I-AM-G: Interest Augmented Multimodal Generator for Item Personalization

Xianquan Wang[1], **Likang Wu**[2], **Shukang Yin**[1], **Zhi Li**[3], **Yanjiang Chen**[1], **Feng Hu**[1]
Yu Su[4], **Qi Liu**[1]*
[1]State Key Laboratory of Cognitive Intelligence, USTC, Hefei, China
[2]College of Management and Economics, Tianjin University, Tianjin, China
[3]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[4]School of Computer and Artificial Intelligence, Hefei Normal University, Hefei, China
{wxqcn, xjtupanda, yjchen, fenghu3}@mail.ustc.edu.cn, wulk@tju.edu.cn,
zhilizl@sz.tsinghua.edu.cn, yusu@hfnu.edu.cn, qiliuql@ustc.edu.cn

Abstract

The emergence of personalized generation has made it possible to create texts or images that meet the unique needs of users. Recent advances mainly focus on style or scene transfer based on given keywords. However, in e-commerce and recommender systems, it is almost an untouched area to explore user historical interactions, automatically mine user interests with semantic associations, and create item representations that closely align with user individual interests. In this paper, we propose a brand new framework called Interest Augmented Multimodal Generator (**I-AM-G**). The framework first extracts tags from the multimodal information of items that the user has interacted with, and the most frequently occurred ones are extracted to rewrite the text description of the item. Then, the framework uses a decoupled text-to-text and image-to-image retriever to search for the top-K similar item text and image embeddings from the item pool. Finally, the Attention module for user interests fuses the retrieved information in a cross-modal manner and further guides the personalized generation process collaborating with the rewritten text. We conducted extensive and comprehensive experiments to demonstrate that our framework can effectively generate results aligned with user preferences, which potentially provides a new paradigm of **Rewrite and Retrieve** for personalized generation.

1 Introduction

If you love adventure movies, which poster in Figure 1 would catch your eyes and make you itch to go to the cinema? Or, if the monotonous uniformity of clothing has led to your aesthetic fatigue, but the platform generates a unique adventure-themed T-shirt based on your past interests, would you be excited and buy it without hesitation? In this era of information homogenization (Lombardo et al., 2019), a personalized T-shirt that resonates with your personal flair, a movie poster that highlights your favorite idol and style, or an illustration from a news article that piques your interest could all spark your desire to explore. In real-scenarios like recommender systems, personalized generation is a strategy to win over a diverse user base, because custom-generated images can precisely cater to user preferences for items. Moreover, it facilitates users to discover the content they truly enjoy.

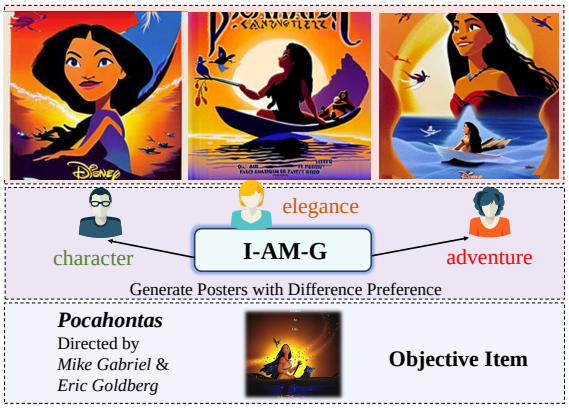


Figure 1: An example illustration for I-AM-G.

Despite the high potential and practical use in online advertisement platforms, exploring user preferences for personalized generation remains largely unexplored. Most works focus on mining user interests and recommending based on existing item information (Wu et al., 2024b), but do not involve content customization for users. Early advances in generation have focused on using keywords for style transfer (Ye et al., 2023; Song et al., 2024) or facial transformations (Wang et al., 2024) on given images, without considering user preference information. An intuitive idea is to let users input keywords based on their preferences to generate images they like. However, users often cannot precisely express their interests, making it difficult for the generator to create personalized representations.

*Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21303–21317

Recent study (Shen et al., 2024) has attempted to address this issue, but it is still tough to extract user interests for satisfactory generation results.

Specifically, the challenges are twofold. **First** is *how to explore user interests*. Due to the presence of preference ambiguity (Maafi, 2011), the interests expressed by users may be subjective and obscure, which is also a common preference elicitation problem in psychology. For example, you might know you like *Fast and Furious*, but it's hard to pinpoint what exactly you like about it. Similarly, you enjoy watching *Forrest Gump*, but it's difficult to summarize what both movies have in common that attracts you. **Second** is *how to let generator better utilize the interests*. Although a user's interest can largely represent their preferences, there is still a gap for creating images that users truly like. Therefore, it is necessary to mine the relationship between user interests and item semantic representations. For instance, *cute* might mean a bow in outfit generation, but in movie posters, it may mean a cartoon style or brilliant color. If user interests can be used to explore semantically similar items as supplementary features for the objective item, then their knowledge could guide the personalized generation process.

To tackle the challenges above, we propose **I**nterest **A**ugmented **M**ultimodal **G**enerator (**I-AM-G**) for item personalization. Given the issue of user preference ambiguity, we propose the Interest Rewrite strategy, which uses pre-trained language models to extract various modal tags for items. For each user, we aggregate these tags from their historical interactions and select the most frequent ones as the user's explicit interests. These interests are then filled into templates to rewrite item descriptions. Moreover, to compensate for the semantic relationships ignorance, we innovatively introduce the Interest Retrieve Attention (IRA) module. This module explores the relevance of interests based on embeddings from different modalities of items. It first uses retrieval techniques to find the top-K most similar items to the rewritten description in text, and top-K images similar to the objective item as references. Then, the Attention mechanism (Vaswani et al., 2017; Zhang et al., 2021) focuses on the relevant parts of text embeddings in the item pool, and integrates image modal information in a cross-modal manner, providing the generator with solid representational basis.

Based on the aforementioned features of rewritten text and the fused semantic information from single and cross-modalities, the generator is trained under these informative condition, and generates image representations that match user interests. We conduct extensive experiments of the generated results, including human study and GPT evaluation, to demonstrate the superiority of our framework. In summary, our contributions are threefold: **1)** We highlighted the significant gap in existing generation models concerning personalized generation based on user interests, and we identified two major challenges: preference ambiguity and semantic correlation ignorance. **2)** We proposed the **I-AM-G** framework for item personalized generation, which includes Interest Rewrite and Interest Retrieve Attention. These components aggregate user preference tags and deeply integrate semantic information, respectively, providing personalized features for model training and inference. **3)** We conducted extensive experiments with various metrics. The generated results demonstrated the effectiveness of our method, which is expected to facilitate further development of personalized generation. Our code is available at https://github.com/xqwustc/I_AM_G.

