100%, summarizing the overall contribution pattern of all modules.

We define the total computational cost per step, $\mathcal{T}_{total}$, as the superposition of three components:

$$\mathcal{T}_{total} = \mathcal{T}_{backbone} + \mathcal{T}_{align} + \mathcal{T}_{reweight} \tag{38}$$

where $\mathcal{T}_{backbone}$, $\mathcal{T}_{align}$, and $\mathcal{T}_{reweight}$ correspond to the costs of the LLM backbone, the modality-invariant alignment module, and the adaptive reweighting module, respectively.

5.1.1 Backbone Complexity. The computational bottleneck of the LLM backbone lies in the self-attention mechanism. For a sequence of length L, the complexity is dominated by the quadratic dependency on L:

$$\mathcal{T}_{backbone} = O(\mathcal{B} \cdot L^2 \cdot d + \mathcal{B} \cdot L \cdot d^2) \tag{39}$$

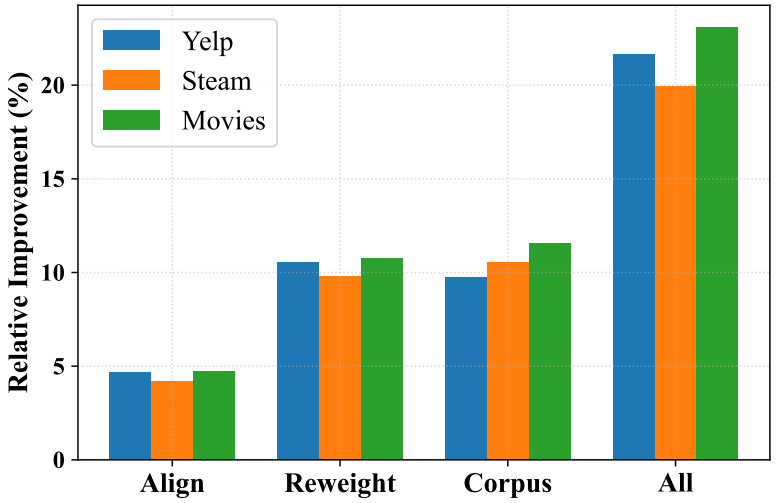

Fig. 6. Relative NDCG@5 improvements contributed by each module of PALER. For each dataset, the gain is computed as the percentage increase of PALER over the corresponding ablated model.

Given the number of layers and width (d) of modern LLMs, this term represents the dominant computational load.

5.1.2 Alignment Module Complexity. The modality-invariant module introduces a contrastive objective (Eq. 2). This necessitates computing the pairwise similarity matrix between the ID representation z^{ID} and text representation z^{txt} within a batch.

- **Similarity Calculation:** Computing the $\mathcal{B} \times \mathcal{B}$ similarity matrix via dot products requires $\mathcal{B}^2$ operations over dimension d.
- **Loss Computation:** The InfoNCE loss involves a summation over the batch dimension.

Thus, the complexity is derived as:

$$\mathcal{T}_{align} = O(\mathcal{B}^2 \cdot d + \mathcal{T}_{enc}) \tag{40}$$

where $\mathcal{T}_{enc}$ is the forward pass of the lightweight ID encoder (e.g., SASRec). Since $\mathcal{T}_{enc} \ll \mathcal{T}_{backbone}$ and typically $\mathcal{B} \ll L$ in LLM training settings (e.g., $\mathcal{B} = 8$ vs. $L = 512$), the term $\mathcal{B}^2 \cdot d$ is negligible compared to the backbone's $\mathcal{B} \cdot L^2 \cdot d$.

5.1.3 Reweighting Module Complexity. The Adaptive Reweighting strategy (Algorithm 1) entails solving for optimal weights β^* and updating them via Exponential Moving Average (EMA).

- **Weight Solution:** The closed-form solution for β^* (Lemma 3.4) requires operations linear to the batch size $\mathcal{B}$.
- **Smoothing:** The EMA update $\beta^{(t)} \leftarrow (1-\eta)\beta^{(t-1)} + \eta\beta^*$ is an element-wise vector operation.

Consequently, the complexity is strictly linear with respect to the batch size:

$$\mathcal{T}_{reweight} = O(\mathcal{B}) \tag{41}$$

This term is computationally trivial relative to the model's forward-backward pass.

