TailorMind: Towards Preference-Aligned Multimodal Content Generation

Anonymous ACL submission

Abstract

Personalized content systems depend on available UGC and struggle when suitable content is absent, delayed, or costly to create. Although multimodal generators can synthesize content on demand, how to translate behavioral traces into generation-ready preferences remains underexplored. We study personalized multimodal content generation: creating user-tailored multimodal content without existing item pools or waiting for matching UGC. We propose **TailorMind**, linking collaborative preference modeling with controllable multimodal generation. TailorMind enriches sparse user histories via hypergraph collaborative filtering and optimizes textual profiles with ranking-error feedback and textual gradient descent. Retrieval-augmented style control grounds outputs in authentic UGC patterns, while cross-modal cohesion reflection reduces semantic drift. We construct **TailorBench**, a benchmark from three mainstream platforms evaluated along five dimensions: coherence, novelty, aesthetic, hallucination, profiling. Experiments show that TailorMind achieves competitive or stronger coherence, improves novelty and aesthetic quality over representative generation baselines and ground-truth UGC, demonstrating advantages over retrieving available content or comparable UGC, while achieving up to 29% Recall gains in reranking.

1 Introduction

User-generated content (UGC) has become the dominant source of online media and remains central to content platforms (Nan et al., 2025; Safonov et al., 2025; Li et al., 2025). Yet relying solely on UGC leaves practical gaps. First, novel or technically demanding multimodal content requires specialized skills, tools, time, and production costs, making high-quality large-scale creation difficult for individual creators. Second, emerging topics depend on whether creators notice trends, collect materials, and publish in time, creating a latency gap between user interest and available content. User-aligned artificial intelligence generated content (AIGC) can complement UGC by synthesizing multimodal content on demand, helping platforms cover fresh topics, reduce creation costs, and seed early engagement in specific scenarios.

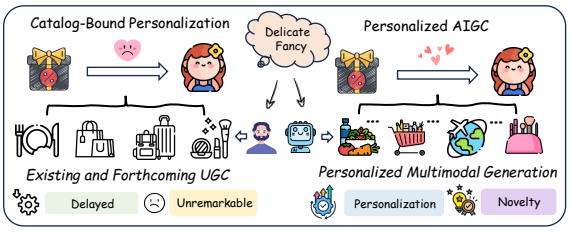


Figure 1: Personalized multimodal content generation.

Recent work increasingly uses large language models (LLMs) as user simulators for synthetic interactions (Chen et al., 2025) or ranking agents for preference modeling over item descriptions (Sun et al., 2023). Multimodal generative models also enhance item representations via cross-modal synthesis (Pan et al., 2025; Zhang et al., 2025a; Fu et al., 2024) or augmented views for contrastive learning (Ren et al., 2024). These studies show the value of generative models for user understanding, but rarely treat generation itself as the primary personalized output. Closer to our goal, PMG (Shen et al., 2024) extracts user preferences from interaction histories as explicit keywords and implicit soft embeddings, jointly conditioning a diffusion model for image generation. Pigeon (Xu et al., 2025) instead operates in the visual token space, filtering noisy history images via token-level masking and incorporating multimodal instructions.

However, existing personalization mechanisms, whether catalog-bound or early generation-oriented, retain two **inherent limitations**. **Personalization.** Catalog-based methods (Zhang et al., 2024; Anand and Maurya, 2025) only select from available items, so outputs typically *intersect* with

user tastes rather than being *fully customized* to unique preference profiles. Recent personalized generation methods condition generators on behavioral descriptions or history-derived visual preferences (Shen et al., 2024; Xu et al., 2025); yet without optimizing interpretable profiles against real interactions or exploiting collaborative signals to enrich sparse evidence, they struggle with fine-grained, cross-modal, and platform-scale preferences. **Novelty and immediacy.** Catalog-based methods face a hard novelty ceiling, as recommendations are confined to existing items. When user preferences point to uncovered content—whether driven by long-tail interests (Tang et al., 2025; Zhang et al., 2025b) or unexplored preference combinations—systems can only fall back on the nearest catalog approximation. A temporal gap compounds this constraint: platforms must wait for creators to notice demand, produce, and publish, leaving emerging interests chronically underserved.

Overcoming these limitations requires generating content on demand, beyond fixed catalogs and without waiting for future UGC supply, while grounding generation in behavior-derived preferences. This raises two challenges: deriving reliable, actionable profiles from sparse interactions and preserving them throughout multimodal generation.

C1. Preference-Profile Alignment. Textual profiles should capture nuanced preferences rather than generic summaries. However, sparse and noisy histories make LLM profiling prone to superficial keyword aggregation, while collaborative signals in interaction patterns are easily missed. A reliable solution should optimize profiles against user behavior, requiring feedback and update mechanisms for discrete natural-language representations.

C2. Aligning Generated Content with Profiles. Even with an accurate profile, multimodal generation can drift during concept planning, text writing, image synthesis, or video creation. Generated components may partially follow the profile while losing authentic UGC style or cross-modal consistency. Thus, profile constraints must be propagated and checked throughout generation to maintain stylistic authenticity and cross-modal coherence.

To address these challenges, we present TailorMind, a framework for personalized multimodal content generation from user preferences. For C1, TailorMind augments sparse histories with hypergraph collaborative filtering and refines textual profiles through textual gradient descent driven by ranking errors on real interactions. For C2, retrieval-augmented style control grounds generation in user-aligned exemplars, while cross-modal cohesion monitoring corrects semantic drift between text and visual modalities. These components generate content that reflects individual preferences while preserving stylistic authenticity.

Validating personalized generation requires benchmarks with rich multimodal content, real user interactions, and unified evaluation across heterogeneous outputs. We introduce TailorBench, a benchmark of real-world interactions and complete multimodal content from mainstream platforms. TailorBench assesses cross-modal coherence, novelty, aesthetic, content safety, and profiling, enabling evaluation of user alignment and generative fidelity.

Our contributions are summarized as follows:

- We introduce personalized multimodal generation for synthesizing tailored content beyond the personalization ceiling, novelty boundaries.
- We propose TailorMind, a personalized generation framework connecting augmented collaborative preference modeling with multimodal generation through data-driven profile optimization and cross-modal alignment mechanisms.
- We construct TailorBench, a personalized generation benchmark with rich multimodal content from mainstream platforms and a unified five-dimensional evaluation framework.
- Comprehensive evaluations demonstrate that TailorMind achieves superior user alignment, novelty, and multimodal generation quality. Our code and datasets are anonymously released at: `https://anonymous.4open.science/r/TailorMind-5843/`.

2 Personalized Multimodal Generation

Personalized multimodal generation bridges two complementary domains: **user preference modeling**, which infers individual interests from behavioral evidence, and **multimodal content generation**, which synthesizes multimodal outputs from creative instructions and references.

Task Definition. Let $\mathcal{U} = \{u_1, \ldots, u_I\}$ be the user set and $\mathcal{V} = \{v_1, \ldots, v_J\}$ the item set. For each user $u \in \mathcal{U}$, the observed interaction history is $\mathcal{D}_u = \{v_{i_1}, \ldots, v_{i_n}\}$. We learn a behavioral preference representation:

$$f_R : (u, \mathcal{D}_u, \mathcal{V}) \rightarrow \mathbf{h}_u, \tag{1}$$

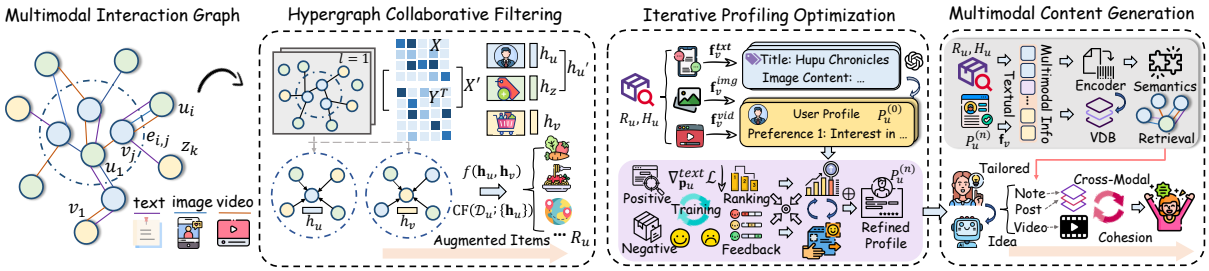


Figure 2: Overall architecture of the proposed TailorMind framework for personalized multimodal generation.

where $\mathbf{h}_u$ encodes latent user preferences from historical interactions. Rather than serving as the final output, $\mathbf{h}_u$ provides behavioral evidence for constructing explicit user profiles and creation instructions to synthesize content beyond $\mathcal{V}$. Let $\mathcal{T}$ denote the space of textual creation instructions and $\mathcal{M} = \{\text{image}, \text{text}, \text{video}\}$ the modality set. The generation process is defined as:

$$f_G : (\mathbf{t}, \{\mathbf{m}_1, \mathbf{m}_2, \ldots\}) \to \mathcal{O}, \tag{2}$$

where $\mathbf{t} \in \mathcal{T}$ is the creation instruction, $\{\mathbf{m}_i\}$ are optional multimodal reference materials, and $\mathcal{O} \notin \mathcal{V}$ is the newly synthesized output.

