Personalized Image Generation with Large Multimodal Models

Yiyan Xu
yiyanxu24@gmail.com
University of Science and Technology of China
China, Hefei

Wenjie Wang
wenjiewang96@gmail.com
University of Science and Technology of China
China, Hefei

Yang Zhang
zyang1580@gmail.com
National University of Singapore
Singapore, Singapore

Biao Tang
biao.tang@meituan.com
Meituan
China, Shanghai

Peng Yan
yanpeng04@meituan.com
Meituan
China, Beijing

Fuli Feng*
fulifeng93@gmail.com
University of Science and Technology of China
China, Hefei

Xiangnan He*
xiangnanhe@gmail.com
MoE Key Lab of BIPC, University of Science and Technology of China
China, Hefei

Abstract

Personalized content filtering, such as recommender systems, has become a critical infrastructure to alleviate information overload. However, these systems merely filter existing content and are constrained by its limited diversity, making it difficult to meet users' varied content needs. To address this limitation, personalized content generation has emerged as a promising direction with broad applications. Nevertheless, most existing research focuses on personalized text generation, with relatively little attention given to personalized image generation. The limited work in personalized image generation faces challenges in accurately capturing users' visual preferences and needs from noisy user-interacted images and complex multimodal instructions. Worse still, there is a lack of supervised data for training personalized image generation models.

To overcome the challenges, we propose a ***P**ersonalized **I**mage **G**enerati**o**n Framework* named Pigeon, which adopts exceptional large multimodal models with three dedicated modules to capture users' visual preferences and needs from noisy user history and multimodal instructions. To alleviate the data scarcity, we introduce a two-stage preference alignment scheme, comprising masked preference reconstruction and pairwise preference alignment, to align Pigeon with the personalized image generation task. We apply Pigeon to personalized sticker and movie poster generation, where extensive quantitative results and human evaluation highlight its superiority over various generative baselines.

CCS Concepts

• **Information systems** → **Recommender systems**; **Personalization**; **Multimedia content creation**.

Keywords

Personalized Image Generation, Large Multimodal Models, Preference Alignment

ACM Reference Format:
Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. 2025. Personalized Image Generation with Large Multimodal Models. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3696410.3714843

1 Introduction

In the era of information overload, individuals are overwhelmed with vast amounts of multimodal content on the Web, underscoring the importance of personalized content delivery [48, 49, 57]. The predominant approach, personalized content filtering like recommender systems [6, 8, 45, 46], relies on user interaction history and contextual information to infer user preferences and filter existing content. However, this approach is constrained by the limited diversity of available content, rendering it inadequate to fully meet users' varied content needs (see an example in Figure 1). To address this limitation, generating personalized new content is becoming increasingly important across various domains, including personalized movie posters [36], advertisements [42, 52], music [4, 27], and fashion designs [51, 54].

Previous works on personalized content generation primarily focus on personalized text generation [15, 31, 34, 35] while personalized image generation receives little attention. Technically,

*Corresponding authors. This work is supported by the National Natural Science Foundation of China (62272437, U24B20180, 62121002) and the advanced computing resources provided by the Supercomputing Center of the USTC.

WWW '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714843

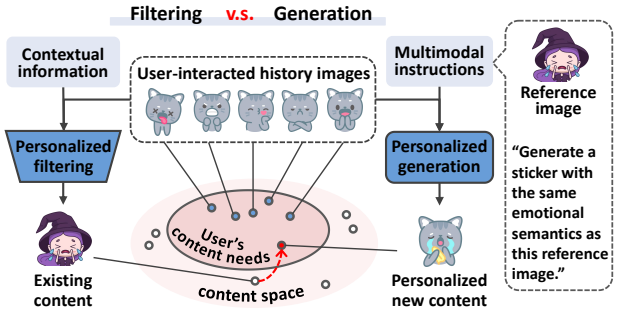


Figure 1: Personalized filtering selects the most relevant existing content while personalized generation creates new and customized ones, more precisely satisfying users' diverse content needs.

personalized image generation aims to capture implicit user preferences from user-interacted history images and then integrate users' explicit needs from multimodal instructions to generate personalized target images, as illustrated in Figure 1. Existing methods mainly rely on Diffusion Models (DMs) or Large Language Models (LLMs) for personalized image generation:

- **DM-based methods** [3, 5, 33, 51, 52] might learn the representations of implicit user preferences from user-interacted history images and combine these representations with explicit user instructions for target images to guide the generation of DMs. However, these methods struggle to accurately capture user preferences from noisy history images, which typically cover diverse and complex user interests.
- **LLM-based Personalized Multimodal Generation (PMG)** [36] converts history images and multimodal instructions into textual descriptions, and then utilizes pre-trained LLMs to encode textual descriptions for guiding image generation. However, the discrete nature of text makes it difficult to convey complex visual information in history images and instructions, leading to imprecise representations.

In this light, the key to personalized image generation lies in accurately inferring implicit user preferences from noisy history images while adhering to explicit multimodal instructions for image generation. This necessitates robust multimodal understanding, reasoning, and instruction-following capabilities, driving the adoption of Large Multimodal Models (LMMs) [7, 13] for personalized image generation. An intuitive approach is to transform history images and multimodal instructions into visual and textual tokens as the input of LMMs for cross-modal understanding and image generation. However, this approach faces critical challenges:

- User-preferred and disliked features (*e.g.,* characters and colors) are often entangled within user-interacted history images, producing fine-grained noise at the feature level. This significantly challenges LMMs to infer implicit user preferences.
- The multimodal instructions may include a reference image alongside textual instructions, *e.g.,* "generate a sticker with the same emotional semantics as this reference image", requiring LMMs to generate the target image with high-level semantic alignment with the reference image.
- Worse still, existing LMMs are not specifically trained for personalized image generation, making it challenging to infer user preferences and align with multimodal instructions. Furthermore, there is a lack of supervised data containing triplets of <user-interacted history images, multimodal instructions, a personalized target image> for LMM training.

