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.

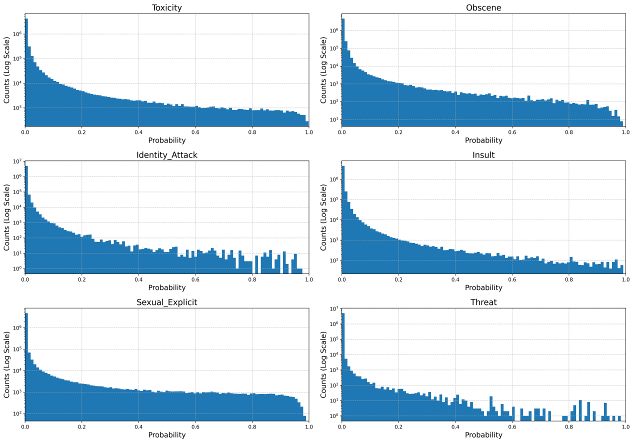


Figure 5. The distribution of prompts based on their predicted probabilites for NSFW content using Detoxify [14]. The y-axis represents the count of propmts in logarithmic scale.

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.

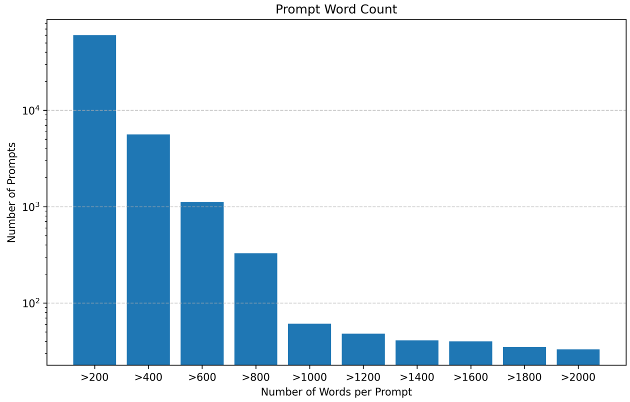


Figure 6. Cumulative Distribution of Prompt Word Counts in Log Scale for Prompts Exceeding 200 Words.

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.

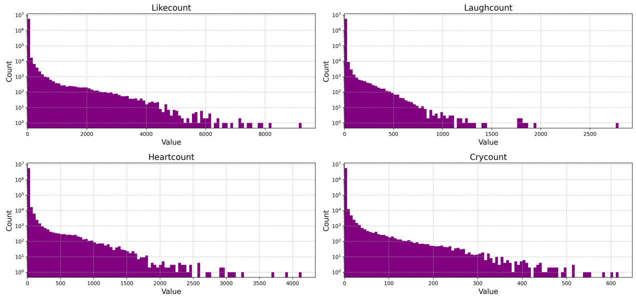


Figure 7. Log-scale distribution of image interactions for each emoji, with interaction values on the x-axis and the number of images on the y-axis.

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.

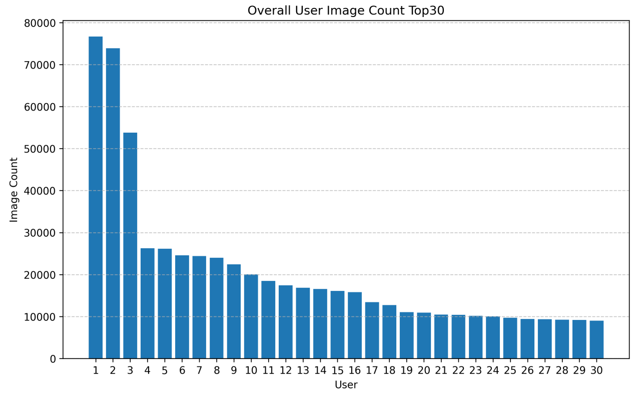


Figure 8. Top 30 uses based on their image count. User names are hidden for privacy.