2 Related Work

2.1 Multimodal Generation

With the rapid advancement of generative models, multimodal generation has gained increasing attention (Yin et al., 2023). Apart from autoregressive models (Ramesh et al., 2021; Yu et al., 2022), early pioneers like Generative Adversarial Networks (GANs) (Goodfellow et al., 2020; Dong et al., 2024) and Variational Autoencoders (VAEs) (Kingma and Welling, 2013) initiated multimodal learning and generation. In GANs, the generator and discriminator engage in adversarial training, enabling the generator to produce increasingly realistic data. VAEs, on the other hand, learn representations of data in a latent space. To bridge the gap between image and text embeddings in the latent space, CLIP (Radford et al., 2021) was trained on 400 million text-image pairs, making it possible to obtain representations of text and images in the same latent space. The advent of Diffusion (Ho et al., 2020) marked a new phase in generative tasks. Utilizing the latent representation learning capability of VAEs and the multimodal feature extraction of CLIP, Diffusion Models (Rombach et al., 2022) learn the distribution of noise in the latent space to generate samples cooperated with the U-Net (Ron-

neberger et al., 2015). Based on Diffusion, many multimodal generation frameworks and fine-tuning techniques have emerged rapidly. Recent advances like Textual Inversion (Gal et al., 2023) and DreamBooth (Ruiz et al., 2023) aim to learn implicit representations of objects or styles (Deng et al., 2024). The former learns a fixed embedding, while the latter fine-tunes the entire Diffusion model. Similarly, Openjourney (Lee et al., 2024) fine-tunes on a large dataset from Midjourney, incorporating more style knowledge and achieving excellent results. As specific applications, IP-Adapter (Ye et al., 2023), InstantID (Wang et al., 2024), and MoMA (Song et al., 2024) succeeded in transferring image style and human face.

Meanwhile, although there have been many improvements to Diffusion Models (Song et al., 2021; Li et al., 2024), none specialize in learning user interests and cannot comprehend user preferences for generation. PMG (Shen et al., 2024) is a concurrent work on item representation generation with user preference; though it is a brave attempt, it does not adequately consider how to utilize item features in a multi-modal context, nor does it consider how to leverage the semantic information of other items to collaboratively guide reliable personalized generation. Our proposed framework addresses these challenges in a harmonious manner.

2.2 Generative Personalized Recommendation

Currently, most personalized recommender systems mine users' historical interactions to recommend items from a set of candidate items (Wu et al., 2024b; Zhang et al., 2024). In traditional recommender systems, all item contents remain identical. They overlook that customizable item representations (*e.g.* text, image) could achieve better attraction. Now, some e-commerce websites have introduced personalized clothing generation services (Zhu et al., 2023), presenting a finer-grained personalization challenge for recommendation. Current works adopting generation for recommendations mainly involve using LLM (Hou et al., 2024), some utilize its zero-shot ability to rank (Zhuang et al., 2023), recommend next POI (Feng et al., 2024), and for conversational recommendation (He et al., 2023), while others use it as a part to predict the next item (Ji et al., 2024), generate personalized news narrative (Gao et al., 2024), summarize item description (Acharya et al., 2023), and recommend jobs (Wu et al., 2024a). Our work focuses on personalized multimodal generation, which imposes higher technique demands.

3 Preliminaries and Problem Setup

We conclude that the goals of personalized item generation are primarily divided into two categories. One category involves generating items that do not actually exist according to the preferences of the users, such as custom clothing and furniture. The other category is to generate different multimodal representations of existing subjects according to the diverse interests of users, such as poster creation and thumbnail generation. The common goal of these two types of generation is to mine and attract the interest of users. For each item i, it is defined as $\mathcal{X}_i = [\mathcal{M}_i^1, \mathcal{M}_i^2, \ldots, \mathcal{M}_i^j, \ldots]$, where $\mathcal{M}_i^j$ represents the the j-th modality of $\mathcal{X}_i$. Specifically, if all items possess two modalities, namely image I and text T, then the i-th item can be represented as $\mathcal{X}_i = \{I_i, T_i\}$.

A user U interacts with numerous items, and for the user's objective item i, we utilize historical item features along with the current objective item information to personalized item representation. The personalized generation results are given by:

$$\tilde{I}_i, \tilde{T}_i = \tilde{\mathcal{X}}_i = \mathrm{Gen}(\mathcal{X}_i | \mathcal{X}_i, \mathcal{X}_{i-1}, \ldots, \mathcal{X}_{i-k}), \quad (1)$$

where k represents the length of the historical interaction with items, Gen represents the generator, and $\tilde{\mathcal{X}}_i$ is the item with personalization.

The ultimate goal of the generation results is to attract a specific user; however, there is currently no definitive metric to gauge whether a user prefers a particular generation result. In practice, the Mean Squared Error (MSE) loss can be employed to guide the training process for the image modality like $\mathcal{L} = \mathrm{MSE}(\tilde{I}_i, I_i)$.

4 Methodology

This section presents our personalized generation framework. We use Large Language Model (LLM) (Touvron et al., 2023) and Vision Language Model (VLM) (Liu et al., 2024) to extract tags for interest rewrite, and detail the training process based on the widely used diffusion model.

4.1 Interest Rewrite

User interests are mirrored in the items they engage with. However, human summaries of language or images can often be inadequate or imprecise. Employing pre-trained language models to extract tags from items can effectively resolve this challenge.