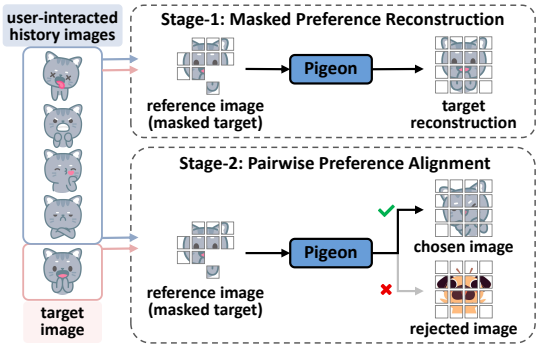


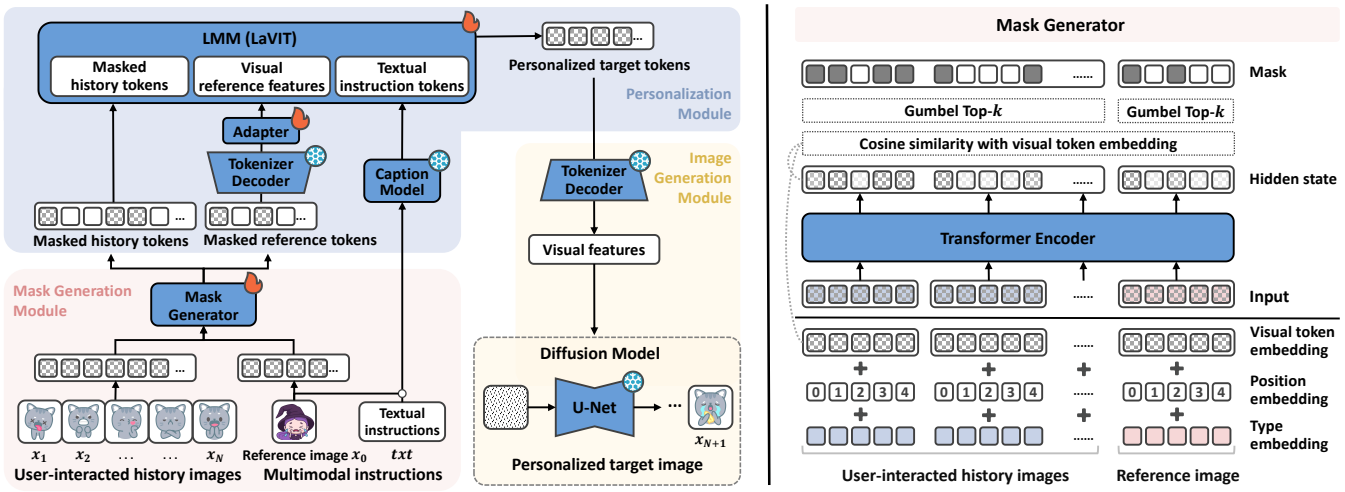
Figure 2: Two-stage preference alignments for Pigeon: given user-interacted images, the last image is treated as the target, with the preceding ones as user history.

To address the challenges, we propose a ***P**ersonalized **I**mage **Ge**neratio**n** Framework* (shorted as Pigeon) for LMMs, comprising three key modules: 1) *Mask generation module* incorporates a mask generator to create token-level masks for reference-aware history filtering, effectively removing noisy signals from the history images at the feature level (*cf.* Section 2.2.1). 2) *Personalization module* integrates masked history tokens and encodes multimodal instructions with the transformed semantic features of the reference image to generate personalized tokens (*cf.* Section 2.2.2). 3) *Image generation module* employs a DM to convert the generated personalized tokens into the personalized target image.

Due to the lack of supervised data, Pigeon adopts a two-stage preference alignment scheme to adapt LMMs for personalized image generation. As shown in Figure 2, the first stage assumes that user-interacted history images, despite some noise, still partially reflect implicit user preferences. Given a sequence of these images, Pigeon treats the last one as the target image and the preceding images as the history images. We then mask the target image as a reference image to construct the user's multimodal instructions and fine-tune Pigeon to reconstruct the target image based on this user's history images and multimodal instructions, regulating Pigeon to infer user preference from history. In the second stage, Pigeon generates multiple target images based on the first-stage alignment and ranks them using a preference reward strategy, thus forming pseudo-labeled preference data pairs of "chosen" and "rejected" images. Pigeon is then optimized with the preference data pairs via Direct Preference Optimization (DPO) [30] to generate more personalized target images, enhancing personalization capabilities.

We validate the effectiveness of Pigeon in two popular scenarios: personalized sticker and movie poster generation. Extensive quantitative evaluation demonstrates that Pigeon outperforms the best baseline in personalization, achieving improvements of 7%~31% while maintaining comparable semantic alignment with the reference image. Notably, human evaluation on Amazon MTurk[1] reveals that, on average, 71% participants rate Pigeon-generated images with superior personalization and semantic alignment.

[1]https://www.mturk.com/.



(a) Overview of Pigeon (b) Mask Generator

Figure 3: Pigeon consists of three key modules: 1) mask generation module creates token-level masks for history and reference images, 2) personalized module encodes multimodal instructions and integrates them with masked history to generate personalized tokens, and 3) image generation module utilizes these tokens to produce personalized images.

Furthermore, we discuss the versatility of Pigeon extending to more domains such as personalized product images, advertisement, and fashion images in Appendix A, highlighting Pigeon's broad applicability and significant economic value. Our code and data are available at https://github.com/YiyanXu/Pigeon.

In summary, the key contributions of this work are as follows:

- We empower LMMs with the capability of personalized image generation by the Pigeon framework, which can infer user preferences from noisy history images and integrate multimodal instructions for personalized image generation. Pigeon offers a wide range of applications, catering to diverse user demands and driving the evolution of content delivery paradigms.
- We introduce a two-stage preference alignment scheme to effectively adapt LMMs for the personalized image generation task, eliminating the need for supervised data.
- We propose multiple quantitative evaluation metrics for personalized image generation and conduct extensive experiments across two scenarios. Both quantitative results and human evaluation validate that Pigeon significantly surpasses all the baselines, effectively aligning with personalized user preferences.

2 Personalized Image Generation

In this section, we first formulate the personalized image generation task, followed by the elaboration of our proposed Pigeon framework and its potential applications across various domains.

2.1 Task Formulation

Personalized image generation aims to synthesize personalized images tailored to implicit user preferences and explicit multimodal instructions. Formally, given a set of user-interacted history images $\mathcal{H} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ and multimodal instructions $\mathcal{R} = \{\boldsymbol{x}_0, \boldsymbol{txt}\}$, where $\boldsymbol{x}_0$ and $\boldsymbol{txt}$ represent the reference image and textual instruction, respectively, this goal is to generate a personalized target image $\boldsymbol{x}_{N+1}$ that not only meets user visual preferences but also adheres to multimodal instructions by high-level semantic alignment with the reference image. This task has broad applications in enhancing user experience across various domains, such as generating personalized product images in e-commerce or creating personalized movie posters and video thumbnails on platforms like Netflix and YouTube.

2.2 Pigeon

To achieve personalized image generation, Pigeon leverages a representative LMM named LaVIT [13] for instantiation[2]. Specifically, LaVIT includes a visual tokenizer that translates images into visual tokens for multimodal understanding, and a tokenizer decoder that transforms generated visual tokens into dense visual features to guide image generation. Built upon LaVIT, as depicted in Figure 3(a), Pigeon comprises three key modules: 1) *mask generation module* employs a mask generator to create token-level masks for both history and reference images. 2) *personalization module* extracts high-level semantic features of multimodal instructions and combines them with the masked history tokens to guide LaVIT to generate personalized tokens that reflect users' content needs. 3) *image generation module* converts these tokens into visual features to generate personalized target images via a DM.

2.2.1 ***Mask Generation Module.*** To discard the noise from user-interacted history images, we introduce a mask generator based on a Transformer encoder [43]. It leverages attention mechanisms to encode both history and reference images, and identifies key history tokens that are more relevant to the reference image and contain more personalized information, producing a history mask to filter out noisy tokens. Besides, the mask generator can also create a token-level mask for the reference image to support the two-stage preference alignments, which will be illustrated in Section 2.2.4.