Table 7. Parameter scale analysis of PALER components.

Module	Architecture	Params	Ratio
Backbone	Llama 3 8B[16]	~8.03 B	100.00%
Alignment	BERT-base[14]	~110 M	~1.37%
ID Encoder	SASRec[26]	~1.65 M	~0.02%

Note: "B" denotes Billion, "M" denotes Million. Ratios are calculated relative to the Backbone size.

5.1.4 Overall Efficiency. Synthesizing the above components, the asymptotic complexity of PALER is:

$$\begin{aligned} \mathcal{T}_{total} &\approx O(\mathcal{B}L^2d) + O(\mathcal{B}^2d) + O(\mathcal{B}) \\ &\approx O(\mathcal{B}L^2d) \approx \mathcal{T}_{backbone} \end{aligned} \tag{42}$$

The additional modules in PALER introduce a marginal computational overhead that does not alter the fundamental time complexity class of the LLM-based recommendation framework.

6 Conclusion

In this work, we proposed PALER, a framework that significantly enhances personalization in LLM-based recommendation systems. By employing a novel strategy of contrastive alignment for cross-modal fusion and adaptive reweighting for user-centric optimization, PALER effectively preserves user-specific semantics. Our experiments demonstrate that this approach achieves state-of-the-art performance on real-world datasets, setting a new benchmark for personalized, LLM-driven recommendation.

Future work will focus on extending PALER's applicability. We plan to explore domain-adaptive training for new verticals such as healthcare and social media, and to modularize the framework into a versatile plugin for seamless integration with existing recommendation architectures. These efforts will position PALER as a foundational and scalable solution for the next generation of personalized recommender systems.

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 119–129.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[3] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*. 1007–1014.

[4] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1007–1014. https://doi.org/10.1145/3604915.3608857

[5] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. ACM, 1007–1014. doi:10.1145/3604915.3608857

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906* (2021).

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple yet effective graph contrastive learning for recommendation. *arXiv preprint arXiv:2302.08191* (2023).

[9] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[10] Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The datasets dilemma: How much do we really know about recommendation datasets?. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 141–149.

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[12] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[13] Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965* (2025).

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

[15] Yingpeng Du, Di Luo, Rui Yan, Hongzhi Liu, Yang Song, Hengshu Zhu, and Jie Zhang. 2023. Enhancing Job Recommendation through LLM-based Generative Adversarial Networks. arXiv:2307.10747 [cs.IR] https://arxiv.org/abs/2307.10747

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[17] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. 2021. Whitening for self-supervised representation learning. In *International conference on machine learning*. PMLR, 3015–3024.

[18] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. A large language model enhanced conversational recommender system. *arXiv preprint arXiv:2308.06212* (2023).

[19] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524* (2023).

[20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[21] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[22] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2023. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). arXiv:2203.13366 [cs.IR] https://arxiv.org/abs/2203.13366

[23] Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging Large Language Models for Sequential Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. ACM, 1096–1102. doi:10.1145/3604915.3610639

[24] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of the ACM Web Conference 2023*. 1162–1171.

[25] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*. Springer, 364–381.

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

[27] Guanghan Li, Xun Zhang, Yufei Zhang, Yifan Yin, Guojun Yin, and Wei Lin. 2025. Semantic convergence: Harmonizing recommender systems via two-stage alignment and behavioral semantic tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12040–12048.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[29] Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. 2023. CTRL: Connect Collaborative and Language Model for CTR Prediction. *arXiv preprint arXiv:2306.02841* (2023).

[30] Xuewei Li, Aitong Sun, Mankun Zhao, Jian Yu, Kun Zhu, Di Jin, Mei Yu, and Ruiguo Yu. 2023. Multi-intention oriented contrastive learning for sequential recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*. 411–419.

[31] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2023. LLaRA: Large Language-Recommendation Assistant. *arXiv preprint arXiv:2312.02445* (2023).

