# GRECF: GENERATIVE RECOMMENDATION VIA CONTINUOUS-SPACE COLLABORATIVE FILTERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generative recommendation aims to create and recommend novel items beyond fixed catalogs, addressing users' evolving or unrepresented preferences that existing catalogs cannot adequately capture. It is important to generate content aligned with users' historical preferences while reliably extrapolating beyond observed history. We formulate this task as collaborative preference-field completion in continuous item space, with user-based and item-based collaborative filtering providing two completion mechanisms. We instantiate this formulation in GReCF, our proposed reference-free generative recommender in the image modality. GReCF represents users with adaptive multi-interest themes, anchors each generation to one theme, and injects user and theme conditions into a pretrained diffusion model through residual preference attention. Shared training transfers evidence across users, while masked-interest training learns relations among co-occurring themes to support extrapolation beyond observed history. To support generative recommendation research, we construct CIGR, a synthetic dataset with one million positive interactions. We also introduce an extrapolation-focused evaluation protocol specifically for the generative recommendation task. Across three datasets, GReCF achieves competitive ranking, strong personalization, favorable image quality, high coverage, and reliable CF extrapolation with few unsupported generations. We further investigate the cold-start generation capability of GReCF and show that it preserves initial interests while enabling CF extrapolation.[1]

## 1 INTRODUCTION

Conventional recommender systems infer user preferences from interaction histories and select items from a fixed catalog (Hu et al., 2008; Rendle et al., 2009). However, they cannot satisfy preferences that are not represented by existing candidates. Rajput et al. (2023); Zheng et al. (2024) generate semantic item identifiers, while Qu et al. (2026) generate continuous item representations. These generative approaches ultimately return items from a fixed catalog. Instead of ranking only existing items, generative recommendation addresses this limitation by creating and recommending new content in a continuous item space (Hou et al., 2026), making it particularly valuable for open-ended visual and creative domains. Similarly, personalized generation is widely studied for adapting generated content to individual preferences (Shen et al., 2024; Chen et al., 2024; Xu et al., 2025a;b). It typically starts from a prompt, target item, or reference image that specifies the main content, while the user's history controls its personalization. By contrast, generative recommendation, which receives no reference, must decide how to create new content that aligns with users' preferences.

A generative recommender should not simply reproduce the user's observed preferences. It should also generate novel items that the user may prefer, with this extrapolation supported by recommendation evidence. Recent research demonstrates generation beyond fixed catalogs (Xu et al., 2024; Patron, 2025; Wang et al., 2025; Liu et al., 2025; Zhou et al., 2026), but does not explicitly formulate collaborative-evidence-grounded extrapolation as a central problem. Collaborative filtering offers a natural basis through population-level patterns among similar users and correlated items (Resnick et al., 1994; Sarwar et al., 2001; Wang et al., 2019). Early research also explores extending this principle beyond finite rating matrices (Abernethy et al., 2009; Guo et al., 2010; Houlsby et al., 2012; Kim et al., 2016). Therefore, it is vital to clarify how this principle applies to continuous space.

We address this problem in two stages. First, we formulate generative recommendation as collaborative preference-field completion in a continuous-item space. We complete this field by combining

[1] Our codes are released at https://anonymous.4open.science/r/GReCF-BFE4/



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

direct historical evidence with user- and item-side collaborative evidence, then use the estimate to define the user-conditional generation distribution. Second, we instantiate this formulation in GReCF, our proposed reference-free generative recommender in the image modality. GReCF represents each user with an adaptive multi-interest theme set, anchors each generation to a selected theme, and uses masked-interest training with residual preference attention to learn personalized generation from a pretrained diffusion model. We further construct a dedicated generative recommendation dataset and an extrapolation-centered evaluation protocol. Across three datasets, GReCF achieves strong overall performance and reliable collaborative extrapolation beyond observed user histories.

Our main contributions are summarized as follows: (1) We formulate generative recommendation as collaborative preference-field completion in a continuous item space, providing a principled formulation of user- and item-based collaborative filtering beyond discrete catalogs. (2) We propose GReCF, instantiating this formulation for reference-free image recommendation with adaptive multi-interest user representation, theme-anchored conditioning, residual preference attention, and masked-interest training, outperforming most existing approaches in personalization and preference extrapolation. (3) We construct a dedicated generative recommendation dataset and an extrapolation-centered evaluation protocol assessing full-pool ranking, personalization, coverage, quality, and collaborative preference extrapolation, providing a framework for future research.

## 2 Related Work

Recommendation evolves from conventional collaborative filtering to generative retrieval over catalog representations (Tay et al., 2022; Zhuang et al., 2022). Conventional collaborative filtering uses user-user similarity (Resnick et al., 1994) or item-item correlation (Sarwar et al., 2001). Early research also extends collaborative filtering beyond finite rating matrices. Abernethy et al. (2009) formulate it as operator estimation over user and item feature spaces, while Guo et al. (2010); Houlsby et al. (2012); Kim et al. (2016) use Gaussian processes and kernels to model multi-user preferences and user-item similarities. Koren et al. (2009); He et al. (2017) learn latent user-item representations, while Wang et al. (2019); He et al. (2020) propagate them on the interaction graph and Sun et al. (2019) model masked sequential histories to capture evolving user preferences.

Hou et al. (2026) adopt broad definitions of generative recommendation and personalized generation that encompass both catalog recommendation and creation of new content. Along the former line, Rajput et al. (2023); Zheng et al. (2024) generate semantic item identifiers, while Yang et al. (2023); Wang et al. (2023); Li et al. (2023); Zhao et al. (2024); Qu et al. (2026) generate or denoise continuous preference representations before catalog retrieval. These approaches still return existing items. In contrast, catalog-free generative recommendation creates and recommends new items directly from user histories (Wang et al., 2025). Xu et al. (2024) jointly generates fashion items under category, compatibility, and user-history conditions, while Patron (2025) generates personalized image representations from user IDs and ratings and decodes them into images with a frozen generator. Liu et al. (2025) generates the next image directly from multimodal history, whereas Zhou et al. (2026) generates instructions for a frozen downstream generator.

Personalized generation uses a user's history to generate a personalized version of a specified item. Shen et al. (2024) represents history with preference keywords and implicit embeddings. Xu et al. (2025a) jointly encodes noisy history images and multimodal instructions with preference-alignment training. Wang et al. (2024) rewrites target descriptions with history tags before retrieving related text and image evidence. For target-conditioned personalization, Ling et al. (2026) retrieves reference-relevant history, Bian et al. (2025) fuses user embeddings with item title and reference-image context under behavior-trained preference rewards, Xu et al. (2025b) disentangles history-derived style from reference semantics, and Chen et al. (2024) rewrites current prompts using prompt history. General customization learns concept tokens or model parameters from references (Gal et al., 2022; Ruiz et al., 2023; Kumari et al., 2023), while IP-Adapter adds lightweight image conditioning (Ye et al., 2023). Preference-aware methods use liked/disliked attributes (Salehi et al., 2024), preference embeddings (Dang et al., 2025; Mo et al., 2025; Wang et al., 2026), encoder attention (Kim et al., 2026), and graph-neighbor DPO (Dunlop et al., 2025).

## 3 The Approach

We propose GReCF in two stages. We first formulate generative recommendation as continuous-space collaborative preference completion. We then instantiate this formulation approximately as

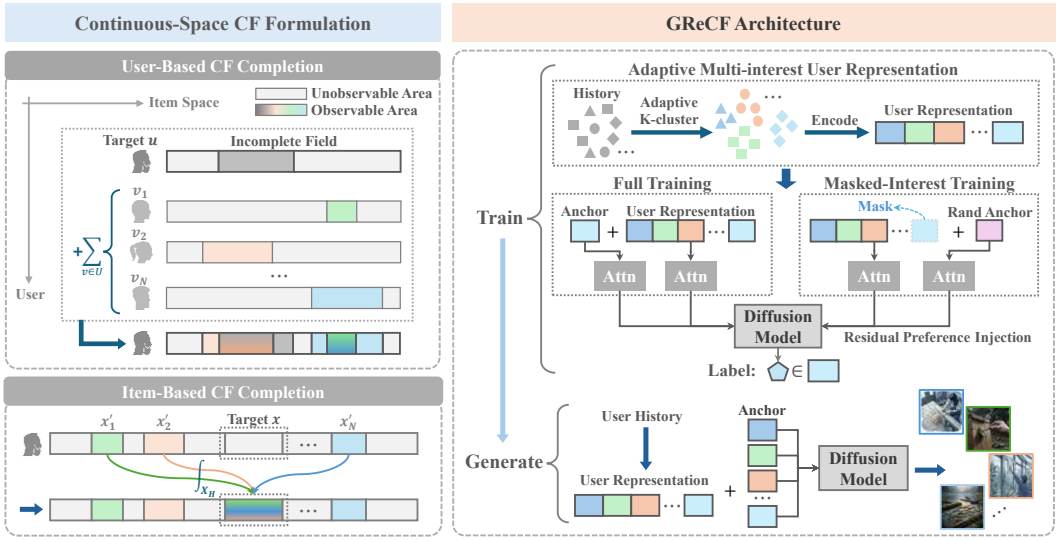


Figure 1: Left: Formulation of continuous-space collaborative filtering, where user-based completion transfers evidence across users (top), and item-based completion captures transitions between co-occurring themes (bottom). Right: Architecture of GReCF. Training adaptively encodes user histories, with labels determining the target theme used as the full-path anchor and masked in the masked-interest path. During generation, each theme is used in turn as an anchor.

GReCF in the image modality. Section 3.1 presents the continuous-space formulation and its collaborative completion mechanisms, while Section 3.2 describes the implementation in GReCF.

### 3.1 Formulation of Continuous-Space Collaborative Filtering

We formulate collaborative filtering in continuous item space as completing the preference field induced by sparse user histories, as shown on the left side of Figure 1. User-based completion transfers evidence across users, while item-based completion propagates evidence between co-preferred item regions, jointly recovering preferences unobserved in the target user's history.

