

RAGAR: Retrieval Augmented Personalized Image Generation Guided by Recommendation

Run Ling^{1*}, Wenji Wang^{1*}, Yuting Liu¹, Guibing Guo^{1†}, Haowei Liu², Jian Lu²,
Quanwei Zhang³, Yexing Xu⁴, Shuo Lu⁵, Yun Wang⁴, Yihua Shao⁵, Linying Jiang¹, Xingwei Wang¹

¹Northeastern University

²Chongqing University of Post and Telecommunications

³Zhejiang University

⁴Sun Yat-sen University

⁵University of Chinese Academy of Sciences

lingr@mails.neu.edu.cn, {wenjiwang, liuyuting}@stumail.neu.edu.cn, guogb@swc.neu.edu.cn

Abstract

Personalized image generation is crucial for improving the user experience, as it renders reference images into preferred ones according to user visual preferences. Although effective, existing methods face two main issues. First, existing methods treat all items in the user’s historical sequence equally when extracting user preferences, overlooking the varying semantic similarities between historical items and the reference item. Disproportionately high weights for low-similarity items distort user visual preferences for the reference item. Second, existing methods heavily rely on consistency between generated and reference images to optimize generation, which leads to underfitting user preferences and hinders personalization. To address these issues, we propose Retrieval Augmented Personalized Image GenerAtion guided by Recommendation (RAGAR). Our approach uses a retrieval mechanism to assign different weights to historical items according to their similarities to the reference item, thereby extracting more refined users’ visual preferences for the reference item. Then we introduce a novel rank task based on the multi-modal ranking model to optimize the personalization of the generated images instead of forcing depend on consistency. Extensive experiments and human evaluations on three real-world datasets demonstrate that RAGAR achieves significant improvements in both personalization and semantic metrics compared to five baselines.

Introduction

Personalized image generation has been widely applied in scenarios such as advertising systems and chat software. It aims to render reference images into preferred ones based on the user’s visual preferences. Existing image generation methods (Nichol et al. 2022; Rombach et al. 2021; Gal et al. 2023) often produce similar outputs for different users with similar input, failing to meet users’ needs for personalization. As a personalized generation method, PMG (Shen et al. 2024) extracts user preferences from user historical

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

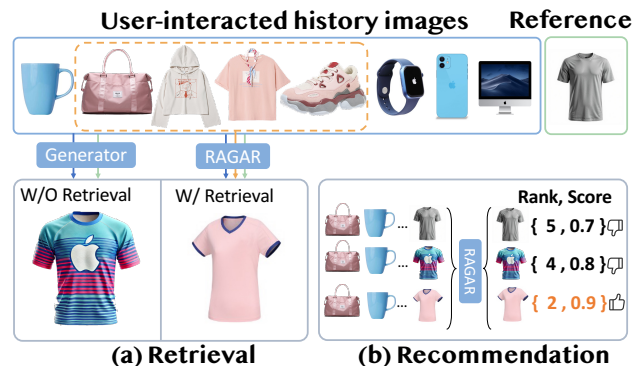


Figure 1: **Effect of retrieval and recommendation on image quality.** Lower rank and higher score of RAGAR-generated images suggest a strong positive correlation between retrieval/recommendation and personalized generation quality.

sequences by treating each item equally, ignoring the varying semantic similarities between the reference item and historical items, such as semantic differences in categories. This disproportionately amplifies the influence of irrelevant items, causing the extracted user preference to deviate from the true visual preferences on the reference item. Fig. 1(a) shows a comparative example, if a user historical sequence includes electronic products and clothes, and the reference item is "short sleeves", an overly simplistic approach might generate a striped T-shirt with the Apple logo—an outcome that misses the user’s true preferences. By incorporating semantic-based retrieval, more relevant visual preferences, such as "clothing, sport, pink", can be captured. Hence, we propose the assumption that filtering the items semantically related to the reference item enhances user preference modeling.

Another challenge lies in defining and evaluating personalization quantitatively. Current methods rely on either human evaluation or large multi-modal models (Cho et al. 2024; Wu et al. 2024), both of which are resource-intensive.

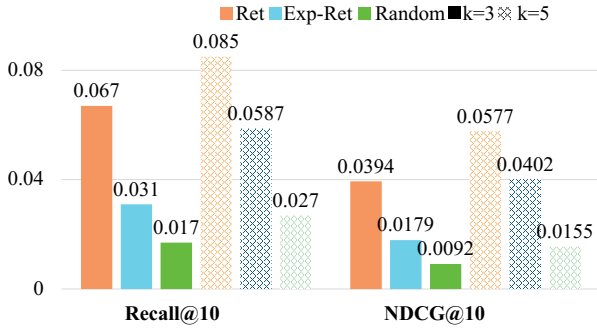


Figure 2: **Comparison of user history sequence construction strategies on POG.** Top- k retrieved items (Ret) consistently outperform lower-ranked (Exp-Ret) and random (Random) ones, validating that high-ranking relevant items better capture user preferences.

