ImageGem: In-the-wild Generative Image Interaction Dataset for Generative Model Personalization

Yuanhe Guo[1*], Linxi Xie[1*], Zhuoran Chen[1], Kangrui Yu[1],
Ryan Po[2], Guandao Yang[2], Gordon Wetzstein[2], Hongyi Wen[1†]
[1]NYU   [2]Stanford
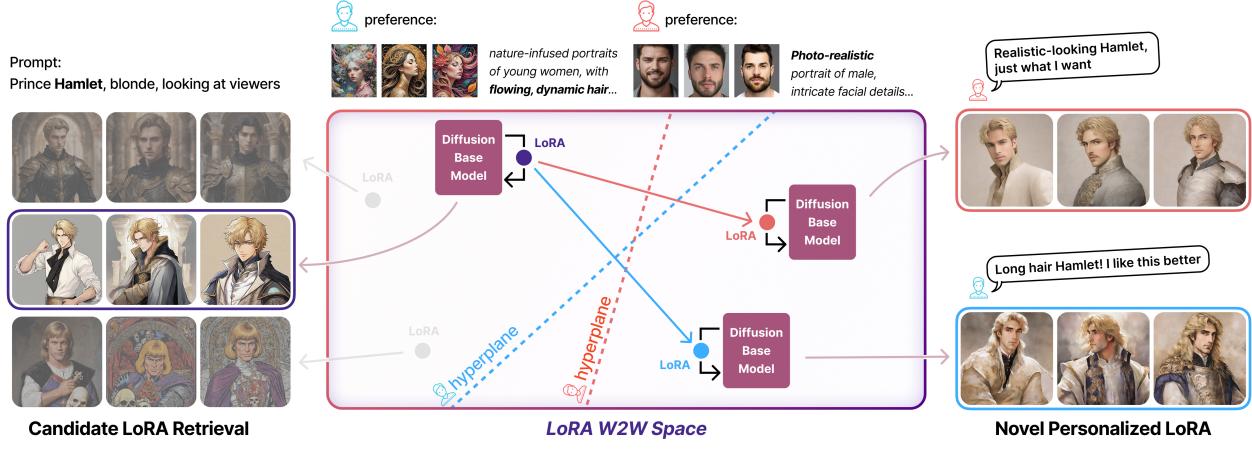https://maps-research.github.io/imagegem-iccv2025/

Figure 1. Our proposed ImageGem dataset and its applications. The left side illustrates image and generative model retrieval. On the right, we demonstrate a novel task of generative model personalization through LoRA weights-to-weights (W2W) space construction.

Abstract

We introduce ImageGem, a dataset for studying generative models that understand fine-grained individual preferences. We posit that a key challenge hindering the development of such a generative model is the lack of in-the-wild and fine-grained user preference annotations. Our dataset features real-world interaction data from 57K users, who collectively have built 242K customized LoRAs, written 3M text prompts, and created 5M generated images. With user preference annotations from our dataset, we were able to train better preference alignment models. In addition, leveraging individual user preference, we investigated the performance of retrieval models and a vision-language model on personalized image retrieval and generative model recommendation. Finally, we propose an end-to-end framework for editing customized diffusion models in a latent weight space to align with individual user preferences. Our results demonstrate that the ImageGem dataset enables, for the first time, a new paradigm for generative model personalization.

1. Introduction

A thousand Hamlets in a thousand people's eyes

Recent advances in text-to-image models [28] have empowered users to generate a portrait of Hamlet, Shakespeare's famous character, from merely a short and underspecified description like "A portrait of Hamlet". However, each user has their imagination about the portrait of Prince Hamlet. Can generative models capture and produce the version that aligns with individual preference?

Current progress toward building personalized textconditioned generative models is mainly driven by data availability. For example, with the presence of datasets that contain different person or object identities, prior works are able to customize diffusion model to generate these userspecified concepts. [26, 29, 30, 36]. These works, however, do not address under-specified inputs that require reasoning about individual preference, such as generating an image of "my favorite dog." Similarly, enabled by datasets with user preference annotations [19, 21, 34, 37], many works are able to create text-to-images models that align with hu-

man preference [11, 32, 38]. These methods, however, focus on aggregated preference, such as generating an image that the general population will favor. How to create a generative model aligned with personal preferences remains under-explored due to the lack of large-scale and fine-grained user preference annotations. Existing efforts toward this end are thus limited to zero-shot approaches [31], which usually require user input during inference time. Such zero-shot approaches find it difficult to leverage similarity among users. As a result, they can be expensive and limited to a few predetermined dimensions of individual preference.

Motivated by the gap between aggregated preference modeling and personalization at the individual level, we propose *ImageGem* dataset, the first large-scale dataset that contains diverse user behaviors from real-world users employing their generative models. We sourced our data from Civitai [1], one of the most popular AIGC platforms where users create and publicly share both their customized models and generated images. In addition to the content filter provided by Civitai, we further evaluated the safety of images and prompts, labeling them accordingly to ensure a reliable dataset for downstream tasks.

We setup a few evaluations on the quality of aggregated and individual preference data from our dataset. Specifically, we train SD1.5 [28] with DiffusionDPO [32] on aggregated preference data and demonstrate improved image quality over a widely-used dataset for preference alignment [19]. We further leverage individual preference data to examine the quality of personalized image retrieval and generative model recommendations with retrieval-based models. In addition to retrieval-based models, we leverage a vision-language model (VLM) [1] for user preference captioning and ranking by prompting the VLM to generate structured descriptions. We demonstrate that integrating VLM enhances ranking interpretability.

Our dataset enables a new application of *Generative Model Personalization*, where customized diffusion models (e.g. LoRA [16]) are created to align with individual preference. We leverage a subset of user-created LoRAs from ImageGem to construct a latent weight space. By capturing individual preferences from historical user-generated images, we learn editing directions in this latent space, enabling progressive model adaptation to user preferences. We summarize our contributions as the following:

- We present the first large-scale dataset consisting of user fine-grained preferences towards generative models and images. Our dataset consists of metadata such as prompts, images, and user feedback. We apply safety checks on metadata and ensure the diversity of our dataset.
- We evaluate the quality of our curated dataset and showcase its preference labels for several downstream applications, including general preference alignment, personal-

ized image retrieval, and generative recommendation.
- We propose an end-to-end framework for editing customized diffusion models toward individual user preference, demonstrating a new application in curating personalized generative models.

2. Related Work

2.1. Dataset for Preference Alignment

Several works investigate how to train better diffusion models that align with human preference [19, 21, 37]. For example, Pick-a-Pic [19] collected ratings on image pairs from about 6K users and demonstrated their PickScore achieved state-of-the-art alignment with human judgments on image generations. RichHF-18K [21] exemplifies the heterogeneous user feedback such as predicted scores and heatmaps can be leveraged to improve RLHF. FiVA [36] curated a fine-grained visual attributes dataset of 1 million generated images with detailed annotations.
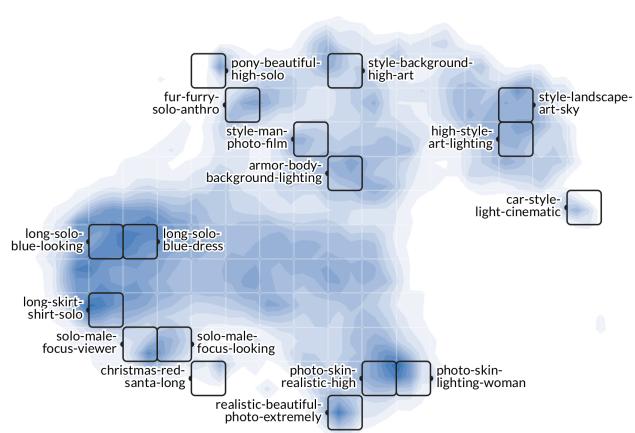
Our dataset connects to this line of research on developing human preference datasets to improve image generation models, but with several key differences. Our dataset contains observational data from a cohort of in-the-wild users, e.g., interaction logs between 57K users and 242K generative models from 2023/09 to 2025/01. Generative models in our dataset are up-to-date and customized by users, reflecting real-world usage and preferences. Moreover, our dataset captures individual-level user preference as opposed to aggregated-level preference.