Benchmark. To support personalized generation research, we construct TailorBench from real-world interactions on **Rednote**, **Bilibili**, and **Hupu**:

$$\mathcal{B} = \{\mathcal{U}, \mathcal{V}, \mathcal{D}, \mathcal{M}, \mathcal{E}\}, \tag{3}$$

where $\mathcal{E}$ is a set of evaluation methods assessing five dimensions: **Coherence** (cross-modal consistency), **Novelty** (originality relative to existing UGC), **Aesthetic** (textual and visual quality), **Hallucination** (absence of fabricated information), and **Profiling** (quality of user profiles). All user data in TailorBench is anonymized. Detailed implementation is discussed in Sec. 4.1.

3 The Proposed TailorMind Framework

This section elaborates the technical details of the proposed TailorMind framework. The overall architecture is depicted in Figure 2.

3.1 Hypergraph Collaborative Filtering

To generate personalized creation instructions $\mathbf{t}$ for user u, VLMs require preference-aligned evidence from behavioral history and multimodal content. Directly profiling from $\mathcal{D}_u$ is often noisy and sparse, so TailorMind augments user evidence via hypergraph collaborative filtering. We treat content tags as hyperedges because they encode semantic item relations and induce multi-way connections, yielding denser connectivity than user-item bipartite graphs and reducing sparsity degradation.

Formally, let $\mathbf{Y} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{Z}|}$ denote the item-tag matrix, where $\mathcal{Z}$ is the tag vocabulary, and $\mathbf{X} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{V}|}$ denote the user-item interaction matrix. We stack $\mathbf{X}$ with $\mathbf{Y}^\top$ to construct the augmented adjacency matrix as:

$$\mathbf{X}' = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y}^\top \end{bmatrix} \in \{0,1\}^{(|\mathcal{U}|+|\mathcal{Z}|)\times|\mathcal{V}|}, \tag{4}$$

which defines a unified hypergraph between the joint node set $\mathcal{U} \cup \mathcal{Z}$ and items in $\mathcal{V}$. We initialize embeddings $\mathbf{h}_{u'}^{(0)}$ for each node $u' \in \mathcal{U} \cup \mathcal{Z}$ and $\mathbf{h}_v^{(0)}$ for each item $v \in \mathcal{V}$, then propagate collaborative information through the hypergraph:

$$\mathbf{h}_{u'}^{l+1} = \text{GNN}\left(\mathbf{h}_{u'}^{l}, \{\mathbf{h}_v^l : v \in \mathcal{N}(u')\}; \Theta^l\right), \tag{5}$$

where $\mathcal{N}(u')$ denotes the neighborhood of node u' and Θ^l are trainable parameters at layer l. After L propagation layers, we rank unseen items by predicted affinity scores:

$$\hat{y}_{uv} = f\left(\mathbf{h}_u^L, \mathbf{h}_v^L\right), \tag{6}$$

$$\text{CF}(\mathcal{D}_u; \{\mathbf{h}_u\}) = \text{TopK}\left(\{\hat{y}_{uv}\}_{v \in \mathcal{V} \setminus \mathcal{D}_u}\right). \tag{7}$$

where f computes user-item preference scores. Top-k items augment $\mathcal{D}_u$ with collaborative signals, providing broader preference context for profiling.

3.2 User Profiling with Iterative Optimization

3.2.1 Multimodal Profile Initialization

Existing profiling methods often process modalities independently or concatenate them directly, missing cross-modal semantics that reveal explicit and implicit interests. We instead distill multimodal content into textual descriptions for coherent profile generation. Given item v with multimodal inputs $\{\mathbf{c}_v^m\}_{m\in\mathcal{M}}$, where $\mathcal{M} = \{\text{vid}, \text{img}, \text{txt}\}$, we extract modality-specific textual features:

$$\mathbf{f}_v^m = \phi_m(\mathbf{c}_v^m), \quad m \in \mathcal{M}, \tag{8}$$

where $\phi_m(\cdot)$ denotes modality-specific feature extractors based on multimodal models that convert visual and textual content into natural language descriptions. Textual features from the augmented evidence set $\mathcal{D}_u^{'} = \mathcal{D}_u \cup \text{CF}(\mathcal{D}_u)$ are fed into the LLM to synthesize the initial profile:

$$\mathbf{p}_u^{(0)} = \text{LLM}\left(\{\mathbf{f}_v^m : v \in \mathcal{D}_u^{'}\};\ \ p_{\text{prof}}\right), \quad (9)$$

where p_{prof} guides the LLM to reason over multimodal textual features and distill them into a coherent user preference profile $\mathbf{p}_u^{(0)}$, which supports subsequent personalized content generation.

3.2.2 Textual-Gradient Profile Refinement

To move beyond heuristic initialization, we formulate profile refinement as a **feedback-driven optimization process** that minimizes ranking errors on the user's validation ground-truth interaction $\mathcal{Z}_u$ through textual gradient descent, thereby grounding profile updates in actual interactions.

Optimization Objective. We optimize the user profile $\mathbf{p}_u$ to maximize ranking quality on interaction groundtruth $\mathcal{Z}_u$ as follows:

$$\min_{\mathbf{p}_u \in \mathcal{P}} \mathbb{E}_{v^+ \sim \mathcal{Z}_u}\left[\ell_{\text{rank}}(\mathbf{p}_u, v^+)\right], \quad (10)$$

where $\ell_{\text{rank}}(\mathbf{p}_u, v^+)$ measures the rank-position loss of positive interaction item v^+ under profile $\mathbf{p}_u$, and $\mathcal{P}$ denotes the profile space of textual descriptions encoding multimodal user preferences.

Ranking Loss Generation. At iteration n, we construct $\mathcal{C}_u$ by sampling one positive item $v^+ \sim \mathcal{Z}_u$ and $M-1$ negative items, then rank them with the current profile $\mathbf{p}_u^{(n)}$ to obtain top-k items:

$$\begin{aligned}\mathcal{C}_u =& \{v^+\} \cup \text{Sample}(\mathcal{V} \setminus \mathcal{Z}_u, M-1),\\ \mathcal{C}_{u,k}^{(n)} =& \text{Rank}_{\text{LLM}}\left(\mathbf{p}_u^{(n)}, \mathcal{C}_u\right), \end{aligned} \quad (11)$$

where $M = 100$ in our implementation and $\mathcal{C}_{u,k}^{(n)}$ denotes the top-k ranked items. We compute $\ell_{\text{rank}}(\mathbf{p}_u^{(n)}, v^+)$ from whether v^+ appears in top-k and its rank position, yielding a **discrete feedback signal** that identifies profile aspects that capture user preferences or require adjustment.

Textual Gradient Descent. Because profiles are textual, conventional numerical gradients are infeasible. Following agent-based optimization (Hu et al., 2024), we use a **textual gradient**: the LLM estimates an update direction from (i) the current profile $\mathbf{p}_u^{(n)}$ and ranking performance, (ii) multimodal features $\{\mathbf{f}_j\}_{j \in \mathcal{C}_{u,k}^{(n)}}$ of top-ranked candidates, and (iii) multimodal features $\{\mathbf{f}_g\}_{g \in \mathcal{Z}_u}$ of groundtruth items. The profile update is:

$$\mathbf{p}_u^{(n+1)} = \mathbf{p}_u^{(n)} \oplus \nabla_{\mathbf{p}_u}^{\text{text}} \mathcal{L}(\mathbf{p}_u^{(n)}), \quad (12)$$

where $\oplus$ denotes the LLM-based refinement operation incorporating the textual gradient, computed:

$$\nabla_{\mathbf{p}_u}^{\text{text}} \mathcal{L} = \text{LLM}\big(\mathbf{p}_u^{(n)}, \{\mathbf{f}_j\}_{j \in \mathcal{C}_{u,k}^{(n)}}, \{\mathbf{f}_g\}_{g \in \mathcal{Z}_u}; p_{\text{grad}}\big), \quad (13)$$

where p_{grad} prompts the LLM to improve profile performance. This approximates $\nabla_{\mathbf{p}_u} \mathbb{E}_{v^+ \sim \mathcal{Z}_u}$ through LLM reasoning over ranking loss, analogous to policy-gradient estimation in reinforcement learning but performed in natural language space.

Convergence Criterion. The optimization terminates when either (1) ranking loss reaches its optimum, *i.e.*, the groundtruth item v^+ ranks first:

$$\text{rank}(v^+; \mathbf{p}_u^{(n+1)}, \mathcal{C}_u) = 1, \quad \forall v^+ \in \mathcal{Z}_u, \quad (14)$$

or (2) the process reaches the maximum iteration $N_{\max}$, ensuring efficiency across users. Thus, TailorMind shifts from **heuristic-based profiling** relying on prompt design to **data-driven optimization** grounded in groundtruth items, directly improving interaction ranking quality while preserving natural-language interpretability. Gradient optimization prompts are provided in Appendix A.5

3.3 Personalized Content Generation

With the optimized user profile $\mathbf{p}_u$, TailorMind employs specialized generation agents for content categories such as image-text posts and videos.