Let $\mathcal{X}$ denote the space of all possible items, where $x \in \mathcal{X}$ represents any item, whether or not it belongs to an existing catalog. Each user $u$ has a partially observable preference field $f_u^\star : \mathcal{X} \to \mathbb{R}$, whose value $f_u^\star(x)$ measures the user's latent preference for item $x$. This field is complete only as a conceptual object, including items the user never encounters and items that remain unobserved.

In practice, user preference is observed only through sparse implicit feedback, so the history of user $u$ is a finite set of positive observations $H_u = \{x_{u,1}, x_{u,2}, \ldots, x_{u,n_u}\}$. These observations provide evidence about $f_u^\star$, but they do not reveal the full field. In particular, the absence of interaction at a location $x$ does not imply that $f_u^\star(x)$ is low. It only means that this preference is unobserved. Accordingly, we denote the incomplete preference field inferred from user $u$'s finite history by $\widehat{f}_u^{\text{his}}(x; H_u)$. Since all observed histories cover only a data-supported portion of the continuous space, let $\mathcal{H} = \{H_w\}_{w \in \mathcal{U}}$ denote all histories and let $\mathcal{X}_\mathcal{H} \subseteq \mathcal{X}$ denote the item domain supported by the dataset. Preference completion is defined on this domain, with the completed estimate satisfying $\widehat{f}_u^{\text{comp}}(x; H_u) \approx f_u^\star(x)$ for $x \in \mathcal{X}_\mathcal{H}$. This estimate approximates the restriction of the latent field $f_u^\star$ to the data-supported domain rather than recovering the entire continuous field.

Collaborative filtering provides the mechanism for this completion, which completes missing entries by exploiting user similarity and item correlation. Continuous-space collaborative filtering preserves the core idea of discrete CF while extending completion beyond finite neighborhoods. Instead of relying only on top-$K$ similar users or items, it aggregates preference evidence from all observed users and item locations in the data-supported space.

The user-based completion operator aggregates history-induced fields from all users:

$$\widehat{f}_u^{\text{user}}(x; H_u) = \sum_{v \in \mathcal{U}} a_{uv} \widehat{f}_v^{\text{his}}(x; H_v). \tag{1}$$

$x$ is a candidate item location for user $u$, $v$ indexes a source user, and $a_{uv}$ weights that user's evidence by similarity. The operator completes $u$'s unobserved preferences using histories from related users. In contrast, the item-based completion relies on co-preference between item locations.

In the discrete setting, items that frequently co-occur in user histories are considered related. We generalize this idea to continuous space by measuring whether many users assign above-average history-induced preference to two item locations. Let the population mean field for item $x$ be:

$$\overline{\widehat{f}^{\mathrm{his}}}(x;\mathcal{H}) = \frac{1}{|\mathcal{U}|}\sum_{w\in\mathcal{U}} \widehat{f}_w^{\mathrm{his}}(x;H_w). \tag{2}$$

The observable item-side co-preference kernel of item $x$ and $x'$ is defined as:

$$k_{\mathrm{item}}(x,x';\mathcal{H}) = \frac{1}{|\mathcal{U}|}\sum_{w\in\mathcal{U}} \Big(\widehat{f}_w^{\mathrm{his}}(x;H_w) - \overline{\widehat{f}^{\mathrm{his}}}(x;\mathcal{H})\Big)\Big(\widehat{f}_w^{\mathrm{his}}(x';H_w) - \overline{\widehat{f}^{\mathrm{his}}}(x';\mathcal{H})\Big). \tag{3}$$

This kernel is the continuous analogue of an item-item similarity matrix. A positive value of $k_{\mathrm{item}}(x,x';\mathcal{H})$ indicates that users who show above-average historical preference for $x'$ also tend to show above-average historical preference for $x$. The centering by the population mean removes global popularity, so the kernel captures relative co-preference. Given this kernel, every historical positive location of the target user provides one item-side collaborative signal for a candidate location through its co-preference with that candidate, and aggregating these signals yields the item-based completion score. Formally,

$$\widehat{f}_u^{\mathrm{item}}(x;H_u) = \int_{\mathcal{X}_{\mathcal{H}}} [k_{\mathrm{item}}(x,x';\mathcal{H})]_+ \, d\mu_u^{H_u}(x'), \qquad \mu_u^{H_u} = \frac{1}{n_u}\sum_{j=1}^{n_u} \delta_{x_{u,j}}. \tag{4}$$

This expression makes the semantics of item-based completion explicit. If the candidate location $x$ is positively co-preferred with many historical positives $x_{u,j}$ of user $u$, then the item-side evidence for $x$ becomes strong, even if $x$ itself never appears in the user's own history.

User- and item-based completion provide complementary evidence. We combine them with the history-induced field to obtain a completed preference estimate:

$$\begin{aligned}\widehat{f}_u^{\mathrm{comp}}(x;H_u) = {} & C_u(x)\widehat{f}_u^{\mathrm{his}}(x;H_u) \\ & + \big(1-C_u(x)\big)\Big(\beta\,\widehat{f}_u^{\mathrm{user}}(x;H_u) + (1-\beta)\,\widehat{f}_u^{\mathrm{item}}(x;H_u)\Big).\end{aligned} \tag{5}$$

$C_u(x) \in [0,1]$ measures confidence in historical evidence, while $\beta$ balances user- and item-based completion. GReCF models this balance implicitly through its user and theme branches, trainable projections, and layer-wise gates instead of estimating these terms as separate scalar functions.

The completed estimate then defines the distribution from which a generative recommender should sample. Instead of selecting a fixed catalog item, the system chooses a user-conditional generation distribution whose probability mass concentrates on high-score regions of $\widehat{f}_u^{\mathrm{comp}}$:

$$p_\theta(x \mid H_u, c_0) \approx \widehat{P}_u^{\mathrm{rec}}(dx), \qquad \widehat{P}_u^{\mathrm{rec}}(dx) = \widehat{Z}_u^{-1}\exp\Big(\widehat{f}_u^{\mathrm{comp}}(x;H_u)/\tau\Big)\,P_0(dx). \tag{6}$$

$P_0$ is the population-level item distribution reflecting general preferences. $c_0$ is a user-independent base generation condition. $\tau$ is a temperature, and $\widehat{Z}_u$ is the normalizing constant.

Conventional discrete recommendation is recovered as a degenerate case. Let $a = i$ denote selecting catalog item $x_i$. Deterministic selection is then represented by $p_\phi(x \mid a = i) = \delta_{x_i}(x)$, where $\delta_{x_i}$ denotes a point mass concentrated at $x_i$. Recommendation therefore reduces to scoring isolated catalog points. This shows that our continuous-space collaborative-filtering formulation subsumes and explains its discrete-space counterpart.

### 3.2 GReCF: An Instantiation in the Image Modality

GReCF instantiates the above formulation approximately as a reference-free image generator. It represents the history-induced preference field with adaptive multi-interest theme prototypes and selects one theme to anchor each generated image. The resulting user and theme conditions are injected into a pretrained diffusion model through residual preference-attention branches. Shared training and masked-interest training approximate user-based and item-based completion, respectively. The architecture of GReCF is shown on the right side of Figure 1.

#### 3.2.1 Adaptive Multi-Interest User Representation

For most users, preference is not concentrated around a single mode. Their positive histories instead consist of several compact theme clusters. Encoding all clusters into one vector mixes semantically distinct interests and discards their local structures. GReCF therefore partitions the history of each

user into an adaptive number $K_u$ of themes. Let $e : \mathcal{X} \to \mathbb{S}^{d-1}$ be a frozen image encoder and let $H_{u,k} \subseteq H_u$ denote the items assigned to theme $k$. We define:

$$p_{u,k} = \mathrm{Norm}\big(\sum\nolimits_{x \in H_{u,k}} e(x)\big), \qquad w_{u,k} = \tfrac{|H_{u,k}|}{|H_u|}. \tag{7}$$

$p_{u,k}$ is the normalized center of theme $k$, and $w_{u,k}$ is the proportion of the user's history assigned to that theme. The number $K_u$ is selected adaptively from the user's history, while low-support themes and redundant themes are removed, as detailed in Appendix A. The resulting weighted theme set $\mathcal{P}_u = \{(p_{u,k}, w_{u,k})\}_{k=1}^{K_u}$ is the user representation used by GReCF. Unlike a single pooled embedding, $\mathcal{P}_u$ preserves each theme cluster as an individual component, so the complete collection of themes jointly characterizes the user's history-induced preference field.

### 3.2.2 Theme-Anchored Dual Conditioning

The full theme set identifies what the user likes, but it does not by itself specify which interest should be expressed in a particular generated image. Conditioning only on the user representation therefore leaves the generator with an ambiguous multi-peak target. GReCF resolves this ambiguity with a theme anchor. After padding the adaptive set to $K_{\max}$ slots with validity mask $m_u$, it constructs a user condition that retains all valid themes:

$$Z_u^{\mathrm{user}}(m) = [m_k\, \phi_{\mathrm{user}}(p_{u,k})]_{k=1}^{K_{\max}}. \tag{8}$$

For a training image $x$, its nearest valid theme determines the anchor index $q(u, x)$, and the corresponding theme condition is obtained by projecting the anchor prototype:

$$q(u, x) = \arg\max_{k: m_{u,k}=1} e(x)^\top p_{u,k}, \qquad Z_u^{\mathrm{theme}}(q) = \phi_{\mathrm{theme}}(p_{u,q}). \tag{9}$$

The actual condition supplied to the generator is therefore the pair

$$C_u(q) = \big(Z_u^{\mathrm{user}}(m_u), Z_u^{\mathrm{theme}}(q)\big). \tag{10}$$

$Z_u^{\mathrm{user}}(m_u)$ preserves the user's overall multi-interest profile, whereas $Z_u^{\mathrm{theme}}(q)$ anchors the current output to one concrete theme direction. At inference time, no target or reference image is given. When generating multiple images, GReCF allocates theme anchors according to their proportions in the user's history, thereby covering different observed interests.