Recent work (Shen et al. 2024) evaluates and rectifies generated images heavily based on their pixel similarity to the reference image, resulting in limited personalization. Inspired by advances in recommendation systems, which excel in learning and assessing personalized preferences, we propose using multi-modal recommendation models to evaluate images, shown in Fig. 1(b). These models provide an efficient and effective way to measure personalization. Moreover, by combining models and metrics from the recommendation domain, we propose and validate two assumptions: 1) items selected through retrieval are more closely aligned with user preferences than randomly selected items; 2) retrieving items semantically related to the reference item enhances the learning of user preferences compared to irrelevant items. We compare three types of user historical sequences on POG (Chen et al. 2019) dataset: **Ret** contains the top- k items selected with the retrieval method described in Sec. . **Exp-Ret** (Expanded Retrieval) contains items ranked between $(k, 2k]$ by the retrieval method. **Random** contains k randomly selected items. The results are shown in Fig. 2. The Random sequence has the lowest values in metrics, demonstrating that randomly selected items are less effective for preference modeling. Our first assumption holds because both Ret and Exp-Ret achieve higher metrics, suggesting that items related to the reference items are crucial for modeling preferences. Moreover, both metrics of Ret are higher than those of Exp-Ret, which implies the second assumption holds. Therefore, both of our assumptions hold in practical applications, motivating us to retrieve higher-ranked items to enhance the preferences modeling. More results can be found in the Appendix.

Hence, we propose **RAGAR** (Retrieval Augmented Personalized Image Generation guided by Recommendation), comprising three key modules. The **Retrieval module** identifies semantically relevant historical items via a correlation unit and aggregates their visual features through a fusion unit to construct a retrieval-augmented preference. The **Generation module** extracts item keywords from the user sequence and integrates both detailed and global preferences,

calibrated by the balance calibrator, to generate personalized images. Finally, the **Reflection module** applies multiple assessment signals to guide the model toward achieving a better balance between personalization and semantic fidelity.

The main contribution of the paper can be summarized as follows:

- To our knowledge, we are the first to emphasise and explore the intrinsic relationship between historical and reference items. We hypothesise this relationship is crucial in capturing preferences more accurately and guiding the personalized generation process. To support this hypothesis, we conduct a comprehensive data analysis on real-world datasets, which reveals significant impact of item-to-item relevance on personalization quality.
- We propose a novel personalized image generation framework, RAGAR, which is specifically designed to address the dual objectives of semantic alignment and personalization. We utilize the retrieval assumption realized by calculating the semantic similarity between items to enhance semantic consistency. Additionally, we incorporate a discriminator to assess and enhance image quality, promoting better personalization.
- Experiments on three diverse and real-world datasets demonstrate that RAGAR outperforms five strong baseline methods across multiple evaluation metrics. The empirical results highlight the effectiveness of our retrieval-based design and the discriminator.

Methods

Task Formulation

We formulate the task of personalized image generation in a multi-modal setting. Let u denote a user, and let their historical interaction sequence be represented as: $\mathbf{S}_u = \{I_1, I_2, \dots, I_{N-1}\}$, where each item $I_j = \{\text{img}_j, \text{txt}_j\}$ is a multi-modal pair consisting of an image img_j and a corresponding textual description txt_j . The final item $I_N = \{\text{img}_N, \text{txt}_N\}$ serves as the reference item, which specifies the semantic content for generation. The goal is to generate a personalized image $\hat{\text{img}}_u$ that reflects the user’s historical preferences captured in \mathbf{S}_u and aligns semantically with the reference item I_N . Formally, the task is to learn a generation function:

$$\hat{\text{img}}_u = \mathcal{G}(\mathbf{S}_u, I_N) \quad (1)$$

such that $\hat{\text{img}}_u$ maximizes personalization fidelity with respect to \mathbf{S}_u and semantic consistency with respect to I_N .

RAGAR

Our proposed method, RAGAR, is illustrated in Fig. 3. To mitigate the influence of irrelevant historical items, the retrieval module computes semantic similarity scores between the reference item and historical items by correlation unit. The fusion unit then fuses the visual features of the most relevant items, weighted by similarity scores, to construct a retrieval-augmented preference. The generation module employs a LLM to extract and summarise item keywords from the sequence. The LLM and CLIP are then used to