3.3.1 Retrieval-Augmented Style Control

To control style, TailorMind uses retrieval-augmented generation grounded in actual UGC. Although $\mathbf{p}_u$ captures preference semantics, directly translating it into generation instructions may miss stylistic nuances from user interactions. We retrieve top-k exemplars by computing cosine similarity between the profile embedding and multimodal feature representations $\mathbf{f}_i$ of items from historical interactions $\mathcal{D}_u$ and recommended candidates $\text{CF}(\mathcal{D}_u; \{\mathbf{h}_u\})$, defined as :

$$\mathcal{E}_u^k = \underset{i \in \mathcal{D}_u \cup \text{CF}(\mathcal{D}_u)}{\arg\text{top-k}} \cos(\mathbf{p}_u, \mathbf{f}_i), \quad (15)$$

where $\mathcal{E}_u^k$ denotes the retrieved exemplar set. These exemplars serve as few-shot references that help the generation agent emulate authentic patterns and stylistic nuances while remaining aligned with preferences encoded in $\mathbf{p}_u$.

Table 1: Statistics of the experimental datasets.

Datasets	#Users	#Items	#Tags	#Text	#Images	#Videos
Rednote	12131	94997	8531	7.87	7.34	2.44
Hupu	23002	126300	278	6.76	11.33	1.83
Bilibili	13349	82704	8709	-	-	6.44

3.3.2 Cross-Modal Cohesion Reflection

The generation framework aims to align produced multimodal content with $\mathbf{p}_u$. However, long-chain generation across heterogeneous modalities can introduce **semantic drift**, where generated components diverge from intended profile semantics as abstract preferences are translated into creative concepts and multimodal realizations.

To bridge abstract profile representations and actionable creative directives, we design $R = 10$ prototypical idea templates covering creative motifs on contemporary social media platforms. The user profile $\mathbf{p}_u$ is mapped to a style subset $\mathcal{S} \subseteq \{1, \ldots, R\}$ to produce candidate creative ideas $\mathbf{I}_u$:

$$\mathbf{I}_u = \Psi(\mathbf{p}_u, \mathcal{S}) \tag{16}$$

where $\Psi(\cdot)$ uses user profiles and selected styles to produce personalized concept prompts. To mitigate semantic drift and enforce profile-content alignment, we employ a **cross-modal cohesion reflection mechanism**: ClipScore measures consistency between generated visual content $\hat{\mathbf{x}}$ and paired text $\hat{\mathbf{t}}$, while Relevance measures whether $\hat{\mathbf{t}}$ remains consistent with the user's interaction history $\mathcal{D}$ via cosine similarity in text embedding space:

$$\mathcal{C}(\hat{\mathbf{x}}, \hat{\mathbf{t}}) = \text{ClipScore}(\hat{\mathbf{x}}, \hat{\mathbf{t}}), \tag{17}$$

$$\mathcal{R}(\hat{\mathbf{t}}, \mathcal{D}) = \text{Relevance}(\hat{\mathbf{t}}, \mathcal{D}), \tag{18}$$

These cohesion signals, together with $\mathbf{p}_u$ and the previous generation $\mathbf{G}_u^{(n)}$, guide iterative reflective optimization to reduce semantic drift:

$$\mathbf{G}_u^{(n+1)} = \Phi\big(\mathbf{p}_u, \mathbf{G}_u^{(n)}, \mathcal{C}(\hat{\mathbf{x}}, \hat{\mathbf{t}}), \mathcal{R}(\hat{\mathbf{t}}, \mathcal{D})\big) \tag{19}$$

where $\Phi(\cdot)$ denotes the VLM-based reflective revision function. By maximizing cross-modal cohesion, this mechanism constrains multimodal generation to remain aligned with user preferences.

4 Evaluation

We conduct extensive experiments to validate TailorMind's effectiveness across six research questions: **RQ1** compares TailorMind with representative baselines, **RQ2** examines human preferences over generated content, **RQ3** analyzes key modules, **RQ4** studies hyperparameter sensitivity, **RQ5** evaluates cost and efficiency, and **RQ6** presents a qualitative case study of generated content.

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Protocols

We construct TailorBench from three mainstream platforms: **Rednote**, featuring lifestyle short videos and graphic-text notes; **Bilibili**, an entertainment-oriented youth video platform; and **Hupu**, a sports community mainly comprising textual discussions. Table 1 reports detailed statistics. We use leave-two-out splitting for profiling, with the last interaction for testing, the penultimate one for validation, and the rest for optimization. We randomly sample 1,000 users from each dataset (Shen et al., 2024), and compare TailorMind's generated content with baseline outputs and the ground-truth test item along five dimensions: **Coherence**, where image–text outputs and video-derived textual descriptions are respectively compared against the equal-weighted caption embedding of the user's history and target item using CLIPScore (Hessel et al., 2021) and cosine similarity, reflecting personalization to some extent. **Novelty**, where VLMs score novelty and interestingness on a normalized $[0, 1]$ scale (Vargas and Castells, 2011); **Aesthetic**, where VLMs assess textual and visual aesthetics within $[0, 1]$ (Kao et al., 2017); and **Hallucination**, where web-search tools detect false or fabricated information, labeling hallucinated instances as 1 and otherwise 0. **Profiling**, where 1:99 negative sampling (Ding et al., 2019) and Recall@N/NDCG@N evaluate reranking performance of user performance modeling.

4.1.2 Baseline Methods

TailorMind is compared with comprehensive baselines: **i) multimodal AI generation methods:** GPT-4o (Hurst et al., 2024), Gemini-2.5-Flash (Team et al., 2023), Nanobanana pro (Google, 2025a), PMG (Shen et al., 2024), Pigeon (Xu et al., 2025), RAGAR (Ling et al., 2026), Veo3.1 (Google, 2025b), Sora2 (OpenAI, 2025), CIPHER (Gao et al., 2024), and PROSE (Aroca-Ouellette et al., 2025); **ii) ID-based representation learning methods:** LightGCN (He et al., 2020), BIGCF (Zhang et al., 2024), and IRLLRec (Wang et al., 2025a); and **iii) LLM-based textual ranking approaches:** MoRE (Qin et al., 2025), Re2LLM (Wang et al., 2025b), LLM4Rerank (Gao et al., 2025), and MLLM-MSR (Ye et al., 2025);

4.1.3 Implementation Details

All methods are implemented following their original papers. For ID-based representation learning

Table 2: Generated Content Comparison on Rednote and Hupu Datasets, averaged over Three Runs.

	Model	Rednote				Hupu			
	Metric	Coherence ↑	Novelty ↑	Aesthetic ↑	Hallucination ↓	Coherence ↑	Novelty ↑	Aesthetic ↑	Hallucination ↓
Image-Text	GPT-4o	0.6148	0.5436	0.7243	0.0110	0.6864	0.4723	0.6332	0.0911
	Gemini-2.5-flash	0.6212	0.4706	0.4391	0.2787	0.6846	0.3826	0.2838	0.5305
	Nanobanana pro	0.6110	0.5931	0.7262	0.0196	0.6996	0.5529	0.6739	0.0549
	PMG	0.6155	0.5682	0.6732	0.0550	0.6200	0.5912	0.6572	0.2250
	PIGEON	0.6050	0.6350	0.6637	0.0124	0.6280	0.6415	0.6706	0.0160
	RAGAR	0.6255	0.5477	0.6420	0.0215	0.6177	0.5905	0.6584	0.0759
	Groundtruth-IT	0.6175	0.3458	0.3707	0.0072	0.6147	0.2731	0.3409	0.0197
	TailorMind-IT	**0.6495**	**0.6643**	**0.8286**	**0.0053**	**0.7132**	**0.6648**	**0.7341**	**0.0100**
Videos	Veo3.1	0.7249	0.6675	0.8345	0.0000	0.7228	0.6735	0.8450	0.0000
	Sora2	0.7437	0.6650	0.8295	0.0000	0.7281	0.6655	0.8410	0.0000
	CIPHER	**0.7503**	0.5940	0.8040	0.0000	0.7391	0.6915	0.8330	0.0500
	PROSE	0.7429	0.6870	0.8345	0.0000	**0.7404**	0.6905	0.8375	0.0000
	Groundtruth-V	0.7049	0.6731	0.7842	0.0000	0.6231	0.5985	0.6768	0.0000
	TailorMind-V	0.7271	**0.6960**	**0.8430**	**0.0000**	0.7230	**0.7240**	**0.8480**	**0.0000**

Table 3: Generated Content comparison on Bilibili.

Data	Bilibili			
Metric	Coherence ↑	Novelty↑	Aesthetic↑	Halluc. ↓
Veo3.1	0.7634	0.6710	0.8335	0.0000
Sora2	0.7932	0.6550	0.8345	0.0000
CIPHER	0.7800	0.6890	0.8250	0.0000
Prose	0.7913	0.6225	0.7925	0.0000
Groundtruth	0.7181	0.6805	0.7802	0.0030
TailorMind	**0.7947**	**0.6975**	**0.8485**	**0.0000**

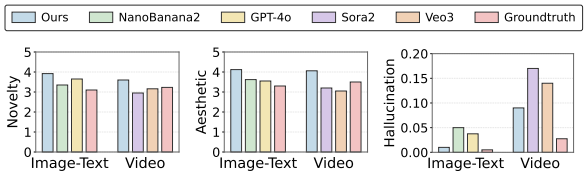


Figure 3: Human evaluation on created content.

methods, we perform grid search to identify optimal hyperparameters. For long video understanding, we analyze the initial and final 8-second segments separately, then aggregate the information to generate a holistic content description. All LLM-based ranking methods and TailorMind's profile refinement and cohesion reflection use 3 iterations. We employ ViT-B/32 for image encoding and text-embedding-3-small for text encoding. User preferences are extracted by selecting the top-1 preference for personalized multimodal content generation. Our default models are Claude-Sonnet-4.5 for ranking and profiling, Gemini-2.5-Flash-Image for image generation, and Veo-3.1 for video generation. For automatic evaluation only, we use Gemini-2.5-pro as the VLM judge and GPT-5-All for web-based factual verification.