### 3.2.3 Residual Preference Injection

A pretrained diffusion model captures a population-level image distribution that supports high-quality generation but is not personalized to a particular user. GReCF preserves this population prior and learns only how user and theme conditions should shift the denoising trajectory toward a personalized region. Specifically, it adds two parallel preference-attention branches to each cross-attention layer of the frozen UNet. For layer $\ell$, the resulting attention response is

$$A^{(\ell)} = A^{(\ell)}_{\mathrm{base}} + g^{(\ell)}_{\mathrm{user}}\, \mathrm{Attn}\Big(Q^{(\ell)}, Z_u^{\mathrm{user}}(m_u)\Big) + g^{(\ell)}_{\mathrm{theme}}\, \mathrm{Attn}\Big(Q^{(\ell)}, Z_u^{\mathrm{theme}}(q)\Big). \tag{11}$$

$A^{(\ell)}_{\mathrm{base}}$ is the original attention response and $Q^{(\ell)}$ denotes the UNet query. The other two attention branches use the full user representation and the theme anchor, respectively. They have independent trainable projections and learnable gates, while the original UNet parameters remain frozen. Their additive form makes the learned preference conditions residual corrections to the population model.

### 3.2.4 Masked-Interest Training for Collaborative Preference Completion

Training only with the label-associated anchor mainly teaches GReCF to reproduce observed interests. We therefore introduce masked-interest training. Masked-interest training applies only to users with at least two themes, whereas single-theme users use the unmasked path. For each training image $x$, let $t = q(u, x)$ denote its target theme, where $q(u, x)$ is the nearest-theme anchor index defined above. With probability $1 - \rho$, the model receives the complete user condition and uses $t$ as the anchor. With probability $\rho$, the target theme is removed from the user condition, and another visible theme $r$ is sampled as the anchor. Both paths retain $x$ as the supervision target:

$$\widetilde{C}_u(x) = \begin{cases} \big(Z_u^{\mathrm{user}}(m_u), Z_u^{\mathrm{theme}}(t)\big), & b = 0, \\ \big(Z_u^{\mathrm{user}}(m_u \odot (1 - \mathbf{1}_t)), Z_u^{\mathrm{theme}}(r)\big), & b = 1, \end{cases} \tag{12}$$
$$b \sim \mathrm{Bernoulli}(\rho), \qquad r \sim \mathrm{Unif}(\mathcal{I}_u \setminus \{t\}), \qquad \mathcal{I}_u = \{k : m_{u,k} = 1\}.$$

$\mathcal{I}_u$ is the set of valid theme indices and $\mathbf{1}_t$ is the one-hot vector marking the target slot. When $b = 1$, the target theme is absent from both the user condition and the anchor, so the model must infer it from the user's remaining themes while reconstructing the original target image.

Let $z_0$ be the latent representation of target image $x$, let $\epsilon \sim \mathcal{N}(0, I)$ be Gaussian noise, and let $z_s = \alpha_s z_0 + \sigma_s \epsilon$ be the noisy state at diffusion step $s$, where $\alpha_s$ and $\sigma_s$ are determined by the noise schedule. The masked and unmasked paths are trained with the same noise-prediction objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\left\|\epsilon - \epsilon_\theta\Big(z_s, s, c_0, \widetilde{C}_u(x)\Big)\right\|_2^2\right]. \tag{13}$$

Training additionally uses a contrastive auxiliary loss $\mathcal{L}_{\text{ce}}$ and a residual regularizer $\mathcal{L}_{\text{reg}}$. Specifically, $\mathcal{L}_{\text{ce}}$ averages two temperature-scaled in-batch cross-entropy losses that align the projected mean user and theme residuals with their corresponding centered, normalized user and theme representations. The residual regularizer $\mathcal{L}_{\text{reg}}$ is the mean squared magnitude of the concatenated preference residual tokens. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{14}$$

Together, these components provide a parametric approximation to the completed preference field introduced in Section 3.1. Shared training approximates user-side completion by jointly training one conditional generator on all user histories, so similar theme sets share conditions and transfer preference evidence across users. A theme observed for one user can therefore receive higher generation probability for another similar user even when it is absent from that user's history. Masked-interest training approximates item-side completion by hiding the target theme and reconstructing its image from the remaining valid themes and a different visible anchor. Repeated co-occurrence leads the shared generator to implicitly learn transitions between preference regions, encoding the co-preference kernel in its parameters without estimating it explicitly. Thus, when a user's history expresses one theme but lacks a correlated theme, GReCF can infer the relationship from other users and increase the generation probability of the unobserved theme.

# 4 Experimental Setup

## 4.1 CIGR: Continuous Item-Space Generative Recommendation Dataset

CIGR is a synthetic user-image interaction dataset that we construct to model real-world recommendation scenarios. It represents users with multiple preference centers and captures population-level clustering among related preferences. During construction, we explicitly shape the interaction process to exhibit collaborative-filtering patterns, including preference similarity among users and recurring co-preference among related item regions. The detailed construction procedure is provided in Appendix B. CIGR contains 10,000 users, 100,000 images, and 1,000,000 positive interactions. It provides a controlled setting for testing whether a generative recommender can reproduce observed preferences and extrapolate toward unobserved interests supported by collaborative evidence. CIGR enables direct empirical evaluation of our formulation and provides a dedicated dataset for future research on continuous-space generative recommendation.

## 4.2 Evaluation Protocol

We evaluate generative recommendation from five perspectives: conventional recommendation performance, personalization, image quality, coverage, and collaborative preference extrapolation. For conventional recommendation performance, each generated image retrieves its $K$ nearest catalog images in CLIP space. We evaluate the top-$K$ results against the test items using HitRate@$K$, Recall@$K$, and NDCG@$K$. CLIP Similarity and LPIPS evaluate semantic and perceptual similarity to held-out preferences. Coverage measures how many adaptive themes in the user's test set are covered by the generated images. The LAION Aesthetic Score measures image quality.

To determine whether generated content reproduces observed preferences, extrapolates toward new interests supported by collaborative evidence, or deviates from both sources, we use a two-stage LLM Judge. It first determines whether a generated image can be explained by the target user's history and then, when necessary, whether it is supported by a CF candidate pool from similar users, classifying each image as Direct History, CF Extrapolation, or Other. Details of candidate retrieval, the two-stage decision process, and the prompts are provided in Appendix C. Due to the subjective nature of LLM judge, we recruit 40 volunteers to verify the accuracy of the LLM attribution.

## 4.3 Implementation Details

We use CIGR, MovieLens-CF5K, and PixelRec50K-GT50 as our evaluation datasets. MovieLens-CF5K contains 5,000 users, 6,194 posters, and 497,770 interactions with ratings of at least 4.0 from

Table 1: Main results on CIGR, MovieLens-CF5K, and PixelRec50K-GT50. We compare GReCF with ViPer (Salehi et al., 2024), REBECA (Patron, 2025), GNR (Liu et al., 2025), and NaviGen (Zhou et al., 2026). The evaluation assesses full-pool ranking, personalization, coverage, image quality, and collaborative preference extrapolation.

**(a) Full-pool ranking under conventional recommendation metrics.**

| Dataset | Approach | HitRate | | | Recall | | NDCG | |
|---|---|---|---|---|---|---|---|---|
| | | @10 | @20 | @50 | @20 | @100 | @20 | @100 |
| CIGR | ViPer | 0.0070 | 0.0137 | 0.0354 | 0.0015 | 0.0076 | 0.0011 | 0.0033 |
| | REBECA | 0.0165 | 0.0325 | 0.0704 | 0.0039 | 0.0169 | 0.0028 | 0.0075 |
| | GNR | 0.0010 | 0.0019 | 0.0040 | 0.0002 | 0.0009 | 0.0002 | 0.0004 |
| | NaviGen | **0.0194** | 0.0379 | 0.0726 | 0.0041 | **0.0172** | 0.0031 | **0.0083** |
| | **Ours (SD1.5)** | 0.0170 | **0.0380** | **0.0760** | **0.0045** | **0.0172** | **0.0032** | 0.0078 |
| | **Ours (SDXL1.0)** | 0.0174 | 0.0315 | 0.0621 | 0.0038 | 0.0142 | 0.0028 | 0.0066 |
| | **Ours (PixArt)** | 0.0099 | 0.0185 | 0.0405 | 0.0021 | 0.0085 | 0.0016 | 0.0040 |
| MovieLens -CF5K | ViPer | 0.0227 | 0.0430 | 0.0986 | 0.0042 | 0.0196 | 0.0035 | 0.0095 |
| | REBECA | 0.0222 | 0.0410 | 0.0945 | 0.0041 | 0.0202 | 0.0033 | 0.0096 |
| | GNR | 0.0208 | 0.0356 | 0.0748 | 0.0035 | 0.0147 | 0.0030 | 0.0073 |
| | NaviGen | 0.0183 | 0.0329 | 0.0753 | 0.0032 | 0.0152 | 0.0027 | 0.0073 |
| | **Ours (SD1.5)** | **0.0398** | **0.0756** | **0.1719** | **0.0080** | **0.0371** | **0.0064** | **0.0178** |
| | **Ours (SDXL1.0)** | 0.0234 | 0.0476 | 0.1055 | 0.0048 | 0.0217 | 0.0038 | 0.0103 |
| | **Ours (PixArt)** | 0.0210 | 0.0413 | 0.0986 | 0.0042 | 0.0200 | 0.0034 | 0.0095 |
| PixelRec50K -GT50 | ViPer | 0.0024 | 0.0038 | 0.0113 | 0.0005 | 0.0030 | 0.0004 | 0.0012 |
| | REBECA | 0.0031 | 0.0049 | 0.0106 | 0.0007 | 0.0036 | 0.0004 | **0.0021** |
| | GNR | 0.0010 | 0.0019 | 0.0045 | 0.0002 | 0.0011 | 0.0001 | 0.0005 |
| | NaviGen | **0.0042** | 0.0063 | 0.0108 | 0.0007 | 0.0020 | **0.0006** | 0.0011 |
| | **Ours (SD1.5)** | 0.0030 | **0.0066** | **0.0168** | **0.0009** | **0.0040** | **0.0006** | 0.0016 |
| | **Ours (SDXL1.0)** | 0.0018 | 0.0038 | 0.0112 | 0.0004 | 0.0028 | 0.0003 | 0.0011 |
| | **Ours (PixArt)** | 0.0018 | 0.0028 | 0.0079 | 0.0004 | 0.0021 | 0.0002 | 0.0009 |