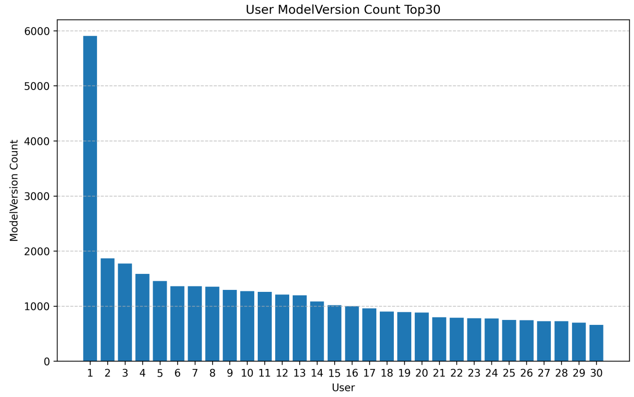


Figure 9. Top 30 users based on their model checkpoint count. User names are hidden for privacy.

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.

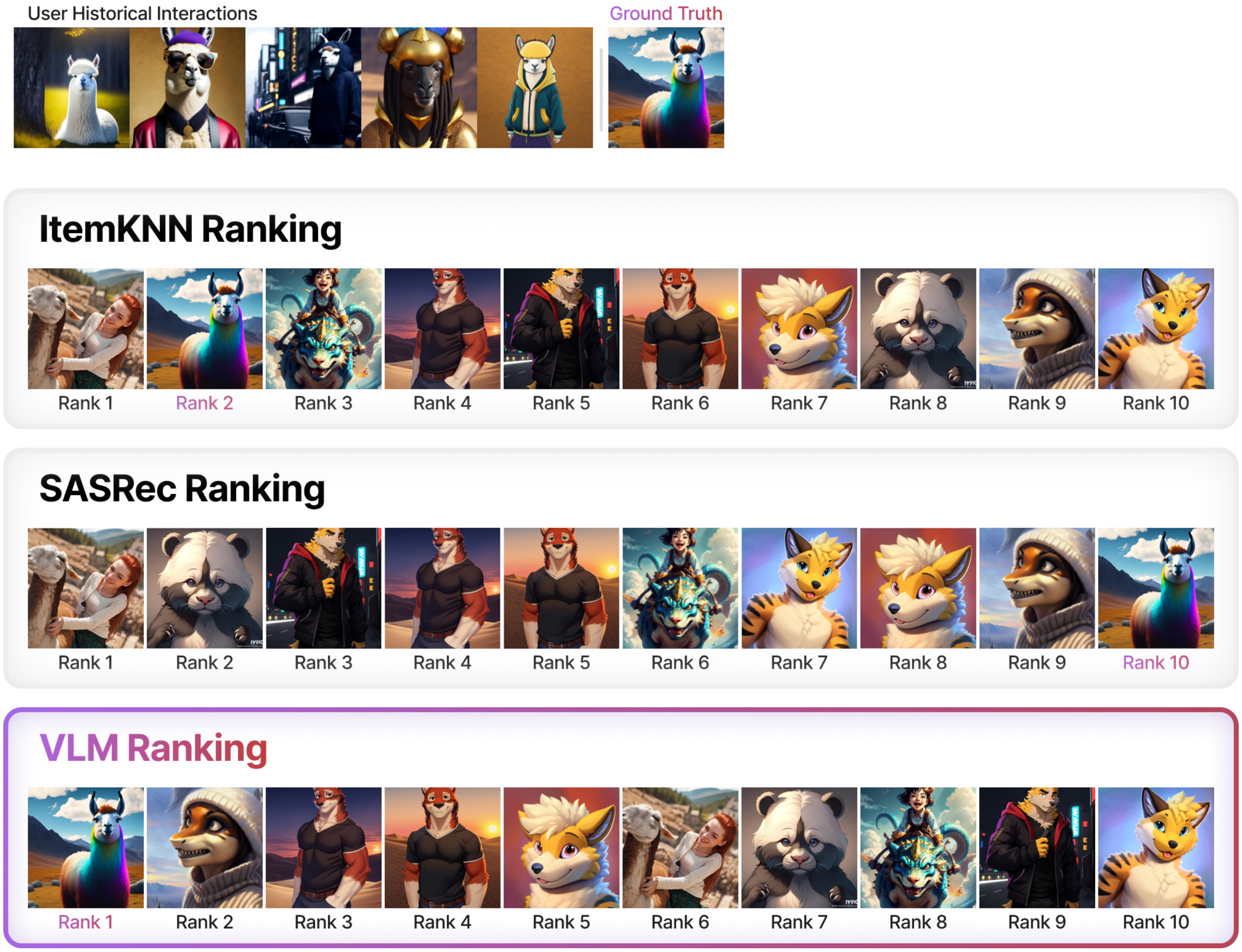


Figure 10. The top row represents the user's historical interactions (training set). The following rows show rankings from three recommendation models: ItemKNN, SASRec, and VLM. Images are ordered by ranking from left to right. The VLM model demonstrates superior performance, as its rankings align most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