4.2 Overall Performance Comparison (RQ1)

We compare TailorMind with baselines from multiple perspectives; results are summarized in Tables 4, 3, and 2. **Stronger User Profiling.** For the profiling dimension, generated user profiles rerank 1:99 candidate sets, and Recall/NDCG measure whether they recover held-out interactions. Table 4 shows that TailorMind consistently improves Recall across all datasets, indicating that textual-gradient profile optimization better captures fine-grained preferences from sparse behavioral evidence. **Higher Novelty and Quality.** Tables 3 and 2 show that TailorMind generates more novel content than generation baselines and UGC while achieving the best aesthetic scores, demonstrating its ability to synthesize fresh and visually appealing user-aligned outputs across modalities. **Better Coherence and Reliability.** TailorMind achieves higher coherence than UGC and competitive or superior results against generation baselines. Since the coherence reference includes both user history and the target item, these gains indicate stronger history-aware alignment while validating cross-modal cohesion reflection.

4.3 Human Evaluation on Content (RQ2)

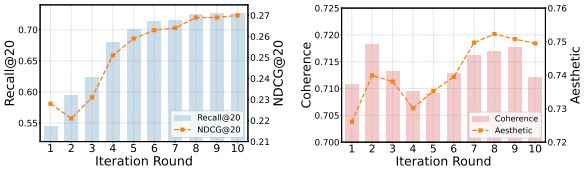
To complement automatic VLM judging and reduce comparison bias with ground-truth UGC, we further conduct a human evaluation of TailorMind's personalized generation quality. Specifically, we randomly selected 20 generated samples from each image-text and video category and asked 18 volunteers to evaluate anonymized and randomized outputs. Each sample was rated on a 5-point scale for Novelty and Aesthetic quality, while Hallucination was labeled as 1 if misinformation was detected and 0 otherwise; the questionnaire contained 80 items and took about 1.5 hours on average. As shown in Figure 3, TailorMind achieves stronger novelty and aesthetic quality than generation baselines and ground-truth UGC, while maintaining near-human hallucination performance, validating

Table 4: *Reranking* performance comparison, in terms of *Recall* and *NDCG*. Boldface marks the best reranking result. Standard deviations are computed over five runs.

Data	**Bilibili**				**Rednote**				**Hupu**			
Metrics	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20
LightGCN	0.3820	0.2554	0.4780	0.2796	0.4460	0.3605	0.4980	0.3736	0.1160	0.0532	0.2190	0.0791
BIGCF	0.3100	0.2100	0.4120	0.2357	0.4190	0.3575	0.4600	0.3678	0.1130	0.0557	0.2210	0.0825
IRLLRec	0.4080	0.2376	0.5240	0.2674	0.4780	0.3446	0.5400	0.3605	0.3320	0.1780	0.4800	0.2155
MoRE	0.1020	0.0490	0.2050	0.0746	0.4390	0.3172	0.5250	0.3436	0.3010	0.1243	0.4740	0.1678
Re2LLM	0.3972	0.2633	0.5260	0.2957	0.5460	0.3979	0.6400	0.4216	0.4320	**0.2100**	0.6040	**0.2532**
LLM4Rerank	0.3360	0.2265	0.4470	0.2548	0.4590	0.3415	0.5430	0.3628	0.3870	0.1586	0.5710	0.1894
MLLM-MSR	0.3820	0.2333	0.5150	0.2665	0.5530	0.3812	0.6420	0.4052	0.4170	0.2085	0.5960	0.2420
TailorMind	**0.5260** ±8.2e-4	**0.2643** ±1.5e-2	**0.6770** ±2.0e-2	**0.3023** ±1.3e-2	**0.6490** ±1.6e-2	**0.4140** ±1.2e-2	**0.6800** ±1.3e-2	**0.4310** ±1.1e-2	**0.4500** ±1.1e-2	0.1919 ±5.1e-3	**0.6230** ±1.4e-2	0.2311 ±5.4e-3

Table 5: Performance of different ablated TailorMind.

Data		Rednote		Hupu	
Metric		R@20	N@20	R@20	N@20
Profiling	w/o historical	0.6028	0.3207	0.5622	0.2232
	w/o rec	0.6371	0.3543	0.5916	0.2470
	w/o TV	0.6443	0.3471	0.5842	0.2115
	w/o IT	0.2091	0.1686	0.1777	0.0876
	w/o IV	0.6503	0.3670	0.5965	0.2289
	Origin	0.6800	0.4310	0.6230	0.2311
Metric		Coherence	Aesthetic	Coherence	Aesthetic
Gen	w/o RAG	0.6276	0.8020	0.6523	0.6882
	w/o Cohesion	0.6238	0.8253	0.6911	0.6971
	Origin	0.6495	0.8286	0.7132	0.7341



(a) Profile optimization. (b) Generation reflection.

Figure 4: Hyperparameter study on Hupu dataset.

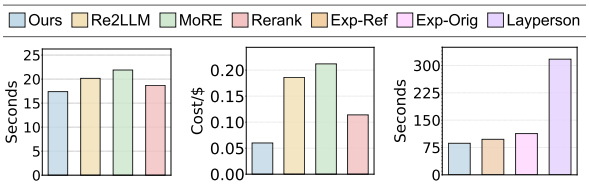
the effectiveness of our personalized generation.

4.4 Ablation Study (RQ3)

Table 5 quantifies each module's contribution. Removing historical records or hypergraph-recommended items consistently lowers Recall@20 and reduces most NDCG@20 scores, showing that observed behaviors and collaborative signals provide complementary evidence for preference modeling. Ablating TV, IT, or IV inputs further weakens preference profiling, confirming the need for multimodal fusion. For generation, removing RAG reduces coherence and aesthetic quality by weakening UGC-style grounding, while removing Cohesion module hurts cross-modal alignment, indicating its role in limiting semantic drift during long-chain personalized generation.

4.5 Hyperparameter Study (RQ4)

Figure 4 reports the effect of profile optimization iterations and generation reflection rounds. For profile optimization, personalization performance rises consistently with more iterations as textual-



(a) Profiling time. (b) Profiling costs. (c) Generation time.

Figure 5: Time and API costs comparison.

gradient updates progressively integrate relevant signals and correct mismatches, but gains plateau beyond a certain number of rounds, indicating a practical optimum that balances preference modeling quality and computational cost. For generation reflection, coherence improves notably in early rounds and then stabilizes, while aesthetic quality shows only minor fluctuations throughout, suggesting that a moderate number of reflection rounds is sufficient to reinforce cross-modal consistency.

4.6 Cost-Efficiency Study (RQ5)

We evaluate the API costs and time efficiency of TailorMind in Figure 5. For the profiling process, TailorMind reduces average computational overhead and latency compared to LLM-based baselines while achieving superior performance. For the content generation process, we compare TailorMind against three human creator groups sampled from 10 Rednote user profiles: a professional team manually operating the same AI tools as TailorMind from scratch (Expert-Orig), the same expert team refining existing groundtruth UGC with those tools (Expert-Ref), and 8 ordinary users without professional creation experience (Layperson). All reported costs and times are averaged across users. TailorMind substantially reduces generation time over all three groups, demonstrating strong suitability for practical personalized content generation.

4.7 Case Study (RQ6)

Figure 6 shows a Rednote user (ID 8001) and several genuine interactions. From these multimodal histories, TailorMind identifies a preference for

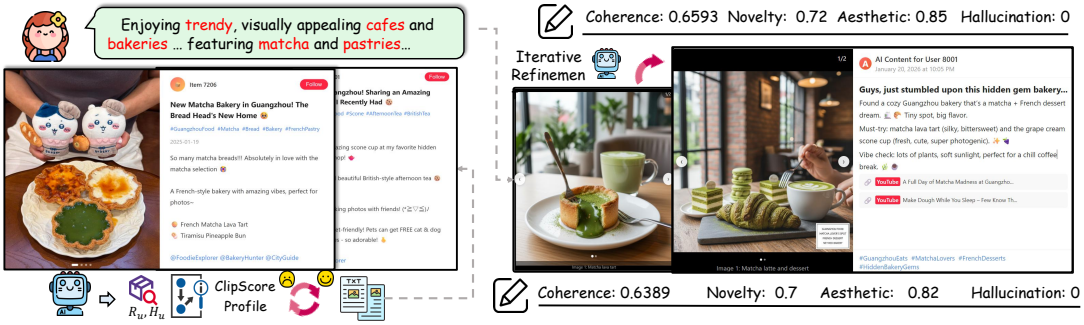


Figure 6: A RedNote post generated by our TailorMind, personalized for a user who enjoys trendy.

street eats, trendy cafes, and visually appealing food photography. The generated image-text post preserves these concrete interests while producing new content beyond the observed UGC items. Guided by cross-modal cohesion reflection, the post improves Coherence, Novelty, and Aesthetic Quality while maintaining zero Hallucination, indicating that iterative reflection corrects cross-modal mismatch without weakening history-aware alignment. The final content exhibits strong visual-textual alignment, supported by textual-gradient profile refinement, retrieval-augmented UGC style control, and cross-modal cohesion reflection.