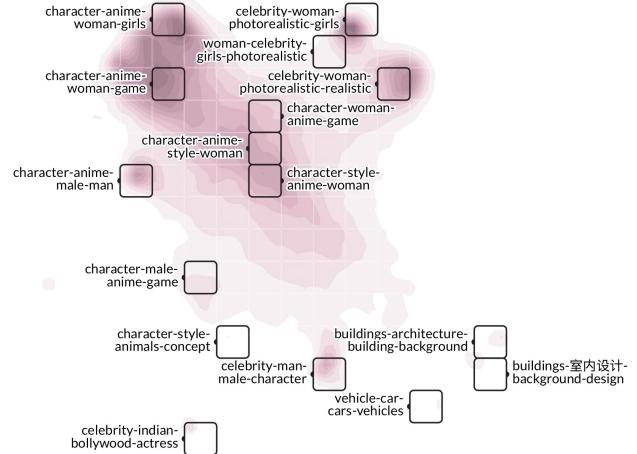### 2.2. Personalizing Generative Models

A group of work personalizes generative models through capturing prompt instructions [3] or other diverse modality of user inputs [40]. Another branch widely used approaches enable efficient fine-tuning given personal preference, starting with textual inversion [12] and DreamBooth [29]. Further works train adapters for preserving specific identities such as human faces [33, 41]. Weights2Weights [10] demonstrates the feasibility of customizing models in a latent weight space towards certain attributes by creating a dataset of 60k LoRAs from human faces. ViPer [31] proposes a zero-shot user preference learning framework via a two-stage process: first capturing the user's general preferences through their comments, and then a vision-language model extracts structured preferences and is used for editing the prompts for text-to-image generation.

Our dataset augments previous work by providing fine-grained user preference data and user-customized diffusion models in large scale. With this dataset, we provide a new perspective to personalizing visual generative models by editing pre-trained models to individual user preference captured from unspecified interaction data, which is a more challenging and realistic task.

(a) Contour plot of 1M images sampled from our dataset, visualized using UMAP to reduce the dimensionality of their CLIP embeddings.

(b) Contour plot of LoRA model checkpoints, where each LoRA is represented by the mean embedding of its corresponding images.

Figure 2. WizMap [35]-based visualization for our ImageGem dataset, divided into two parts. The left panel shows a UMAP embedding of 1M images sampled from the dataset, while the right panel illustrates a contour plot of LoRA model checkpoints. Both visualizations use grid tiles to display key words extracted from image prompts or model tags.

3. The ImageGem Dataset

We constructed the **ImageGem dataset** by sourcing data from Civitai, an open-source platform for sharing fine-tuned model weights and images. This dataset captures real-world interactions between users and image generation models, offering a unique opportunity to study personalized preferences in diffusion-based systems. Below, we detail its construction, curation, and key characteristics.

3.1. Metadata and Relational Database

Civitai serves as a comprehensive source of user-generated content, featuring personalized diffusion models, images, and associated metadata. To build ImageGem, we leveraged Civitai's public API, which provides information about licenses and NSFW (Not Suitable For Work) classifications. Prior to data collection, we obtained institutional IRB approval to ensure ethical compliance.

Our dataset captures three core components: LoRA models (light-weight adapters for fine-tuning diffusion models), images generated using these LoRAs, and users who upload images and models. To enable flexible querying and analysis, we established a ternary relationship between images, LoRAs, and users, allowing for efficient retrieval of user-specific preferences and model interactions.

3.2. Safety Check

Given the open nature of Civitai, we implemented rigorous safety checks to ensure the dataset's reliability for downstream tasks. While Civitai categorizes images based on NSFW levels [2], prompts and user-labeled LoRA tags lack

---

[2] https://education.civitai.com/civitais-guide-to-content-levels/

explicit ratings. To address this, we used Detoxify [14], a multilingual toxic text classifier, to estimate NSFW probabilities for prompts. A detailed distribution of NSFW probability in each aspect is shown in Appendix Fig. 5. Images whose prompt's unsafe probabilities above 0.2 were excluded. These steps ensured that the final dataset balances diversity with safety, making it suitable for research on preference learning and model personalization.

3.3. Dataset Overview

Tab. 1 shows the essential numbers in our dataset before and after safety filtering. All 4,916,134 filtered images, which have associated prompts recorded in the metadata and are accessible from the Civitai website as of March, 2025. In the following paper, we focus on analyzing the safety checked dataset.

Images. We computed the CLIP [27] embedding for all images, and used UMAP [23] to reduce the dimension to 2D for visualization. The distribution shown in Fig. 2a illustrates the wild variety of topics covered in our dataset. Recent methods, such as Compel [5], encode long prompts within 77 tokens with CLIP [27] using prompt weighting techniques, which makes token counting less accurate. As a result, we count the number of words in each prompt, with the average word count being 48.5. Additionally, we observe some prompts with exceptionally large word counts. The distribution of prompts with word counts exceeding 200 is shown in Appendix Fig. 6. Image feedback are captured by various emojis, including thumbs-up, heart, laugh, and cry. The distribution of each type of user feedback is shown in Appendix Fig. 7.

	Images	Unique Prompts	LoRA Model Checkpoints	Unique Model Tags	Total Users	Model Uploaders *	Avg Images Per Uploader †	Avg Models Per Uploader †	Avg Images Per Model ‡
Raw	5,658,107	2,975,943	242,889	105,788	57,245	19,003	49	12	62
Filtered	4,916,134	2,895,364	242,118	97,434		18,889	48	13	54

Table 1. Statistics of our ImageGem dataset. * While every user generated at least one image, not all users uploaded LoRAs. † Excluded highest-uploader counts for unbiased averages. ‡ Many-to-many image-model relationships may cause image double-counting.

LoRA Models. In our dataset, LoRA models are fine-tuned based on 37 different base model structures, with 41% being SD 1.5 [28], 31% Pony [3], 12% SDXL 1.0 [25] and 9% Flux.1 [20]. Fig. 2b shows the distribution of LoRA models. We represent each LoRA by averaging the embeddings of its image embeddings, and the text labels are tags labeled by model uploaders.

User Interactions. Our dataset includes two types of user interaction data: (1) *Individual level user-model interactions*, which capture user-specific image generation configurations (e.g., prompts) to analyze individual-level preference, usage patterns, and prompting strategies across LoRA. This include $1,739,947$ in-house images created by LoRA up-loaders to showcase their model capability. The remaining $3,176,187$ images serve as historical records of user preferences, enabling tasks like image and model recommendation. (2) *Aggregate level user-image feedback*, which provide aggregated emoji feedback (e.g. like/dislike) for content filtering and benchmarking preference alignment methods.

4. Applications and Methods

We now present three applications followed by the creation of the ImageGem dataset, focusing on the aggregate-level and individual-level user interaction data accompanied by other metadata to enable various applications.

4.1. Aggregated Preference Alignment

With in-the-wild user preference annotations (e.g., likes, crys) towards prompt-image pairs from our dataset, a direct application of our dataset is to train preference alignment models. Following DiffusionDPO [32], for preference pairs $(\mathbf{c}, \mathbf{x_0^w}, \mathbf{x_0^l})$ of prompt $\mathbf{c}$, and image preference label $\mathbf{x_0^w} \succ \mathbf{x_0^l}$, the training objective is as following:

$$\max_{p_\theta} \quad \mathbb{E}_{c \sim \mathcal{D}_c, \mathbf{x}_{0:T} \sim p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}[r(\mathbf{c}, \mathbf{x}_0)] \\ - \beta \mathbb{D}_{KL}[p_\theta(\mathbf{x}_{0:T}|\mathbf{c})\|p_{ref}(\mathbf{x}_{0:T}|\mathbf{c})]. \quad (1)$$

Here, p_{ref} is a pre-trained diffusion model, and p_θ is the updated model to align with preferences with trainable parameters θ, and T being the diffusion timestep. The reward function $r(\mathbf{c}, \mathbf{x}_0)$ is defined as:

---
3 AstraliteHeart/pony-diffusion

$$r(\mathbf{c}, \mathbf{x}_0) = \mathbb{E}_{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}[R(\mathbf{c}, \mathbf{x}_{0:T})], \quad (2)$$

with $R(\mathbf{c}, \mathbf{x}_{0:T})$ being the reward on the whole chain.