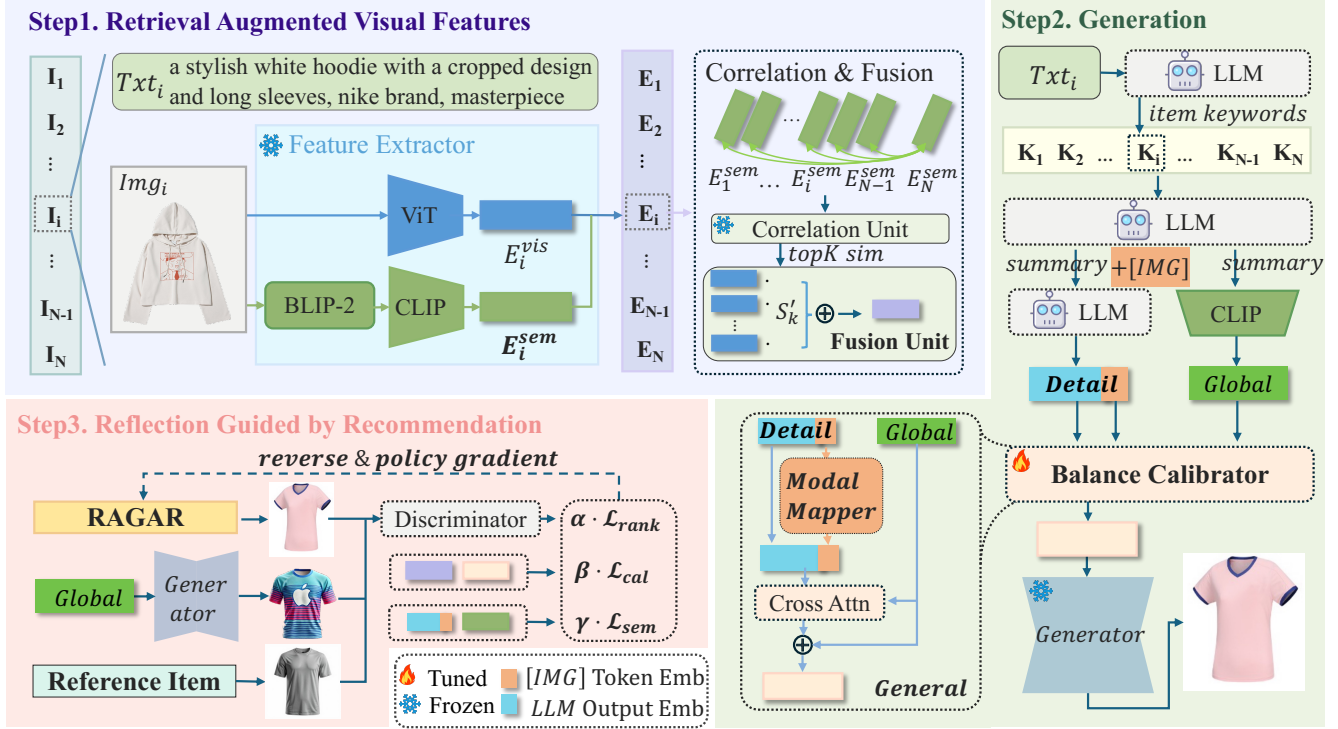


Figure 3: **Overview of the RAGAR framework.** The framework consists of three stages. In the **Retrieval** stage, semantic and visual features are extracted to identify relevant historical items, whose visual features are fused into the retrieval-augmented preference. The **Generation** stage uses a LLM, CLIP and the balance calibrator to extract and integrate detailed and global preferences, enabling personalized image generation. In the **Reflection** stage, the generated images are evaluated through ranking, calibration, and semantic alignment, with policy gradients guiding the model to balance personalization and semantic fidelity.

derive detailed and global user preferences, respectively, which are subsequently integrated via a balance calibrator. A diffusion-based generator then synthesizes personalized images based on the general preference. Finally, the reflection module evaluates the generated images and updates the model by balancing semantic consistency and personalization through multi-dimensional assessments, thereby ensuring high-quality results.

Retrieval Module

Given the user’s interaction history \mathbf{S}_u and a reference item I_N , we compute the semantic relevance of each historical item. First, we transform items’ image into the caption cap_i using a pretrained image-to-text model (e.g., BLIP-2 (Li et al. 2023)). Each caption is then encoded into a high-dimensional semantic embedding via a text encoder such as CLIP (Radford et al. 2021):

$$\mathbf{E}_i^{sem} = \mathbf{Enc}^{sem}(cap_i), \quad i = 0, 1, \dots, N \quad (2)$$

We compute the semantic similarity between each historical item and the reference item using cosine similarity:

$$s_i = \frac{\sum \mathbf{E}_i^{sem} \mathbf{E}_N^{sem}}{\sqrt{\sum \mathbf{E}_i^{sem2}} \sqrt{\sum \mathbf{E}_N^{sem2}}}, \quad i = 0, 1, \dots, N-1 \quad (3)$$

The top- k most similar items form the retrieval sequence \mathbf{S}_u^{ret} , capturing the most relevant preference signals from the user history. To form a retrieval-augmented preference, we extract visual features of the selected items using a visual encoder (e.g., ViT (Dosovitskiy et al. 2021)):

$$\mathbf{E}_i^{vis} = \mathbf{Enc}^{vis}(img_i), \quad i = 0, 1, \dots, N-1 \quad (4)$$

These top- k features are then aggregated using the normalized similarity scores s'_i as weights. The resulting vector \mathbf{P}^{ret} captures a preference representation that emphasizes semantically aligned visual cues while suppressing irrelevant information.

$$\mathbf{P}^{ret} = \sum_{i=1}^k \frac{\exp(s'_i)}{\sum_{j=1}^k \exp(s'_j)} \cdot \mathbf{E}_i^{vis} \quad (5)$$

Generation Module

The generation module bridges global preference with detailed preference to produce personalized images. To capture user preferences from interactions, we transform interacted items into structured textual descriptions suitable for LLM analysis. Specifically, we summary items with concise keywords to enhance interpretability and reduce extraneous information. We construct a prompt p_k (all prompts are detailed in Appendix) populated with item text to extract

concise keywords with an LLM (e.g., Qwen2.5 (Bai et al. 2023)), denoted as ϕ_k :

$$w_i = \phi_k(p_k(\text{cap}_i, \text{txt}_i)), i = 0, 1, \dots, N - 1 \quad (6)$$