5 Related Work

Personalized Generation. Personalized generation adapts content to users, styles, or preference descriptions. PMG (Shen et al., 2024) derives personalized multimodal generation instructions with LLMs, while Pigeon (Xu et al., 2025) aligns image generation with visual preferences inferred from histories and instructions. Related preference-alignment methods infer user intent from edits or demonstrations to steer generated outputs (Gao et al., 2024; Aroca-Ouellette et al., 2025). Yet existing methods often rely on descriptive preference signals, with limited correction for semantic drift. TailorMind instead optimizes profiles with interaction feedback and maintains profile-content alignment through retrieval-augmented control and cross-modal cohesion reflection.

Multimodal Generative Models. Recent AIGC models seamlessly synthesize diverse, high-quality text, images, and videos from free-form natural-language instructions. Large vision-language foundation models increasingly support multimodal reasoning and cross-modal content creation (Hurst et al., 2024; Jin et al., 2025), while image and video generation models steadily improve visual fidelity, prompt following, temporal consistency, and motion control (Sun et al., 2025; Seedance et al., 2026). These advances enable interactive user-facing content, but are still typically optimized for prompt-conditioned generation rather than behavior-grounded personalization, making noisy, sparse histories difficult to translate into reliable generation conditions.

Recommendation Systems. Recommendation systems model user preferences from behavioral data. Traditional recommenders capture collaborative signals through graph or intent-based representations (He et al., 2020; Zhang et al., 2024; Zhao et al., 2025), while sequential and multimodal recommenders model temporal dynamics and heterogeneous content features (Ren et al., 2024; Fu et al., 2025; Lin et al., 2025). Recent LLM-based ranking methods use language models to understand descriptions, rerank candidates, simulate users, or refine recommendations (Sun et al., 2023; Qin et al., 2025; Ye et al., 2025). These methods mainly select from existing catalogs; our work instead uses recommendation signals to build generation-ready profiles and synthesize new personalized content.

6 Conclusion

This paper studies personalized multimodal content generation beyond existing UGC catalogs and proposes TailorMind, which connects behavior-grounded preference modeling with controllable generation through hypergraph collaborative filtering, textual-gradient profile optimization, retrieval-augmented style control, and cross-modal cohesion reflection. We further construct TailorBench, a benchmark with user interactions, multimodal content from three mainstream platforms, and five-dimensional evaluation metrics. Experiments show that TailorMind improves preference alignment and generates higher-quality personalized content than generation and retrieval baselines. Future work will explore interactive generation, stronger safety control and more efficient profile optimization.

Limitations

Current social platforms are still largely driven by user-generated content, and creators may not have access to explicit or sufficiently structured follower preferences when producing personalized multimodal content. This limits how directly AIGC systems can support creators in matching audience interests. If platforms can disclose partial, aggregated, high-level follower preference signals to creators within acceptable privacy, consent, and platform-governance constraints, such signals may better facilitate AIGC-assisted creation and reduce the time and effort required for content production.

Ethical Considerations

The datasets used in this study are collected from publicly available user-generated content on three mainstream platforms. We anonymize all user-related information, remove personally identifiable information, and retain only internal identifiers needed to link interaction histories with items. All data are used solely for aggregate-level, privacy-preserving research evaluation, and we do not attempt to re-identify or target individual users.

References

Vineeta Anand and Ashish Kumar Maurya. 2025. A survey on recommender systems using graph neural network. *ACM Transactions on Information Systems*, 43(1):1–49.

Stéphane Aroca-Ouellette, Natalie Mackraz, Barry-John Theobald, and Katherine Metcalf. 2025. Aligning llms by predicting preferences from user writing samples. *arXiv preprint arXiv:2505.23815*.

Jia Chen, Qian Dong, Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, and 1 others. 2025. Qilin: A multimodal information retrieval dataset with app-level user sessions. In *SIGIR*, pages 3670–3680.

Jingtao Ding, Yuhan Quan, Xiangnan He, Yong Li, and Depeng Jin. 2019. Reinforced negative sampling for recommendation with exposure data. In *IJCAI*, pages 2230–2236. Macao.

Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M Jose. 2024. Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft. In *SIGIR*, pages 687–697.

Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Kaiwen Zheng, Yongxin Ni, and Joemon M Jose Joemon. 2025. Efficient and effective adaptation of multimodal foundation models in sequential recommendation. *IEEE TKDE*.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning llm agents by learning latent preference from user edits. *Advances in neural information processing systems*, 37:136873–136896.

Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Wanyu Wang, Huifeng Guo, and Ruiming Tang. 2025. Llm4rerank: Llm-based auto-reranking framework for recommendations. In *Proceedings of the ACM on Web Conference 2025*, pages 228–239.

Google. 2025a. Nano banana pro (gemini 3 pro image): Ai image generation model. Image generation model with reasoning capabilities, supporting 1K to 4K resolution outputs.

Google. 2025b. Veo 3.1: Ai video generation model. Video generation model with improved audio generation and narrative control capabilities.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pages 639–648.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528.

Shengran Hu, Cong Lu, and Jeff Clune. 2024. Automated design of agentic systems. *arXiv preprint arXiv:2408.08435*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhipeng Jin, Wen Tao, Yafei Li, Yi Yang, Cong Han, Shuanglong Li, and Lin Liu. 2025. Large vison-language foundation model in baidu aigc image advertising. In *KDD*, pages 2303–2312.

Yueying Kao, Ran He, and Kaiqi Huang. 2017. Deep aesthetic quality assessment with semantic information. *TIP*, 26(3):1482–1495.

Xin Li, Xijun Wang, Bingchen Li, Kun Yuan, Yizhen Shao, Suhang Yao, Ming Sun, Chao Zhou, Radu Timofte, and Zhibo Chen. 2025. Ntire 2025 challenge on short-form ugc video quality assessment and enhancement: Kwaisr dataset and study. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1126–1136.

Xixun Lin, Rui Liu, Yanan Cao, Lixin Zou, Qian Li, Yongxuan Wu, Yang Liu, Dawei Yin, and Guandong Xu. 2025. Contrastive modality-disentangled learning for multimodal recommendation. *ACM TOIS*, 43(3):1–31.

Run Ling, Wenji Wang, Yuting Liu, Guibing Guo, Haowei Liu, Jian Lu, Quanwei Zhang, Yexing Xu, Shuo Lu, Yun Wang, and 1 others. 2026. Ragar: retrieval augmented personalized image generation guided by recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 15278–15286.

Haonan Nan, Zixiao Wang, Yu Zhang, Xuantao Zhang, and Saixing Zeng. 2025. Responsight: explainable and collaborative moderation approach for responsible video content in ugc platform. In *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems*, pages 1–12.

OpenAI. 2025. Sora 2: A video generation model with synchronized audio. Video generation model supporting 15-25 second videos.

Yinchen Pan, Jun Zeng, Ziwei Wang, Haoran Tang, Junhao Wen, and Min Gao. 2025. Hgdrec: Next poi recommendation based on hypergraph neural network and diffusion model. *IEEE Transactions on Services Computing*.

Weicong Qin, Yi Xu, Weijie Yu, Chenglei Shen, Xiao Zhang, Ming He, Jianping Fan, and Jun Xu. 2025. More: A mixture of reflectors framework for large language model-based sequential recommendation. In *Recsys*, pages 299–308.

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 *WWW*, pages 3464–3475.

Nickolay Safonov, Alexey Bryntsev, Andrey Moskalenko, Dmitry Kulikov, Dmitriy Vatolin, Radu Timofte, Haibo Lei, Qifan Gao, Qing Luo, Yaqing Li, and 1 others. 2025. Ntire 2025 challenge on ugc video enhancement: Methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1503–1513.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, and 1 others. 2026. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 3833–3843.

Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. 2025. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency. *arXiv preprint arXiv:2502.04076*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Zuoli Tang, Zhaoxin Huan, Zihao Li, Xiaolu Zhang, Jun Hu, Chilin Fu, Jun Zhou, Lixin Zou, and Chenliang Li. 2025. One model for all: Large language models are domain-agnostic recommendation systems. *ACM TOIS*, 43(5):1–27.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Recsys*, pages 109–116.

Yu Wang, Lei Sang, Yi Zhang, and Yiwen Zhang. 2025a. Intent representation learning with large language model for recommendation. In *SIGIR*, pages 1870–1879.

Ziyan Wang, Yingpeng Du, Zhu Sun, Haoyan Chua, Kaidong Feng, Wenya Wang, and Jie Zhang. 2025b. Re2llm: reflective reinforcement large language model for session-based recommendation. In *AAAI*, volume 39, pages 12827–12835.

Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. 2025. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, pages 264–274.

Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025. Harnessing multimodal large language models for multimodal sequential recommendation. In *AAAI*, volume 39, pages 13069–13077.

Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. 2025a. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE TKDE*.

Yi Zhang, Lei Sang, and Yiwen Zhang. 2024. Exploring the individuality and collectivity of intents behind interactions for graph collaborative filtering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1262.

Zijian Zhang, Shuchang Liu, Ziru Liu, Rui Zhong, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Qidong Liu, and Peng Jiang. 2025b. Llm-powered user simulator for recommender system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13339–13347.