**(b) Personalization, quality, coverage, and preference sources of generated content.**

| Dataset | Approach | CLIP Sim↑ | LPIPS↓ | LAION Aesthetic↑ | Coverage↑ | LLM Judge (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | His | CF | Other↓ |
| CIGR | ViPer | 0.6359 | 0.5923 | 5.0739 | 0.8737 | 49.68 | 24.45 | 25.87 |
| | REBECA | **0.7245** | 0.5371 | 6.5404 | 0.8393 | 84.95 | 11.49 | 3.56 |
| | GNR | 0.6597 | 0.5716 | 5.1647 | 0.8889 | 31.95 | 27.82 | 40.23 |
| | NaviGen | 0.7018 | 0.5726 | 6.3380 | 0.8305 | 86.12 | 11.74 | 2.14 |
| | **Ours (SD1.5)** | 0.7038 | **0.5047** | **6.8596** | **0.9464** | 78.65 | 16.53 | 4.82 |
| | **Ours (SDXL1.0)** | 0.6851 | 0.5971 | 6.0134 | 0.9424 | 77.95 | 13.92 | 8.13 |
| | **Ours (PixArt)** | 0.6752 | 0.5463 | 6.6831 | 0.9432 | 68.49 | 18.69 | 12.82 |
| MovieLens -CF5K | ViPer | 0.5099 | 0.6215 | 5.3150 | **1.0000** | 13.86 | 39.03 | 47.11 |
| | REBECA | 0.6478 | 0.5585 | 4.8153 | 0.9980 | 71.87 | 27.85 | 0.28 |
| | GNR | 0.5906 | 0.5539 | 3.7043 | 0.9970 | 11.23 | 55.37 | 33.40 |
| | NaviGen | 0.5607 | 0.6192 | **5.7319** | 0.9980 | 52.67 | 42.51 | 4.82 |
| | **Ours (SD1.5)** | **0.6508** | **0.5521** | 4.2260 | 0.9990 | 67.35 | 31.78 | 0.87 |
| | **Ours (SDXL1.0)** | 0.6006 | 0.7085 | 4.2999 | **1.0000** | 41.09 | 38.03 | 20.88 |
| | **Ours (PixArt)** | 0.5771 | 0.6502 | 4.5415 | 0.9990 | 18.19 | 41.73 | 40.08 |
| PixelRec50K -GT50 | ViPer | 0.4691 | 0.5969 | 3.3925 | 0.9873 | 23.71 | 52.16 | 24.13 |
| | REBECA | 0.6594 | **0.5718** | 4.7993 | 0.9843 | 60.18 | 38.78 | 1.04 |
| | GNR | 0.6473 | 0.5980 | 3.2367 | 0.9883 | 14.66 | 67.33 | 18.01 |
| | NaviGen | 0.6007 | 0.5740 | **6.3389** | 0.9795 | 47.00 | 52.75 | 0.25 |
| | **Ours (SD1.5)** | **0.6799** | 0.6102 | 4.8651 | **0.9902** | 47.50 | 46.93 | 5.57 |
| | **Ours (SDXL1.0)** | 0.6481 | 0.7420 | 4.5401 | 0.9873 | 44.29 | 34.07 | 21.64 |
| | **Ours (PixArt)** | 0.6015 | 0.6538 | 4.1559 | 0.9892 | 51.21 | 36.67 | 12.12 |

MovieLens Latest[2]. Its construction is detailed in Appendix D. PixelRec50K-GT50 retains PixelRec50K[3] users with more than 50 interactions, resulting in 2,433 users, 55,607 video covers, and 190,297 implicit positive interactions. We use Stable-Diffusion-v1.5[4], Stable-Diffusion-XL-Base-1.0[5], and PixArt-XL-2-512x512[6] as pretrained diffusion backbones. We use SD1.5 and SDXL1.0 as primary backbones, while PixArt-XL-2-512x512 uses a fine-tuned caption-prefix adapter to inject

[2] https://grouplens.org/datasets/movielens/latest/
[3] https://github.com/westlake-repl/Pixelrec
[4] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5
[5] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[6] https://huggingface.co/PixArt-alpha/PixArt-XL-2-512x512

Table 2: Ablation study on CIGR. We evaluate replacing the adaptive multi-interest user representation with item averaging, removing the theme anchor, and removing masked-interest training.

| Approach | CLIP Sim | LPIPS | Coverage | LLM Judge (%) His | CF | Other | HitRate @10 | Recall @20 | NDCG @20 |
|---|---|---|---|---|---|---|---|---|---|
| **Full** | 0.7038 | 0.5047 | 0.9464 | 78.65 | 16.53 | 4.82 | 0.0170 | 0.0045 | 0.0032 |
| Item Average | 0.6913 | 0.5094 | 0.9392 | 71.67 | 21.42 | 6.91 | 0.0078 | 0.0015 | 0.0012 |
| w/o Anchor | 0.7037 | 0.5033 | 0.9137 | 77.80 | 17.86 | 4.34 | 0.0070 | 0.0014 | 0.0011 |
| w/o Mask | 0.7363 | 0.5017 | 0.9696 | 94.43 | 4.66 | 0.91 | 0.0258 | 0.0052 | 0.0040 |

the same user and theme conditions into its Diffusion Transformer (Appendix E). GReCF uses at most eight adaptive themes, and the masked-interest probability is set to 0.50. Training is conducted on eight NVIDIA A100 80GB GPUs, using AdamW with a learning rate of $5 \times 10^{-4}$. For $\mathcal{L}_{\text{ce}}$, we use a contrastive temperature of $0.07$, with loss weights $\lambda_{\text{ce}} = 0.05$ and $\lambda_{\text{reg}} = 10^{-7}$. Only the preference adapter and additional attention projections are optimized. We generate 10 images per user using 20–30 denoising steps. The LLM Judge is mainly based on Qwen3-VL-32B[7].

# 5 Results and Analyses

## 5.1 Overall Results

Table 1 compares GReCF with four representative generative recommendation approaches on CIGR, MovieLens-CF5K, and PixelRec50K-GT50 across ranking, personalization, coverage, quality, and preference-source attribution. GReCF with SD1.5 achieves the best results on five of the seven full-pool ranking metrics on CIGR, ranks first on all seven metrics on MovieLens-CF5K, and leads or ties on five metrics on PixelRec50K-GT50. This shows that images generated outside a fixed catalog still retain user-specific signals for retrieving held-out positive items. On generation metrics, the same variant achieves competitive CLIP Similarity and the lowest LPIPS on CIGR and MovieLens-CF5K, while obtaining the highest CLIP Similarity and Coverage on PixelRec50K-GT50. It also obtains the highest aesthetic score on CIGR, although its lower score on MovieLens-CF5K reveals a remaining domain-specific visual-quality limitation. Coverage remains high for SD1.5, SDXL, and PixArt. The results confirm the cross-backbone applicability of our conditioning mechanism, although neither alternative backbone consistently outperforms SD1.5.

Source attribution shows that 95.18%, 99.13%, and 94.43% of outputs on CIGR, MovieLens-CF5K, and PixelRec50K-GT50 are supported by History or CF evidence, respectively. CF accounts for 16.53%, 31.78%, and 46.93%, respectively, while Other remains low. A high CF rate usually coincides with many Other outputs. The balanced results indicate reliable extrapolation. Across all three datasets, GReCF maintains substantial CF evidence with few unsupported outputs, whereas several baselines achieve higher CF rates only alongside many Other outputs.

To assess the robustness of source attribution, we repeat the two-stage classification with two additional LLMs and conduct a human review. Appendix F reports cross-judge and human-validation results. Across different LLM judges, Other proportions remain relatively consistent, while most variation arises from the ambiguous semantic boundary between History and CF classifications. Nevertheless, their approach-level results remain positively correlated across judges, preserving broadly consistent comparisons among the evaluated methods. We further sample 1,000 classified outputs from each experimental group and distribute them among 40 volunteers for verification. Only 3.82% of the primary-judge labels are judged incorrect. Since visual-interest relatedness is inherently subjective, perfect agreement is not expected. Nevertheless, this low error rate shows that the LLM Judge closely approximates aggregate human assessments and provides reliable evaluation.

## 5.2 Ablation Study

Table 2 reports three ablations on CIGR with SD1.5. Item Average replaces the adaptive multi-interest representation with the mean of all liked-item features, verifying whether adaptive clustering and explicit multi-interest structure improve upon a pooled history vector. w/o Anchor retains the fused user representation but removes the single-theme anchor, verifying whether assigning each generation to one theme improves coverage. w/o Mask removes masked-interest training to verify preference extrapolation beyond observed history. Item Average lowers CLIP Similarity from 0.7038 to 0.6913, and w/o Anchor lowers Coverage from 0.9464 to 0.9137. Both reduce HitRate@10 from 0.0170 to 0.0078 and 0.0070, respectively. w/o Mask shifts History and CF from

[7] https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct




Figure 2: A qualitative case on CIGR. The left shows three themes extracted from the target user's history. The upper right presents ten generated recommendations. The bottom right provides collaborative evidence from similar users for the CF-supported generations.

78.65% and 16.53% to 94.43% and 4.66%, respectively. Together, these results support the full design: multi-interest representation and anchoring can preserve structured, diverse preference coverage, while masked-interest training promotes extrapolation rather than history reproduction.

### 5.3 Cold-start Generation

Cold-start capability is vital to practical recommender systems. For new users with limited interactions, the system should generate content aligned with their observed preferences while making evidence-based extensions beyond sparse histories. We evaluate GReCF on 100 held-out CIGR users, each initialized with only 5–10 liked images. For five successive rounds, the model generates 10 images per user and feeds them into the next history. This setup evaluates iterative personalization and evidence-based expansion, while also capturing the risk that feedback from generated content gradually reinforces dominant themes across rounds. Detailed per-round results and a concrete cold-start generation example are provided in Appendix G. Across the five rounds, the average CLIP Similarity, LPIPS, and Coverage are 0.7533, 0.4631, and 0.9182, respectively, while the average source distribution is 85.48% History, 10.92% CF, and 3.60% Other. These results show that iterative generation preserves the original interests while still performing reliable extrapolation.