With the keyword list w_i for item I_i , we filter out low-frequency keywords and retain the top- n keywords as w' . Next, we construct a prompt p_g to capture the general user preference. To enhance the expressivity of the LLM, denoted as ϕ_g , for image generation, we expand its vocabulary by incorporating L additional special tokens $[\text{IMG}\{i\}]_{i=1}^L$. The tokens and keywords w' are then populated to p_g to extract preferences. Given the prompt p_g , we divide the output into text-related embedding \mathbf{E}^{txt} and $[\text{IMG}]$ -related features $\mathbf{E}^{\text{img}} \in \mathbb{R}^{L \times d}$:

$$[\mathbf{E}^{\text{txt}}; \mathbf{E}^{\text{img}}] = \phi_g(p_g) \quad (7)$$

To bridge the gap between textual description and visual representation in LLMs, we adopt the Modal Mapper, a 4-layer encoder-decoder transformer with trainable queries and two linear layers, to align \mathbf{E}^{img} with the image space following Gill (Koh, Fried, and Salakhutdinov 2023). Then we concatenate the text-related embedding and the image-related embedding and obtain the detailed multi-modal feature.

$$\mathbf{E}^d = [\mathbf{E}^{\text{txt}}; \phi_m(\mathbf{E}^{\text{img}})] \quad (8)$$

To maximize keyword utilization and capture the global features, we use the text encoder to generate keyword feature \mathbf{E}^g . We employ a cross-attention layer and a residual connection to integrate \mathbf{E}^d with \mathbf{E}^g . In this way, we obtain the general preference feature \mathbf{P}^{gen} . The generator is then used to generate personalized image.

$$\mathbf{P}^{\text{gen}} = [\text{softmax}(\frac{\mathbf{E}^d \mathbf{E}^g \top}{\sqrt{d_{\mathbf{E}^g}}}) \mathbf{E}^g; \mathbf{E}^g] \quad (9)$$

Reflection Module

As the diffusion-based generator, propagating gradients backwards from the generated images is challenging (Shen et al. 2024). To address this issue, we design a multi-assessment reflection module based on policy gradient (Sutton et al. 1999), ensuring both user alignment and semantic consistency.

Recommendation Reflection To provide personalized feedback on generated images, we leverage a pre-trained multi-modal ranking model (e.g., MICRO (Zhang et al. 2023)) to evaluate images, reducing the reliance on expensive manual labeling. More details about the ranking model (RM) can be found in Appendix. Given a set of images $\{v^{\text{ref}}, v^{\text{glob}}, v^{\text{gen}}\}$, where v^{ref} represents the reference image, v^{glob} represents the image generated with the global preference and v^{gen} represents the image generated by RAGAR. By substituting visual features from this set into the sequence of original items, the RM assigns the scores $\rho \in \{\rho^{\text{ref}}, \rho^{\text{glob}}, \rho^{\text{gen}}\}$ and rankings $rk \in \{1, 2, 3\}$ derived from the score for each item:

$$\{\rho, rk\} = \mathbf{RM}(S_u, v) \quad (10)$$

Intuitively, images that better align with the user’s personalization preference are expected to achieve higher scores and lower ranks.

To enable gradient propagation from image-level feedback, we adopt a policy gradient-inspired approach. The underlying formulation is based on a Markov Decision Process (MDP), defined as $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ (state/action spaces, reward, transition probability, discount factor). In RAGAR, we simplify the multi-step decision process into a single-step generation by setting $\mathcal{P} = \gamma = 1.0$, while retaining the core policy gradient framework. Model parameters are treated as states, image generation conditioned on user preference as actions, and rank-based rewards as action values. Although the states do not transition explicitly, parameter updates implicitly induce state changes. To support multi-action exploration, we sample r perturbations $\epsilon_t, t \in [1, r]$, from a Gaussian distribution over user preferences and compute the final reward as the mean rank score of the generated images. Since policy gradients require maximizing reward while other losses are minimized, we incorporate the reward as a penalty term by negating it.

Specifically, we employ a reward function to guide the model toward generating images that reflect personalization preferences. We then accumulate rewards for each noise ϵ_t and calculate the $\mathcal{L}_{\text{rank}}$.

$$\mathcal{R}_t = \sum_{c \in \{\text{ref}, \text{glob}\}} \max(0, \rho^c - \rho_t^{\text{gen}}) + \delta \quad (11)$$

$$\mathcal{L}_{\text{rank}} = -\frac{1}{r} \sum_{t=1}^r \log_{\pi_{\theta}}(\mathbf{P}^{\text{gen}} + \epsilon_t) \cdot \mathcal{R}_t \quad (12)$$

To encourage the model to generate higher-quality images we add a margin δ . And $\log_{\pi_{\theta}}$ computes the probability density of the generated embedding under the Gaussian distribution.

Calibrator Reflection As mentioned in Sec. , the retrieval-augmented preference focuses primarily on item features associated with the reference item. To narrow the gap between the general preference feature \mathbf{P}^{gen} and the retrieval-augmented preference feature \mathbf{P}^{ret} , we minimize the calibrator loss between them:

$$\mathcal{L}_{\text{cal}} = \|\mathbf{P}^{\text{gen}} - \mathbf{P}^{\text{ret}}\|_2^2 \quad (13)$$

This adjustment ensures that \mathbf{P}^{gen} reflects both the general and retrieval-augmented preferences, and guide the model to learn from retrieval.