• **Identification of important visual tokens.** Given a set of user-interacted history images $\mathcal{H}$ and a reference image $\boldsymbol{x}_0$, we first tokenize these images into visual token sequences:

$$E_i = \textbf{Visual\_Tokenizer}(\boldsymbol{x}_i), i = 0, \ldots, N, \quad (1)$$

[2]Pigeon can also be applied to more LMMs, which is left for future exploration.

where $\boldsymbol{E}_i = [\boldsymbol{e}_{i1}, \ldots, \boldsymbol{e}_{iL_i}]$ represents the visual token embedding sequence of each image $\boldsymbol{x}_i$ with length L_i, and $\textbf{Visual\_Tokenizer}(\cdot)$ refers to the visual tokenizer with a visual embedding layer from the pre-trained LaVIT. This process is omitted in Figure 3(a) for brevity. The mask generator, as shown in Figure 3(b), combines position and type embeddings with the visual token embeddings via element-wise addition to form the input, which allows the Transformer encoder to distinguish between history and reference tokens and capture the token sequence order within each image. The encoding process is formulated as follows:

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N, \boldsymbol{Z}_0 = \textbf{Encoder}(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_N, \boldsymbol{E}_0), \tag{2}$$

where $\boldsymbol{Z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iL_i}]$ represents the hidden states of each token sequence $\boldsymbol{E}_i$, and $\textbf{Encoder}(\cdot)$ encapsulates both the element-wise addition and the encoding process. During the encoding process, the attention mechanism allows the visual tokens from both history and reference images to attend to each other, prioritizing important information while reducing the impact of outlier noise. To quantify the importance of each token, we compute the cosine similarity between the hidden states and the original visual token embeddings:

$$s_{ij} = cosine(\boldsymbol{z}_{ij}, \boldsymbol{e}_{ij}), j = 1, \ldots, L_i, \tag{3}$$

where s_{ij} denotes the importance score of the j-th token in each visual token sequence $\boldsymbol{E}_i$. Intuitively, a higher score indicates more key information is retained in the token.

• **Reference-aware history filtering.** We create a multi-hot binary mask $\boldsymbol{m}_h$ to mask the low-score tokens according to the history mask ratio $\alpha_h \in [0, 1]$. This mask filters out noisy or reference-irrelevant history tokens, yielding the filtered token embeddings for each history image: $[\tilde{\boldsymbol{E}}_1, \ldots, \tilde{\boldsymbol{E}}_N] = \boldsymbol{m}_h \odot [\boldsymbol{E}_1, \ldots, \boldsymbol{E}_N]$, where $\tilde{\boldsymbol{E}}_i$ denotes masked history token embeddings. For gradient backpropagation in this discrete sampling process, the Gumbel-Softmax trick [24] is applied to the non-differentiable binary mask.

2.2.2 ***Personalization Module.*** To effectively handle multi-modal instructions, this module first encodes them to extract essential high-level semantic features, then combines these features with masked history tokens into a hybrid prompt, which serves as the input to LMM, enabling the generation of personalized tokens.

• **Multimodal instructions encoding.** When directly utilizing the reference image to guide target image generation, LMMs often duplicate the reference image, failing to effectively incorporate personalized information (see empirical results in Section 3.4.2). This highlights the necessity to extract high-level semantics from the reference image for image generation. To enrich the semantics of the reference image $\boldsymbol{x}_0$ and enhance the comprehension of multimodal instructions in LMMs, we utilize a caption model (*e.g.,* BLIP-2 [18] and LLaVA [22]) to generate a textual description of the reference image, which is then tokenized into textual tokens:

$$\boldsymbol{r}_t = \textbf{Text\_Tokenizer}(\textbf{Caption}(\boldsymbol{x}_0)), \tag{4}$$

where $\boldsymbol{r}_t$ refers to the high-level textual semantic features extracted from the reference image, and $\textbf{Text\_Tokenizer}(\cdot)$ denotes the text tokenizer with the word embedding layer from LaVIT.

For visual semantics, we transform the low-level reference token embedding sequence $\boldsymbol{E}_0$ into high-level dense visual features. Here, we utilize the pre-trained tokenizer decoder of LaVIT for the transform to avoid introducing extra parameters, followed by average pooling to aggregate the multiple feature vectors from the tokenizer decoder:

$$\boldsymbol{v} = \textbf{AvgPooling}(\textbf{Tokenizer\_Decoder}(\boldsymbol{E}_0)). \tag{5}$$

Next, an adapter layer is introduced to align the feature dimension of $\boldsymbol{v}$ with the LaVIT embeddings, *i.e.,* $\boldsymbol{r}_v = \textbf{Adapter}(\boldsymbol{v})$, where $\boldsymbol{r}_v$ denotes the extracted high-level visual semantic features.

• **Hybrid prompt for LMM.** To integrate these encoded semantic features with filtered history into prompts for LMM instruction tuning, we propose a hybrid prompt that is structured as follows:

$$\boldsymbol{p} = \textbf{Prompt}(\tilde{\boldsymbol{E}}_1, \ldots, \tilde{\boldsymbol{E}}_N, \boldsymbol{r}_t, \boldsymbol{r}_v). \tag{6}$$

Instruction: You are a helpful personalized assistant. You will receive a list of user-liked images that reflect the user's visual preferences. By analyzing user preferences, please generate a personalized image that aligns with the user's aesthetic taste and the semantics in a specified reference image.

Input: The user likes the following images: $\tilde{\boldsymbol{E}}_1, \ldots, \tilde{\boldsymbol{E}}_N$. The reference image: $\boldsymbol{r}_t, \boldsymbol{r}_v$.

Response: <Personalized Target Tokens $\boldsymbol{E}_{N+1}$ >

By using a hybrid prompt similar to the above one, LMMs can adapt to various scenarios to generate personalized target tokens.

2.2.3 ***Image Generation Module.*** With personalized target tokens $\boldsymbol{E}_{N+1}$, the pre-trained tokenizer decoder of LaVIT converts these discrete tokens into dense visual features, which can guide the generation of the personalized target image $\boldsymbol{x}_{N+1}$ in DM.

2.2.4 ***Two-stage Preference Alignments.*** To optimize Pigeon for personalized image generation, an intuitive strategy is maximizing the generation likelihood of the target tokens $\boldsymbol{E}_{N+1}$, based on the prompt $\boldsymbol{p}$ in Eq. (6). However, since there is no supervised dataset containing triplets of <user-interacted history images, multimodal instructions, personalized target image>, we propose a two-stage preference alignment process for effective instruction tuning.

• **Stage-1: Masked Preference Reconstruction.** We assume that user-interacted history images, despite containing some noise, still reflect user visual preferences. Based on this, as shown in Figure 2, given a sequence of user-interacted images $\{\boldsymbol{x}_i\}_{i=1}^{N+1}$, the last one $\boldsymbol{x}_{N+1}$ is considered the personalized target image, while the preceding images are treated as history images $\mathcal{H} = \{\boldsymbol{x}_i\}_{i=1}^{N}$.

Supervised dataset construction. Considering the lack of multimodal instructions, we adopt the target image as the reference to construct multimodal instructions $\mathcal{R} = \{\boldsymbol{x}_{N+1}, \boldsymbol{txt}\}$. A token-level reference mask is then applied to corrupt the reference image, encouraging the model to extract user preferences from history images for target reconstruction. Specifically, we utilize the importance score defined in Eq. (3) to rank all the reference tokens and create the token-level mask for the reference image.

Unlike the history mask, which filters out noise by discarding low-score tokens, we introduce a dual-phase mask scheme for the reference image. During training, we mask high-score reference tokens, which contain more personalized information (as discussed in Section 2.2.1), forcing the model to rely on history images to recover the target. During inference, low-score tokens are masked instead, utilizing the preference reconstruction capability

to generate more personalized content. Formally, the dual-phase mask $\boldsymbol{m}_r$ with a reference mask ratio $\alpha_r \in [0, 1]$ is applied to the reference tokens by $\tilde{\boldsymbol{E}}_0 = \boldsymbol{m}_r \odot \boldsymbol{E}_0$. We then replace $\boldsymbol{E}_0$ in Eq. (5) with $\tilde{\boldsymbol{E}}_0$ to derive the modified visual features for the hybrid prompt $\boldsymbol{p}$ in Eq. (6), optimizing the model to reconstruct the target token sequence $\boldsymbol{E}_{N+1}$. In this way, we could construct a supervised prompt-response dataset $\mathcal{D} = \{(\boldsymbol{p}, \boldsymbol{E}_{N+1})^k\}_k$ from the available interaction sequences for masked preference reconstruction.