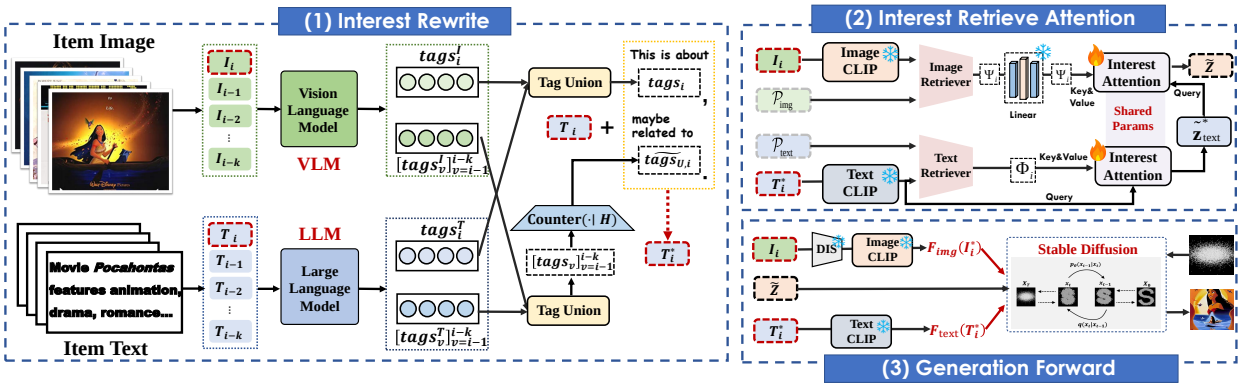


Figure 2: The whole pipeline of I-AM-G. In the Stable Diffusion model, the image attention provided by IP-Adapter is **tuned together** during training, while other parts of SD like U-Net, VAE and text/image encoders are frozen.

We focus on extracting tags from the two most prevalent modalities of items: image and text. The process of tag extraction and item information summarization can be formulated as:

$$\text{tags}_i^I = \text{VLM}(p_{I,\phi}, I_i), \tag{2}$$

$$\text{tags}_i^T = \text{LLM}(p_{T,\phi}, T_i). \tag{3}$$

In these equations, $p_{S,\phi}$ denotes the prompt for modality S tailored to the task or scenario ϕ (*e.g.*, a movie poster), which helps to guide the generation of tags. These tags embody the extraction and summarization of information regarding the style, color, and type, etc., of the given item. The detailed prompts can be found in Appendix E, below is a template example of $p_{S,\phi}$:

`This is a` S `for` ϕ`, please conclude the tags of it: ..... Note that the number of tags should be less than <a pre-set number>.`

The advantage of using language models to extract tags for different modalities lies in their ability to fully capture characteristic information. By aggregating these tags from cross-modalities, we obtain the tag information for item i as follows:

$$\text{tags}_i = \bigcup_{j \in \{I,T\}} \text{tags}_i^j, \tag{4}$$

where $\bigcup$ denotes the union of the sets.

For the objective item i that user U intends to generate, we utilize the item's own information along with the user's previous interactions with items $i-1$ to $i-k$ to rewrite user U's interest in item i. Initially, we obtain the tags from all k historical interaction items, then tally their frequency, and select the top-H tags in descending order of frequency, which is expressed as:

$$\tilde{\text{tags}}_{U,i} = \text{Counter}([\text{tags}_v]_{v=i-k}^{i-1} \mid H), \tag{5}$$

where $\text{Counter}(\cdot \mid H)$ represents the H most frequent tags, and $\tilde{\text{tags}}_{U,i}$ represents the tags derived from the interests of user U for rewriting item i. It should be noted that the set $\tilde{\text{tags}}_{U,i}$ does not include the tags of item i itself, but only the tags from the historical interaction items. The paradigm of rewriting description of item i is as follows:

$$T_i^* = T_i \texttt{ + "This } \phi \texttt{ is about tags}_i\texttt{, maybe related to } \tilde{\text{tags}}_{U,i}\texttt{."},$$

where T_i^* represents the rewritten text description of the item i. Here, *about* is used to denote the inherent attributes of item i, while *maybe related to* indicates the influence of user U's historical interactions on item i.

4.2 Interest Retrieve Attention (IRA)

The above discussion outlined the integration of user preferences and interests in the way of text. However, we hold that integrating interests solely through text is insufficient. For instance, the tag *happy* for a generator does not provide detailed guidance on which implicit image representations should embody the concept of *happy*. Therefore, the rewritten item text must search for implicit semantic similarities within the item pool and integrate additional image elements. Meanwhile, to avoid significant noise from irrelevant items in the pool, the attention should be diverted away from them. We will detail how IRA achieves the aforementioned objectives.

To achieve information fusion and mutual enhancement, first we need to encode each item's original text and image information into a latent representation space. With the powerful zero-shot capability of CLIP (Radford et al., 2021), we obtain latent embeddings for each item's text and image separately, then store them in a vector database. In our experiments, due to the relatively small size of the item pool, we can store them directly, as:

$$\mathbf{z}_{\text{text}} = F_{\text{text}}(T_i), \mathbf{z}_{\text{img}} = F_{\text{img}}(I_i), \quad (6)$$

where F represents the text and image encoder of CLIP. For $\mathbf{z}_{\text{text}}$ and $\mathbf{z}_{\text{img}}$ of each item, we store them in the text and image embedding pools, denoted as $\mathcal{P}_{\text{text}}$ and $\mathcal{P}_{\text{img}}$, respectively.

Based on this, the proposed IRA can integrate users' potential interests and preferences in the latent space. Specifically, we derive latent representations from the rewritten item text and retrieve the K most similar original item text from the item pool. Similarly, for the item's image, we retrieve the K most similar original item images. Various metrics can be used to measure the similarity; here, we use cosine similarity in the vector space:

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \cos < \mathbf{a}, \mathbf{b} >= \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}, \quad (7)$$

where $\mathbf{a}$ and $\mathbf{b}$ are the text embeddings or image embeddings. We encode the **rewritten text** T_i^* (not the original text T_i) as $\mathbf{z}_{\text{text}}^* = F_{\text{text}}(T_i^*)$, then the retrieved top-K embeddings of text will be:

$$\Phi_i = \arg \min_{\mathbf{e} \in \mathcal{P}_{\text{text}}}^{K} \text{dist}(\mathbf{z}_{\text{text}}^*, \mathbf{e}), \quad (8)$$

and the top-K image embeddings are:

$$\Psi_i = \arg \min_{\mathbf{e} \in \mathcal{P}_{\text{img}}}^{K} \text{dist}(F_{\text{img}}(I_i), \mathbf{e}). \quad (9)$$

Ψ_i is then processed by an additional linear layer (pre-trained by IP-Adapter), *i.e.*, $\Psi_i \leftarrow Proj_{\text{img}}(\Psi_i)$, to align the shapes of Ψ_i with the text embedding. Next, the IRA extracts the parts of the embeddings related to the objective item's interests and then performs fusion. The Attention module first fuses $\mathbf{z}_{\text{text}}^*$ with the retrieved top-K ones, which can be expressed as:

$$\tilde{\mathbf{z}}_{\text{text}}^* = \text{Softmax}\left(\frac{Q_1 K_1^T}{\sqrt{d}}\right) V_1, \quad (10)$$

where $Q_1 = \mathbf{z}_{\text{text}}^* W_q$, $K_1 = \Phi_i W_k$, $V_1 = \Phi_i W_v$, and d represents the dimension of the text embeddings. Then, we fuse the information of image embeddings in a cross-modal way, to get the interest embedding:

$$\tilde{\mathbf{z}} = \text{Softmax}\left(\frac{Q_2 K_2^T}{\sqrt{d}}\right) V_2, \quad (11)$$

where $Q_2 = \tilde{\mathbf{z}}_{\text{text}}^* W_q$, $K_2 = \Psi_i W_k$, $V_2 = \Psi_i W_v$. Note that the projection matrices in Equation 10 and Equation 11 are **shared**, as CLIP and the linear layer have already projected the text and image embeddings in the same latent space.

