

In this supplementary material, we provide comprehensive additional resources to further support our research. These include representative training samples, additional qualitative results to illustrate the model’s behavior. Furthermore, we provide an in-depth description of experimental setups for reproducibility to offer deeper insights into the implications and potential improvements of our approach.

A MORE QUALITATIVE ANALYSIS RESULTS

We present a comparison between our model and ViPer (Salehi et al., 2024), supported by qualitative results in Fig. 9, where target images with green borders indicate preferences aligned with the user. Unlike ViPer, which primarily relies on explicit features from reference images, our method leverages Multimodal Large Language Models (MLLMs) to capture deeper semantic relationships in user preferences. By leveraging learnable preference tokens, our approach captures both shared and individual preferences, enhancing prediction accuracy and robustness.

B MATHEMATICAL ANALYSIS OF AMBIGUOUS DECISION BOUNDARY CHALLENGES

The ambiguous decision boundary problem is most prominent when comparing user preference choices. As established in the main text, when $\mathcal{M}^+(\mathcal{S}, z) \approx \mathcal{M}^-(\mathcal{S}, z)$, the prediction function yields $Q(\mathcal{S}, z) \approx 0.5$, creating unstable decision boundaries. We now provide a detailed mathematical analysis of how this ambiguity leads to violations of both intra-group consistency and inter-group discrimination requirements in Assumption 1.

B.1 INTRA-GROUP DECISION INCONSISTENCY AT AMBIGUOUS BOUNDARIES

Consider a scenario where users $i, j \in \mathcal{U}_k$ from the same group evaluate an image pair (z_1, z_2) . When the model’s predictions approach the ambiguous boundary of 0.5, the following problematic situation can occur.

For the predicted scores of user i :

$$Q(\mathcal{S}_i, z_1) = 0.5 + \delta_1, \quad Q(\mathcal{S}_i, z_2) = 0.5 - \delta_1 \quad (10)$$

For the predicted scores of user j :

$$Q(\mathcal{S}_j, z_1) = 0.5 - \delta_2, \quad Q(\mathcal{S}_j, z_2) = 0.5 + \delta_2 \quad (11)$$

where δ_1 and δ_2 are small perturbations. Although these predictions satisfy the score similarity constraint from the assumption:

$$|Q(\mathcal{S}_i, z_k) - Q(\mathcal{S}_j, z_k)| \leq \epsilon_k \quad (12)$$

However, due to minute differences around the 0.5 boundary, the pairwise decisions of the two users become completely inconsistent. Specifically, the decision for user i is:

$$D_i(z_1, z_2) = \mathbf{1}[Q(\mathcal{S}_i, z_1) > Q(\mathcal{S}_i, z_2)] = 1 \quad (13)$$

whereas the decision for user j is:

$$D_j(z_1, z_2) = \mathbf{1}[Q(\mathcal{S}_j, z_1) > Q(\mathcal{S}_j, z_2)] = 0 \quad (14)$$

This inconsistency directly violates the intra-group decision consistency requirement from the assumption:

$$\mathbb{P}[D_i(z_1, z_2) = D_j(z_1, z_2)] \geq 1 - \alpha_k \quad (15)$$

B.2 INTER-GROUP DECISION CONSISTENCY AT AMBIGUOUS BOUNDARIES

Similarly, when users from different groups \mathcal{U}_k and \mathcal{U}_l both exhibit ambiguous predictions near 0.5, their preference scores become unexpectedly similar, leading to undesired inter-group consistency.

Consider the case where both users have predictions close to the ambiguous boundary:

$$\mathcal{Q}(\mathcal{S}_i, z) = 0.5 + \delta_i, \quad \mathcal{Q}(\mathcal{S}_j, z) = 0.5 + \delta_j \quad (16)$$

where $|\delta_i|, |\delta_j| \ll 0.5$ are small perturbations. This results in:

$$|\mathcal{Q}(\mathcal{S}_i, z) - \mathcal{Q}(\mathcal{S}_j, z)| = |\delta_i - \delta_j| \leq |\delta_i| + |\delta_j| \quad (17)$$

This proximity violates the inter-group score divergence constraint, as the expected difference can be arbitrarily small:

$$\mathbb{E}[|\mathcal{Q}(\mathcal{S}_i, z) - \mathcal{Q}(\mathcal{S}_j, z)|] \approx 0 < \max(\epsilon_k, \epsilon_l) \quad (18)$$