Chu Zhao, Enneng Yang, Yuliang Liang, Jianzhe Zhao, Guibing Guo, and Xingwei Wang. 2025. Symmetric graph contrastive learning against noisy views for recommendation. *ACM TOIS*, 43(3):1–28.

A Appendix

A.1 Baseline Methods

To ensure a comprehensive study, we compare TailorMind against 17 baseline methods including various techniques for personalized content generation.

Multimodal AI Generation Methods

- **GPT-4o** (Hurst et al., 2024): It adopts an end-to-end omni-architecture for low-latency multimodal understanding and generation.
- **Gemini-2.5-Flash** (Team et al., 2023): Built on native multimodal pretraining and MoE scaling, it supports efficient cross-modal reasoning.
- **Nanobanana pro** (Google, 2025a): It uses reasoning-enhanced diffusion with controllable visual grounding for coherent high-resolution image generation from 1K to 4K outputs.
- **PMG** (Shen et al., 2024): LLM-derived user signals are used to guide personalized multimodal content generation effectively.
- **Pigeon** (Xu et al., 2025): Retrieval-augmented preference inference guides frozen agents without model fine-tuning.
- **RAGAR** (Ling et al., 2026): Semantic retrieval prioritizes relevant historical records, whereas ranking feedback jointly accounts for personalized adaptation and faithful representation.
- **Veo3.1** (Google, 2025b): Diffusion-based video generation is combined with audio synthesis and narrative-aware conditioning.
- **Sora2** (OpenAI, 2025): A spatiotemporal diffusion transformer enables synchronized video-audio generation with consistent motion.
- **CIPHER** (Gao et al., 2024): Historical user edits are retrieved to infer preferences and align generated outputs with user intent.
- **PROSE** (Aroca-Ouellette et al., 2025): Iterative refinement and consistency checks infer user preferences from demonstrations.

ID-based Representation Learning Methods

- **LightGCN** (He et al., 2020): Simplified graph convolution propagates user-item embeddings for collaborative filtering.

Table 6: Comparison on different backbones.

Backbone	Gemini-3-flash		DeepSeek-V3.2		GPT-5.2	
Metric	R@20	N@20	R@20	N@20	R@20	N@20
Re2LLM	0.4060	0.2512	0.4610	0.2655	0.5190	0.2891
MLLM-MSR	0.5800	0.3524	0.5320	0.2865	0.6050	0.3553
TailorMind	0.7760	0.3754	0.7780	0.3548	0.7900	0.3608
Backbone	IRLLRec		RLMRec		LightGCN	
Metric	R@20	N@20	R@20	N@20	R@20	N@20
TailorMind	0.7210	0.2960	0.7350	0.3021	0.6770	0.3023

- **BIGCF** (Zhang et al., 2024): Behavior-aware graph modeling enhances collaborative filtering with richer user-item signals.
- **IRLLRec** (Wang et al., 2025a): A dual-tower design aligns LLM-derived textual intents with interaction-based preferences.

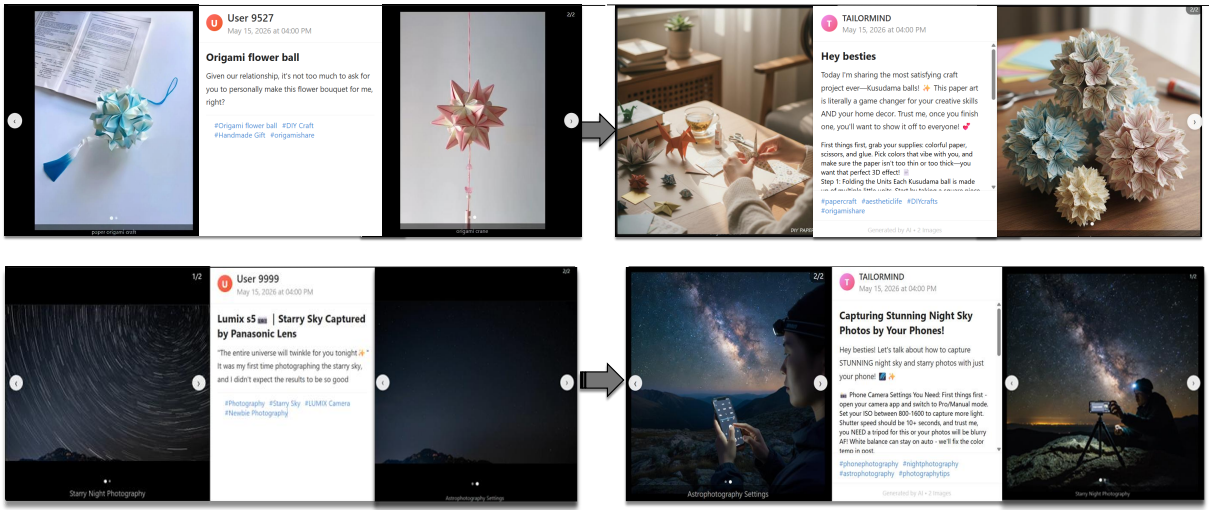
LLM-based Textual Ranking Approaches

- **MoRE** (Qin et al., 2025): Multiple LLM reflectors capture explicit, implicit, and collaborative preferences, while a meta-reflector selects the most suitable perspective.
- **Re2LLM** (Wang et al., 2025b): LLM self-reflection generates hints, while PPO-based retrieval selects useful guidance.
- **LLM4Rerank** (Gao et al., 2025): Aspect-graph reasoning with Chain-of-Thought supports personalized item reranking.
- **MLLM-MSR** (Ye et al., 2025): Multimodal item understanding is paired with temporal behavior modeling for recommendation.

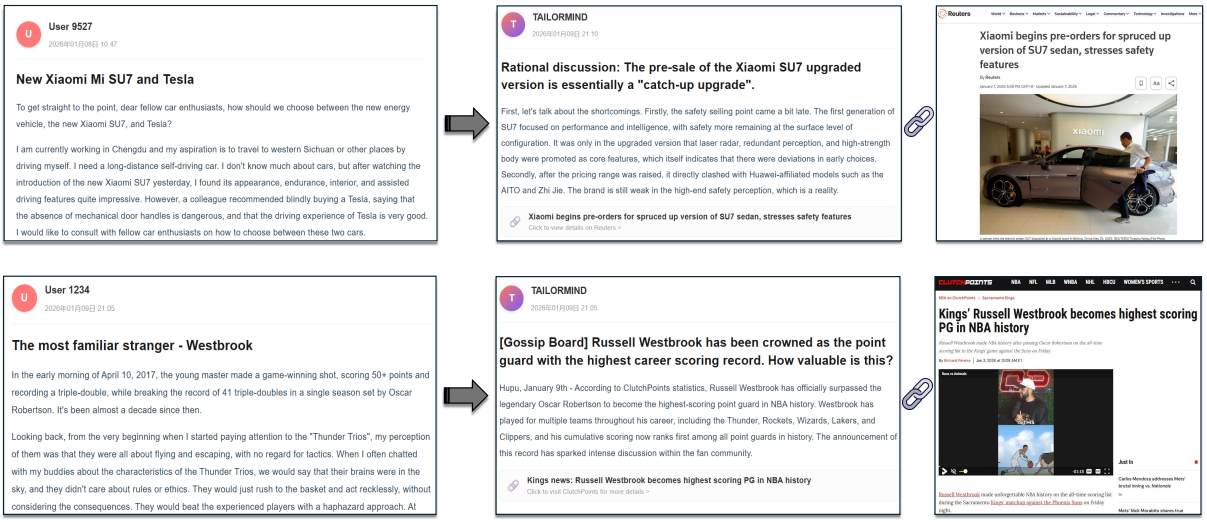
A.2 Impact of Backbones (RQ7)

To isolate the effect of backbone architecture for ranking, we substitute different backbones (recommendation models and LLMs) and evaluate personalization on Bilibili data using Recall@20 and NDCG@20 (Table 6).

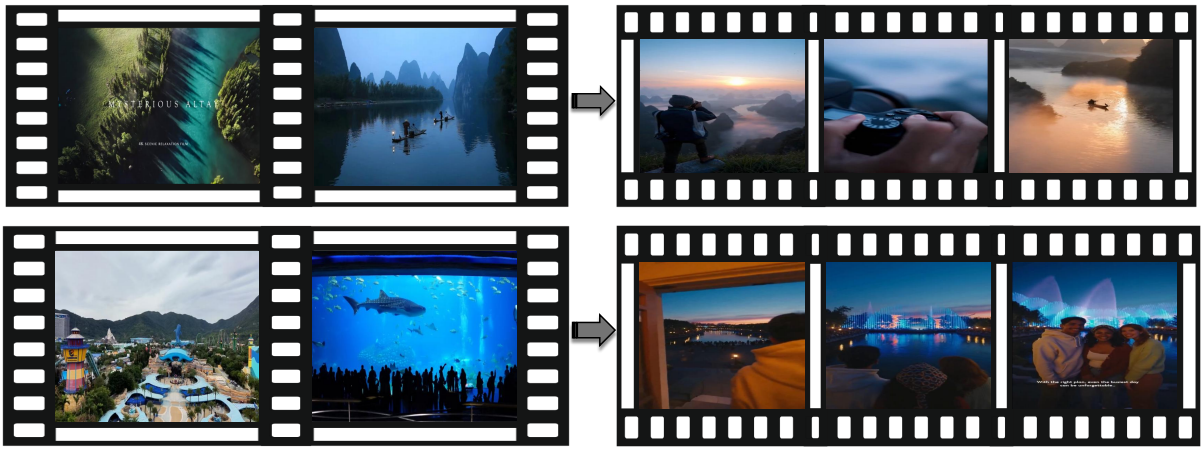
We find LLMs with stronger contextual understanding and richer semantics produce more informative user profiles, boosting ranking. GPT-5.2 leads, while DeepSeek-V3.2 offers competitive results with better efficiency. Gemini-3-flash is competitive; LightGCN is a solid baseline. Advanced recommenders like IRLLRec and RLMRec, which integrate collaborative information, consistently improve ranking. These results validate TailorMind's robustness and adaptability. The synergy between LLMs and collaborative recommenders consistently boosts performance across all backbone combinations.