4.3 Training the Generator

We have obtained the rewritten text embedding $F_{\text{text}}(T_i^*)$ for user U on item i, along with the implicit interest representation $\tilde{\mathbf{z}}$ based on IRA. These two components will guide denoise process. For image guidance, to ensure the generator focuses more on the preference of the rewritten sentence and implicit interest, rather than the image background, we use the DIS model (Qin et al., 2022) to extract the foreground I_i^* from the image I_i. This process is very fast and we can use it in preprocessing, incurring very little additional overhead. At the t-th step of training, for the given latent noise $\mathbf{x}_t$, the noise generated by the model is:

$$\epsilon_\theta(\mathbf{x}_t, F_{\text{text}}(T_i^*), F_{\text{img}}(I_i^*), \tilde{\mathbf{z}}, t), \quad (12)$$

where θ represents the model trainable parameters.

During the generation process using the given embeddings, for the query features $\mathbf{Z}$ (originated from $\mathbf{x}_t$) in each layer of the U-Net, it queries the latent embeddings $F_{\text{text}}(T_i^*)$ and $F_{\text{img}}(I_i^*)$. Following the IP-Adapter's approach of using decoupled text and image attention, we also use a similar cross-attention structure: setting cross-attention for the text embedding, while the image embedding and interest embedding use the same cross-attention since both are image features. Note that the attention here in denoise process is for fusing the item features with the rewritten text, foreground image, and interests, which is entirely different from the purpose of the IRA mentioned previously.

$$\begin{aligned}\mathbf{Z}' &= \text{Attn}(\mathbf{Z}W_q', F_{\text{img}}(I_i^*)W_k', F_{\text{img}}(I_i^*)W_v') \\ &+ \lambda_1 \cdot \text{Attn}(\mathbf{Z}W_q'', F_{\text{text}}(T_i^*)W_k'', F_{\text{text}}(T_i^*)W_v'') \\ &+ \lambda_2 \cdot \text{Attn}(\mathbf{Z}W_q', \tilde{\mathbf{z}}W_k', \tilde{\mathbf{z}}W_v'), \end{aligned} \quad (13)$$

where $\mathbf{Z}'$ will be further processed by the model (like linear projection), form the output of each U-Net layer, and ultimately be ϵ_θ. The W_q'' is **identical** to W_q', because they are used for the same

query features. Specifically, λ_1 and λ_2 control how strongly the results of rewritten interests and IRA guide the learning of noise. **Attn** refers to Multi-Head Attention, which is similar in form to Equation 10. Then, the loss can be formulated as:

$$\mathcal{L} = \mathbb{E} \left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, F_{\text{text}}(T_i^*), F_{\text{img}}(I_i^*), \tilde{\mathbf{z}}, t) \right\|_2^2, \tag{14}$$

where ϵ is a random Gaussian noise with the same shape as ϵ_θ. This is essentially derived from the original MSE loss between original image and the generated one, thus the MSE is converted into the loss between the learned noise and Gaussian noise.

To enable classifier-free guidance during inference, a certain proportion of conditions are randomly dropped during training as:

$$\begin{aligned} \hat{\epsilon}_\theta = & \tau \cdot \epsilon_\theta(\mathbf{x}_t, F_{\text{text}}(T_i^*), F_{\text{img}}(I_i^*), \tilde{\mathbf{z}}, t) \\ & + (1 - \tau) \cdot \epsilon_\theta(\mathbf{x}_t, t), \end{aligned} \tag{15}$$

where τ controls the condition ratio. This enables the model to flexibly use or ignore conditional information, thereby improving the model's robustness and generalization ability.

5 Experiment

Our experiment focuses on the five questions.

- **RQ1:** How does I-AM-G framework perform evaluated by human and ChatGPT?
- **RQ2:** What is the effect of the generation on quantitative metrics?
- **RQ3:** How are the I-AM-G generation results?
- **RQ4:** What role do the interest rewrite and IRA play in the generation process?
- **RQ5:** How do hyperparameters affect I-AM-G?

Due to space constraints, we add two interesting experiment details included in Appendix B.

5.1 Experiment Setup, Datasets and Metrics

Specifically, we adopt LLaMA and LLaVA as LLM and VLM. The datasets we have selected, including MovieLens (Harper and Konstan, 2015) for movies, MIND for news (Wu et al., 2020), and POG (Chen et al., 2019) for outfits, all encompass image information for each item. However, some data lack the original text information. We use the VLM to provide a brief description of each item as T_i. The details of the three datasets are in Appendix A.1.

Method	MovieLens	MIND	POG
Original Item	3.375	2.667	3.308
Openjourney	2.358	2.717	2.333
DreamBooth	2.383	**2.300**	2.367
I-AM-G	**1.883**	2.317	**1.992**

Table 1: Human-evaluated average scores for results.

Method	MovieLens	MIND	POG
Original Item	3.100	2.583	3.267
Openjourney	2.475	2.733	2.392
DreamBooth	2.575	2.350	2.342
I-AM-G	**1.850**	**2.333**	**2.000**

Table 2: ChatGPT 4o-evaluated average ranks.

For image quality, we use **(1)** Structural Similarity (**SSIM**) (Wang et al., 2004) and **(2)** Learned Perceptual Image Patch Similarity (**LPIPS**) (Zhang et al., 2018) to measure the similarity between two images. This ensures the generated results are not too away from the objective item. Moreover, we use **(3) CLIP Score** (Hessel et al., 2021) to calculate the similarity between the text description of the item and the generated result as the text-image fidelity. Since there is little work on item personalized generation and other existing generation works focus on transferring style for a given object (away from our task), we use Openjourney fine-tuned model and Dreambooth fine-tuning techniques as baselines* based on three datasets. We input users' historical interactions and the objective item for them to fine-tune, and then compare the generated results. The experiment details of the three methods can be found in Appendix A.2.