Semantic Reflection To enhance the semantic consistency, we minimize the distance between the modal mapper’s output in the generation module, \mathbf{E}^d , and the semantic features of the reference image, $\mathbf{E}_N^{\text{sem}}$, using the semantic loss. Combined with the semantic loss, generated images align more semantically with the reference.

$$\mathcal{L}_{\text{sem}} = \|\mathbf{E}^d - \mathbf{E}_N^{\text{sem}}\|_2^2 \quad (14)$$

Reflection Loss We update the Balance Calibrator in the generation module. The overall loss is defined with three adjustable weighting hyper-parameters α, β and γ :

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{rank}} + \beta \cdot \mathcal{L}_{\text{cal}} + \gamma \cdot \mathcal{L}_{\text{sem}} \quad (15)$$

Experiment

We evaluate RAGAR on three different scenarios: commodity, movie poster, and sticker. Our experiments aim to answer the following research questions:

- **RQ 1:** How does RAGAR perform compared to other generative methods in quantitative evaluation metrics?
- **RQ 2:** Does RAGAR outperform baseline methods in human evaluation?
- **RQ 3:** What is the impact of each module of RAGAR on performance? Specifically, how do the retrieval module and the reflection module contribute to the generation?
- **RQ 4:** Can personalized images generated by RAGAR improve the performance in recommendation systems?

Experimental Settings

Datasets. We utilize three real-world datasets to generate personalized images across different scenarios: POG, ML-latest, SER30K. POG is a multi-modal dataset of fashion clothing with user interaction history. Due to the large size, we randomly select a subset of users and interactions. ML-latest is a benchmark movie dataset with user ratings. We collect corresponding posters from IMDB and split user historical sequences to ensure each contains 20 interactions. SER30K is a large-scale sticker dataset where each sticker is categorized by theme and annotated with an associated emotion label. We merge two random themes into user historical sequences and randomly select a reference sticker from other themes. Dataset statistics are summarized in Tab. 2.

Comparison Methods. We compare RAGAR with five generative baselines, including three DM-based models Glide (Nichol et al. 2022), SD (Rombach et al. 2021), and TI (Gal et al. 2023), and two LLM-based models LaVIT (Jin et al. 2024) and PMG (Shen et al. 2024).

Evaluation Metrics. We compare RAGAR with baselines through the following quantitative metrics: To evaluate **semantic alignment**, we calculate CLIP Score (CS), CLIP Image Score (CIS), LPIPS (Zhang et al. 2018) and SSIM (Wang et al. 2004) for reference images. To evaluate **personalization**, we calculate the CLIP Personalization Score (CPS) and the CLIP Personalization Image Score (CPIS), which measure the similarity between the generated images and the text description or images representing user preferences, respectively. In addition, we calculate the LPIPS (Zhang et al. 2018) and SSIM (Wang et al. 2004) to quantify the perceptual similarity. Furthermore, we compute the rank change ΔR between the original and generated images. This metric emphasizes improvements for higher-ranked items by diminishing the weight for lower-ranked ones. A larger ΔR indicates greater personalization in the image.

$$\Delta R = \frac{rk_{ori} - rk_{gen}}{1 + rk_{ori}} \quad (16)$$

Parameter Settings. To make a fair comparison, all baselines are tuned with a fixed learning rate of $1e^{-5}$ and Stable

Diffusion 1.5 is used as the image generator. For RAGAR, we set the learning rate at $1e^{-5}$. The number of retrieval items is fixed to 5. The number of noise is set at 3. The hyperparameters for joint reflection are $\alpha = 0.2, \beta = 0.5, \gamma = 0.3$. More training details can be found in the Appendix. All experiments are conducted on a single NVIDIA-A100 GPU.

Performance Comparison (RQ1)

Tab. 1 summarizes the performance of RAGAR and competing methods. Traditional DM-based models (Glide, SD, TI) exhibit limited personalization, as CLIP-based generation struggles to capture deeper semantic relations between user history and reference items. Although TI improves personalization by introducing stylized word embeddings (e.g., ΔR : -25.64 vs. -54.02), these methods mainly reconstruct the reference image, with Glide achieving the highest SSIM. LLM-based approaches show stronger overall performance due to richer textual reasoning. LaVIT enhances visual-text alignment but remains weak in personalization (e.g., ΔR : -51.09 on POG), while PMG leverages LLMs to extract user preferences, outperforming DM-based baselines but still constrained by excessive reliance on reference-image consistency. RAGAR outperforms all baselines across datasets, owing to its retrieval-augmented preference modeling and ranking-based training strategy. By weighting semantically relevant historical items and optimizing toward recommendation-aligned metrics, RAGAR achieves superior personalization and semantic fidelity, particularly on ML-latest, where richer interaction histories amplify the benefits of retrieval-guided learning.

Human Evaluation (RQ2)

To evaluate RAGAR’s effectiveness in personalization and semantic alignment, we conducted a human study comparing it with PMG and original images. Fifty participants completed two ranking tasks (50 cases per dataset): one on personalization—ranking generated images by their likelihood of being clicked based on 5 historical items—and one on semantics—ranking images by how well they convey the meaning of a reference image. As shown in Tab. 3, RAGAR consistently outperforms baselines on both tasks, demonstrating its superior ability to capture user preferences while preserving semantic content.