### 5.4 Case Study

Figure 2 presents a representative CIGR user with distinct interests in botanical structures, robotic maintenance, and urban monuments. The history panel shows separate semantic regions, while the generated outputs cover multiple themes rather than collapsing onto one dominant mode. They preserve recognizable botanical and robotic content while varying objects, compositions, and visual details, demonstrating diversity beyond reproducing individual historical images. Four outputs introduce CF-supported activities absent from the displayed history, including collaborative planning, woodworking, and food preparation. Retrieved examples from similar users provide supporting evidence for these extensions, making them traceable rather than arbitrary. Generation cases of several other users are shown in the Appendix H.

## 6 Conclusion

We formulate generative recommendation as collaborative preference-field completion in continuous item space, extending user- and item-based CF beyond fixed catalogs to unobserved preference regions. We instantiate this formulation in GReCF using adaptive multi-interest representation, theme anchoring, residual preference attention, and masked-interest training. Across three datasets, GReCF achieves competitive ranking, strong personalization, high coverage, and evidence-based CF extrapolation with few unsupported generations. Cold-start results further show that it preserves users' initial interests while progressively expanding preferences from evolving histories.

## AI USE STATEMENT

The authors used AI-assisted tools solely for language polishing and improving the presentation of figures. These tools were not used to develop the research ideas, design the methodology or experiments, analyze the results, or draw scientific conclusions. All AI-assisted content was carefully reviewed and verified by the authors, who take full responsibility for the final manuscript.

## ETHICS STATEMENT

This work is conducted solely for academic research purposes. The authors use publicly available datasets in accordance with their intended research use and do not collect or disclose any private or sensitive personal information. All volunteers participate in the review process voluntarily. The authors have carefully considered the potential ethical implications of generative recommendation and believe that this work does not introduce significant ethical risks beyond those generally associated with generative models and recommender systems.

## REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training objectives, evaluation protocol, and implementation settings in Sections 3–4. Additional details on adaptive theme selection, dataset construction, evaluation procedures, backbone adaptation, and human validation are provided in the appendices. Code and datasets will be released to facilitate reproducibility.

## REFERENCES

Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(29):803–826, 2009.

Zhipeng Bian, Jieming Zhu, Qijiong Liu, Wang Lin, Guohao Cai, Zhaocheng Du, Jiacheng Sun, Zhou Zhao, and Zhenhua Dong. ICG: Improving cover image generation via MLLM-based prompting and personalized preference alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12268–12278, 2025.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7727–7736. IEEE, 2024.

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8020–8030. IEEE, 2025.

Connor Dunlop, Matthew Zheng, Kavana Venkatesh, and Pinar Yanardag. Personalized image editing in text-to-image diffusion models via collaborative direct preference optimization. *Advances in Neural Information Processing Systems*, 38:167742–167771, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Shengbo Guo, Scott Sanner, and Edwin V. Bonilla. Gaussian process preference elicitation. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

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

Min Hou, Le Wu, Yuxin Liao, Yonghui Yang, Zhen Zhang, Yu Wang, Changlong Zheng, Han Wu, and Richang Hong. A survey on generative recommendation: Data, model, and tasks. *AI Open*, 2026.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and José Hernández-Lobato. Collaborative gaussian processes for preference learning. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272. IEEE, 2008.

Hyungjin Kim, Seokho Ahn, and Young-Duk Seo. Foundation encoders are all you need for preference-aware personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14692–14701, 2026.

Hyunjik Kim, Xiaoyu Lu, Seth Flaxman, and Yee Whye Teh. Collaborative filtering with side information: A gaussian process perspective. *arXiv preprint arXiv:1605.07025*, 2016.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

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

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

Jiongnan Liu, Zhicheng Dou, Ning Hu, and Chenyan Xiong. Generate, not recommend: Personalized multimodal content generation. *arXiv preprint arXiv:2506.01704*, 2025.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Ligong Han, Ying Ba, and Dimitris N Metaxas. PrefGen: Multimodal preference learning for preference-conditioned image generation. *arXiv preprint arXiv:2512.06020*, 2025.

Gabriel Alfonso Patron. Personalized image generation for recommendations beyond catalogs. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 1492–1498, 2025.

Haohao Qu, Shanru Lin, Yujuan Ding, Yiqi Wang, and Wenqi Fan. Diffusion generative recommendation with continuous tokens. In *Proceedings of the ACM Web Conference 2026*, pp. 7259–7270, 2026.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, 1994.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. ViPer: Visual personalization of generative models via individual preference learning. In *European Conference on Computer Vision*, pp. 391–406. Springer, 2024.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.

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

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21831–21843, 2022.

Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 832–841, 2023.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Generative recommendation: Towards personalized multimodal content generation. In *Companion Proceedings of the ACM Web Conference 2025*, pp. 2421–2425, 2025.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 165–174, 2019.

Xianquan Wang, Likang Wu, Shukang Yin, Zhi Li, Yanjiang Chen, Feng Hu, Yu Su, and Qi Liu. I-AM-G: Interest augmented multimodal generator for item personalization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21303–21317, 2024.

Zihao Wang, Yuxiang Wei, Xinpeng Zhou, Tianyu Zhang, Tao Liang, Yalong Bai, Hongzhi Zhang, and Wangmeng Zuo. Premier: Personalized preference modulation with learnable user embedding in text-to-image generation. *arXiv preprint arXiv:2603.20725*, 2026.

Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. Diffusion models for generative outfit recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 1350–1359, 2024.

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

Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. DRC: Enhancing personalized image generation via disentangled representation composition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9667–9676, 2025b.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. *Advances in Neural Information Processing Systems*, 36:24247–24261, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jujia Zhao, Wenjie Wang, Yiyan Xu, Teng Sun, Fuli Feng, and Tat-Seng Chua. Denoising diffusion recommender model. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1370–1379, 2024.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1435–1448. IEEE, 2024.

Hengji Zhou, Yufeng Liu, Ye Liu, Yong Xu, Lianghao Xia, and Liqiang Nie. Navigating user behavior toward personalized multimodal generation. *arXiv preprint arXiv:2606.24196*, 2026.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128*, 2022.

# A Adaptive Selection of User-Specific Theme Counts

GReCF does not impose one fixed number of themes on every user. Instead, it estimates a user-specific count $K_u$ exclusively from the positive interactions in the user's training history. The procedure first uses cross-validated predictive coverage to select the smallest number of themes that captures most of the useful multi-interest structure, and then removes weak or redundant themes through support pruning and similarity-based merging. This separation prevents the theme count from growing merely because additional clusters improve the in-sample fit.

For user $u$, let $H_u = \{x_{u,1}, \ldots, x_{u,n_u}\}$ be the training history and let $g(x) \in \mathbb{R}^d$ be the frozen image-encoder representation. All representations are normalized before clustering:

$$y_{u,j} = \frac{g(x_{u,j})}{\|g(x_{u,j})\|_2}. \tag{A.1}$$

Let $K_{\max}$ be the global upper bound and let $m_{\min}$ be the minimum absolute cluster support. The largest candidate considered for user $u$ is

$$K_u^{\max} = \min\left(K_{\max}, \max\left(1, \left\lfloor \frac{n_u}{m_{\min}} \right\rfloor\right)\right). \tag{A.2}$$

This bound avoids evaluating a number of clusters that cannot receive even the minimum nominal support. Candidate counts are therefore $K \in \{1, \ldots, K_u^{\max}\}$ and can differ across users before cross-validation begins.

The $n_u$ historical items are partitioned into $F$ folds $\{V_{u,f}\}_{f=1}^{F}$. For every candidate $K$ and fold $f$, spherical $K$-means is fitted on the remaining $F-1$ folds. Denote the resulting unit-normalized centers by $\{p_{u,k}^{(-f,K)}\}_{k=1}^{K}$. The held-out score of an item is its maximum cosine similarity to these centers, and the cross-validated coverage is

$$s_u(K) = \frac{1}{n_u} \sum_{f=1}^{F} \sum_{j \in V_{u,f}} \max_{1 \le k \le K} y_{u,j}^{\top} p_{u,k}^{(-f,K)}. \tag{A.3}$$

Unlike the spherical $K$-means training objective, this score is evaluated on items excluded from the corresponding clustering fit. It therefore measures how well $K$ themes predict unseen interactions from the same user's history rather than how tightly they fit the observations used to construct the centers. Let $K_u^{\text{best}}$ be the count with the highest coverage. Rather than always choosing this potentially more complex count, GReCF selects the smallest count that retains a fraction $\gamma$ of the improvement over the single-theme baseline:

$$\begin{aligned} K_u^{\text{best}} &= \arg\max_{1 \le K \le K_u^{\max}} s_u(K), \\ \widetilde{K}_u &= \min\left\{K : s_u(K) \ge s_u(1) + \gamma\big(s_u(K_u^{\text{best}}) - s_u(1)\big)\right\}. \end{aligned} \tag{A.4}$$

The intermediate count $\widetilde{K}_u$ is the cross-validated candidate, not yet the final theme count. The retained-gain rule favors a parsimonious representation whenever several values of $K$ provide similar predictive coverage.

Spherical $K$-means with $\widetilde{K}_u$ clusters is next fitted to the complete training history of user $u$. Each item is assigned to its nearest center, producing cluster counts $n_{u,k}$. The required support is

$$m_u^{\text{req}} = \max(m_{\min}, \lceil \alpha n_u \rceil), \tag{A.5}$$

where $\alpha$ is the minimum support fraction. If any cluster has fewer than $m_u^{\text{req}}$ assigned items, the least-supported center is removed. All historical items are then reassigned among the remaining centers, so pruning removes a theme center rather than discarding interactions.