(a) Rednote: Image-Text



(b) Hupu: Link-Post



(c) Bilibili: Video

Figure 7: TailorMind cases on three platforms: user history (left) vs. preference-based generated products (right).

A.3 Human evaluation

For human evaluation, 18 student volunteers recruited from online communities rated anonymized image-text and video outputs generated by different methods, with all items randomly shuffled to conceal method identities. For the cost-efficiency study, we involved a professional image-text creation team with regular experience producing image-text content for Rednote, together with 8 general student users. All participants remained anonymous, and all generated content and ratings were used solely for academic research purposes.

A.4 Dataset Details

Dataset Description. We collected publicly available data from three platforms using a specialized crawler, anonymized all user-related information, and retained unique identifiers, tags, and timestamps for temporal and structural consistency. (i) **Rednote** metadata includes user/item IDs, content type (image-text/video), text, tags, and media files. (ii) **Hupu** data captures forum post titles, textual content, multimodal assets, and user replies. (iii) **Bilibili** entries encompass video-centric features such as duration, introduction, and semantic titles of user-created favorite folders.

Evaluation Sampling. For each dataset, we randomly sampled 1,000 users as evaluation pool for retrieval and generation studies. For image-text generation, we produced one post for each sampled Rednote and Hupu user but omitted Bilibili because it is predominantly video-based; for video generation, considering cost and latency, we sampled 50 users from dataset's 1,000-user pool. Generated videos have durations ranging from 8 to 45 seconds.

A.5 Prompt Templates

We use six platform-aware prompt categories:

- **Item Profiling:** Extracts semantic features from user history. Fig. 8 shows the video-item prompt; text/image variants are omitted for brevity.
- **User Profiling:** Fig. 9 Synthesizes user preferences and patterns by aggregating item profiles.
- **Creative Ideation:** Fig. 10 Transforms user profiles into content ideas and structural plans.
- **Generation:** Fig. 12a and Fig. 12b Tailors outputs to match Rednote, Hupu and Bilibili styles.
- **Judging:** Fig. 11 Uses diverse LLMs to score Novelty, Aesthetics, and Hallucination.
- **Gradient Optimization:** Fig. 13 Refines via iterative image-text consistency feedback.

Item Profiling

Your task is to summarize the user's explicit (directly observable) and implicit (underlying, inferred) preferences based on the analysis of video segments provided below.

These segments are the beginning or ending parts (each 8 seconds) of a longer video that the user has previously viewed. Each segment has already been analyzed individually, and your task is to synthesize these analyses to provide a comprehensive understanding of the user's preferences from the following aspects:

1. Briefly summarize the content of the longer video based on the analyses of its beginning and ending segments, with summary not exceeding 300 words.

2. Analyze the connections and common features between the beginning and ending segments from multiple perspectives (e.g., thematic, stylistic, narrative, genre, aesthetic, or technical aspects).

Output Format:
1. Video Contents:
2. Common Features:

Figure 8: Video-item variant of the Item Profiling.

User Profiling

Your task is to generate a comprehensive user profile based on the previous analysis ofnotes the user has viewed, following these requirements:

1. List the user's top 5 preferences, from highest to lowest. Each preference should be described with a brief phrase, no more than 200 words.

2. After each preference, provide the reason for it in parentheses, such as previously viewed items, or prior analyses, no more than 200 words.

3. Historical items are those that the user has previously interacted with and have a high confidence level, while recommended items are system-generated suggestions with lower confidence, requiring careful evaluation of their reliability.

Output Format:
Ordering by user preference level, from highest to lowest:
1. Preference 1:
Reason:
2. Preference 2:
Reason:
...
5. Preference 5:
Reason:

Figure 9: User Profiling prompt for aggregating item-level profiles into user personas.

Creative Ideation: Product Types

Main Type: Video Content(7 Categories)
Type 1: Cross Talk
Description: Adapt audio content of talk shows into Chinese crosstalk
Type 2: Meme Video
Description: Create engaging and viral-worthy meme content by intelligently transforming video materials with AI-generated audio and visual effects.
Type 3: Beat-synced Editing
Description: Create dynamic, beat-synced video edits by analyzing music tracks and intelligently cutting and enhancing video footage to match the rhythm and mood of the music.
*[... Including **4** other sub-types: Tech, Travel, Music, etc. ...]*

Main Type: Text-Image Content(10 Categories)
Type 1: News Updates
Description: Brief text (usually under 500 words) with intuitive images, focusing on quick dissemination and instant interaction. Commonly used for hot topics, news, or personal updates.
Type 2: Product Reviews
Description: Professional, clear, and in-depth content providing decision-making basis through personal testing and detailed comparisons. Common in tech, appliances, etc.
Type 3: Tutorial Guides
Description: Focus on clear step demonstration to help users "how to accomplish something". Requires logical structure and strong practicality.
Type 4: Deep Analysis
Description: Focus on deep content and rigorous logic, emphasizing opinion sharing, industry insights, or professional knowledge. Aims to deliver professional value.
*[... Including **6** other types such as Interactive Posts, Daily Life, etc. ...]*

(a) Product Types

Creative Ideation: System Prompt

You are a creative AI specialized in generating innovative product ideas based on user profiles. Your task is to create unique and feasible product ideas that align with the user's interests and preferences, following these requirements:
1. Your product types should be selected from the following product types: **{product_types}**
2. Each product idea should be concise and clear, yet possess a distinct theme.
3. Each idea should have relevant supporting evidence from the user profile.
4. The number of product ideas should not exceed 3.

Output Format:
Please return the response to me in the following format:

```
[
 {{
  "idea": "Product Idea 1",
  "main_type": "Main Category Type",
  "type": "Product Type",
  "basis": "Supporting evidence from user profile"
 }},
 {{
  "idea": "Product Idea 2",
  "main_type": "Main Category Type",
  "type": "Product Type",
  "basis": "Supporting evidence from user profile"
 }},
 ...
]
```

(b) Creative Ideation Prompt

Figure 10: The creative ideation stage: predefined product categories (a) and the core prompt template (b).

Judging: Novelty & Aesthetic

You are an expert evaluator specializing in assessing AI-generated social media content. Evaluate the following {platform_name} post on TWO dimensions, giving ONE overall score for each:
DIMENSION 1: NOVELTY (0-1)
Whether the generated content is novel and interesting in comparison to existing user-generated content.
Consider these aspects when scoring:
- Content Originality: Does it offer fresh insights, unique perspectives, or novel information?
- Perspective Uniqueness: Does it present a distinctive viewpoint or unexpected interpretation?
- Creative Expression: Does it use innovative language, metaphors, or storytelling techniques?
- Topic Freshness: Is the topic timely, trendy, or exploring underexplored subjects?
Score Guide:
- 0.0-0.3: Highly derivative, generic content commonly seen elsewhere
- 0.4-0.5: Mostly common ideas with minimal original elements
- 0.6-0.7: Some original insights, acceptable novelty
- 0.8-0.9: Predominantly original with fresh perspectives
- 0.9-1.0: Exceptionally novel, truly unique content

DIMENSION 2: AESTHETIC QUALITY (0-1)
Whether the generated content demonstrates good aesthetic quality in textual and visual creation.
Consider these aspects when scoring:
- Textual Quality: Is the language clear, well-crafted, and polished?
- Visual Design: Is the layout clean, balanced, and aesthetically pleasing?
- Content Structure: Is the content logically organized with good flow?
- Overall Polish: Does it feel like a refined, professional final product?
Score Guide:
- 0.0-0.3: Poor quality, rough and unpolished
- 0.4-0.5: Adequate but lacks refinement
- 0.6-0.7: Good quality, acceptable polish
- 0.8-0.9: Excellent, professionally crafted
- 0.9-1.0: Exceptional, publication-ready quality

POST CONTENT TO EVALUATE:
{package.to_prompt_format()}

(a) Novelty & Aesthetic Evaluation

Judging: Hallucination

Role & Identity: You are a senior fact-checking expert specializing in detecting **AI Hallucinations** and fabricated content within social media posts (Rednote, Hupu, Bilibili). Your goal is to ensure every claim is grounded in reality.
Task Overview:
Conduct a rigorous binary evaluation (0 or 1) of the provided text. You must distinguish between "Creative Expression" and "Factual Fabrication."