[32] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. LLaRA: Large Language-Recommendation Assistant. arXiv:2312.02445 [cs.IR] https://arxiv.org/abs/2312.02445

[33] Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. 2024. LLM-ESR: Large Language Models Enhancement for Long-tailed Sequential Recommendation. arXiv:2405.20646 [cs.IR] https://arxiv.org/abs/2405.20646

[34] Wensheng Lu, Jianxun Lian, Wei Zhang, Guanghua Li, Mingyang Zhou, Hao Liao, and Xing Xie. 2024. Aligning Large Language Models for Controllable Recommendations. arXiv:2403.05063 [cs.IR] https://arxiv.org/abs/2403.05063

[35] Wensheng Lu, Jianxun Lian, Wei Zhang, Guanghua Li, Mingyang Zhou, Hao Liao, and Xing Xie. 2024. Aligning Large Language Models for Controllable Recommendations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8159–8172. doi:10.18653/v1/2024.acl-long.443

[36] Sichun Luo, Bowei He, Haohan Zhao, Wei Shao, Yanlin Qi, Yinya Huang, Aojun Zhou, Yuxuan Yao, Zongpeng Li, Yuanzhang Xiao, et al. 2025. Recranker: Instruction tuning large language model as ranker for top-k recommendation. *ACM Transactions on Information Systems* 43, 5 (2025), 1–31.

[37] Thanh Toan Nguyen, Nguyen Quoc Viet Hung, Thanh Tam Nguyen, Thanh Trung Huynh, Thanh Thi Nguyen, Matthias Weidlich, and Hongzhi Yin. 2024. Manipulating recommender systems: A survey of poisoning attacks and countermeasures. *Comput. Surveys* 57, 1 (2024), 1–39.

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

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

[40] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. 2023. Recommender Systems with Generative Retrieval. arXiv:2305.05065 [cs.IR] https://arxiv.org/abs/2305.05065

[41] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation Learning with Large Language Models for Recommendation. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*. ACM, 3464–3475. doi:10.1145/3589334.3645458

[42] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[43] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[44] Zhu Sun, Hui Fang, Jie Yang, Xinghua Qu, Hongyang Liu, Di Yu, Yew-Soon Ong, and Jie Zhang. 2022. Daisyrec 2.0: Benchmarking recommendation for rigorous evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2022), 8206–8226.

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

[46] Jianling Wang, Haokai Lu, Yifan Liu, He Ma, Yueqi Wang, Yang Gu, Shuzhou Zhang, Shuchao Bi, Lexi Baugher, Ed Chi, et al. 2024. LLMs for User Interest Exploration: A Hybrid Approach. *arXiv preprint arXiv:2405.16363* (2024).

[47] Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112* (2023).

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[49] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. LLMRec: Large Language Models with Graph Augmentation for Recommendation. arXiv:2311.00423 [cs.IR] https://arxiv.org/abs/2311.00423

[50] Li Yang, Anushya Subbiah, Hardik Patel, Judith Yue Li, Yanwei Song, Reza Mirghaderi, and Vikram Aggarwal. 2024. Item-Language Model for Conversational Recommendation. *arXiv preprint arXiv:2406.02844* (2024).

[51] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems* 33 (2020), 5824–5836.

[52] Zhenrui Yue, Sara Rabhi, Gabriel de Souza Pereira Moreira, Dong Wang, and Even Oldridge. 2023. Llamarec: Two-stage recommendation using large language models for ranking. *arXiv preprint arXiv:2311.02089* (2023).

[53] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*. PMLR, 12310–12320.

[54] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1817.

[55] Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024. Agentcf: Collaborative learning with autonomous language agents for recommender systems. In *Proceedings of the ACM on Web Conference 2024*. 3679–3689.

[56] Junjie Zhang, Ruobing Xie, Yupeng Hou, Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024. Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach. *ACM Trans. Inf. Syst.* (Dec. 2024). https://doi.org/10.1145/3708882 Just Accepted.

[57] Kaike Zhang, Qi Cao, Yunfan Wu, Fei Sun, Huawei Shen, and Xueqi Cheng. 2025. LoRec: Large Language Model for Robust Sequential Recommendation against Poisoning Attacks. arXiv:2401.17723 [cs.IR] https://arxiv.org/abs/2401.17723

[58] Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. 2025. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2025).

[59] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering* 36, 11 (2024), 6889–6907.

[60] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. arXiv:2311.09049 [cs.IR] https://arxiv.org/abs/2311.09049

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

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009