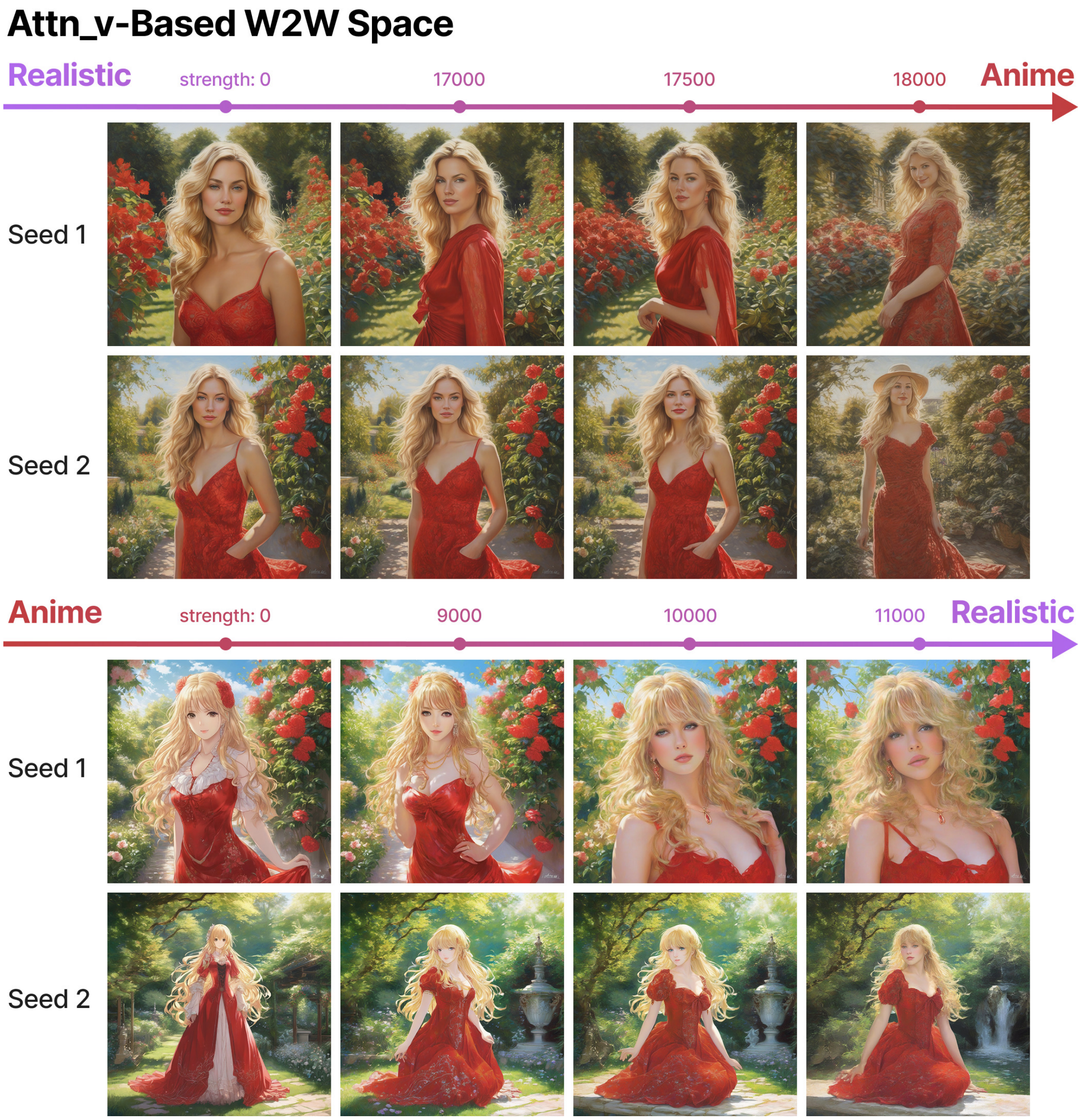


Figure 11. Editing results using the W2W space constructed from `attn_v` layers. Top: transformation from *realistic* to *anime*. Bottom: transformation from *anime* to *realistic*. The first column shows outputs from the unedited base model; subsequent columns show results with increasing tuning strength. Each row shares the same generation seed. While the *ani-real* direction produces coherent transitions, the reverse *real-ani* direction is less effective.