Ablation Study (RQ3)

We study the effectiveness of key components of RAGAR in the POG dataset, including the retrieval and reflection module. Experiments on more datasets can be found in the Appendix. **Effect of retrieval module.** We assess the impact of the retrieval module on improving semantic consistency and capturing user preferences by excluding it during training. Fig. 5(a) shows that: 1) noisy items in the historical sequence diminish the model’s ability to capture user preferences and interfere with semantics; 2) the retrieval module selects items relevant to user preferences, thereby improving the performance. Furthermore, we investigate the effect of varying the retrieval number k during training. The results reveal that $k = 5$ yields the best results when the historical

Datasets	Methods	Personalization					Semantic Alignment			
		$\Delta R \uparrow$	CPS \uparrow	CPIS \uparrow	LPIPS \downarrow	SSIM \uparrow	CS \uparrow	CIS \uparrow	LPIPS \downarrow	SSIM \uparrow
POG	GLIDE	-54.02	13.90	59.01	<u>49.01</u>	17.24	17.52	59.17	<u>58.35</u>	27.14
	SD-v1.5	-25.88	15.02	62.97	55.46	12.53	<u>24.18</u>	63.72	63.50	12.39
	TI	-25.64	<u>15.64</u>	<u>63.79</u>	53.68	15.67	24.14	67.83	62.40	15.23
	LaVIT	-51.09	15.40	<u>63.46</u>	48.77	22.65	24.07	<u>72.04</u>	56.42	<u>24.79</u>
	PMG	<u>-22.96</u>	14.82	56.74	55.76	5.19	18.48	57.00	63.89	5.02
	RAGAR	-19.77	15.79	66.88	55.93	<u>17.99</u>	24.28	74.55	59.29	19.08
ML-latest	GLIDE	-3.20	13.27	29.52	60.34	<u>17.99</u>	18.21	42.31	61.88	26.22
	SD-v1.5	-1.97	12.85	31.07	59.74	14.02	18.38	53.59	58.58	14.51
	TI	-2.58	14.42	32.85	59.33	14.29	25.25	<u>54.17</u>	59.35	14.79
	LaVIT	-1.54	14.80	34.78	57.01	18.16	<u>21.04</u>	57.56	56.35	<u>19.64</u>
	PMG	<u>-0.11</u>	<u>15.07</u>	<u>41.80</u>	53.83	7.34	14.30	43.85	53.51	7.29
	RAGAR	0.01	16.20	43.45	<u>56.56</u>	15.14	19.04	53.24	<u>56.30</u>	15.75
Sticker	GLIDE	-0.86	12.27	49.52	59.78	18.00	16.47	48.97	67.88	28.23
	SD-v1.5	-0.93	12.48	51.08	59.45	17.38	18.46	49.61	68.41	16.43
	TI	-0.83	13.24	50.85	59.48	16.76	17.26	50.14	67.17	15.72
	LaVIT	-0.75	12.64	<u>53.05</u>	<u>57.82</u>	<u>24.41</u>	21.12	<u>58.99</u>	<u>62.27</u>	24.43
	PMG	<u>-0.73</u>	<u>13.31</u>	50.97	59.93	4.94	<u>18.72</u>	50.92	68.81	4.73
	RAGAR	0.99	14.97	55.25	56.71	25.61	17.69	64.41	61.38	<u>27.48</u>

Table 1: Quantitative performance comparison on three datasets in terms of personalization and semantic alignment. The best performance is **bold** while the second-best is underlined.

Features	POG	ML-latest	SER30K
#Users	1000	4689	2230
#Items	19242	9742	30739

Table 2: Characteristics of experimental datasets.

Methods	POG		ML-latest		Sticker	
	Per.	Sen.	Per.	Sen.	Per.	Sen.
ORI	2.26	/	2.16	/	2.18	/
PMG	2.08	1.62	2.10	1.94	2.24	1.88
RAGAR	1.66	1.46	1.74	1.66	1.58	1.12

Table 3: Human evaluation of RAGAR, PMG and origin images. RAGAR consistently outperforms both ORI and PMG in all metrics, achieving up to 29.46% improvement in personalization and 40.43% in semantic alignment.

Methods	POG		ML-latest	
	R@10	N@10	R@10	N@10
ORI	0.1765	0.1544	0.3229	0.2653
PMG	0.1804	0.1550	0.3213	0.2682
RAGAR	0.1927	0.1668	0.3299	0.2882

Table 4: Application of personalized generation in recommendation domain. Compared to PMG and original images (ORI), RAGAR leads to an average improvement of approximately 7% in recommendation metrics.

sequence length is 20. At $k = 0$ (no retrieval), performance is poor, while increasing k initially improves performance but declines beyond $k = 5$ due to reintroduced noise and redundancy. **Effect of rank rewards.** We investigate the role of rank rewards by excluding it while retaining the other loss functions during training, denoted as w/o. The results in Fig. 5(b) demonstrate the model’s performance decreases without rank rewards. Adding the amount r of sample noise ϵ boosts the model’s performance. $r = 5$ strikes optimal balance between personalization and semantic alignment.

Auxiliary Generation (RQ4)

Generating items beyond the training distribution helps uncover user interests unseen in existing data (Li et al. 2024). We evaluate RAGAR-generated images using the multi-modal recommendation model MICRO (Zhang et al. 2023) on POG and ML-latest by replacing original reference images with outputs from RAGAR, PMG, and ORI. As shown in Tab. 4, both generative methods enhance recommendation performance over ORI, with RAGAR achieving the highest Recall@10 and NDCG@10 gains (up to 7.6%), confirming its effectiveness in modeling user preferences.