Different from other large-scale datasets such as Pick-a-Pic that rely on human annotations of explicit preference over image pairs, the preference pairs in our dataset are from natural observations and are implicit. To curate high-quality preference pairs, we first cluster the prompts' CLIP embeddings in our dataset with HDBScan [22]. Within each cluster, we construct preference pairs using min-max pairing over Human Preference Score v2 [37]. Through comparisons with Pick-a-Pic on various metrics [15, 19, 37], we demonstrate that our dataset achieves improved aggregate-level preference alignment in Sec. 5.1.

4.2. Retrieval and Generative Recommendation

With abundant individual-level preference data in our dataset, we explore personalized image retrieval [6] and generative model recommendations [13]. For both image and generative model items, we formulate the recommendation task on the large item corpus size of millions following a two-stage retrieval-ranking paradigm [4], where we use collaborative filtering (CF) to retrieve a subset of top-k next interacted items given user, and then use a complex visual-language model (VLM) for ranking on that subset.

Candidate Retrieval. For large-scale image retrieval, we employ FAISS [9] for Approximate Nearest Neighbors (ANN) search, enabling efficient and scalable vector search across millions of items. To mitigate the sparsity of the training data, we initialize the ID embedding of each image by encoding each image with a pre-trained ViT [8]. We evaluate the performance of ItemKNN [7], Item2Vec [2], and a two-tower model [39] for the retrieval task.

As for candidate model retrieval, we evaluate various approaches that capture diverse facets of user preferences for generative models. UserKNN and ItemKNN compute user/item similarity from historical interaction data using cosine similarity and aggregate similar user/item ratings on the target item. SASRec [18] utilizes self-attention mechanisms to model the sequential nature of user interactions, considering order and temporal dynamics.

Generative Recommendation. We explore generative recommendation by building a VLM-based recommendation workflow that generates structured user preference descriptions as representations of user interests. We selected

Pixtral-12B [1] as our VLM due to its ability to process multiple images and texts within one prompt, making it convenient for multi-item captioning and ranking tasks.

Our workflow consists of two stages: item captioning and ranking. In the captioning stage, given the user's historical preferences, we prompt the VLM to generate textual representations of user's visual preference profile, denoted as q_i. For image items, the VLM extracts common features from user-generated images. For model items, we select the most-liked user-provided prompt for each model, and prompt the VLM to summarize its key attributes.

In the ranking stage, we construct a prompt that instructs the VLM to compare q_i with each item in C_i and generate a similarity score along with an explanation, where C_i denotes the CF candidate set from the retrieval stage. However, VLM ranking exhibits instability, as item scores may vary across different inference requests or ranking orders. To mitigate this, we designed templates with detailed ranking criteria and adopted a randomized scoring strategy under VLM input constraints. Details in VLM prompting and our scoring strategy are included in Appendix B.

4.3. Generative Model Personalization

Given the limitations of the retrieval-based recommendation paradigm, we propose a new framework of *Generative Model Personalization*, which generates personalized LoRA models aligned with individual preferences. Building on previous work that explores a LoRA Weights2Weights (W2W) space for identity editing [10], we adapt this method to model personalization. Specifically, we use a set of user-created LoRAs to construct a latent LoRA weight space and learn editing directions that reflect user preferences. These directions can then be used to transform any LoRA model within the space, producing a personalized version without requiring re-training.

To create a W2W space, we first need to reduce and standardize the set of user-created LoRAs, we apply singular value decomposition (SVD) to each LoRA weight matrix, and retain only the top-1 component. We then flatten and concatenate the reduced matrices from all layers to obtain a vector representation $\theta_i \in \mathbb{R}^d$ for each LoRA. This yields a dataset $D = \{\theta_1, \theta_2, \ldots, \theta_N\}$, where each point represents a distinct preference of an individual. To reduce dimensionality and identify meaningful subspaces, we applied Principal Component Analysis (PCA) on the dataset, retaining the top m principal components. This process established a basis of vectors $\{w_1, w_2, \ldots, w_m\}$, where each basis vector inherently encodes user preference, ensuring that all modifications remain within the user preference space.

To generate personalized LoRAs, we sought a direction $v \in \mathbb{R}^d$ in the weight space that captures individual preference. Using binary labels for each user-model pair (e.g. preferred/not preferred) obtained from our dataset, we trained

linear classifiers with model weights as input features. The hyperplane determined by the classifier separates the models according to whether a target user likes it or not, and the normal vector v to this hyperplane serves as the traversal direction. Given a model weight θ, tuning is achieved by moving orthogonally along the direction v. The edited weights are calculated as $\theta_{\text{edit}} = \theta + \alpha v$, where α is a scalar controlling the strength of the tuning operation. This adjustment modifies the model to approximate individual preferences while preserving other features.

5. Experiments

5.1. Aggregated Preference Alignment

We fine-tuned Stable Diffusion 1.5 (SD1.5) [28], using three subsets sampled by specific key words shown in Tab. 2, from both our ImageGem and pick-a-pic [19]. We use the original SD1.5 checkpoint, as well as the checkpoints fine-tuned with pick-a-pic subsets for baseline comparison. All checkpoints were trained with $4\times$A100 GPUs, with batch size 1 and gradient accumulation 128 for 2000 steps. The remaining hyperparameters were configured as described in DiffusionDPO [32]. For evaluation, we sampled 200 Out-of-Distribution (OOD) prompts from DiffusionDB [34] per topic and generated 600 images per checkpoint using three random seeds.

As shown in Tab. 4 and Fig. 3, model checkpoints fine-tuned with subsets sampled from our datasets outperform those from Pick-a-Pic in all three topics. For the scenery topic, as we scale up the subset, improvements in all metrics are observed, but the CLIP score remains lower than the original SD1.5. We speculate that the DiffusionDPO train-

Dataset (Subset)	Key Words	#Pairs
Pick-a-pic Cars	"cars", "car", "vehicle", "vehicles"	13,436
ImageGem(Ours) Cars Small	"cars", "car", "vehicle", "vehicles"	13,436
ImageGem(Ours) Cars Large	"cars", "car", "vehicle", "vehicles"	27,837
ImageGem(Ours) Dogs	"dog", "dogs", "puppy", "puppies"	8,764
Pick-a-pic Dogs	"dog", "dogs", "puppy", "puppies"	10,184
Pick-a-pic Scenery	"scenery", "landscape"	12,024
ImageGem(Ours) Scenery Small	"scenery", "landscape"	9,498
ImageGem(Ours) Scenery Large	"scenery", "landscape"	85,473

Table 2. Subsets of three topics sampled from Our dataset and Pick-a-Pic at comparable scales, with images filtered by key words and excluding human character-related prompts. For cars and scenery, we found significantly more image pairs than Pick-a-Pic, so we split it into a small set and a large set for ablation study.