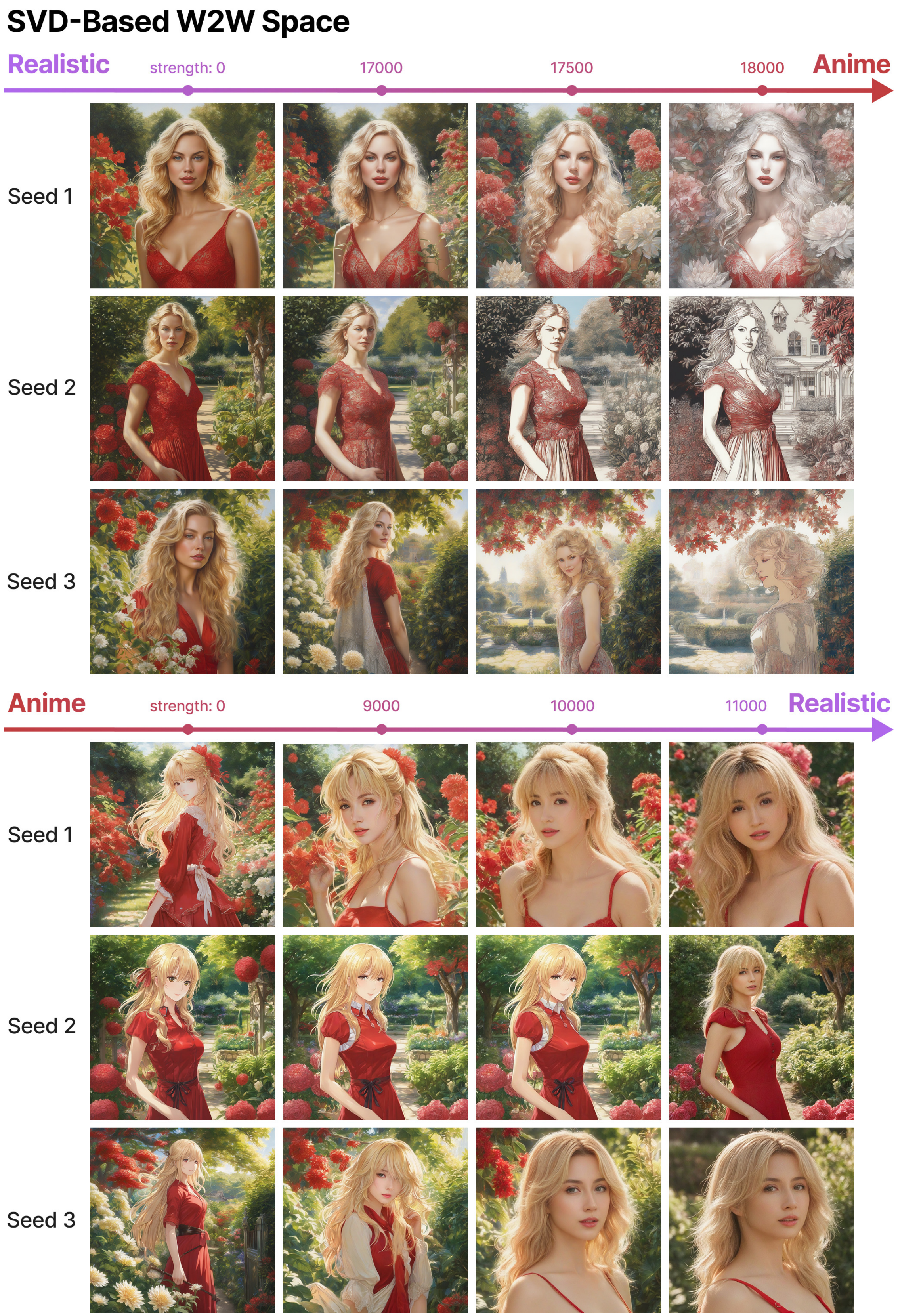