5.2 Overall Performance (RQ1)

To investigate whether users truly appreciate the generated movie posters, news posters, and outfits by I-AM-G, we conducted a human study (see Appendix F for questionnaire) where participants were asked to score the generation results with ranks. Each questionnaire presented the three generated images and the original image for a specific item in a completely shuffled order.

To ensure a fair evaluation, we assigned 15 participants to rank the generated items with movies, 15 with news, and 15 with outfits. Each participant was asked to evaluate the results of 8 items. In detail, participants were required to score the

*Concurrent work PMG did not provide models or code during our work, thus we could not compare with it.

Table 3: Personalized outfit generation by I-AM-G.

Table 4: Personalized movie poster by I-AM-G.

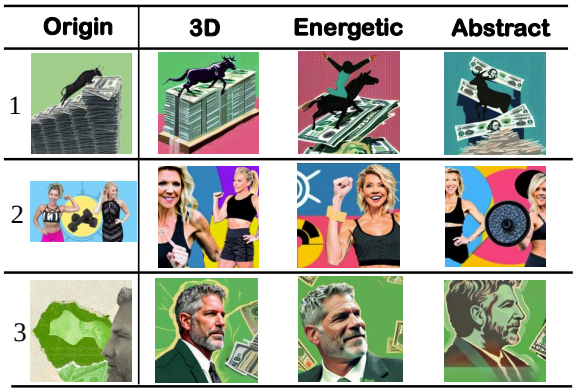


Table 5: Personalized news poster generation by I-AM-G.

Method	MovieLens	POG
Openjourney	0.2682	0.2614
DreamBooth	0.2664	0.2623
I-AM-G	**0.2714**	**0.2631**

Table 6: CLIP Score↑ for personalized results by 3 methods on MovieLens and POG.

item images from 1 to 4, and score lower ranks to those they deemed better. From Table 1, we can conclude that I-AM-G achieved the overall best performance in three scenarios, with an average score of around 2. The results on the MovieLens dataset were the most satisfying. Detailed questionnaire data showed that among votes for it, 48.3% of the best-scored results were produced by I-AM-G, and 74.2% of the top two results were generated by I-AM-G. This demonstrates that our method effectively considers user preferences.

Leveraging the powerful capabilities and aesthetic judgment of ChatGPT 4o, we make it rank the generated results. We used a prompt-based approach (details in Appendix C) to ask it to rank the input images. The average rank results are shown in Table 2. This essentially lets GPT act as a participant. As observed, the overall evaluation results from ChatGPT 4o are largely consistent with human judgments. Specifically, on MIND, ChatGPT 4o considers our generated results to be slightly better than those of DreamBooth.

5.3 Metric Evaluation (RQ2)

Table 6 shows the results generated from the image-text similarity perspective. We used Openjourney and DreamBooth as the baseline methods. It can be observed that the images generated by I-AM-G match the text description well in the CLIP semantic space. These results quantitatively prove the superiority of the method from image-text fidelity.

The results of Openjourney and DreamBooth show inconsistent performance on the MovieLens and POG datasets, likely due to their different characteristics. Openjourney, fine-tuned on Midjourney's large dataset, exhibits more creativity and is better suited for generating movie posters reflecting the text. In contrast, DreamBooth has higher fidelity to the original image layout, which explains its advantage in outfit generation tasks.

5.4 Generation Comparison (RQ3)

Combining the users' click histories for generation, we carefully selected results with some **representative styles** for POG, MovieLens, and MIND datasets to directly show the effects of personalized generation. Table 3 displays the outfit generation results on POG with several common interests (sim-

Method	MovieLens		POG	
	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
w/o Foreground Extraction	0.2547	0.4092	0.1625	0.5442
w/o Interest Rewrite	0.2518	0.4113	0.1596	0.5497
w/o Interest Retrieve Attention (IRA)	0.2536	0.4099	0.1617	0.5449
I-AM-G	**0.2558**	**0.4087**	**0.1632**	**0.5433**

Table 7: Ablation study on I-AM-G core components.

ple, vivid, cool, and cartoon style). It is evident that the personalized generation results could reflect various styles. In outfits 1, 2, and 4, the generation results for the cartoon column incorporate cartoon figures, while in outfit 3, the cartoon style is highlighted by the addition of small star patterns. For each outfit's different styles, outfit 2 of simple style gives a clean and neat feeling, while its cool style features a monkey with an extended arm, showing a cool vibe. Similarly, outfit 4 of vivid style includes a large photo of a young man, whereas the cool style shows a blue, black, and white illustration.

Furthermore, the poster generation results are commendable. Table 4 and Table 5 respectively present the personalized generation results for movie posters on MovieLens and news posters on MIND. On MovieLens, it is observed that the cartoon style predominantly features animations or a light-hearted aesthetic, while the horror style incorporates elements of mystery, suspense, and even dread. For each movie item, such as movie 1, the original poster depicted an elephant, yet it was adapted into various styles like cartoon, adventure, and horror, tailored to possibly different interests. Similarly, for movie 3, originally a war film, the posters generated under the cartoon and horror styles are distinctly different. However, they all contain war elements, indicating that our generated posters remain real to the original movie topic. Using the MIND dataset combined with user preferences, I-AM-G generates news posters (*a.k.a.* thumbnails) as shown in Table 5 including three interests (3D, energetic, and abstract style). The detailed analysis can be found in Appendix B.1.

5.5 Ablation Study (RQ4)

We detail the impact of I-AM-G core components on the generated results: Foreground Extraction, Interest Rewrite, and Interest Retrieve Attention.

Table 7 shows the results of the ablation study, where we separately explore the impact of removing Foreground Extraction, Interest Rewrite, and the IRA module on the results. It shows that using all three components together is most effective.

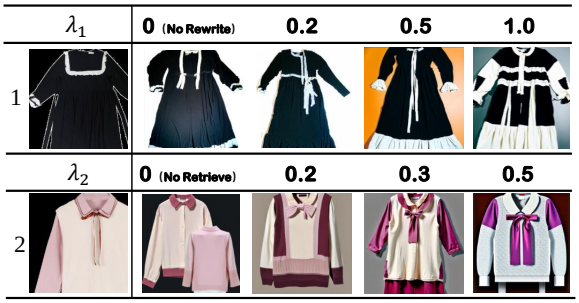


Table 8: Case study for λ_1 and λ_2.