Dataset (Subset)	#Users	#Items	#Interactions	#Avg.Seq	Sparsity
Images - 1M	15,917	1,002,796	1,002,796	63.00	99.99%
Models - 200K	10,364	53,590	205,160	19.80	99.96%

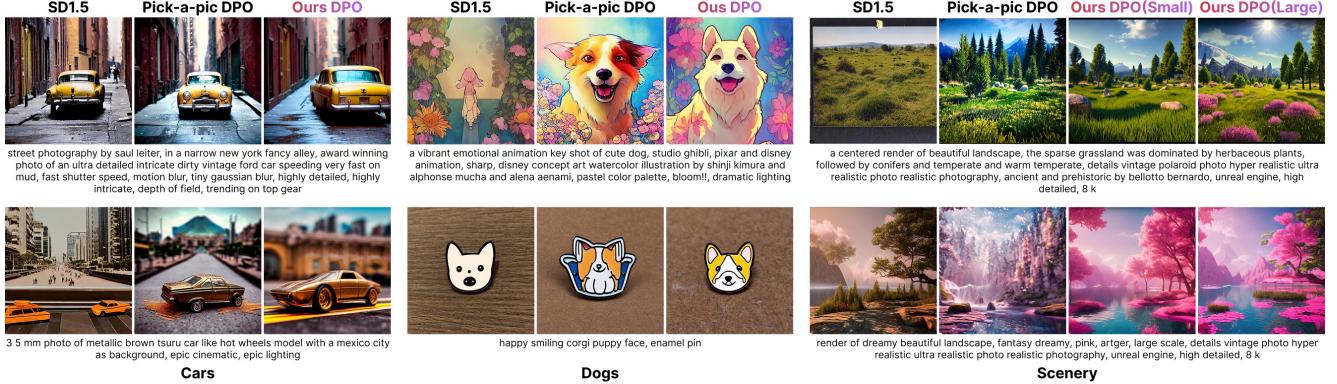Table 3. Dataset overview for the image retrieval and generative recommendation tasks.

Figure 3. Qualitative DiffusionDPO results comparison of images generated with OOD prompts in three topics sampled from DiffusionDB. For each prompt, random seed and all other hyperparameters are kept the same. Zoom in for the best view.

Dataset (Subset)	Pick Score ↑	HPSv2 ↑	CLIP Score ↑
Original SD1.5	0.1977	0.2637	0.3581
Pick-a-pic Cars	0.1993	0.2690	0.3607
ImageGem(Ours) Cars Small	0.2004	**0.2741**	**0.3745**
ImageGem(Ours) Cars Large	**0.2007**	0.2738	0.3710
Original SD1.5	0.2010	0.2646	0.3560
Pick-a-pic Dogs	0.2058	0.2739	0.3617
ImageGem(Ours) Dogs	**0.2069**	**0.2789**	**0.3683**
Original SD1.5	0.1954	0.2640	**0.3446**
Pick-a-pic Scenery	0.1936	0.2676	0.3289
ImageGem(Ours) Scenery Small	0.1949	0.2730	0.3403
ImageGem(Ours) Scenery Large	**0.1961**	**0.2747**	0.3427

Table 4. Quantitative DiffusionDPO results comparing average scores: Pick Score [19] and HPSv2 [37] for human preference alignment, and CLIP Score [15] for image-prompt alignment.

Model	Rec@10000 ↑	Rec@5000 ↑	Rec@1000 ↑	Rec@100 ↑
ItemKNN	0.4705	0.4190	0.3298	0.2342
Item2Vec	0.5032	0.4425	0.3431	0.2399
Two-Tower	**0.5157**	**0.4479**	**0.3501**	**0.2402**

Table 5. Comparisons of retrieval performance on *Images-1M*. "Rec@k" denotes Recall at rank k.

Model	Rec@100 ↑	Rec@50 ↑	Rec@10 ↑	NDCG@10 ↑
ItemKNN	0.1282	0.1036	0.0773	0.057
UserKNN	0.232	0.1818	0.1023	0.0705
SASRec	**0.2845**	**0.2451**	**0.1839**	**0.1239**

Table 6. Transposed comparison of ranking performance on *Models-200K*. "Rec@k" denotes Recall at rank k.

ing objective tends to prioritize models following human preference over prompt alignment.

5.2. Retrieval and Generative Recommendation

We sample another data subset for recommendation experiments, whose overall statistics are shown in Tab. 3. For model recommendation, as user may interact with the same LoRA models multiple times, we select the last timestamp of the user interaction in our dataset. We filter out users with less than 3 interactions for both image and model recommendation to eliminate cold-start scenarios. For evaluation, we use leave-one-last [24] with Recall@k and NDCG@k as retrieval and ranking metrics [17], where Recall@k measures how often relevant items appear in the top-k list, and NDCG@k considers both presence and position, rewarding relevant items with higher ranking to assess ranking quality.

5.2.1. Image and Model Retrieval

For image retrieval, by explicitly modeling user interest as an embedding vector, the two-tower model yields the highest performance (Tab. 5). For generative model retrieval, by introducing the structure of self-attention over user se-

quences, SASRec successfully captured temporal information in how each user's interest evolves, significantly outperforming traditional collaborative filtering methods such as ItemKNN and UserKNN (Tab. 6). These results serve as baseline performances on our dataset, leaving the study of more sophisticated retrieval models for future work.

5.2.2. Generative Recommendation

We randomly selected 20 users to test the feasibility of VLM-based ranking for generative recommendation. Each user's test item, denoted as x_i, appears in the top-10 retrieved list. To construct the VLM input for each user, we retain their latest $H = 5$ historical interactions and the top $M = 10$ retrieved items.

From Tab. 7, we observe that VLM shows promising potential in capturing user preference and items ranking in recommendation systems. In rankings of both image recommendation and model recommendation, VLM outperforms ItemKNN and SASRec in ranking quality. Compared to these traditional embedding-based methods, VLM provides human-readable and explainable rankings by generating textual justifications for its scores (Appendix B.2).

Method	Avg Rank ↓	Rank Std ↓	Rec@5 ↑	NDCG@5 ↑
Image Recommendation				
ItemKNN	2.9000	3.1439	0.7500	**0.7065**
SASRec	3.4500	2.1145	0.8500	0.5494
VLM	**2.4000**	**1.5355**	**0.9500**	0.6745
Model Recommendation				
ItemKNN	7.9500	3.8179	0.2500	0.1509
SASRec	5.3500	**2.7004**	0.5000	0.2795
VLM	**3.9444**	2.7965	**0.7222**	**0.4981**

Table 7. Ranking performance of ItemKNN, SASRec, and VLM on image and model recommendation tasks.

These results demonstrate that the captioning of user preferences by VLM effectively guides the ranking, improving both the ranking performance and the interpretability in recommendation systems. Furthermore, the success of VLM-based ranking highlights that structured textual representations can effectively capture user intent, which may further inspire advancements in generative recommendation tasks.

5.3. Generative Model Personalization

5.3.1. Aggregated Preference Editing

To assess the effectiveness of LoRA editing in the W2W space constructed with user-created LoRAs, we conducted a study to learn an editing direction from *anime* to *realistic* style, denoted as *ani-real*, within the W2W space. This serves as a preliminary study to validate the W2W pipeline under a clear semantic shift before applying it to personalized preference alignment.

Learning Tuning Direction. We curated 23K SDXL-based LoRA models with diverse visual styles. Noting a predominance of human-focused styles in the model metadata, we chose to focus on learning an edit direction from *anime* to *realistic* within the human figure domain. Initially, we used the "tags" field for binary labeling, but this approach proved noisy, as models tagged *anime* often lacked anime characteristics in their sample outputs.

To address this issue, we employed CLIP [15] to compute the similarity between the models' example images and textual descriptions of the target styles. To build a reliable W2W space, we filtered models with a *person* CLIP score ≥ 0.2. We then computed *anime* and *realistic* CLIP scores, excluding models with high values in both to avoid ambiguity. Among the rest, those with a *realistic* CLIP score ≥ 0.26 were labeled 1, and those with an *anime* score ≥ 0.24 were labeled -1, ensuring a clear editing direction from *anime* to *realistic* within the human figure domain.