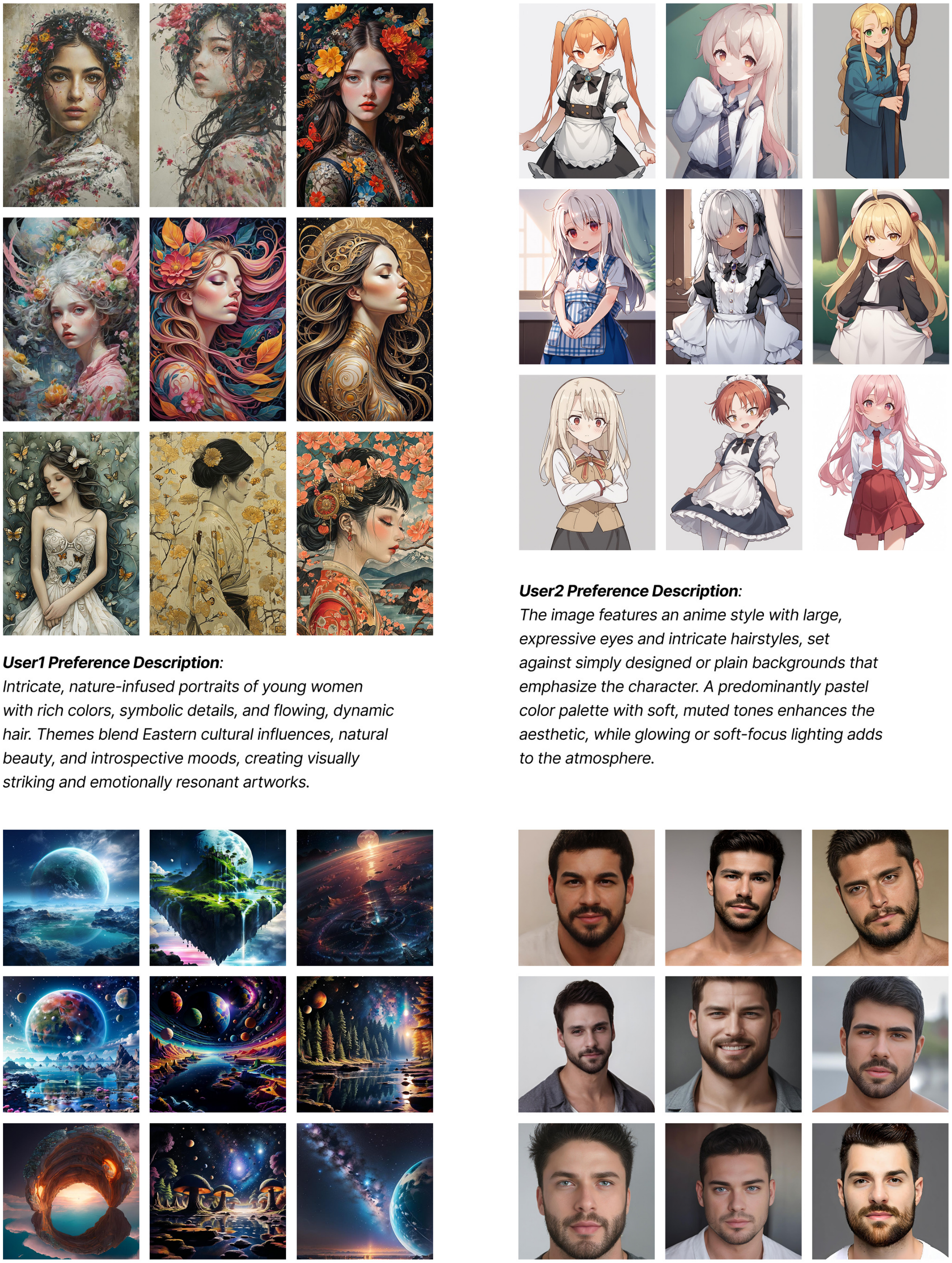