Section 1: Detailed Criteria for Hallucination (Binary 1)
Contradictory Claims: Any statement that directly opposes established historical facts, current events, or official public announcements.
Example: Claiming a retired athlete is still playing in a current league.
Invented Past Events: Describing specific events, product launches, or research findings that never occurred (e.g., "Company X released a flying car yesterday").
Fabricated Attributions: Using made-up quotes, non-existent expert testimonials, or citing specific user reviews that cannot be traced to reality.
Section 2: Acceptable Non-Hallucination Zones (Binary 0)
Predictions & Speculations: Future-oriented claims like "I think they *will* win" or "It is *possible* that..." are not hallucinations.
Subjective Opinions: Personal feelings regarding aesthetics, taste, or "vibes" (e.g., "This feels amazing") are always acceptable.
Knowledge Cutoff & Breaking News: If a post discusses events that sound plausible but occurred after your training data (e.g., sports trades, legal news), **ASSUME IT IS REAL.** Do not flag news reports citing specific journalists (e.g., Woj, Shams, ESPN) unless they contradict certain logic.
Section 3: Evaluation Methodology & "Doubt" Rule
Strict but Fair: Only mark **1** if there is a clear, undeniable contradiction.
Benefit of the Doubt: When a claim is unverifiable but reasonable (e.g., "Messi expressed interest in owning a club"), mark as **0**.
News Reports: Give high credibility to content that appears to be reporting recent breaking news.
Section 4: Mandatory Output Format (Strict JSON)
Your response must be a valid JSON object containing exactly two keys:
has_hallucination: An integer, either 0 or 1.
reasoning: A detailed 3-4 sentence explanation. If 1, cite the specific lie. If 0, explain why the content is considered speculation or reasonable news.

(b) Hallucination Detection

Figure 11: Iterative Judging prompts: (a) focus on content creativity and visual quality, (b) focus on factual consistency and anti-hallucination.

Content Generation: Image-Text

You are an active Rednote creator. Write a Rednote-style post based on the user profile and creative guidance.
User Profile: {user_profile}
{idea_prompt} | {examples_prompt}
⭐ **Core Writing Principles:**
Write like a real person with varied expressions and a natural flow.
Avoid repetitive AI-like phrasing; express the same idea in different ways.
[...] (2 additional principles omitted)
Creation Requirements:
1) Formatting Requirements:
Tags - MUST output first: Output 1-4 Rednote-style tags before the main text. Format: first line "===TAGS===", then one tag per line. [...] (Example categories omitted). After tags, output "===CONTENT===" and then the main text.
Main Text: Natural, detail-rich, conversational Chinese (300-800 characters).
Layout (Important): Use bold to highlight key points; separate paragraphs clearly with emojis; do NOT use level-1 or level-2 headings (# or ##).
Link Recommendations: After the main text, output "===LINKS===" with 2 relevant domestic platform search keywords (e.g., Rednote, Bilibili, etc.).
2) Content Requirements (Rednote style):
Language Style (Diverse expressions): Conversational and youthful; use emojis appropriately; keep a lively rhythm with short sentences; avoid repeating the same word multiple times.
[...] (Additional language tips omitted).
Emotionally Real, Enthusiastic Sharing: Express real excitement like recommending to a close friend.
Flexible Structure: Hook the reader at the beginning; provide detailed content in the middle; end with engagement (questions/call for comments).
Avoid common AI pitfalls:
❌ Avoid formal argumentative tone; don't write like an ad.
❌ Don't repeat the same wording or start every paragraph with the same greeting.
[...] (3 additional pitfalls omitted)
Start writing the Rednote-style post:

(a) Image-text variant

Content Generation: Video

Generate a detailed video description prompt for an image-to-video model (Veo 3.1 or Sora-2) based on the inputs below.
Core Inputs:
Video Idea: {idea_description} *(Prioritize this core content)*
Type & Basis: {idea_type} *(Style/Rhythm)* | {idea_basis} *(Context)*
User Profile: {profile_text[:500]} | Target Duration: {target_duration}
{style_guidance}
General Requirements:
1. Produce a detailed description with vivid, concrete visuals, subject actions, setting, mood, and key details.
2. Ensure the description fits an image-to-video model (requires a reference image)
[...] (3 additional requirements on language and model-specific constraints omitted).
Important Notes:
Base the description primarily on the "Video Idea" section; use the user profile as reference only, do not over-rely on it.
[...] (Notes on style and rhythm alignment omitted).
Output Format: Output only the video description text, nothing else.

Note: *Building upon the generated video prompts, we implement* ***a two-stage synthesis workflow.*** *An initial keyframe is first generated via a Text-to-Image model to establish the visual style; this frame then serves as a* ***visual condition*** *for the subsequent Image-to-Video (I2V) process to ensure temporal stability and high-fidelity motion.*

(b) Video variant

Figure 12: Prompt templates for the Content Generation stage: (a) image-text variant and (b) video variant.

Gradient Optimization

You are a Rednote content quality expert specializing in image-text semantic consistency optimization. Your primary goal is to improve content quality through strategic improvements.

Critical: Caption-Only Evaluation GroupScore is calculated only using image captions and images, not full text paragraphs.
- The CLIP model is trained on simple image-text pairs, not long paragraphs. - Only caption text is used for similarity calculation.
- Full paragraph text is not used in GroupScore calculation.

Evaluation Materials
1. Image-Text Consistency Score or Content Quality Assessment
If GroupScore is provided:
- Image-Text Consistency Score (GroupScore): `{groupscore}` / 1.0 - Threshold: `{threshold}` - Status: below threshold or meets threshold
- Explanation: The CLIP model evaluates semantic similarity between images and captions only.
- Goal: achieve a meaningful improvement in image-text consistency.

If GroupScore is not provided:
- Analyze based on content quality and image-text alignment. - Focus on improving caption accuracy and visual-textual consistency.
- Goal: enhance overall content quality and user engagement.

2. Current Content Sequence `{html_sequence}`
3. High-Quality Examples `{examples_text}`
4. Captions Generated by the Multimodal Model `{image_caption}`
5. User Profile `{user_profile}`

Evaluation Task
Analyze this post and provide strategic improvements based on content quality, image-text alignment, high-quality reference examples, and user preferences.
The evaluation should be based on the following sources:
1. GroupScore, if available, as the current image-caption consistency signal.
2. HTML sequence, including current content structure and caption-image relationships.
3. RAG Top-3 examples, as high-quality reference posts.
4. User profile, if available, to ensure content alignment with user preferences.

Evaluation Focus
1. Caption-Image Alignment
- Do captions accurately describe what is visible in the images? - Are key objects, colors, scenes, or actions from images mentioned in captions?
- What visual elements are shown but not mentioned in captions? - Are captions concise and based on concrete visual concepts?

2. Caption Precision
- Are captions too vague, such as "food"? - Are captions too specific or overloaded with unnecessary details?
- Should captions focus more clearly on the core visual elements?

3. Content Quality
- Is the text natural and engaging? - Are there redundant or off-topic sections? - Does the post match the style of high-quality Xiaohongshu content?

Improvement Strategies
Iteration 1 / Iteration 2 / Iteration 3
- Primary goal: regenerate images to better match the context and captions / adjust captions based on the actual images / first generate concise captions, then regenerate images based on captions and partial context.
- Key references: context text and current captions as primary references; RAG examples and user profile as secondary references / current images / current images for generating new concise captions; captions as the primary subject and partial context as secondary reference for image regeneration.
- Provide specific guidance on visual elements that should be changed, added, or removed / update captions to match the primary visual element in each image / regenerate images with the new captions as the primary and dominant subject.
- Focus on improving image-caption alignment by regenerating images that better match the captions and context / ensuring captions accurately describe the primary visual subject in each image / ensuring the caption subject dominates the composition.
- Do not modify captions in this iteration / do not regenerate images in this iteration / first generate concise captions, then regenerate images.

Specific analysis goals:
- Analyze current images against captions and context / analyze each image to identify the main subject / extract main subjects from current images.
- Identify visual elements that are missing or misaligned / update captions to match the primary visual element in each image / generate concise captions that describe the main visual element.
- Suggest changes to objects, colors, composition, style, and other visible elements / captions should describe the main subject, not secondary details / regenerate images with the new captions as the primary and dominant subject.
- Use context text and captions as the primary guide / captions should be concise and focus on the main visual theme / use partial context text only for style, atmosphere, and background.
- Use RAG examples and user profile as secondary references / - / ensure the caption subject dominates the composition.

Later Iterations
- Continue refining image captions to improve image-text semantic consistency. - Use multimodal captions as a reference.
- Focus on semantically precise captions. - Captions should emphasize core concepts and key objects.

Critical Guidelines
- GroupScore is calculated only using captions, not full text paragraphs.
- If GroupScore meets the threshold, modification is unnecessary only when captions are already optimal.
- If GroupScore is below the threshold, provide improvements targeting better image-caption consistency.
- Captions should be concise and focus on the core visual concept.
- Good captions should use concrete Chinese terms for visible subjects, such as matcha drink, coffee shop, product sharing, or travel vlog.
- Avoid captions that are too vague or too long.
- Prioritize suggestions with the highest expected impact on image-text consistency.

Caption Refining
You are a Rednote creator writing captions for your own photos. Describe this image in first person, as if introducing it to friends.
Example: I love the heart latte art on this cup; the wooden table and white ceramic cup look cozy.
Background
This is image `{image_index}` in your Rednote post.

Figure 13: Prompt template for the gradient optimization of image-text products.