Limiting Input Size. Applying Principal Component Analysis (PCA) to the weight space requires constraints on input matrix size. Given the variability in publicly generated LoRA models, we experiment with two alternative reduction strategies: (i) applying singular value decomposi-
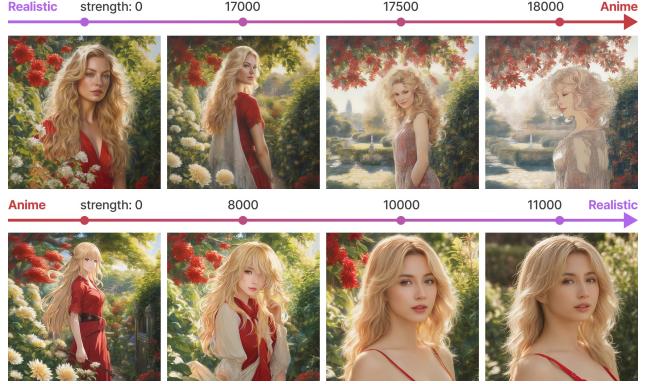

Figure 4. Perform tuning along the direction *real-ani* (top) and direction *ani-real* (bottom). The first column displays images generated by the unedited SDXL base model, while the subsequent columns show images generated by progressively edited models with increasing tuning strength. Each row shares the same generation seed for consistency.

tion (SVD), and (ii) selecting only LoRAs of a fixed rank and extracting specific layers.

Unlike the original W2W framework, which used self-trained rank-1 LoRAs [10], Civitai LoRAs vary in rank and structure. User-created LoRAs are often high-rank and vary widely in rank. To address this, we first experiment with input standardization by applying singular value decomposition (SVD) to each LoRA and retaining only the top-1 singular component (Appendix C.1). This significantly reduces weight size and ensures consistent dimensions across models, regardless of their original rank. As an alternative, we also experiment with selecting specific layers from fixed-rank LoRAs. We filtered for models with rank 16, yielding 857 LoRAs. To further reduce weight size, we selected only feed-forward (FF) and attention value (attn_v) layers based on their significant impact on the base model (Appendix C.2). Since FF layers contain three times more parameters than attn_v layers, we explored both to balance efficiency and effectiveness.

Results. Our results show that the SVD-based strategy yields the most robust transformations, enabling smooth and coherent edits in both the *ani-real* and *real-ani* directions (Figure 4). In contrast, the W2W space constructed from attn_v layers performs well primarily in the *ani-real* direction but fails to generalize to the reverse (Appendix C.3). We also observe that using only FF layers leads to poor performance in both directions, suggesting that FF may not capture semantically aligned features necessary for effective editing. Since SVD operates on LoRAs of arbitrary rank and supports bidirectional, model-level editing with stronger consistency, we adopt it as primary approach.

5.3.2. Individual Preference Learning

Building upon the *ani-real* transformation, we extend our approach to learn personalized editing directions within the W2W space in the human figure domain.

Preference Labeling. To capture individual preferences for user P_i, we compute CLIP embeddings for all their generated images, then apply HDBScan clustering to identify a representative preference cluster. To describe the user's stylistic preference, we select the top-9 images closest to the cluster mean and use a Vision-Language Model (VLM) to generate textual descriptions of their common features. Following the approach in the *ani-real* experiment, we compute CLIP similarity between LoRA models' example images and these descriptions. Models with higher similarity score are labeled 1, and those with lower similarity score are labeled -1, allowing us to learn a unique hyperplane to separate preferred and non-preferred models for each user.

Multi-Direction LoRA Editing. We first demonstrate the effectiveness of learned preference directions by editing a single LoRA model M_0 that initially lies in the "not-preferred" region for two different users. Let $\vec{d_1}$ and $\vec{d_2}$ represent the learned editing directions for each user. We traverse the W2W space by updating along each direction:$M_1 = M_0 + \lambda_1 \vec{d_1}, M_2 = M_0 + \lambda_2 \vec{d_2}$.

Multi-User Preference Alignment. To generalize preference learning, we select three users and construct three distinct preference directions $\left\{ \vec{d_i} \right\}_{i=1}^{3}$. For each direction, we choose two initial LoRA models M_{i1} and M_{i2}, neither initially aligned with the respective user's preference. We then update these models along their corresponding preference directions:$M'_{ij} = M_{ij} + \lambda \vec{d_i}, j \in \{1, 2\}$, where M'_{ij} represents the preference-aligned model. For evaluation, we generate two images per model: before editing (M_{ij}) and after editing (M'_{ij}). We rank these images using CLIP and VLM: (1) We compute CLIP similarity between each image and the user preference cluster mean, ranking images by similarity; (2) We prompt VLM to compare each image to the user's top-9 preferred images and rank by similarity.

Results. As shown in Figure 1, the initial LoRA model M_0 lies in the "not-preferred" region for both users P_1 and P_2. By traversing the W2W space along their preference directions, we obtain two modified models, M_1 and M_2, that generate images better aligned with each user's stylistic preference. Similarly, Figure 14 (Appendix C.5) demonstrates preference alignment for three users, where initial misaligned models were adjusted along learned directions. Both CLIP and VLM rankings confirm improved alignment in images generated by the adjusted LoRA models.

6. Discussion

Given the promising results from the three applications, we discuss several limitations and future work directions.

Preference Data for DPO. We curated preference sets for DPO based on HPS [37] within each semantic cluster. Future experiments could explore ways to leverage the implicit feedback data from user interactions available in our dataset. Additionally, a possible extension is to conduct human preference alignment for larger, up-to-date diffusion models, such as Flux [20], leveraging the entire dataset.

Generative Model Retrieval and Personalization. We evaluated classical models across several retrieval and ranking tasks based on user interaction data, highlighting the space for improvement on generative model retrieval. Our dataset thus provides a testbed to further study this new task formulation in large-scale, where generative models are treated as "items" to be retrieved, and abundant user implicit feedback on prompts, tags and images are associated with these models. We also presented a first look into the generative model personalization paradigm by directly editing a pre-trained LoRA according to user preference data. We demonstrated the promise of our approach from different image domains, but future work can explore how to generate models that align with various types of implicit user preference across multiple domains and data modalities.

Constraint of PCA-based Weight Space. The reliance on PCA restricts model selection to low-rank (e.g., rank 8, rank 16), limiting the diversity of the models. Consequently, the available models within a given domain are constrained, which might lead to a lack of alignment with certain users' preferences. The limited number of models prohibits effective learning of W2W space for less popular domains (e.g., scenery). With a more diverse set of models, it would be possible to learn a wider range of meaningful directions in W2W space. Future work could explore alternative methods for learning LoRA weight spaces.

7. Conclusion

We propose ImageGem, a large-scale dataset consisting of in-the-wild user interactions with generative models and images. We show that our dataset empowers the study of various tasks related to preference alignment and personalization with generative models. We demonstrate for the first time a generative model personalization paradigm by customizing diffusion models in a latent weight space aligned with individual user preference. Our dataset opens a few new research directions on generative models for fine-grained preference learning and image generations.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. 2, 5

[2] Oren Barkan and Noam Koenigstein. Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2016. 4

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2022. 2

[4] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016. 4

[5] Damian0815. Compel, 2024. GitHub repository, version 2.0.2. 3

[6] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008. 4

[7] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025. 4

[10] Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 5, 7

[11] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023. 2

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[13] Yuanhe Guo, Haoming Liu, and Hongyi Wen. Gemrec: Towards generative model recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024. 4

[14] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020. 3, 11

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 4, 6, 7

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[17] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. 6

[18] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation, 2018. 4

[19] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 1, 2, 4, 5, 6

[20] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 4, 8

[21] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. 1, 2

[22] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of open source software*, 2(11):205–, 2017. 4

[23] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. 3

[24] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Exploring data splitting strategies for the evaluation of recommendation models. In *Proceedings of the 14th acm conference on recommender systems*, pages 681–686, 2020. 6

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[26] Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Omni-id: Holistic identity representation designed for generative tasks. *arXiv preprint arXiv:2412.09694*, 2024. 1

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 4, 5

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 2

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6527–6536, 2024. 1

[31] Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. Viper: Visual personalization of generative models via individual preference learning. In *European Conference on Computer Vision*, pages 391–406. Springer, 2024. 2

[32] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2, 4, 5

[33] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[34] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 1, 5

[35] Zijie J. Wang, Fred Hohman, and Duen Horng Chau. WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings. *arXiv 2306.09328*, 2023. 3

[36] Tong Wu, Yinghao Xu, Ryan Po, Mengchen Zhang, Guandao Yang, Jiaqi Wang, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Fiva: Fine-grained visual attribute dataset for text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:31990–32011, 2025. 1, 2