However, the impact of each component varies. Interest Rewrite has the greatest effect on the generated results because the rewritten text includes both the user's historical interests and the original item tags, which are crucial for guiding the denoising process. Notably, the IRA enhances the similarity of the generated images, indicating that our strategy effectively guides the image generation process by integrating semantic information from the item pool. Besides, using the foreground as the image embedding improves the generation quality, likely because it helps the U-Net's attention focus more on the main subject.

5.6 Case Study & Hyperparameter Analysis (RQ5)

In this section, we adjust the strength of **interest rewriting**, **IRA** and **the number of used tags**, represented by λ_1, λ_2 and H, to find how changes in these hyperparameters affect the generated results.

Table 8 shows the results on the POG dataset. Outfit 1 is generated for a user interested in *white*, *tiny*, and *cool* styles. As λ_1 increases, the generated outfit gradually reflects the user's preferences, evolving from the original clothing to a cool black and white style at $\lambda_1 = 1.0$. Outfit 2 is generated for a user interested in *purple*, *cute*, and *neat* styles. At $\lambda_2 = 0$, the result attempts to align with purple (deep color) and neat but lacks clear guidance, leading to unsatisfactory results. When $\lambda_2 = 0.2$ or 0.3, the generated outfit better reflects the dark tone and includes a bow near the collar, indicating consideration of the *cute* aspect. However, at $\lambda_2 = 0.5$, the results become bizarre, showing that a high λ_2 can cause some inconsistencies with the original item. In practice, λ_2 value between $0.1 \sim 0.3$ could usually yield great results.

Figure 3 shows the SSIM changes between the generated images and the original item image for different numbers of tags H. As H increases, the

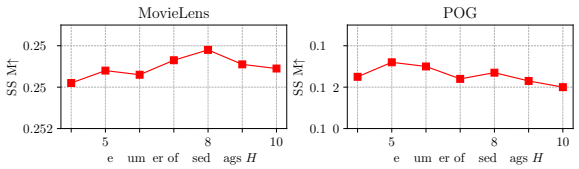


Figure 3: The relationship between the maximum number of used tags H and SSIM.

SSIM value first reaches a peak and then tends to decrease. For the MovieLens dataset, the SSIM is highest when $H = 8$, indicating that using 8 tags typically results in generated images most similar to the original under personalized conditions. For the POG dataset, the SSIM peaks at $H = 5$ and then decreases. This suggests that the outfit generation is highly sensitive to the number and quality of tags. When there are too many tags, the generated images might become too casual and lose their ability to reflect the original details accurately.

6 Conclusion

For the goal of item personalized generation, we developed the I-AM-G framework. First, we proposed the **Interest Rewrite** strategy to tackle the preference ambiguity issue. This strategy leverages pre-trained language models to extract tags for each item's multiple modalities, and uses the tags to rewrite the description of a given item for the user. Furthermore, we introduced **IRA**, an attention mechanism that retrieves the most relevant other items, and adaptively integrates cross-modal features based on their semantic similarity to the rewritten sentence. Extensive experiments demonstrated the effectiveness of our approach, potentially offering new insights for the field of item personalization in online advertisement.

Limitations

Although our work is a pioneer in the field of personalized generation that incorporates user interests, it still has some shortcomings. Overall, the limitations are primarily threefold: **First** is the overhead of data preprocessing. Our work utilizes a series of tags to amalgamate keywords, which decouples the interest reflected by each item from the user's interest. Although it does not necessitate the use of large language models for inferring the interests of each user, the summarization of each item by a language model remains a relatively costly process. Therefore, the exploration of how to obtain the tags of each item more accurately with lower overhead is a research direction for future work. **Second** is the issues with detail generation. When dealing with images that involve details of human body parts (*e.g.* fingers and toes), or when there is text or word present in the image, the generated results often exhibit flaws, which is defined as as hallucinations (Gunjal et al., 2024). For instance, the body parts may not conform to human anatomy (like poster 2 in Table 5), or the generated text may resemble a mere rearrangement of the existing text in the image, losing its original meaning. Hence, it is a upcoming research focus to maintain the fidelity of the original image details during the personalized generation process. **Moreover**, our fine-tuned data is relatively sparse. With higher quality interaction data and original item information, the generation results of our framework are expected to improve more.

About the future work, we plan to address these limitations by using NLP techniques for more efficient and quick extraction of tags, and use text classification methods to merge similar tags at a finer granularity. Moreover, we will analyze the layers of U-Net to identify which parts are crucial for generating detailed human features, aiming to improve the fidelity of the generated results and alleviate the hallucination issues.

Ethics Statement

We are dedicated to upholding the highest ethical standards in our research. While following ethical guidelines, we strictly respect the copyrights of the text and images in the dataset items, ensuring careful data management and stringent informed consent processes. Our commitment safeguards research integrity, while privacy and security of participant & data remain a top priority. Through these practices, we aim to advance the field of personalized generation in a responsible manner, contributing to scientific progress and societal well-being.

Acknowledgement

Sincere thanks to Kunxi Li (Zhejiang University), Guoliang Li (Nankai University) for their insightful suggestions. The work is partially supported by the National Natural Science Foundation of China (No. 62206155), the China Postdoctoral Science Foundation (No. 2024T170497), Anhui Provincial Natural Science Foundation (No. 2308085QF229), and the Fundamental Research Funds for the Central Universities.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207.

Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2662–2670.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. 2024. Z*: Zero-shot style transfer via attention reweighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6934–6944.

Zhijun Dong, Likang Wu, and et al. 2024. Fzr: Enhancing knowledge transfer via shared factors composition in zero-shot relational learning. In *CIKM*.

Shanshan Feng, Haoming Lyu, Fan Li, Zhu Sun, and Caishun Chen. 2024. Where to move next: Zero-shot generalization of llms for next poi recommendation. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1530–1535. IEEE.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*.

Shen Gao, Jiabao Fang, Quan Tu, Zhitao Yao, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2024. Generative news recommendation. In *Proceedings of the ACM on Web Conference 2024*, pages 3444–3453.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 720–730.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

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

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. 2024. Genrec: Large language model for generative recommendation. In *European Conference on Information Retrieval*, pages 494–502. Springer.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. 2024. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36.

Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. 2024. On the scalability of diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9400–9409.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Michael V Lombardo, Meng-Chuan Lai, and Simon Baron-Cohen. 2019. Big data approaches to decomposing heterogeneity across the autism spectrum. *Molecular psychiatry*, 24(10):1435–1450.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *ICLR 2019*.