Case Study

We showcase personalized image generation results from RAGAR and PMG on the POG, ML-latest, and Sticker datasets (Fig. 4). Each example includes eight historical items, a reference image, and generated outputs, with orange dashed boxes marking items retrieved by RAGAR. In the clothing case, RAGAR produces red-and-blue shoes that match user preferences while filtering out irrelevant styles like patent leather, outperforming PMG. In the movie

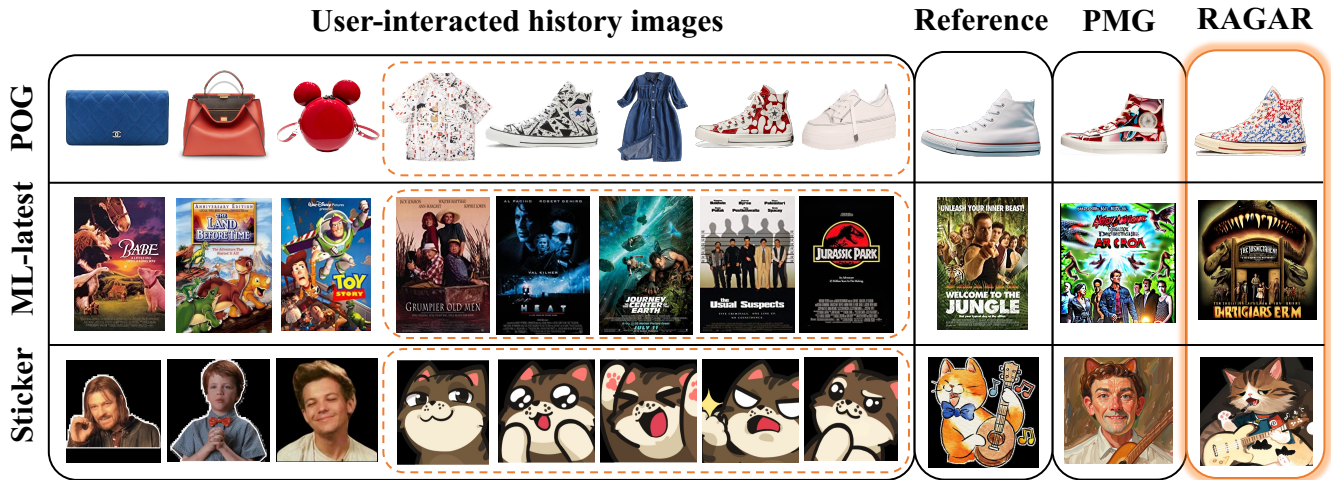


Figure 4: **Qualitative comparison of different methods.** Three representative cases highlight how RAGAR incorporates retrieved features into the generation. Left: user interaction history (dashed boxes indicate retrieved items). Right: images generated by RAGAR and PMG.

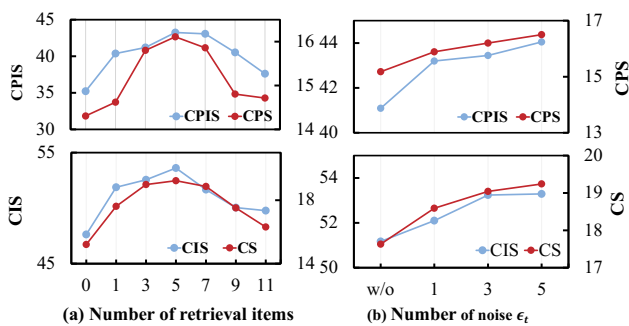


Figure 5: **The impact of the retrieval module and rank reward mechanism.** Performance improves as the number of retrieval items and noise samples increases, peaking at moderate values (around 5 for retrieval and noise), after which excessive inputs introduce noise and degrade results.

poster case, RAGAR reflects the user’s preference for multiple characters and avoids unrelated anime elements, which PMG fails to do. In the sticker case, RAGAR maintains semantic consistency (e.g., cat and guitar) and style (e.g., pink ears, brown fur), while excluding unwanted realistic styles. More examples are in the Appendix.

Related Works

Diffusion models (DMs). (Rombach et al. 2021) have enabled high-quality image generation (Nichol et al. 2022; Xu et al. 2024b; Wei et al. 2024; Zeng et al. 2024), but personalization remains underexplored. Early methods (Gal et al. 2023; Ruiz et al. 2023) rely on text prompts, while others (Xu et al. 2024a; Yang et al. 2024; Shilova et al. 2023) combine user interactions and prompts to generate personalized content. Some approaches connect DMs with large lan-

guage models (LLMs) (OpenAI et al. 2024; Touvron et al. 2023; Bai et al. 2023), such as (Koh, Fried, and Salakhutdinov 2023), and (Shen et al. 2024). However, (Shen et al. 2024) depends solely on consistency loss, often overfitting the reference image and ignoring user preferences. Multi-modal LLMs (Liu et al. 2023; Anil et al. 2023; Jin et al. 2024) generate personalized images from text but lack fine-grained preference modeling. **Recommendation systems.** aim to deliver content aligned with user preferences, using retrieval (Borgeaud et al. 2022; Cui et al. 2024) and ranking methods (Zhao et al. 2024; Geng et al. 2023; Zhang et al. 2023). Multi-modal approaches like (Cui et al. 2024) fuse text and image features for better reasoning, while (Geng et al. 2023) adds visual understanding. (Zhang et al. 2023) further models cross-modal user-item relationships. More detailed related work provide in the Appendix.