[37] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 1, 2, 4, 6, 8

[38] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 2

[39] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM conference on recommender systems*, pages 269–277, 2019. 4

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 2

[41] Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. Flashface: Human image personalization with high-fidelity identity preservation. *arXiv preprint arXiv:2403.17008*, 2024. 2

A. Dataset

A.1. Image Prompts Safety Check

Fig. 5 shows the predicted the probability of NSFW content with Detoxify [14] for six aspects: toxicity, obscenity, identity attack, insult, threat, and sexual explicitness.
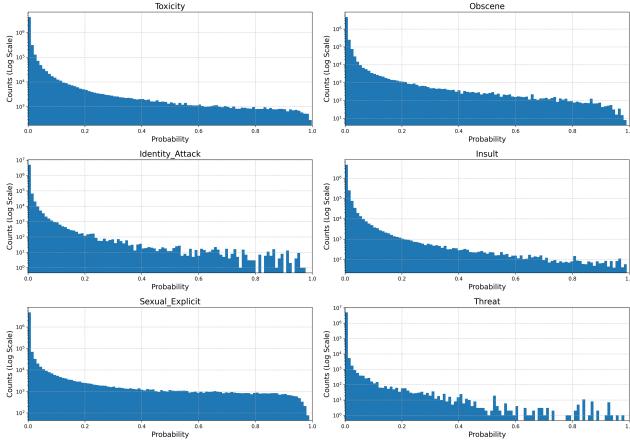

Figure 5. The distribution of prompts based on their predicted probabilites for NSFW content using Detoxify [14]. The y-axis represents the count of propmts in logarithmic scale.

A.2. Dataset Details

Prompt Word Count. We observed some exceptionally long prompts in our dataset. Fig. 6 shows the distribution of word counts for prompts with more than 200 words.

Figure 6. Cumulative Distribution of Prompt Word Counts in Log Scale for Prompts Exceeding 200 Words.

User-Image Feedback. Civitai enables users to respond to images with emojis anonymously, including "Heart", "Like" (Thumbs Up), "Laugh", "cry". Fig. 7 shows the distribution of these user-image interactions, which could serve as an indicator of popularity biases.
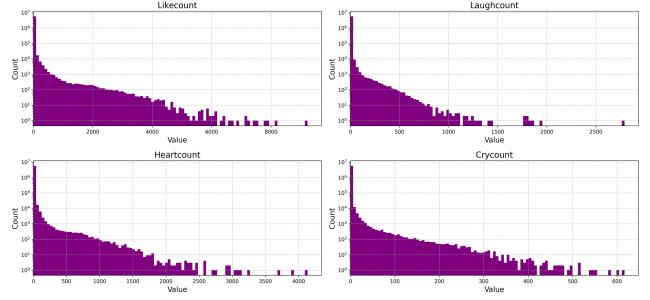

Figure 7. Log-scale distribution of image interactions for each emoji, with interaction values on the x-axis and the number of images on the y-axis.

User Interactions. We observe that the distribution of both user-image interactions and user-model interaction follows a long-tail manner. Fig. 8 plots the top30 users for image count and Fig. 9 shows the top30 users for model count.
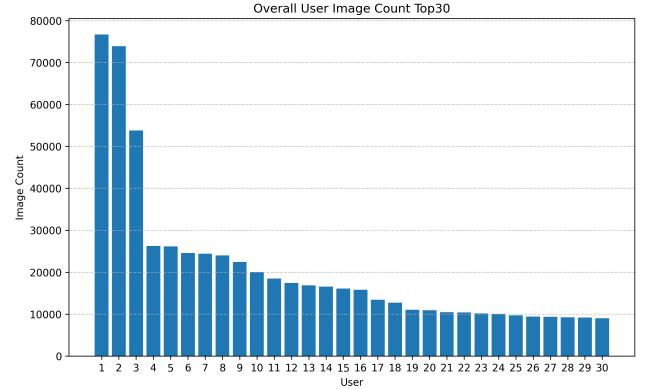

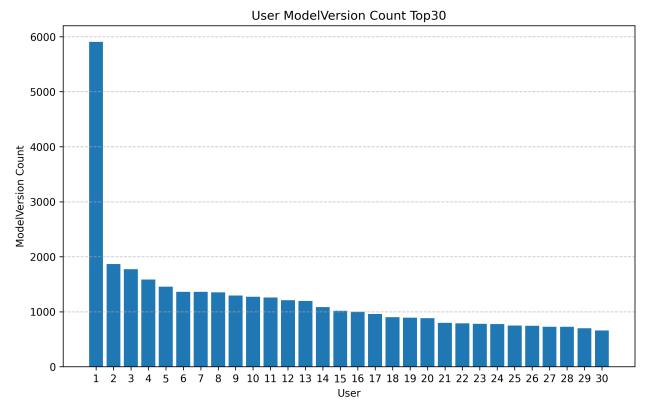Figure 8. Top 30 uses based on their image count. User names are hidden for privacy.

Figure 9. Top 30 users based on their model checkpoint count. User names are hidden for privacy.

B. VLM Captioning and Ranking

B.1. VLM Prompting Strategies and Ranking Demonstration

This appendix presents the structured prompts used in our VLM recommendation system, categorized into image recommendation and model recommendation, each with captioning and ranking tasks.

B.1.1. Image Captioning

```
Analyze these images and generate a structured
    description focusing on:

1. Primary Subject Type (e.g., human, fantasy
    creature, landscape).

2. Defining Visual Features (facial structure,
    clothing details, body posture).

3. Artistic Style (anime, realistic, digital
    painting).

4. Background Elements (futuristic city, ancient
    palace, foggy forest).
```

B.1.2. Image Ranking

```
Rank images based on similarity to the visual
    preference profile.

1. Overall Similarity (60 pts)
   - Primary Subject Match (20 pts): Does it
       belong to the same category? (Human,
       anthropomorphic, animal, scenery, object)
   - Artistic Style (15 pts): Matches reference?
       (Anime, realistic, digital painting, etc.)
   - Color Palette & Mood (15 pts): Similar tones
       , lighting, contrast?
   - Background & Setting (10 pts): Same
       environment (indoor, nature, fantasy, city
       , etc.)?

2. Detail Similarity (40 pts)
   - Key Features (20 pts):
       - Humans: Hair, clothing, accessories.
       - Animals: Fur color, body shape, eye
           design.
       - Scenery/Objects: Texture, materials,
           lighting effects.
   - Pose & Expression (10 pts): Consistency with
       visual preference profile.
   - Fine Details (10 pts): Composition, small
       artistic elements.

Return a JSON object:

{
    "image\_id": ID,
    "similarity\_score": score,
    "explanation": "Brief reason"
}
```

B.1.3. Model Captioning

```
Summarize the common features, themes, and styles
    across these descriptions in detail.
```

B.1.4. Model Ranking

```
Extract a detailed description of the user's
    visual style preferences.

Compare prompts based on:

1. Primary Subject (e.g., architecture, people,
    nature, abstract).

2. Artistic Style & Features (e.g., brushwork,
    realism, shading).

3. Color, Composition, Lighting (e.g., soft
    pastels, dark cyberpunk,
contrast).

Scoring:

90-100: Perfect match with all key preferences
70-89: Strong match with most preferences
50-69: Moderate match with some preferences
30-49: Weak match with few preferences
10-29: Very weak match with preferences
0-9: No match with preferences

Return a JSON object:
{
    "version\_id": Version ID,
    "similarity\_score": score,
    "explanation": "Brief reason"
}
```
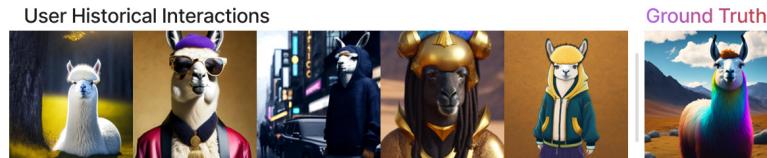
B.1.5. Randomized Scoring Strategy

To address the instability of VLM ranking results, we randomly sample a subset $C_i^{(k)} \subseteq C_i$ of k items, repeat the VLM scoring process T times with different sampled subsets, and compute the final score $s(x)$ for each item $x \in C_i$ as the expectation over multiple trials. This strategy ensures more consistent evaluations rather than relying on a single inference pass.