Hela Maafi. 2011. Preference reversals under ambiguity. *Management science*, 57(11):2054–2066.

Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. 2022. Highly accurate dichotomous image segmentation. In *ECCV*.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*, pages 22500–22510.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, and et al. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on NIPS Datasets and Benchmarks Track*.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. WWW '24, page 3833–3843, New York, NY, USA. Association for Computing Machinery.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. 2024. Moma: Multimodal llm adapter for fast personalized image generation. *arXiv preprint arXiv:2404.05674*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. 2024. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3597–3606.

Likang Wu, Zhaopeng Qiu, Zhi Zheng, Hengshu Zhu, and Enhong Chen. 2024a. Exploring large language model for graph data understanding in online job recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9178–9186.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024b. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2(3):5.

Haotian Zhang, Shuanghong Shen, Bihan Xu, Zhenya Huang, Jinze Wu, Jing Sha, and Shijin Wang. 2024. Item-difficulty-aware learning path recommendation: From a real walking perspective. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4167–4178, New York, NY, USA. Association for Computing Machinery.

Kai Zhang, Hao Qian, Qing Cui, Qi Liu, Longfei Li, Jun Zhou, Jianhui Ma, and Enhong Chen. 2021. Multi-interactive attention network for fine-grained feature learning in ctr prediction. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 984–992.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad

Norouzi, and Ira Kemelmacher-Shlizerman. 2023. Tryondiffusion: A tale of two unets. In *CVPR 2023*, pages 4606–4615.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8807–8817, Singapore. Association for Computational Linguistics.

A Experiment Details

A.1 Dataset Details

We selected three scenarios to validate our method, corresponding to the MovieLens, MIND, and POG datasets. Below, we detail these datasets and our preprocessing approach. For each dataset, we first downloaded image information by web scraping, thus creating an item image pool. We then combined these with the corresponding historical interaction records, sorting the items in the image pool by user interaction time from latest to earliest. The latest item served as the objective item for personalized generation, and the previous k items were used as historical records, forming an interaction history entry. Users with a history length of fewer than the number k were excluded. If the collected user interaction history entries exceeded 1500 for any dataset, no new entries would be added.

The MovieLens[1] dataset contains numerous movies, users, and their interactions (*e.g.*, rating records). We consider ratings of 4 or above as positive interactions and processed the data accordingly. However, some image URLs in the original dataset were inaccessible. Therefore, we used available poster images and user interaction histories to generate personalized movie posters.

The MIND dataset[2], released by Microsoft, contains news click records covering various types and topics. The dataset includes user behavior logs, news content information, and user personal information. We scraped 4536 news images based on the news URLs.

The POG dataset[3], released by Alibaba, contains 1.01 million outfits, 583 thousand fashion items with rich contextual information, and 0.28 billion user click actions from 3.57 million users. Due to limited download capabilities, we scraped 7253 outfit pictures based on the provided URLs.

[1] https://grouplens.org/datasets/movielens/

[2] https://www.kaggle.com/datasets/arashnic/mind-news-dataset

[3] https://github.com/wenyuer/POG

A.2 Experiment Settings

The I-AM-G is based on the pre-trained SD v1.5[4] with IP-Adapter[5]. We reuse the decoupled attentions from IP-Adapter in Equation 13. In Figure 2, the frozen linear layer is also adopted from IP-Adapter. We use OpenCLIP ViT-L/14 (Schuhmann et al., 2022) as the text encoder, the same as in SD v1.5, and adopt ViT-H/14 as the image encoder.

I-AM-G can be trained on a single NVIDIA-A100 80G GPU for 2000 steps with a batch size of 16. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a fixed learning rate of 0.0001 and weight decay of 0.01. Similar to IP-Adapter, during training, we resize the shortest side of the image to 512 and then center crop the image to a 512×512 size. In Equation 15, we use a probability of 0.05 to drop the condition guidance. During inference, we adopt the DDIM sampler with 50 steps and set the guidance scale to 7.5. For these common hyperparameters, the settings of the baseline (Openjourney and DreamBooth) are identical to those of I-AM-G.

We use DIS model[6] to segment the image thus obtaining the image foreground as guidance. For tags extraction, we use LLaMA 3-8B[7] as the language model and LLaVA 1.5-7B[8] as the vision model, to summarize the text and visual features of a given item. For the user's history length k, we set it to 4 or 5. Depending on the dataset, the number of used tags H is set between 5 and 8. Setting H too small fails to reflect user preferences, while setting it too large can cause hallucinations and exceed CLIP's text encoding length. We set retrieval module $K = 5$, and use multi-head attention with 4 heads for fusing and integrating the retrieved results. For the strength of controlling Interest Rewrite and IRA (λ_1 and λ_2 respectively), we set $\lambda_1 = 1$ and λ_2 is around 0.2.

B Additional Experiment

B.1 How about the generation results on MIND in Table 5?

In this subsection, we analyze the generation results on MIND dataset. For abstract style posters, the

[4] https://huggingface.co/runwayml/stable-diffusion-v1-5

[5] https://ip-adapter.github.io/

[6] https://xuebinqin.github.io/dis/index.html

[7] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[8] https://huggingface.co/llava-hf/llava-1.5-7b-hf

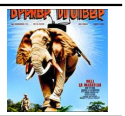
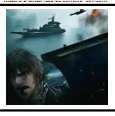
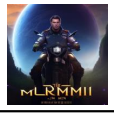
	Origin	Openjourney	DreamBooth	I-AM-G
1				
2				
3				
4				

Table 9: Comparison of generation results by different models.

composition elements incorporate more geometric shapes and lines (such as sketch-drawn entity shadows) when entities are included. In the energetic style, all generated posters convey a positive and uplifting attitude. The 3D style posters emphasize realism and depth. Horizontally, as shown in news 3, the 3D style enriches the colors and details, the protagonist's smile reflects the energetic style, and the character's profile is depicted in a sketch, representing the abstract style.

B.2 How about the generation results with different generators (Openjourney, DreamBooth)?

In addition to the evaluation results presented in the main text, we also show different personalized generation results with other fine-tuning models and techniques in Table 9. We selected the MovieLens dataset to showcase the results for users who like adventure-type movies. Each row corresponds to the objective item of the same user's historical interaction records. It can be observed that the results generated by I-AM-G are generally more vibrant and appealing. For instance, the depiction of an elephant's expedition with a suitcase and a blue sky backdrop, or the portrayal of exploration spirit against the vast ocean, are more engaging. However, other methods significantly lag behind ours. For example, results of Openjourney tend to have more anime and virtual elements, likely due to its prior fine-tuning data. Similarly, although DreamBooth's results are relatively close to the original objective items, they have a darker overall tone and fail to effectively incorporate the adventure style.