Supervised fine-tuning. For parameter-efficient fine-tuning, we introduce a LoRA [10] module into the pre-trained LaVIT, which keeps the LaVIT parameters frozen and imports trainable low-rank decomposition matrices for updates. As shown in Figure 3(a), we only fine-tune specific components of Pigeon, namely the mask generator, adapter, and LoRA for LaVIT, while freezing all the other parameters. During training, we randomly sample the reference mask ratio $\alpha_r \in [0, 1]$ and fine-tune Pigeon for target reconstruction, aiming to capture more robust user preferences. Formally, the loss function is defined as the negative likelihood of the target token sequence via an auto-regressive manner:

$$\mathcal{L}_{sft} = -\sum_{\substack{(\boldsymbol{p}, \boldsymbol{E}_{N+1}) \in \mathcal{D} \\ \alpha_r \sim \mathcal{U}(0,1)}} \sum_{j=1}^{L_{N+1}} \log \left(P_\Theta(\boldsymbol{e}_{N+1,j} | \boldsymbol{p}(\alpha_r), \boldsymbol{e}_{N+1,<j}) \right), \quad (7)$$

where $\boldsymbol{e}_{N+1,j}$ is the j-th token in the sequence $\boldsymbol{E}_{N+1}$ of length L_{N+1}, $\boldsymbol{p}(\alpha_r)$ is the hybrid prompt with a uniformly sampled reference mask ratio, and $\boldsymbol{\Theta}$ includes all the learnable parameters of Pigeon.

• **Stage-2: Pairwise Preference Alignment.** After the first-stage fine-tuning, Pigeon is capable of following the instructions for personalized image generation. To further enhance its personalization capability, we adopt DPO [30] for pairwise preference alignment, which utilizes preference pairs of chosen and rejected responses to optimize the model to produce the chosen one.

Preference dataset construction. To construct the preference data pairs for DPO, we first generate multiple personalized target token sequences for each prompt $\boldsymbol{p}(\alpha_r)$ with varying reference mask ratios $\alpha_r \in \{0.0, 0.1, \ldots, 1.0\}$ based on the first-stage alignment. These tokens are then transformed into images $\boldsymbol{x}(\alpha_r)$ via the image generation module. To identify the best and worst personalized images, we introduce a preference reward strategy to rank all generated images. Following [36], we compute the CLIP similarity between each generated image and the history images:

$$s(\alpha_r) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{CLIPSim}(\boldsymbol{x}(\alpha_r), \boldsymbol{x}_i), \quad (8)$$

where $s(\alpha_r)$ is the preference score of image $\boldsymbol{x}(\alpha_r)$. We rank the generated images based on these scores to form the pseudo-labeled preference dataset $\bar{\mathcal{D}} = \{(\boldsymbol{p}, \boldsymbol{E}', \boldsymbol{E}'')^k\}_k$, where $\boldsymbol{E}'$ and $\boldsymbol{E}''$ denote the chosen and rejected token sequences for DPO, corresponding to images with the highest and lowest preference scores.

Preference optimization. In this stage, we continue updating the LoRA weights while keeping all the other parameters frozen. With the preference dataset, the loss function can be formulated as:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{\substack{(\boldsymbol{p}, \boldsymbol{E}', \boldsymbol{E}'') \sim \bar{\mathcal{D}} \\ \alpha_r \sim \mathcal{U}(0,1)}} \left[\log \sigma \left(\beta \frac{P_{\Theta_l}(\boldsymbol{E}' | \boldsymbol{p}(\alpha_r))}{P_{\hat{\Theta}_l}(\boldsymbol{E}' | \boldsymbol{p}(\alpha_r))} - \beta \frac{P_{\Theta_l}(\boldsymbol{E}'' | \boldsymbol{p}(\alpha_r))}{P_{\hat{\Theta}_l}(\boldsymbol{E}'' | \boldsymbol{p}(\alpha_r))} \right) \right], \quad (9)$$

where Θ_l denotes the learnable parameters of the LoRA module, and β is a parameter controlling the deviation from the reference model $\hat{\Theta}_l$ obtained in the first-stage alignment.

2.2.5 ***Inference.*** To manage the trade-off between personalization and semantic alignment with the reference image, users could adjust the reference mask ratio to control how much reference information is retained in the generated images. During inference, given history images $\mathcal{H} = \{\boldsymbol{x}_i\}_{i=1}^N$, multimodal instructions $\mathcal{R} = \{\boldsymbol{x}_0, \boldsymbol{txt}\}$ and a user-specified reference mask ratio α_r, Pigeon can mask the low-score reference tokens accordingly to generate an image $\boldsymbol{x}_{N+1}$ that aligns with the user's visual preferences and multimodal instructions.

3 Experiments

We evaluate Pigeon in sticker and movie poster scenarios to validate its superiority by answering the following research questions:

- **RQ1:** How does Pigeon perform compared with DM-based, LLM-based, and LMM-based personalized image generation methods, based on quantitative evaluation?
- **RQ2:** Can Pigeon surpass the baselines in human evaluation?
- **RQ3:** How do the special designs of Pigeon (*e.g.,* history mask, multimodal instruction encoding strategy, and two-stage preference alignment process) affect the performance?

3.1 Experimental Settings

3.1.1 ***Datasets.*** We conduct experiments on two datasets, focusing on sticker and movie poster scenarios: 1) **SER30K**[3] is a large-scale dataset of stickers, each categorized by theme and annotated with an associated emotion label; and 2) **ML-Latest**[4], a benchmark dataset containing user ratings on movies. Details regarding data processing and dataset statistics are available in Appendix B.

3.1.2 ***Baselines.*** We compare Pigeon with various generative baselines, including methods based on DMs, LLMs, and LMMs: **1) Textual Inversion (TI)** [5] introduces a word embedding to learn user preference representation, which is then combined with textual instructions to guide the text-to-image generation process in DMs. **2) PMG** [36] transforms user-interacted and reference images into textual descriptions, using pre-trained LLMs to extract user preferences through keywords and implicit embeddings to condition the image generator. **3) LLaVA** [22] is an LMM designed to extract dense image features for visual reasoning, generating text by default but capable of producing images when integrated with an external text-to-image generator. **4) LaVIT** [13] is another LMM that converts images into discrete visual tokens for reasoning and generates visual tokens to guide the image generation process.

Additionally, we include two results for reference: 5) **Recon**, which utilizes the visual tokenizer, tokenizer decoder, and DM of the pre-trained LaVIT for image reconstruction without personalization; and 6) **Grd**, representing the evaluation results of the reference images. The performance gap between Recon and Grd reflects the difference between generated and real-world images.

3.1.3 ***Evaluation Metrics.*** We employ various quantitative evaluation metrics for performance comparison. Following [36, 38], we mainly focus on **personalization** and **semantic alignment** with the reference image by measuring the semantic and perceptual similarity between generated and history/reference images.

[3] https://github.com/nku-shengzheliu/SER30K.
[4] https://grouplens.org/datasets/movielens.

Table 1: Quantitative performance comparison between Pigeon and the baselines in both scenarios. Baselines labeled with "$*$" indicate the pre-trained models. The best results are highlighted in bold, while the second-best results are underlined.