B.2. Example of VLM Ranking

The Table 8 and Figure 10 presents VLM ranking results from the same user. Table 8 presents the ranked images along with their similarity scores and explanations. These rankings correspond directly to the visual results in Figure 10, demonstrating VLM's interpretability—each ranked image is accompanied by a justification. Additionally, the ground truth (GT) image is ranked relatively high, showcasing VLM's promising performance. This example further illustrates how VLM-generated user preferences effectively guide ranking, contributing to more personalized and explainable recommendations.

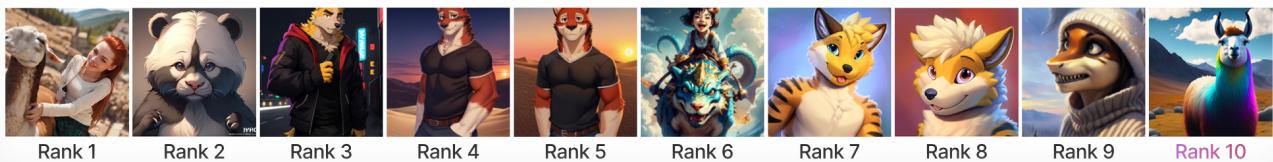| Image ID	Similarity Score	Explanation
242811	82.5	High similarity with primary subject match, artistic style, color palette, and key facial features.
173182	79.5	Good match with similar facial features and similar anime style.
660727	78.3	High similarity with key features, but difference in clothing and background.
244921	76.3	Decent match with feminine features but less intricate in background details.
244821	76.0	High overall similarity, similar style and key features but slight difference in color palette.
173226	72.7	Moderate match with some preferences but weaker in details and artistic style compared to the highest matches.
173227	70.0	Moderate similarity with key features but significant difference in style and color palette.
456861	69.3	Weak match with key preferences; differences in artistic style, color palette, and less pronounced facial features.
523827	68.3	Moderate match overall, slightly weaker because of hybrid eye color and differences in artistic style and setting.
456856	62.7	Weak match due to differences in artistic style, background, and slight disparity in key facial features.

Table 8. VLM assigns higher scores to images that closely match key visual features. Lower-ranked images often exhibit differences in background details, artistic style, or facial attributes, highlighting VLM's ability to provide an interpretable ranking explanation.

Figure 10. The top row represents the user's historical interactions (training set). The following rows show rankings from three recommendation models: ItemKNN, SASRec, and VLM. Images are ordered by ranking from left to right. The VLM model demonstrates superior performance, as its rankings align most closely with the user's ground truth interaction.

C. Generative Model Personalization

C.1. SVD Preliminary Study

To evaluate the effectiveness of SVD-based rank reduction, we decompose each LoRA into singular vectors and retain only the top-1 component. Using the same seed and prompt, we generate images from three models: the base SDXL model, the user-created full-rank LoRA, and the corresponding rank-1 reduced LoRA. We compute CLIP similarity between the base model's image and each LoRA-generated image to assess fidelity. As shown in Tab. 9, rank-1 LoRA shows only a slight increase in average CLIP similarity compared to the full-rank version, suggesting that the top-1 singular direction captures most of the useful information. This experiment is conducted across 10178 SDXL LoRAs with an average rank of 23.95.

Model Type	Avg. CLIP Score	Std Dev
Rank-1 LoRA	0.8114	0.1151
Full-Rank LoRA	0.7563	0.1215

Table 9. CLIP similarity between images generated by the unedited SDXL base model and those generated using the original high-rank LoRA and its SVD-reduced rank-1 version.

C.2. Significance of Different Layers

To assess the significance of different LoRA layers, we conducted experiments by injecting weight residuals from individual layers into a base model. Using identical seeds, we generated images and computed CLIP scores to measure the difference between these images and those from the base model. The results in Tab. 10 showed that feed-forward (FF) and attention value (attn_v) layers had the most significant impact on image generation

C.3. *ani-real* and *real-ani* Editing Results

To evaluate the effectiveness of different W2W space construction strategies, we compare the performance of the SVD-based and attn_v-based approaches on both the *ani-real* and *real-ani* directions. As shown in Fig. 12, the SVD-based W2W space enables smooth and coherent transformations in both directions. In contrast, the attn_v-based W2W space performs well for *ani-real* but fails to generalize to *real-ani* (Fig. 11). These results underscore the superior bidirectional editing capability of the SVD-based approach.

C.4. User Preference Description

Fig.13 shows Top 9 preference images of user P_1, P_2, P_3, P_4, along with their corresponding textual descriptions.

Layer Type	Average CLIP Score
attn_v	0.8851
attn	0.8433
ff	0.8319
ff+attn_v	0.7774

Table 10. Comparison of CLIP scores across different layer types. Scores are averaged over 24 models.

C.5. Multi-User Preference Alignment Results

Fig.14 demonstrates preference alignment for four users, where initial misaligned models were adjusted along learned directions. Beyond visual improvements, both the CLIP score and VLM-based rankings are higher for these edited images compared to the original outputs, confirming enhanced alignment after editing.

C.6. Image Generation Implementation

Tab. 11 provides a comprehensive overview of the image generation settings for different users. It outlines the model versions used, specific prompts, seeds, and key parameters such as edit strength. All images for generative model personalization were generated as 1024×1024px, with 30 inference steps, guidance scale 5, and LoRA scale 1.
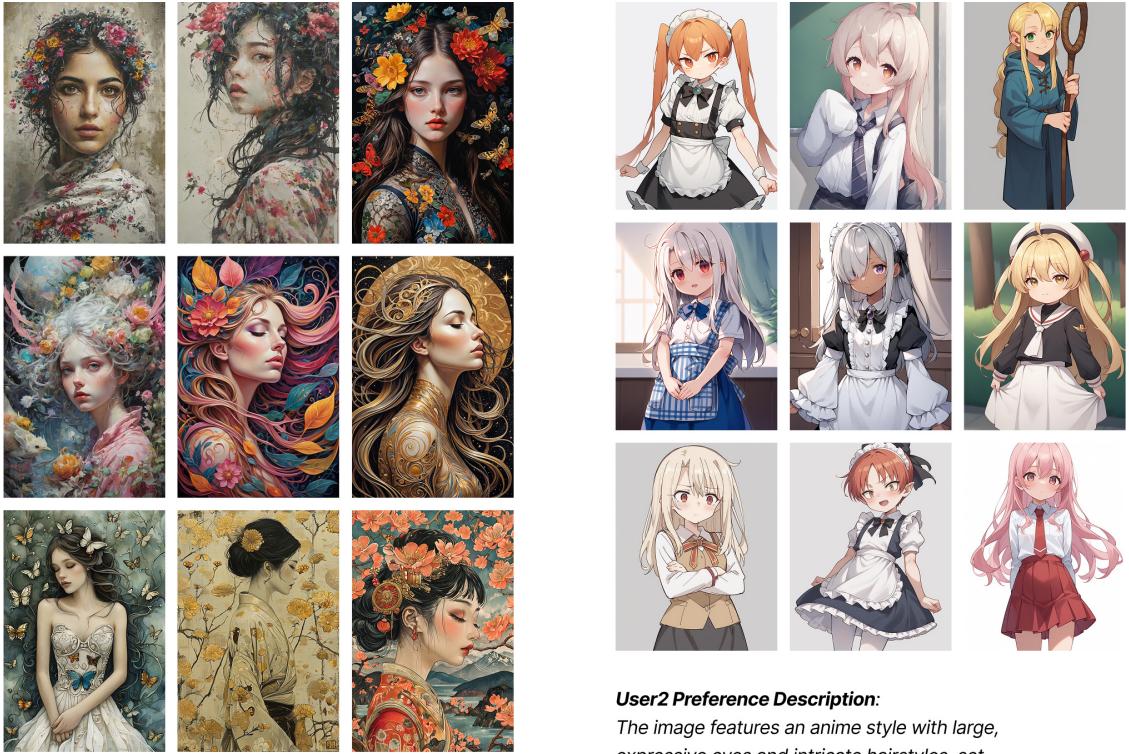
Figure 11. Editing results using the W2W space constructed from `attn_v` layers. Top: transformation from *realistic* to *anime*. Bottom: transformation from *anime* to *realistic*. The first column shows outputs from the unedited base model; subsequent columns show results with increasing tuning strength. Each row shares the same generation seed. While the *ani-real* direction produces coherent transitions, the reverse *real-ani* direction is less effective.
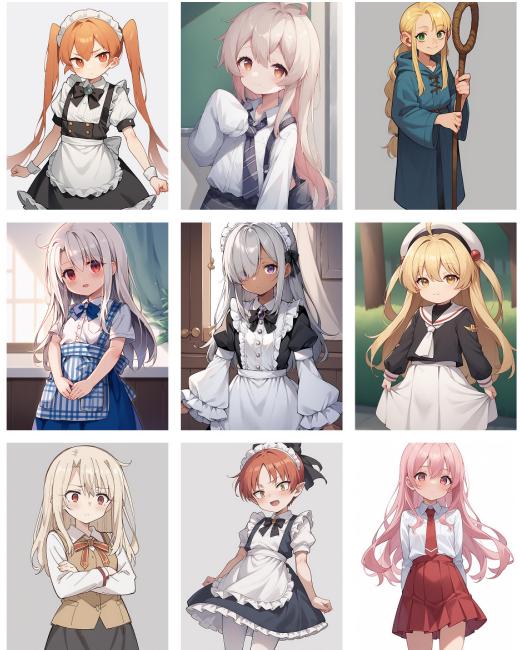
SVD-Based W2W Space