Evaluation Prompt for GPT

If you are a <user info if provided> who has previously clicked on and viewed <item history>, and you are now browsing a <task/scenario>: <item brief information>, please rank the images from your top choice to your least favorite, with the best ranked first and the worst last.

<fig 1>, <fig 2>, <fig 3>, …

Figure 4: Evaluation prompt for GPT.

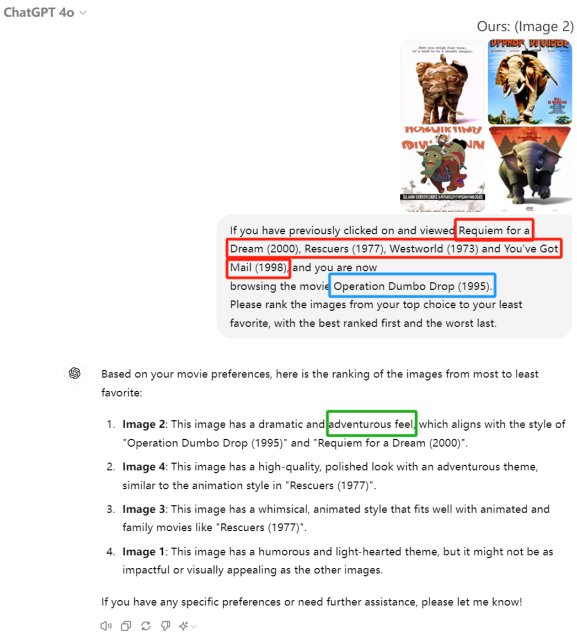


Figure 5: An example of generation results evaluation by ChatGPT 4o.

This indicates that the generation results with this fine-tuning method alone does not adequately reflect the user's preferences.

C ChatGPT 4o Evaluation Details

In our experiments, we used ChatGPT 4o[9] to evaluate the results of different generators. We utilized the Plus subscription, enabling the model to understand and assess multimodal data. Figure 4 shows the prompt used for evaluation with GPT. We only needed to use this prompt and upload the generated results. This process can be fully automated with script, which also automatically tallies the results. To facilitate result collection, we added an extra prompt: *Please first output the sorted results directly, then explain the reasons.* Figure 5 provides an evaluation example, where the red section repre-

[9] https://chat.openai.com/, the GPT responses are from June 2024.

sents **historical click information**, the blue section represents the **objective item text information**, and the green section represents the **evaluation of the generated results**. As shown, the generation results of our framework satisfy ChatGPT 4o.

D Time Consumption Analysis

During the preparation stage, the extraction speed for foreground information from images is ∼7 items/s. The speed for extracting tags from item images using LLaVA/LLaMA is ∼4-5s/item, and the encoding speed for text and images using PLMs is 120 items/s and 10 items/s, respectively. Note that these preprocessing steps are persistent and only required during the preparation phase. Before inference, the text and image embedding pools need to be preloaded. The loading speed for text is ∼280 items/s, and for images, it is ∼4800 items/s.

In inference, it takes ∼**3s** to generate 4 personalized images for each user's objective item, which is **roughly the same** as without using rewrite and IRA. This is affordable for personalized services.

E Prompt for LLM/VLM to Get Tags

To obtain more precise tags for various scenarios, it is often necessary to further refine the prompts for items in the text and image modalities in a scene-specific manner. Figure 6 shows the prompts for the scenarios of movie poster (MovieLens), news poster (MIND) and outfit (POG). We manually specify the key points the model should focus on. For posters, common aspects that reflect user interests include color, tone, theme, and style. For outfits, particular attention should be given to color, style, pattern, and occasion. We could find that, for different scenarios and different modalities, the prompt should be different and adaptive.

F Human Studies Questionnaire

In the human study section of the main text, we asked participants to score the generated results with ranks based on user history interests. We provided a unified template for scoring across three dataset scenarios, as shown in Figure 7 (in the next page). To avoid participant bias towards the order during the scoring process, the positions of images within the same questionnaire can be randomly shuffled when creating the questionnaire, in order to further improve reliability.

E.1 MovieLens

Prompt for VLM

<image> is a **movie poster**, please conclude the tags of it: ..... Focus on color, tone, character, expression, setting, symbolism, artistic style, cultural elements, note that the number of tags should be less than 8, separate them by commas.

(a) Prompt to VLM for MovieLens.

Prompt for LLM

<text> is a **movie description**, please conclude the tags of it: ..... Focus on tone, character, expression, setting, cultural elements, note that the number of tags should be less than 8, separate them by commas.

(b) Prompt to LLM for MovieLens.

E.2 MIND

Prompt for VLM

<image> is a **news poster**, please conclude the tags of it: ..... Focus on the news theme, location, figures, color, mood, style, note that the number of tags should be less than 8, separate them by commas.

(c) Prompt to VLM for MIND.

Prompt for LLM

<text> is a **news abstract including title**, please conclude the tags of it: ..... Focus on the news theme, location, mood, style, note that the number of tags should be less than 8, separate them by commas.

(d) Prompt to LLM for MIND.

E.3 POG

Prompt for VLM

<image> is an **outfit figure**, please conclude the tags of it: ..... Focus on the outfit color, style, material, pattern, occasion, size, seasonality, note that the number of tags should be less than 8, separate them by commas.

(e) Prompt to VLM for POG.

Prompt for LLM

<text> is an **outfit description**, please conclude the tags of it: ..... Focus on the outfit style, pattern, occasion, seasonality, note that the number of tags should be less than 8, separate them by commas.

(f) Prompt to LLM for POG.

Figure 6: Prompts for VLM & LLM on MovieLens, MIND, and POG.

Template for User Questionnaire

If you are a <user info if provided> who has previously clicked on and viewed <item history>, and you are now browsing a <task/scenario>: <item brief information>, please select the image you find most appealing below.

You should score the images from your top choice to your least favorite within 1-4, with the better one scored lower rank (from 1) and the worse one scored higher rank (up to 4).

We have automatically organized the item images below, please score them based on the preference:

<fig 1>, <fig 2>, <fig 3>, …

Figure 7: Questionnaire template of human study.