#Sticker			Personalization					Semantic Alignment			Fidelity
Methods		Overall	CS↑	CIS↑	DIS↑	LPIPS↓	MS-SSIM↑	CS↑	CIS↑	DIS↑	FID↓
DM-based	TI	36.91	18.67	40.90	36.58	0.7654	0.0887	**32.91**	53.67	48.50	105.48
LLM-based	PMG	32.83	19.16	47.34	39.15	0.7383	0.0827	18.31	45.45	37.80	84.91
LMM-based	LLaVA*	32.40	17.88	47.26	42.59	0.7575	0.0966	17.54	42.65	39.25	93.23
	LLaVA	32.23	18.72	37.44	33.19	0.7552	0.0851	27.02	49.15	43.88	95.19
	LaVIT*	34.56	18.77	53.63	50.96	0.6855	0.1376	15.49	40.76	39.09	107.53
	LaVIT	33.15	16.39	40.56	40.84	0.7377	0.1128	25.74	**70.80**	**69.93**	**83.39**
	Pigeon	**44.38**	**23.69**	**67.65**	**62.23**	**0.6814**	**0.1568**	21.10	47.44	45.44	89.43
Reference	Recon	33.22	16.30	40.60	40.76	0.7370	0.1126	25.84	71.09	70.14	83.57
	Grd	36.98	16.93	45.00	43.71	0.6443	0.1349	28.95	100.00	100.00	-
#Movie poster			Personalization					Semantic Alignment			Fidelity
Methods		Overall	CS↑	CIS↑	DIS↑	LPIPS↓	MS-SSIM↑	CS↑	CIS↑	DIS↑	FID↓
DM-based	TI	31.07	12.41	28.29	19.18	0.7721	0.0399	**33.84**	43.53	39.81	79.77
LLM-based	PMG	20.36	13.61	25.11	**22.73**	0.7692	0.0261	15.60	27.29	25.15	77.25
LMM-based	LLaVA*	22.08	12.24	29.60	19.73	0.7607	0.0373	14.55	31.76	21.99	73.77
	LLaVA	30.59	12.62	30.64	19.33	0.7690	0.0370	30.53	48.50	41.45	54.55
	LaVIT*	23.81	12.64	28.23	17.50	0.7546	0.0458	19.39	36.93	37.71	50.08
	LaVIT	27.82	13.86	30.49	19.95	0.7548	0.0370	25.15	46.02	**60.07**	**33.53**
	Pigeon	**33.31**	**15.41**	**40.16**	21.29	**0.7508**	**0.0464**	26.45	**49.66**	44.07	47.79
Reference	Recon	27.81	13.85	30.33	19.95	0.7548	0.0367	25.29	46.08	60.52	33.74
	Grd	41.58	10.94	51.34	20.75	0.7502	0.0402	31.81	100.00	100.00	-

- **Semantic similarity.** We adopt CLIP [29] and DINO [26] to extract image features from generated and history/reference images, and compute the cosine similarity between them to obtain the CLIP Image Score (CIS) and DINO Image Score (DIS). Additionally, the CLIP Score (CS) measures the similarity between generated images and textual descriptions of the history/reference images. To assess the overall performance, we also calculate a unified F1-score, combining the history CIS and reference CS.
- **Perceptual similarity.** To evaluate finer-grained visual personalization, we apply LPIPS [55] and MS-SSIM [47] to quantify the perceptual similarity between generated and history images.
- **Fidelity.** We also employ the widely-used FID metric to assess the fidelity of the generated images.

3.1.4 ***Implementation Details.*** All the baselines are tuned with a fixed learning rate of $1e^{-5}$. We implement PMG following its default model designs, while other baselines are implemented with Stable Diffusion XL [28] as the image generator for fair comparisons. Detailed hyper-parameter settings and computational overhead of Pigeon are summarized in Appendix C.

3.2 Quantitative Evaluation (RQ1)

The comparison between Pigeon and the baselines is shown in Table 1. The observations are summarized as follows:

- DM-based TI outperforms most baselines in semantic alignment by directly using the textual description of the reference image for text-to-image generation. However, noisy signals in interaction history hinder its ability to precisely capture user preferences, resulting in inferior personalization.
- PMG converts images into textual descriptions and uses LLMs to infer user preferences for guiding image generation. The image-to-text conversion may overlook critical visual details, leading to inaccurate preference modeling and multimodal instruction understanding. As a result, PMG presents moderate performances in both personalization and semantic alignment.
- The decent performance of the pre-trained LLaVA and LaVIT in personalization validates the strength of advanced instruction-following and visual understanding capabilities in LMMs for personalized image generation. Among them, LLaVA relies on personalized text to guide image generation, which can cause misalignments between expressed textual preferences and actual visual preferences, resulting in relatively lower performance.
- After fine-tuning in each scenario, both LLaVA and LaVIT tend to reconstruct reference images rather than generate personalized ones, as evidenced by significant improvements in semantic alignment alongside a decline in personalization. This is mainly due to the lack of supervised data for model training.
- Pigeon exhibits superior performance in most personalization metrics across two scenarios, while maintaining comparable semantic alignment and fidelity. These results underscore the effectiveness of Pigeon in capturing user visual preferences from noisy history images and accurately understanding multimodal instructions to produce personalized images.

3.3 Human Evaluation (RQ2)

To assess the qualitative performance of Pigeon in personalization and semantic alignment, we conduct a human evaluation on Amazon MTurk[5], comparing it against Grd and two representative baselines: 1) TI, which exhibits the second-best overall performance in Table 1 , and 2) PMG, designed for personalized image generation. The evaluation adopts binary-choice tests across sticker and movie poster scenarios, each with 50 cases. For personalization, we present

[5]https://www.mturk.com/.

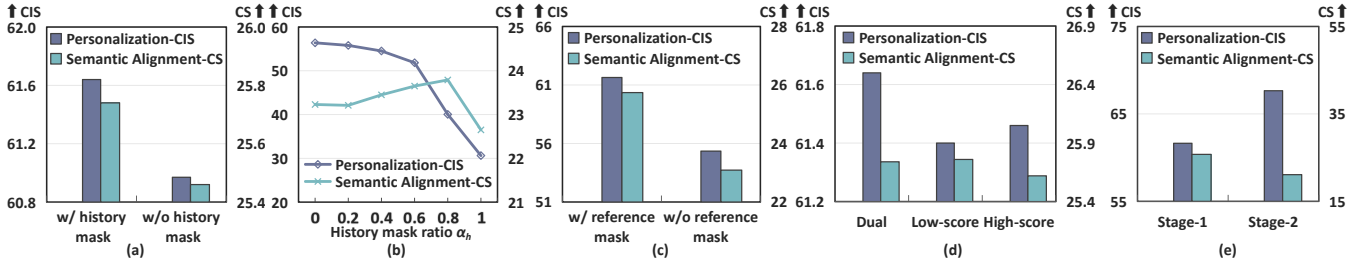


Figure 4: In-depth analysis of the history mask and the two-stage preference alignment process.

Table 2: The human evaluation results, where "±" denotes 95% confidence interval. Pigeon is consistently preferred (≥ 50%) over the baselines across sticker and movie poster scenarios.