Figure 12. Editing results using the SVD-based W2W space. Top: transformation from *realistic* to *anime*. Bottom: transformation from *anime* to *realistic*. The base model outputs are shown in the first column, followed by results with increasing tuning strength. Each row uses a fixed generation seed.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

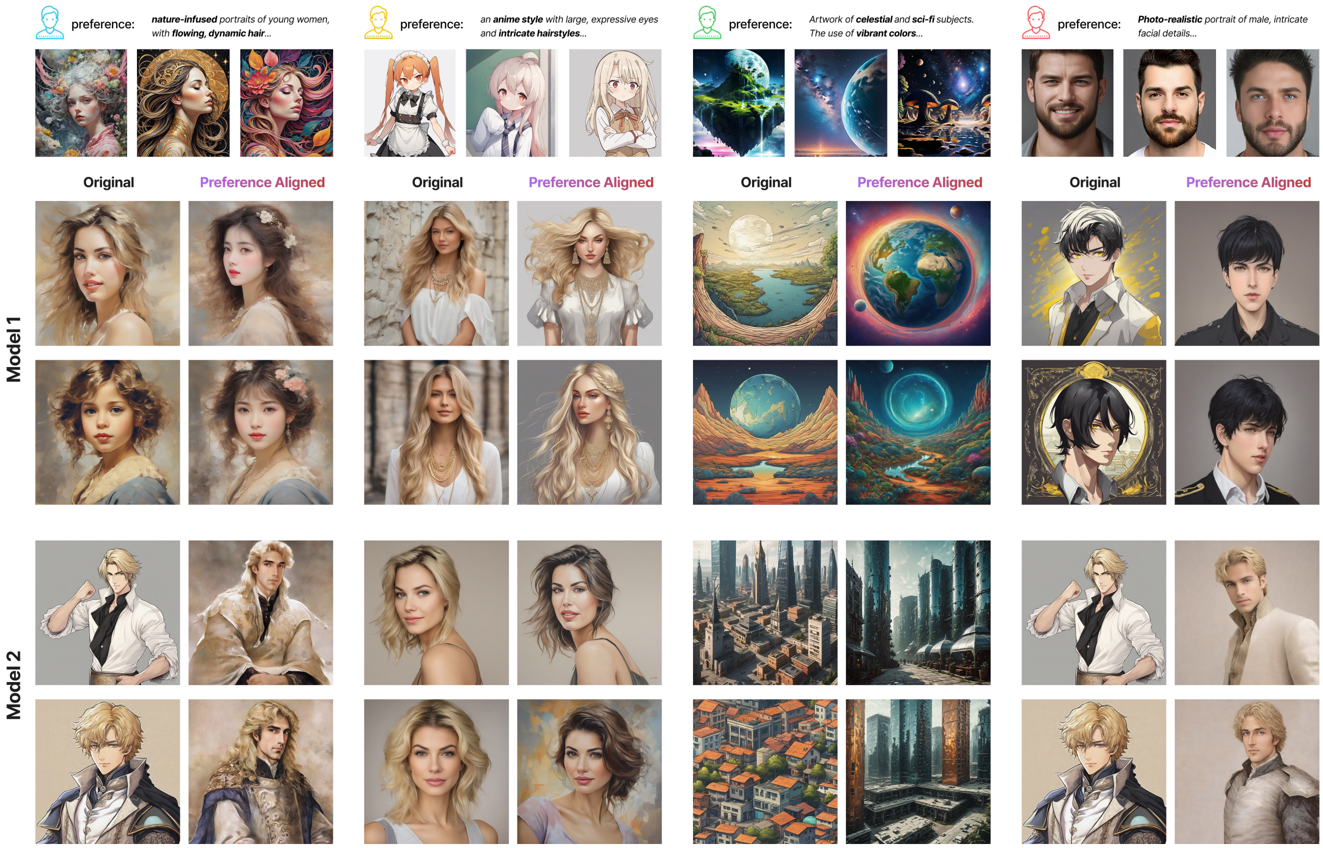

preference:
nature-infused portraits of young women, with flowing, dynamic hair...
preference:
an anime style with large, expressive eyes and intricate hairstyles...
preference:
Artwork of celestial and sci-fi subjects. The use of vibrant colors...
preference:
Photo-realistic portrait of male, intricate facial details...
Original
Preference Aligned
Original
Preference Aligned
Original
Preference Aligned
Original
Preference Aligned
Model 1
Model 2

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