If every remaining cluster has sufficient support, the procedure checks redundancy. Let $\zeta$ be the merge threshold. If the largest pairwise center similarity is at least $\zeta$, the most similar pair $(i, j)$ is merged using its support-weighted center:

$$p_{u,i \oplus j} = \text{Norm}(n_{u,i} p_{u,i} + n_{u,j} p_{u,j}). \tag{A.6}$$

After every removal or merge, spherical assignments and centers are refined again on the complete history. Support pruning is checked before redundancy merging at each iteration. The procedure stops when all clusters satisfy the support constraint and every pair of centers has cosine similarity below $\zeta$, or when only one cluster remains. The number of surviving centers is the final user-specific count $K_u$. The surviving clusters are sorted by decreasing support. Their centers and normalized weights form the adaptive theme set:

$$\mathcal{P}_u = \{(p_{u,k}, w_{u,k})\}_{k=1}^{K_u}, \qquad w_{u,k} = \frac{n_{u,k}}{\sum_{l=1}^{K_u} n_{u,l}}. \tag{A.7}$$

For batched model input, the centers and weights are zero-padded to $K_{\max}$ slots and accompanied by a binary validity mask. Padding does not change $K_u$ and does not participate in preference attention; it only provides a common tensor shape across users.

# B Construction of the CIGR Dataset

CIGR is designed to approximate a real-world recommendation environment in which users have multiple interests and related preferences form population-level clusters. The dataset combines a structured image catalog, shared preference archetypes, user-specific multi-interest profiles, and a stochastic positive-feedback generator. The resulting interactions contain collaborative-filtering signals without assigning users with similar preferences the same fixed image pool.

The image catalog is based on a pre-existing collection and contains 100,000 images organized into 20 macro topics, 100 middle-level topics, 500 fine-grained topics, and 10,000 image clusters. Each cluster contains 10 images, and complete clusters are selected so that all fine-grained topics are represented equally. Images are also associated with style, palette, lighting, composition, medium texture, mood, shot scale, and viewpoint attributes. The public data provide images and identifiers, while semantic, aesthetic, and interaction-generation information is stored as oracle metadata.

We define 100 shared preference archetypes, one for each middle-level topic. Each archetype specifies macro- and middle-level semantic anchors, a broad distribution over fine-grained topics within its macro topic, and preferred values for the eight aesthetic attributes. It represents a semantic community rather than a fixed set of image IDs. Each user receives two to four archetypes with different macro-topic anchors. The primary archetype weight is sampled from $[0.42, 0.58]$, while the remaining weights are sampled with a Dirichlet distribution. User-specific fine-grained preferences are independently sampled within each archetype, so users sharing an archetype have related broad preferences but need not prefer the same images.

For every user-item pair, we compute a hierarchical semantic score and combine it with aesthetic compatibility, item quality, and a small user-item noise term. The semantic score is

$$s_{\text{sem}}(u,i) = 0.65 s_{\text{fine}}(u,i) + 0.25 s_{\text{mid}}(u,i) + 0.10 s_{\text{macro}}(u,i). \tag{B.1}$$

The hidden utility is

$$\begin{aligned} U_u(i) = {} & 0.72 s_{\text{sem}}(u,i) + 0.18 s_{\text{aes}}(u,i) \\ & + 0.08 s_{\text{sem}}(u,i) s_{\text{aes}}(u,i) + 0.02 q_i + \epsilon_{u,i}, \qquad \epsilon_{u,i} \sim \mathcal{N}(0, 0.005^2). \end{aligned} \tag{B.2}$$

The exposure score is

$$R_u(i) = 0.92 U_u(i) + 0.08 r_i + g_{u,i}, \qquad g_{u,i} \sim \text{Gumbel}(0, 0.09), \tag{B.3}$$

where $r_i$ is normalized item popularity. The top $8n_u$ items form the exposure set, and positive interactions are sampled from this set using $U_u(i)$ with Gumbel noise of scale 0.04. The selection is constrained to at most two images per cluster and at most 60% of interactions from one macro topic. Selected interactions are randomly ordered for splitting, and unobserved pairs remain unknown rather than negative feedback.

This construction yields user-side collaboration because users sharing an archetype tend to have related interaction histories. It yields item-side collaboration because different users can repeatedly prefer different images from the same broader semantic regions. The activity levels are 50, 100, and 150 positive interactions per user, allocated in proportions of 25%, 50%, and 25%. Each history is split into 80% training, 10% validation, and 10% test interactions, resulting in 800,000, 100,000, and 100,000 interactions. Model development uses only public user, image, interaction, and split files. Hidden archetypes, user profiles, item attributes, utilities, exposures, and source labels are reserved for auditing and evaluation design.

CIGR also includes users for evaluating cold-start recommendation. These users are absent from model training and provide only a small initial interaction history. Generated images may be appended to their histories in later rounds to evaluate few-shot personalization and iterative preference expansion. The cold-start users and their initial interactions are maintained separately from the original seen-user splits.

# C LLM JUDGE FOR PREFERENCE-SOURCE ATTRIBUTION

The LLM Judge determines whether each generated image is explained by the target user's observed history, supported only by collaborative evidence, or supported by neither source. Passing every historical image and all images from similar users to the LLM is computationally impractical and may introduce substantial irrelevant evidence. We therefore use CLIP only for candidate recall and reserve semantic classification for the LLM.

Let $h_u$ denote the normalized representation of user $u$, $e(x)$ denote the normalized CLIP representation of image $x$. We first select the $k$ users with the highest cosine similarity to the target user. Their training interactions are pooled after removing images already observed by the target user:

$$\begin{aligned} \mathcal{N}_k(u) &= \mathrm{TopK}_{v \neq u}\ h_u^\top h_v, \\ \mathcal{P}_u^{\mathrm{CF}} &= \left( \bigcup_{v \in \mathcal{N}_k(u)} H_v^{\mathrm{train}} \right) \setminus H_u^{\mathrm{train}}. \end{aligned} \tag{C.1}$$

For each generated image $g$, we retrieve the $m$ most similar images from the target user's training history and the $m$ most similar images from the CF pool:

$$\begin{aligned} \mathcal{R}_H(g,u) &= \mathrm{TopM}_{x \in H_u^{\mathrm{train}}}\ e(g)^\top e(x), \\ \mathcal{R}_{\mathrm{CF}}(g,u) &= \mathrm{TopM}_{x \in \mathcal{P}_u^{\mathrm{CF}}}\ e(g)^\top e(x). \end{aligned} \tag{C.2}$$

For the CIGR evaluation of GReCF, we set $k = 5$ and $m = 5$. CLIP scores are used only to reduce the candidate space and do not determine the final label.

The recalled images are evaluated by Qwen3-VL-32B in two stages. Stage 1 receives $g$ and $\mathcal{R}_H(g,u)$. Historical matching is deliberately broad. A match succeeds when the generated image has any reasonable connection to at least one historical image in subject, scene, activity, function, semantic theme, visual theme, or artistic style. A successful match immediately produces Direct History. Only a failed history match activates Stage 2, which receives $g$ and $\mathcal{R}_{\mathrm{CF}}(g,u)$. A successful CF match produces CF Extrapolation, while failure in both stages produces Other. This history-first design prevents generations that can already be explained by observed preferences from being counted as collaborative extrapolation.

The following prompt is designed for input into Qwen3-VL-32B to serve as an LLM judge on CIGR. Since the prompts for the other models and datasets are semantically identical, only one is shown here. Image identifiers are instantiated as H01–H05, C01–C05, and G01 for each evaluated sample.

**Stage 1: History matching**

**Role.** You are a history-interest recall judge.

**Task.** Decide whether the generated image G01 can be explained by the target user's historical interests.

**Input.**
- H: five historical images liked by the target user, numbered as H01–H05.
- G: one generated image, numbered G01.

**Decision rule.** Historical matching must be broad. Answer YES if G has any reasonable connection to any H image at the broad-interest level.

**A YES decision may be based on:**
- same or similar broad subject category;
- same or similar object, animal, person, place, or scene category;
- same or similar activity, function, or usage;
- same or similar visual theme, semantic theme, or artistic style;
- natural extension, variant, combination, or neighboring interest of the user's historical interests.

Answer NO only when G has no reasonable connection to all H images in subject, scene, activity, function, semantic theme, or visual interest. If uncertain but G can reasonably be interpreted as an extension or neighboring interest of the user's history, answer YES.

**Output.** Output valid JSON only.

```
{
  "history_related": true,
  "matched_history_images": ["H01"]
}
```

**Field constraints.**
- `history_related` must be true or false.
- `matched_history_images` may only contain IDs from H01–H05.
- If `history_related` is true, include at least one matched history image.
- If `history_related` is false, `matched_history_images` must be an empty list.

**Stage 2: CF matching**

**Role.** You are a collaborative-filtering visual judge.

**Task.** Decide whether the generated image G01 is related to the collaborative-filtering candidate images C.

**Input.**
- C: five candidate images from similar users, numbered as C01–C05.
- G: one generated image, numbered G01.

**Decision rule.** CF matching must be broad. Answer YES when G is related to any C image in broad visual-interest terms, including same or similar subject category, related scene category, related activity or object function, similar semantic theme, natural extension of the candidate interest, or same visual interest family. Do not require exact object identity, exact scene, exact activity, exact number of objects, or exact visual composition. Only answer NO when G is clearly unrelated to all C images at the broad-interest level. If unsure, answer YES.

**Output.** Output valid JSON only.

```
{
  "cf_related": true,
  "matched_cf_images": ["C01"]
}
```

**Field constraints.**
- `cf_related` must be true or false.
- `matched_cf_images` may only contain IDs from C01–C05.
- If `cf_related` is true, include at least one matched CF image.
- If `cf_related` is false, `matched_cf_images` must be an empty list.

## D CONSTRUCTION OF MOVIELENS-CF5K

MovieLens-CF5K is constructed from the ratings and movie metadata in MovieLens Latest together with the corresponding movie posters. The objective is to obtain a tractable positive-feedback dataset that retains strong user-level collaborative structure without being dominated by globally popular movies. The construction proceeds as follows.