Pigeon		Grd	TI	PMG
Personalization	Sticker	$0.91^{\pm 2.19\%}$	$0.91^{\pm 2.19\%}$	$0.89^{\pm 1.79\%}$
	Movie	$0.62^{\pm 2.85\%}$	$0.66^{\pm 2.16\%}$	$0.57^{\pm 2.51\%}$
Semantic Alignment	Sticker	-	$0.54^{\pm 2.67\%}$	$0.67^{\pm 3.65\%}$
	Movie	-	$0.58^{\pm 2.58\%}$	$0.73^{\pm 2.22\%}$

five user-interacted history images and the generated images, with the question: "When provided with someone's five previously liked stickers (movies), please select the next sticker (movie poster) that is more attractive to her/him." For semantic alignment, we display the reference and generated images with the question: "Which image aligns more closely with the semantics of the reference image?" As shown in Table 2, Pigeon consistently surpasses (≥ 50%) the baselines, even the Grd, in personalization and maintains decent results in semantic alignment with reference images. These findings emphasize its superiority in capturing user preferences from noisy history images and effectively integrating multimodal instructions for image generation, which aligns with the quantitative analysis. More detailed information can be found in Appendix D.

3.4 In-depth Analysis (RQ3)

In this section, we conduct additional experiments in the sticker scenario to further investigate the effects of various Pigeon designs, including the history mask, multimodal instruction encoding strategy, and the two-stage preference alignment process. To reduce resource costs, we mainly focus on the results after first-stage preference alignment for fair comparisons.

3.4.1 ***Effect of history mask.*** To assess the effectiveness of the history mask in managing noisy history images, we exclude it during training and present the results on two key metrics in Figure 4(a). The findings show that: 1) noise in the history images prevents the model from accurately capturing user preferences and even disrupts the semantic alignment with the reference image. 2) The history mask could effectively filter out the noisy signals, thereby enhancing model performance.

Additionally, we vary the history mask ratio α_h during inference, with the reference mask ratio fixed at 0.5. The results in Figure 4(b) reveal that increasing α_h discards both noise and useful personalized information in history images, causing Pigeon to rely more on the reference image, thus slightly improving the semantic alignment. However, this also makes it harder for Pigeon to extract user preferences, reducing the performance in personalization.

Table 3: Effects of multimodal instruction encoding.

	Personalization		Semantic Alignment
	CIS↑	LPIPS↓	CS↑
Pigeon	**61.64**	**0.6800**	**25.74**
- w/o visual feature	55.46	0.6828	23.53
- w/o textual tokens	65.73	0.6731	20.37
- w/o encoding	55.35	0.6976	24.72

3.4.2 ***Effect of multimodal instruction encoding.*** To validate the necessity to extract high-level semantics via multimodal instruction encoding, we perform three ablation studies during the training phase. Specifically, we remove the encoded visual features and textual instruction tokens, referred to as "w/o visual features" and "w/o textual tokens", respectively. We also disable the encoding process by directly inputting the masked reference tokens into LaVIT, denoted as "w/o encoding". Results on three key metrics, reported in Table 3, reveal the following insights: 1) removing the visual features reduces the performance, highlighting the importance of high-level visual semantics for understanding the reference image and enhancing personalization. 2) Excluding textual tokens improves personalization while significantly reducing semantic alignment, indicating that the model over-prioritizes user preferences when textual semantics are absent. 3) Disabling the encoding process leads to simple duplication of the reference image rather than true personalization, as evidenced by a notable drop in personalization and an increase in semantic alignment.

3.4.3 ***Effect of two-stage preference alignments.***

• **Stage-1: masked preference reconstruction.** To evaluate the impact of the first-stage masked preference reconstruction, we perform additional experiments that analyze the effect of the reference mask and explore alternative masking schemes: 1) removing the reference mask, as shown in Figure 4(c), leads to a notable performance decline, underscoring the importance of masked preference reconstruction, which allows Pigeon to effectively integrate user preferences with reference semantics for personalization. 2) Exploring alternative masking schemes for the reference tokens: "Low-score" refers to masking low-score tokens during both training and inference, while "High-score" masks high-score tokens in both phases. These schemes are compared to the dual mask scheme of Pigeon, with results presented in Figure 4(d). The significant decline in personalization suggests that masking either high-score or low-score tokens during both phases causes the model to over-focus on preference reconstruction, limiting its ability to generalize this reconstruction for broader personalization.

• **Stage-2: pairwise preference alignment.** We evaluate the effect of the second-stage pairwise preference alignment by comparing

Figure 5: Examples of generated images in sticker and movie poster scenarios, along with four user-interacted history images and one reference image.

the performance after the first and second stages of alignments, as shown in Figure 4(e). Despite a slight decline in semantic alignment, the second-stage preference alignment further enhances personalization. This demonstrates the effectiveness of DPO in aligning the generation process more closely with user preferences, ultimately resulting in more personalized image generation. Analysis of different preference reward strategies during the second-stage alignment is presented in Appendix E.

3.5 Case Study

In this section, we present two examples of Pigeon-generated images in sticker and movie poster scenarios, along with four user-interacted history images and one reference image. We compare Pigeon with two competitive baselines, TI and PMG, as shown in Figure 5. In the sticker scenario, Pigeon effectively captures the user's visual preference for Yoda and integrates it with the high-level semantics of the reference sticker, such as "drinking coffee", achieving impressive personalization and semantic alignment with the reference image. In the movie poster scenario, Pigeon-generated poster for the movie "Rise of the Planet of the Apes" showcases high semantic alignment with the reference poster by emphasizing an intense central ape figure, evoking a similar sense of power and conflict. Meanwhile, it matches the user's preference for character-centered movie posters with a dark and dramatic color palette. More examples are provided in Figure 6 and Figure 7 in Appendix F.

4 Related Work

• **Personalized Content Filtering.** Traditional filtering-based personalized content delivery approaches, such as recommender systems [2, 20, 21, 45, 56], typically rank existing content based on user interaction history and contextual information, delivering the top-ranked content. However, constrained by the limited diversity of available content, they often fall short of meeting users' diverse needs [44, 51, 54], motivating the emergence of personalized content generation across various domains.

• **Personalized Content Generation.** The rise of powerful generative models, such as DMs [28, 32], LLMs [41], and LMMs [13, 22], has sparked increasing interest in their potential for personalized content generation. Most previous work focuses on personalized text generation [15, 31, 34, 39]. For example, the LaMP benchmark [35] is developed to train and evaluate LLMs in various personalized text scenarios like personalized news headline generation and tweet paraphrasing. Further work, such as RSPG [34], studies the retrieval-augmented solutions to personalize LLM outputs, while PER-PCS [39] introduces a parameter-sharing framework to enable more efficient and fine-grained personalization.

In contrast, personalized image generation has received relatively less attention. Current research mainly adopts DMs and LLMs for this task: 1) DM-based methods [9, 16, 19, 37], such as TI [5] and DreamBooth [33], focus on aligning image generation with users' explicit multimodal instructions, without consideration of user implicit visual preferences. Other approaches like DiFashion [51], CG4CTR [52], and AdBooster [38], integrate user data (*e.g.,* interaction history and user features) with multimodal instructions to guide personalized fashion and product image generation. However, these methods often struggle with the noisy signals in user-interacted history images, leading to inaccurate preference modeling. 2) LLM-based PMG [36] translates images into texts for the LLM to extract user visual preferences, while the limitations of text in conveying complex visual details hinder its effectiveness. In this work, we leverage the notable visual understanding and reasoning capabilities of LMMs, along with dedicated modules, to develop the Pigeon framework that effectively handles noisy history images for accurate, tailored image generation.