Figure 12. Editing results using the SVD-based W2W space. Top: transformation from *realistic* to *anime*. Bottom: transformation from *anime* to *realistic*. The base model outputs are shown in the first column, followed by results with increasing tuning strength. Each row uses a fixed generation seed. The SVD-based representation supports smooth, bidirectional editing with semantically coherent outputs in both directions.

User1 Preference Description:
Intricate, nature-infused portraits of young women with rich colors, symbolic details, and flowing, dynamic hair. Themes blend Eastern cultural influences, natural beauty, and introspective moods, creating visually striking and emotionally resonant artworks.

User2 Preference Description:
The image features an anime style with large, expressive eyes and intricate hairstyles, set against simply designed or plain backgrounds that emphasize the character. A predominantly pastel color palette with soft, muted tones enhances the aesthetic, while glowing or soft-focus lighting adds to the atmosphere.

User3 Preference Description:
Photo-realistic portrait with lifelike proportions, intricate facial details, and subtle skin/hair textures.lighting that enhances natural contours.

User4 Preference Description:
Photo-realistic portrait with lifelike proportions, intricate facial details, and subtle skin/hair textures. Well-balanced lighting enhances natural contours, emphasizing realism and depth. The subject is male.

Figure 13. User TOP 9 preference images along with the textual descriptions
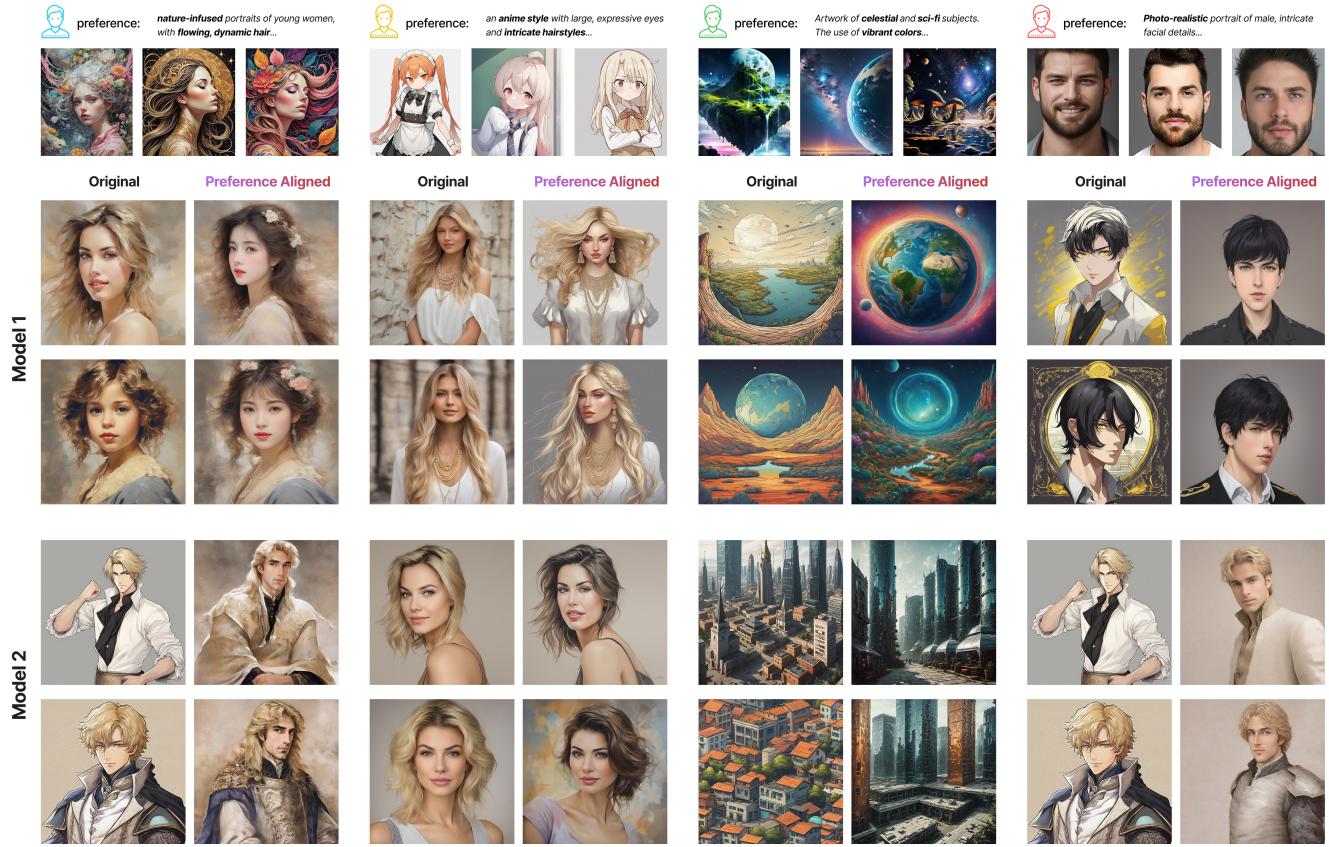
Figure 14. This figure illustrates the alignment of generative models to individual user preferences. Each user's visual preference is shown at the top, with generated samples below. Left images are from the unedited SDXL base model; right images are from the edited models.

User	Model Version ID	Prompt	Seeds	Edit Strength
User1	315523	portrait of a girl, high quality, ftsy-gld.	[2, 900]	6000
	150333	a man, Prince Hamlet, blonde, cessa style, looking at viewers, half-body, simple background, simple outfit.	[2, 37480]	7500
User2	480560	Dasha, with her blonde hair cascading over her shoulders and a delicate necklace accentuating her long hair.	[900, 7892]	6500
	802411	portrait of a women, high quality, J4ck13RJ.	[2, 50]	7500
User3	179603	view of planet earth from distant, cartooneffects one.	[2, 24]	7000
	565887	view of some buildings, from a distant, high quality, detailed, secretlab.	[23, 37480]	6000
User4	577810	portrait of a boy, high quality, linden de romanoff, black hair, yellow eyes, short hair, hair between eyes, bangs, simple background.	[10, 285891]	6000
	150333	A man, Prince Hamlet, blonde, cessa style, looking at viewers, half-body, simple background, simple outfit.	[2, 3]	6000

Table 11. Generation settings for preference alignment