Conclusion

In the paper, we propose and validate two key assumptions in Sec. . Based on these insights, we introduce RAGAR, a novel framework that enhances personalization via retrieval-augmented preference and recommendation while preserving semantic consistency. We propose the Balance Calibrator to combine global and detailed preferences for guiding personalized generation. To address gradient propagation at the image level, we integrate recommendation signals and policy gradients into training. In future work, we aim to inject user preferences directly into the diffusion process to better align generation with user intent.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grant No. 62576083.

References

- Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.; and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR*.
- Bai, J.; Bai, S.; Chu, Y.; and et.al. 2023. Qwen Technical Report. arXiv:2309.16609.
- Borgeaud, S.; Mensch, A.; Hoffmann, J.; and et.al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proc. of ICML*, 2206–2240.
- Chen, W.; Huang, P.; Xu, J.; Guo, X.; Guo, C.; Sun, F.; Li, C.; Pfadler, A.; Zhao, H.; and Zhao, B. 2019. POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In *Proc. of KDD*, 2662–2670.
- Cho, J.; Hu, Y.; Baldridge, J. M.; Garg, R.; Anderson, P.; Krishna, R.; Bansal, M.; Pont-Tuset, J.; and Wang, S. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation. In *Proc. of ICLR*.
- Cui, W.; Bi, K.; Guo, J.; and Cheng, X. 2024. MORE: Multi-mODal REtrieval Augmented Generative Commonsense Reasoning. In *Proc. of ACL Findings*, 1178–1192.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Hounsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proc. of ICLR*.
- Geng, S.; Tan, J.; Liu, S.; Fu, Z.; and Zhang, Y. 2023. VIP5: Towards Multimodal Foundation Models for Recommendation. In *Proc. of EMNLP Findings*, 9606–9620.
- Jin, Y.; Xu, K.; Xu, K.; Chen, L.; Liao, C.; Tan, J.; Huang, Q.; Chen, B.; Song, C.; Meng, D.; Zhang, D.; Ou, W.; Gai, K.; and Mu, Y. 2024. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. In *Proc. of ICLR*.
- Koh, J. Y.; Fried, D.; and Salakhutdinov, R. 2023. Generating Images with Multimodal Language Models. In *Proc. of NeurIPS*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc of ICML*, 19730–19742.
- Li, Y.; Lin, X.; Wang, W.; Feng, F.; Pang, L.; Li, W.; Nie, L.; He, X.; and Chua, T. 2024. A Survey of Generative Search and Recommendation in the Era of Large Language Models. *CoRR*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Proc. of NeurIPS*.
- Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proc. of ICML*, 16784–16804.
- OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; and et.al. 2024. GPT-4 Technical Report. arXiv:2303.08774.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *CoRR*.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proc. of CVPR*, 22500–22510.
- Shen, X.; Zhang, R.; Zhao, X.; Zhu, J.; and Xiao, X. 2024. PMG: Personalized Multimodal Generation with Large Language Models. In *Proceedings of the ACM Web Conference 2024*, 3833–3843.
- Shilova, V.; Santos, L. D.; Vasile, F.; Racic, G.; and Tanielian, U. 2023. AdBooster: Personalized Ad Creative Generation using Stable Diffusion Outpainting. *CoRR*.
- Sutton, R. S.; McAllester, D. A.; Singh, S.; and Mansour, Y. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proc. of NeurIPS*, 1057–1063.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 600–612.
- Wei, F.; Zeng, W.; Li, Z.; Yin, D.; Duan, L.; and Li, W. 2024. Powerful and Flexible: Personalized Text-to-Image Generation via Reinforcement Learning. In *Proc. of ECCV*, 394–410.
- Wu, T.; Yang, G.; Li, Z.; Zhang, K.; Liu, Z.; Guibas, L. J.; Lin, D.; and Wetzstein, G. 2024. GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation. In *Proc. of CVPR*, 22227–22238.
- Xu, Y.; Wang, W.; Feng, F.; Ma, Y.; Zhang, J.; and He, X. 2024a. Diffusion Models for Generative Outfit Recommendation. In *Proc. of SIGIR*, 1350–1359.
- Xu, Y.; Xu, X.; Gao, H.; and Xiao, F. 2024b. SGDM: An Adaptive Style-Guided Diffusion Model for Personalized Text to Image Generation. *IEEE Trans. Multim.*, 9804–9813.
- Yang, H.; Yuan, J.; Yang, S.; Xu, L.; Yuan, S.; and Zeng, Y. 2024. A New Creative Generation Pipeline for Click-Through Rate with Stable Diffusion Model. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, 180–189.
- Zeng, Y.; Patel, V. M.; Wang, H.; Huang, X.; Wang, T.; Liu, M.; and Balaji, Y. 2024. JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation. In *Proc. of CVPR*, 6786–6795.

Zhang, J.; Zhu, Y.; Liu, Q.; Zhang, M.; Wu, S.; and Wang, L. 2023. Latent Structure Mining With Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 9154–9167.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. of CVPR*, 586–595.

Zhao, Z.; Fan, W.; Li, J.; Liu, Y.; Mei, X.; Wang, Y.; Wen, Z.; Wang, F.; Zhao, X.; Tang, J.; and Li, Q. 2024. Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Trans. Knowl. Data Eng.*, 6889–6907.