• **Multimodal Content Generation.** A lot of prior studies utilize pre-trained generative models for content generation across various modalities, including image [11, 14], text [18, 53], video [12, 23], and audio [17, 25, 50]. From these works, we have witnessed the impressive capabilities of the pre-trained LMMs, such as GPT-4 [1], LLaVA [22], LaVIT [13], and Gemini [40] in instruction-following, multimodal content understanding, and generation. However, despite their success, these models primarily generate content conditioning on text prompts or other given modalities, without incorporating users' personalized information. When directly applied to personalized content generation, these models often exhibit suboptimal performance (*cf.* the empirical results of pre-trained LaVIT and LLaVA in Table 1) due to their limited understanding of user preferences. Therefore, we propose the Pigeon framework, empowering the pre-trained LMMs with personalization capabilities.

5 Conclusion and Future Work

In this work, we proposed a novel framework named Pigeon, which integrates a pre-trained LMM with specialized modules to infer implicit user preferences from noisy user history and incorporate explicit multimodal instructions for personalized image generation. To alleviate data scarcity, Pigeon adopts a two-stage preference alignment scheme with masked preference reconstruction and pairwise preference alignment, enhancing the personalization capabilities of LMMs. Both quantitative results and human evaluation highlight Pigeon's effectiveness in generating personalized images.

This work marks an initial attempt to align pre-trained LMMs with implicit user visual preferences, paving the way for several promising directions: 1) adapting Pigeon to consider evolving user preferences; 2) developing efficient strategies to manage lifelong user history for superior personalization; 3) integrating personalized content generation and filtering to construct more powerful personalized content delivery systems.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774* (2023).

[2] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *RecSys*. ACM, 1007–1014.

[3] Ádám Tibor Czapp, Mátyás Jani, Bálint Domián, and Balázs Hidasi. 2024. Dynamic Product Image Generation and Recommendation at Scale for Personalized E-commerce. In *RecSys*. ACM, 768–770.

[4] Shuqi Dai, Xichu Ma, Ye Wang, and Roger B Dannenberg. 2022. Personalised popular music generation using imitation and structure. *J New Music Res* 51, 1 (2022), 69–85.

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*. OpenReview.net.

[6] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2023. CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System. *TOIS* 42, 1, Article 14 (2023).

[7] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024. Making LLaMA SEE and Draw with SEED Tokenizer. In *ICLR*. OpenReview.net.

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

[9] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. 2024. Imagine yourself: Tuning-Free Personalized Image Generation. *arXiv:2409.13346* (2024).

[10] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*. OpenReview.net.

[11] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. 2024. Instruct-Imagen: Image generation with multi-modal instruction. In *CVPR*. IEEE, 4754–4763.

[12] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. 2024. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv:2402.03161* (2024).

[13] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. 2024. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. In *ICLR*. OpenReview.net.

[14] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *NeurIPS* 36 (2024).

[15] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. LongLaMP: A Benchmark for Personalized Long-form Text Generation. *arXiv:2407.11016* (2024).

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*. IEEE, 1931–1941.

[17] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. 2024. Efficient neural music generation. *NeurIPS* 36 (2024).

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

[19] Xiaoming Li, Xinyu Hou, and Chen Change Loy. 2024. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *CVPR*. IEEE, 2187–2196.

[20] Xinyu Lin, Wenjie Wang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Bridging Items and Language: A Transition Paradigm for Large Language Model-Based Recommendation. In *KDD*. ACM, 1816–1826.

[21] Fan Liu, Shuai Zhao, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. 2024. Cluster-Based Graph Collaborative Filtering. *TOIS* 42, 6, Article 167 (2024).

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *NeurIPS* 36 (2024).

[23] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv:2402.17177* (2024).

[24] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *ICLR*. OpenReview.net.

[25] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning Diffusion-based Text-to-Audio Generative Models through Direct Preference Optimization. In *MM*. ACM.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *TMLR* (2024).

[27] Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis. 2024. Investigating Personalization Methods in Text to Music Generation. In *ICASSP*. IEEE, 1081–1085.

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*. OpenReview.net.

[29] 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 *ICML*. PMLR, 8748–8763.

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS* 36 (2024).

[31] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *NeurIPS* 36 (2024).

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.

[33] 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 *CVPR*. IEEE, 22500–22510.

[34] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. In *SIGIR*. ACM, 752–762.

[35] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *ACL*. ACL, 7370–7392.

[36] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. PMG: Personalized Multimodal Generation with Large Language Models. In *WWW*. ACM, 3833–3843.

[37] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*. IEEE, 8543–8552.

[38] Veronika Shilova, Ludovic Dos Santos, Flavian Vasile, Gaëtan Racic, and Ugo Tanielian. 2023. AdBooster: Personalized Ad Creative Generation using Stable Diffusion Outpainting. *arXiv:2309.11507* (2023).

[39] Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts. *arXiv:2406.10471* (2024).

[40] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805* (2023).

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288* (2023).

[42] Shanu Vashishtha, Abhinav Prakash, Lalitesh Morishetti, Kaushiki Nag, Yokila Arora, Sushant Kumar, and Kannan Achan. 2024. Chaining text-to-image and large language model: A novel approach for generating personalized e-commerce banners. In *KDD*. ACM, 5825–5835.

[43] A Vaswani. 2017. Attention is all you need. *NeurIPS* (2017).

[44] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv:2304.03516* (2023).

[45] Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. 2023. Diffusion recommender model. In *SIGIR*. ACM, 832–841.

[46] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*. ACM, 165–174.

[47] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *ACSSC*. IEEE, 1398–1402.

[48] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *TOIS* 41, 1 (2023), 1–50.

[49] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. 2013. Data mining with big data. *TKDE* 26, 1 (2013), 97–107.

[50] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent

aligners. In *CVPR*. IEEE, 7151–7161.
[51] Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. 2024. Diffusion Models for Generative Outfit Recommendation. In *SIGIR*. ACM, 1350–1359.
[52] Hao Yang, Jianxin Yuan, Shuai Yang, Linhe Xu, Shuo Yuan, and Yifan Zeng. 2024. A New Creative Generation Pipeline for Click-Through Rate with Stable Diffusion Model. In *Companion WWW*. ACM, 180–189.
[53] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. 2024. Glyphcontrol: Glyph conditional control for visual text generation. *NeurIPS* 36 (2024).
[54] Cong Yu, Yang Hu, Yan Chen, and Bing Zeng. 2019. Personalized fashion design. In *ICCV*. IEEE, 9046–9055.
[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. IEEE, 586–595.
[56] Jujia Zhao, Wenjie Wang, Yiyan Xu, Teng Sun, Fuli Feng, and Tat-Seng Chua. 2024. Denoising diffusion recommender model. In *SIGIR*. ACM, 1370–1379.
[57] Zhengbang Zhu, Rongjun Qin, Junjie Huang, Xinyi Dai, Yang Yu, Yong Yu, and Weinan Zhang. 2024. Understanding or Manipulation: Rethinking Online Performance Gains of Modern Recommender Systems. *TOIS* 42, 4 (2024), 1–32.

A Domain Applications of Pigeon

Pigeon empowers LMMs with the capability to generate personalized images, which is applicable in various scenarios such as personalized stickers on social media platforms like Twitter and personalized movie posters on platforms like Netflix (see demonstration in Section 3). Beyond these, we showcase the potential of Pigeon in other representative domains.

• **E-commerce: personalized product images.** In e-commerce, compelling product images are crucial for drawing attention and driving purchase decisions. Pigeon can analyze user visual preferences from their behaviors to generate personalized product images that match individual tastes in personalized display style and background, delivering a more customized shopping experience.

• **Advertising: personalized advertisements.** Pigeon can assist advertisers in creating highly customized and context-aware multimodal advertisements based on user behaviors, which are more likely to improve user engagement and conversion rates.