1. **Positive-feedback extraction.** We retain interactions with ratings of at least 4.0 and treat them as positive feedback. This produces 16,865,394 interactions from 324,389 users over 53,627 movies. Unobserved user–movie pairs remain unknown rather than being treated as negative feedback.
2. **Initial activity filtering.** We retain users with 50–150 positive interactions, resulting in 62,864 users and 5,433,700 interactions. This range removes extremely sparse and highly active users while keeping histories comparable to those in CIGR.
3. **Popularity and support filtering.** Within this candidate population, we remove 166 globally dominant movies liked by more than 10% of users. We then remove movies supported by fewer than five users and re-filter users whose remaining histories contain fewer than 50 interactions. These operations reduce popularity-driven similarity while preserving item-level collaborative evidence.
4. **Temporal data splitting.** Each user's interactions are ordered by $(\text{timestamp}, \text{movieId})$ and divided into approximately 80% training, 10% validation, and 10% test data. All neighbor construction and CF-opportunity statistics use only the training split.
5. **Train-only neighbor construction.** Candidate users are restricted to those with 80–130 complete-history interactions. We represent each user by a binary training-interaction vector weighted with BM25-style inverse document frequency, compute cosine similarity, and retain the top 50 neighbors for each user.

6. **CF-opportunity screening.** A user must have at least 92 interactions, at least 10 reliable neighbors sharing 10 or more training movies, a median top-10 overlap of at least 10, and at least 90% of training interactions shared with other candidates. The top-10 neighbors must additionally provide at least 30 unseen movies, including at least 20 supported by two or more neighbors. In total, 5,370 users satisfy these constraints.
7. **Cohort selection.** Eligible users are ranked using history-length balance, neighbor overlap, IDF-weighted similarity, reliable-neighbor count, shared-interaction coverage, and multi-neighbor CF support. We apply a minimum-degree-15 core constraint to the reliable-neighbor graph and prune it to 5,000 users while preserving this constraint.
8. **Fixed-point refinement.** After selecting the cohort, we repeatedly recompute the temporal splits, remove movies with fewer than five training users, remove users with fewer than 50 remaining interactions, and update support counts. The process converges after four rounds to 5,000 users, 6,194 movies, and 497,770 interactions, split into 396,189 training, 49,439 validation, and 52,142 test interactions. Every retained movie has a verified poster path, and the dataset references the original poster files without duplicating them.

# E PixArt Backbone Adaptation

PixArt-XL-2-512x512 uses a Diffusion Transformer rather than the UNet architecture used by SD1.5 and SDXL. The high-level GReCF design remains unchanged, but its conditioning interface must be adapted to the caption-conditioning input expected by PixArt. We therefore freeze the PixArt transformer and train only a lightweight caption-prefix adapter.

Let $E_0(c_0) \in \mathbb{R}^{L\times d_c}$ denote the frozen caption embedding of the user-independent base condition, where $d_c$ is PixArt's caption dimension. For the padded theme prototypes, the adapter maps each prototype into this space and forms the user prefix

$$U_u = [m_{u,k}\, \sigma(g_u)\psi_{\text{user}}(p_{u,k})]_{k=1}^{K_{\max}}. \tag{E.1}$$

For a selected anchor theme $q$, a separate projection produces a sequence of theme prefix tokens:

$$D_u(q) = \sigma(g_t)\psi_{\text{theme}}(p_{u,q}) \in \mathbb{R}^{D_t\times d_c}. \tag{E.2}$$

The resulting PixArt caption condition and attention mask are

$$\widetilde{E}_u(q) = [E_0(c_0); U_u; D_u(q)], \qquad \widetilde{m}_u = [m_0; m_u; \mathbf{1}_{D_t}], \tag{E.3}$$

where $g_u$ and $g_t$ are learnable scalar gates, $m_0$ is the mask of the original caption, $m_u$ is the validity mask of the user themes, and $\mathbf{1}_{D_t}$ marks the theme prefix as valid. The appended tokens are consumed through PixArt's existing caption-attention pathway, allowing preference information to affect every denoising block without modifying the transformer itself. The full multi-theme user representation therefore remains visible to the transformer, while $D_u(q)$ selects the interest expressed by the current generation. During training, $q$ is selected from the target-image routing label, and the masked-interest rule in Section 3.2 removes the target theme when the masked path is sampled. During inference, each valid theme is used in turn as $q$ to generate the multi-image set.

Given a noisy latent state $z_s$, the frozen transformer predicts noise from the prefixed condition:

$$\widehat{\epsilon}_\theta = \epsilon_\theta^{\text{PixArt}}\left(z_s, s, \widetilde{E}_u(q), \widetilde{m}_u\right). \tag{E.4}$$

Only the parameters of $\psi_{\text{user}}$, $\psi_{\text{theme}}$, and their scalar gates are optimized; the PixArt transformer and the base caption embeddings remain fixed. The final theme projection is initialized at zero, so the theme path is introduced gradually during training.

The auxiliary contrastive loss is transferred to the new caption space. Let $\overline{U}_u$ be the validity-masked mean of the user prefix and let $\overline{D}_u$ be the mean of the theme prefix. Projection heads produce $\widehat{y}_u$ and $\widehat{y}_t$ from these summaries, while the targets are the centered user representation and the selected theme prototype:

$$\begin{aligned} \widehat{y}_u &= \text{Norm}(W_u\overline{U}_u), & \widehat{y}_t &= \text{Norm}(W_t\overline{D}_u),\\ y_u &= \text{Norm}(a_u - \overline{a}), & y_t &= \text{Norm}(p_{u,q}). \end{aligned} \tag{E.5}$$

Here, $a_u$ is the auxiliary representation of user $u$, $\overline{a}$ is its population mean, and $W_u$ and $W_t$ are the two projection heads. For a batch of $B$ distinct users, define the in-batch contrastive loss

$$\ell_{\text{NCE}}(\widehat{Y}, Y) = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\widehat{y}_i^\top y_i/T)}{\sum_{j=1}^{B}\exp(\widehat{y}_i^\top y_j/T)}. \tag{E.6}$$

Table 3: Human validation of LLM source classifications. Each dataset–approach group contains 1,000 binary judgments from 100 users with 10 generated images per user. Human accuracy is the percentage of classifications marked as correct.

| Dataset | Approach | Human Accuracy↑ |
|---|---|---|
| CIGR | ViPer | 97.3% |
| | REBECA | 97.8% |
| | GNR | 96.4% |
| | NaviGen | 99.0% |
| | **Ours (SD1.5)** | 97.7% |
| | **Ours (SDXL1.0)** | 95.2% |
| | **Ours (PixArt)** | 98.1% |
| MovieLens -CF5K | ViPer | 96.9% |
| | REBECA | 97.5% |
| | GNR | 94.8% |
| | NaviGen | 95.6% |
| | **Ours (SD1.5)** | 96.1% |
| | **Ours (SDXL1.0)** | 92.3% |
| | **Ours (PixArt)** | 95.9% |
| PixelRec50K -GT50 | ViPer | 96.6% |
| | REBECA | 93.9% |
| | GNR | 95.5% |
| | NaviGen | 94.4% |
| | **Ours (SD1.5)** | 96.7% |
| | **Ours (SDXL1.0)** | 95.9% |
| | **Ours (PixArt)** | 96.1% |

Here, $T$ is the contrastive temperature. The PixArt auxiliary loss averages the user- and theme-level terms, $\mathcal{L}_{\text{ce}} = \frac{1}{2}\big(\ell_{\text{NCE}}(\widehat{Y}_u, Y_u) + \ell_{\text{NCE}}(\widehat{Y}_t, Y_t)\big)$. The residual regularizer is the mean squared magnitude of the concatenated prefixes, $\mathcal{L}_{\text{reg}} = \text{mean}\big([U_u; D_u(q)]^2\big)$. The diffusion, contrastive, and residual terms are combined using the objective in Eq. (14). This adaptation preserves GReCF's adaptive multi-interest representation, theme anchoring, and masked-interest completion while replacing UNet residual attention injection with caption-prefix conditioning for the Transformer.

## F Human Validation and Cross-Judge Robustness

Human validation verifies the correctness of the LLM-assigned source label rather than asking annotators to produce a new label from scratch. For each dataset–approach group, we sample 100 users and evaluate 10 generated images per user, yielding 1,000 binary judgments. The sampled cases are distributed among 40 volunteers. Annotators answer Yes when the LLM classification is correct and No when it is incorrect. We report the percentage of Yes judgments as human-validation accuracy. Table 3 reports group-level human-validation accuracy across all three datasets. Across all 21 dataset–approach groups, the mean human-validation accuracy is 96.18%, corresponding to an error rate of 3.82%.