• **Fashion: personalized fashion designs.** Pigeon can infer users' fashion preferences to generate personalized designs for fashion products like clothing, shoes, and jewelry. Besides, both fashion designers and users can provide their preferred fashion images with explicit multimodal instructions for Pigeon to customize designs, fostering an interactive and collaborative design experience.

B Datasets

For the sticker scenario, we exclude low-quality themes or those with fewer than six stickers, constructing user interaction sequences where each user interacts with a single theme. For the movie scenario, we adopt the small version of the dataset, retaining user interactions with ratings of four or higher, sorted by the timestamps. We apply a sliding window of six interactions, moving one step at a time to create data samples for each user in both scenarios. Each sample treats the first five interactions as the user history images and the last as the target image. We split the samples into training, validation, and testing sets with a ratio of 8:1:1. In the sticker testing set, we randomly select one sticker from a different theme than the user history as the reference image, while in the movie poster scenario, the target image is used as the reference. Dataset statistics are summarized in Table 4, where each "sample" consists of user-interacted history images and one reference image.

Table 4: Overview of dataset statistics.

	#Users	#Items	#Samples
Stickers	725	14,345	10,719
Movie posters	594	6,961	31,058

C Implementation Details of Pigeon

In Pigeon, the learning rate is set to $1e^{-5}$ and $5e^{-6}$ for the first and second stage alignment, respectively. The history mask ratio α_h is fixed at 0.2. During inference, we select the optimal reference mask ratio $\alpha_r \in \{0.0, 0.1, \ldots, 1.0\}$ for each reference image by averaging the history CIS and reference CS.

All experiments are conducted using a single NVIDIA-A100 GPU. As shown in Table 5, while the total number of parameters in Pigeon is substantial, the trainable components represent only a small fraction, leading to relatively low computational overhead. The training process costs about 20 hours and 5 hours for the first and second stage alignment, respectively. For inference, each sample takes about 7 seconds for LaVIT and 9 seconds for SDXL.

Table 5: Model parameters and trainable ratio of Pigeon.

		Parameters	Trainable Ratio
Total: LaVIT + SDXL		11,468,249,325	-
Trainable	**Mask Generator**	100,726,784	0.878%
	Adapter Layer	3,150,336	0.027%
	LoRA	4,194,304	0.037%

D Human Evaluation

In the human evaluation, we conduct binary-choice tests across both sticker and movie scenarios, each consisting of 50 cases. To ensure diversity, the sticker cases involve 49 distinct themes and 6 different emotion labels, including anger, fear, happiness, neutral, sadness, and surprise. On the other hand, the movie cases include 21 different movie genres, such as action, animation, horror, romance, and sci-fi. As shown in Table 2, we perform a total of 10 binary tests, comparing Pigeon with Grd, TI, and PMG. For each test, 50 participants were recruited for evaluation. The qualitative results validate the effectiveness of Pigeon in personalized image generation across these diverse scenarios and contexts.

E Analysis of Preference Reward Strategy

To evaluate the impact of different preference reward strategies during the second-stage alignment, we conduct additional experiments to compare the strategy outlined in Eq.(8) with the approach proposed in [36]. From the results shown in Table 6, we observe the following key findings:

- The second-stage preference alignment significantly improves personalization, despite a slight decline in semantic alignment. This highlights the effectiveness of DPO in better aligning the generation process with user preferences.
- The reward strategy employed by Pigeon achieves superior effectiveness in enhancing personalization, with only a minor and acceptable trade-off in semantic alignment. This indicates that the strategy successfully prioritizes user-specific preferences while maintaining a reasonable degree of semantic consistency.

Table 6: Effect of different preference reward strategies during the second-stage alignment.

#Sticker	Personalization					Semantic Alignment		
Pigeon	CS↑	CIS↑	DIS↑	LPIPS↓	MS-SSIM↑	CS↑	CIS↑	DIS↑
Stage-1	22.03	61.64	57.26	**0.6800**	0.1467	**25.74**	**50.66**	**48.34**
- Stage-2 [36]	**24.15**	64.85	59.43	0.6803	0.1398	25.57	50.16	47.92
- Stage-2-ours	23.69	**67.65**	**62.23**	0.6814	**0.1568**	21.10	47.44	45.44
#Movie	**Personalization**					**Semantic Alignment**		
Pigeon	**CS↑**	**CIS↑**	**DIS↑**	**LPIPS↓**	**MS-SSIM↑**	**CS↑**	**CIS↑**	**DIS↑**
Stage-1	15.19	37.42	17.70	**0.7496**	**0.0493**	27.30	47.79	41.35
- Stage-2 [36]	15.28	38.67	19.03	0.7509	0.0454	**27.80**	49.34	43.03
- Stage-2-ours	**15.41**	**40.16**	21.29	0.7508	0.0464	26.45	**49.66**	**44.07**

F Case Study

We present five additional Pigeon-generated examples for both sticker and movie poster scenarios, respectively, along with several user-interacted history images and one reference image.

Sticker scenario. As illustrated in Figure 6, Pigeon effectively captures users' visual preferences for character figures and styles in stickers, and combines these preferences with the high-level semantics of the reference sticker to generate personalized stickers. The generated stickers exhibit high semantic alignment with the reference image, including the conveyed emotions, facial expressions, character actions, and elements like hearts.

Figure 6: Examples of generated stickers, along with user-interacted history stickers and one reference sticker.

Movie poster scenario. As depicted in Figure 7, each user shows a distinct set of visual preferences, ranging from action and sci-fi to historical drama and crime thrillers. Pigeon-generated posters effectively mirror these preferences through character-centered designs, dynamic compositions, and color palettes that align with each user's unique taste. By tailoring its designs to the emotional tone, genre, and thematic focus of the reference posters, Pigeon creates personalized posters that strongly resonate with individual users' past interactions and preferences. For instance, the user in the first row shows a strong preference for action-heavy, explosive films with a focus on dramatic visuals and blockbuster-style presentations. Pigeon matches the user's love for explosive visuals, with characters taking center stage and environments filled with dynamic elements like fire, destruction, and warfare.

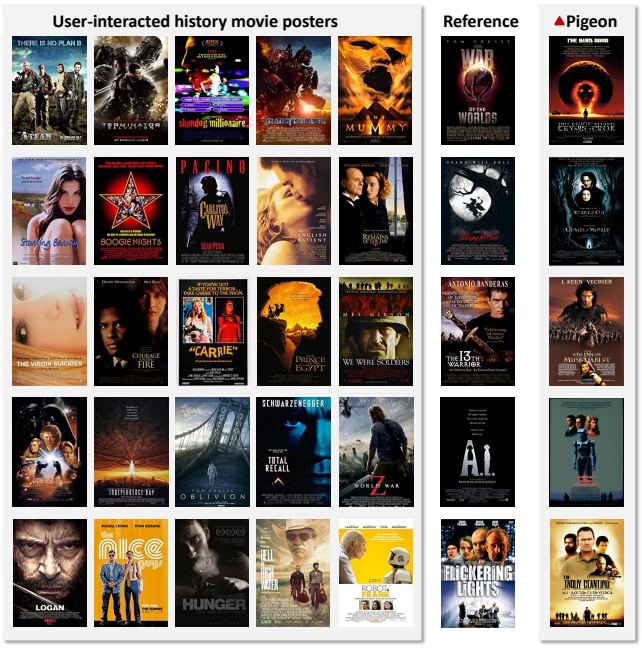


Figure 7: Examples of generated movie posters, along with user-interacted history posters and one reference poster.