To evaluate cross-judge robustness, we repeat the full source-classification evaluation with Qwen3.5-27B[8] and Qwen3.8-27B[9] while keeping the generated images, CLIP-recalled candidates, prompts, and two-stage decision rules unchanged. Table 4 compares their source distributions with those produced by the primary Qwen3-VL-32B judge. The absolute History and CF proportions differ across judges, reflecting different levels of semantic tolerance, but their approach-level assessments remain positively associated. More importantly, the discrepancy mainly reflects a redistribution between History and CF rather than a change in whether an output is supported by either source. Macro-averaged over all 21 experimental groups, the combined History+CF rates are 84.39%, 87.14%, and 84.80% under Qwen3-VL-32B, Qwen3.5-27B, and Qwen3.8-27B, respectively, while the corresponding Other rates remain close at 15.61%, 12.86%, and 15.20%. This stability indicates that the three judges largely agree on supported versus unsupported generation, even though they place the boundary between direct historical support and collaborative support differently. Relative approach rankings are also stable. Against Qwen3-VL-32B, the History, CF, and Other Spearman correlations of Qwen3.5-27B are 1.000, 0.964, and 1.000 on CIGR, 1.000, 0.893, and 1.000 on MovieLens-CF5K, and 1.000, 0.857, and 0.955 on PixelRec50K-GT50. The corresponding correlations of

[8] https://huggingface.co/Qwen/Qwen3.5-27B [9] https://huggingface.co/Qwen/Qwen3.8-27B

Table 4: Cross-judge comparison of preference-source attribution. Each judge column group reports the percentages of History, CF, and Other outputs under the same candidate evidence.

| Dataset | Approach | Qwen3-VL-32B (%) | | | Qwen3.5-27B (%) | | | Qwen3.8-27B (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | His | CF | Other | His | CF | Other | His | CF | Other |
| CIGR | ViPer | 49.68 | 24.45 | 25.87 | 71.61 | 13.44 | 14.95 | 61.89 | 18.35 | 19.76 |
| | REBECA | 84.95 | 11.49 | 3.56 | 95.59 | 3.30 | 1.11 | 93.38 | 4.94 | 1.68 |
| | GNR | 31.95 | 27.82 | 40.23 | 53.26 | 24.17 | 22.57 | 35.13 | 28.80 | 36.07 |
| | NaviGen | 86.12 | 11.74 | 2.14 | 96.09 | 3.41 | 0.50 | 92.77 | 6.17 | 1.06 |
| | **Ours (SD1.5)** | 78.65 | 16.53 | 4.82 | 90.76 | 6.84 | 2.40 | 85.58 | 10.54 | 3.88 |
| | **Ours (SDXL1.0)** | 77.95 | 13.92 | 8.13 | 85.67 | 8.43 | 5.90 | 81.93 | 10.39 | 7.68 |
| | **Ours (PixArt)** | 68.49 | 18.69 | 12.82 | 82.14 | 9.56 | 8.30 | 75.32 | 13.46 | 11.22 |
| MovieLens -CF5K | ViPer | 13.86 | 39.03 | 47.11 | 13.41 | 40.84 | 45.75 | 6.65 | 45.43 | 47.92 |
| | REBECA | 71.87 | 27.85 | 0.28 | 68.15 | 31.69 | 0.16 | 48.22 | 51.72 | 0.06 |
| | GNR | 11.23 | 55.37 | 33.40 | 5.67 | 65.73 | 28.60 | 1.62 | 70.53 | 27.85 |
| | NaviGen | 52.67 | 42.51 | 4.82 | 38.40 | 58.65 | 2.95 | 20.65 | 76.03 | 3.32 |
| | **Ours (SD1.5)** | 67.35 | 31.78 | 0.87 | 64.45 | 34.48 | 1.07 | 40.70 | 58.60 | 0.70 |
| | **Ours (SDXL1.0)** | 41.09 | 38.03 | 20.88 | 32.87 | 50.54 | 16.59 | 22.04 | 61.54 | 16.42 |
| | **Ours (PixArt)** | 18.19 | 41.73 | 40.08 | 14.37 | 46.97 | 38.66 | 5.06 | 51.75 | 43.19 |
| PixelRec50K -GT50 | ViPer | 23.71 | 52.16 | 24.13 | 19.85 | 53.72 | 26.43 | 11.46 | 60.61 | 27.93 |
| | REBECA | 60.18 | 38.78 | 1.04 | 56.37 | 42.85 | 0.78 | 39.18 | 58.38 | 2.44 |
| | GNR | 14.66 | 67.33 | 18.01 | 9.84 | 69.60 | 20.56 | 3.75 | 73.13 | 23.12 |
| | NaviGen | 47.00 | 52.75 | 0.25 | 38.67 | 60.55 | 0.78 | 22.41 | 74.67 | 2.92 |
| | **Ours (SD1.5)** | 47.50 | 46.93 | 5.57 | 43.16 | 49.02 | 7.82 | 27.68 | 69.07 | 3.25 |
| | **Ours (SDXL1.0)** | 44.29 | 34.07 | 21.64 | 35.72 | 48.84 | 15.44 | 24.37 | 51.77 | 23.86 |
| | **Ours (PixArt)** | 51.21 | 36.67 | 12.12 | 45.63 | 45.63 | 8.74 | 32.18 | 52.92 | 14.90 |

Table 5: Five-round cold-start evaluation on 100 users with theme-balanced input histories. The final three columns report Qwen3-VL-32B source attribution.

| Round | History Size | CLIP Sim | LPIPS | Coverage | History (%) | CF (%) | Other (%) |
|---|---|---|---|---|---|---|---|
| 1 | 5–10 | 0.6811 | 0.5191 | 1.0000 | 51.80 | 36.20 | 12.00 |
| 2 | 15–20 | 0.7538 | 0.4621 | 0.9459 | 91.30 | 6.40 | 2.30 |
| 3 | 16–30 | 0.7658 | 0.4496 | 0.9042 | 93.50 | 4.70 | 1.80 |
| 4 | 16–40 | 0.7793 | 0.4437 | 0.8763 | 94.70 | 3.80 | 1.50 |
| 5 | 16–50 | 0.7863 | 0.4407 | 0.8647 | 96.10 | 3.50 | 0.40 |
| **Average** | | 0.7533 | 0.4631 | 0.9182 | 85.48 | 10.92 | 3.60 |

Qwen3.8-27B are $0.964$, $1.000$, and $1.000$ on CIGR, $0.929$, $0.571$, and $1.000$ on MovieLens-CF5K, and $0.964$, $0.929$, and $0.964$ on PixelRec50K-GT50.

## G Cold-Start Iterative Generation

We evaluate SD1.5-based GReCF on 100 users sampled with a fixed seed from a separate pool of 1,000 CIGR cold-start users. Neither these users nor their interactions appear during model training, and each user initially provides only 5–10 liked images from one to three fine-grained topics. The model generates 10 images per user in each of five rounds, producing 5,000 images in total. All earlier generations are treated as positive feedback and retained in a complete history archive.

Directly conditioning on this growing archive can cause a frequently generated theme to dominate the user representation. Before each round, we therefore reconstruct adaptive themes from the complete archive and deterministically subsample overrepresented themes so that the largest theme contains at most three times as many images as the smallest theme. The full archive remains available for reclustering in subsequent rounds, but only the balanced subset is re-encoded and used to condition the current generation.

Each round is evaluated against the exact balanced history used for conditioning. CLIP Similarity is the maximum similarity between a generated image and the input-history images, LPIPS is the minimum perceptual distance to that history, and Coverage measures the fraction of reconstructed history themes represented by the 10 new outputs. For the Qwen3-VL-32B judge, the first stage receives the five input-history images most similar to each generated image. The normalized mean of the same history also serves as a virtual user representation for retrieving five seen users, whose training images form the collaborative candidate pool used by the second stage. Thus, History denotes direct support from the current input, whereas CF denotes support from similar users but not from the current history. The results are displayed in Table 5.

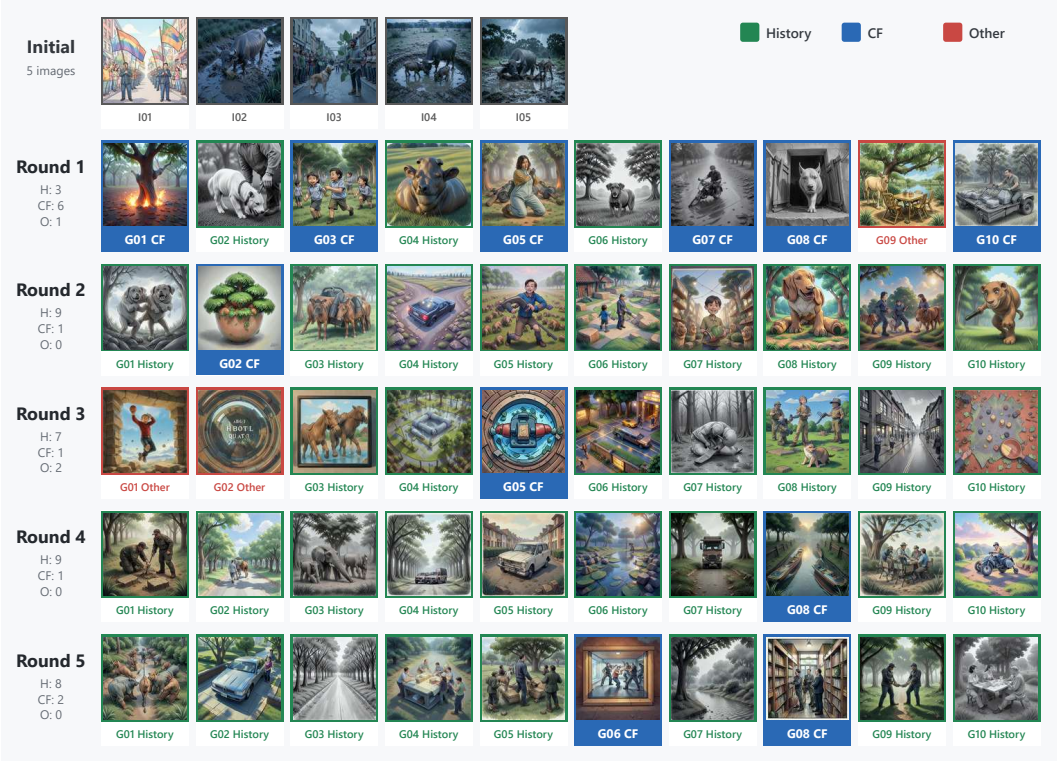


Figure 3: Five-round cold-start generation for user10947. The five initial images are followed by 10 generated images in each round, with each round's outputs added to the history for the next round. Green, blue, and red denote History, CF, and Other.

CF reaches 36.20% in Round 1 but falls to 3.50% in Round 5, while History rises from 51.80% to 96.10%. Compared with directly using the full archive, balancing increases the CF rate from 1.40% to 3.50% and Coverage from 0.8095 to 0.8647. This indicates that controlling theme frequency mitigates late-round interest collapse and retains more collaborative exploration. Nevertheless, the sharp CF decline from Round 1 to Round 2 shows that repeatedly treating every generated image as positive feedback still creates a strong self-reinforcing effect. Figure 3 shows the five-round cold-start trajectory for user10947. All previously generated images are added to the next history.

## H OTHER GENERATION CASES




Figure 4: Display of image generation cases and source attribution of seven other users.