Furthermore, when both users' predictions hover around 0.5, their pairwise decisions for an image pair (z_1, z_2) may coincidentally align:

$$D_i(z_1, z_2) = \mathbf{1}[\mathcal{Q}(\mathcal{S}_i, z_1) > \mathcal{Q}(\mathcal{S}_i, z_2)] = D_j(z_1, z_2) \quad (19)$$

This unexpected agreement between users from different groups violates the inter-group decision divergence requirement:

$$\mathbb{P}[D_{u_i}(z_1, z_2) \neq D_{u_j}(z_1, z_2)] < 1 - \beta_{kl} \quad (20)$$

To resolve this fundamental issue, our contrastive preference learning method enforces clear preference discrimination through the following constraints.

For a positive sample, the positive logit is forced to be significantly greater than the negative logit:

$$\mathcal{M}^+(\mathcal{S}, z_{\text{pos}}) \gg \mathcal{M}^-(\mathcal{S}, z_{\text{neg}}) \quad (21)$$

For a negative sample, the negative logit is forced to be significantly greater than the positive logit:

$$\mathcal{M}^-(\mathcal{S}, z_{\text{neg}}) \gg \mathcal{M}^+(\mathcal{S}, z_{\text{pos}}) \quad (22)$$

B.3 INTRA-GROUP CONSISTENCY ANALYSIS

When the contrastive constraints are satisfied, clear decision boundaries are established that ensure intra-group consistency.

For a positive sample z_{pos} , we have:

$$\mathcal{Q}(\mathcal{S}, z_{\text{pos}}) = \frac{\exp(\mathcal{M}^+(\mathcal{S}, z_{\text{pos}}))}{\exp(\mathcal{M}^+(\mathcal{S}, z_{\text{pos}})) + \exp(\mathcal{M}^-(\mathcal{S}, z_{\text{pos}}))} \gg 0.5 \quad (23)$$

Similarly, for a negative sample z_{neg} :

$$\mathcal{Q}(\mathcal{S}, z_{\text{neg}}) = \frac{\exp(\mathcal{M}^-(\mathcal{S}, z_{\text{neg}}))}{\exp(\mathcal{M}^+(\mathcal{S}, z_{\text{neg}})) + \exp(\mathcal{M}^-(\mathcal{S}, z_{\text{neg}}))} \ll 0.5 \quad (24)$$

Let $\tau > 0$ be a confidence margin. The contrastive constraints then ensure:

$$\mathcal{Q}(\mathcal{S}, z_{\text{pos}}) \geq 0.5 + \tau, \quad \mathcal{Q}(\mathcal{S}, z_{\text{neg}}) \leq 0.5 - \tau \quad (25)$$

In this scenario, for users $i, j \in \mathcal{U}_k$ within the same group, even with score perturbations ϵ_k , decision consistency is guaranteed when $\epsilon_k < \tau$:

$$|\mathcal{Q}(\mathcal{S}_i, z_{\text{pos}}) - \mathcal{Q}(\mathcal{S}_j, z_{\text{pos}})| \leq \epsilon_k < \tau \quad (26)$$

This ensures that the decisions of both users on the same image pair remain consistent:

$$D_i(z_{\text{pos}}, z_{\text{neg}}) = D_j(z_{\text{pos}}, z_{\text{neg}}) = 1 \quad (27)$$

This satisfies the intra-group decision consistency constraint.

B.4 INTER-GROUP DISCRIMINATION ANALYSIS

For users from different groups $i \in \mathcal{U}_k$ and $j \in \mathcal{U}_l$ where $k \neq l$, the contrastive constraints create distinct preference distributions that ensure proper inter-group discrimination.

Under the contrastive learning framework, users from different groups develop distinct preference patterns for the same images. Consider the case where user u_i from group \mathcal{U}_k has learned to prefer certain visual patterns, while user u_j from group \mathcal{U}_l has learned different preferences.

For an image z that group \mathcal{U}_k generally likes but group \mathcal{U}_l dislikes, we have:

$$\mathcal{Q}(\mathcal{S}_i, z) \geq 0.5 + \tau, \quad \mathcal{Q}(\mathcal{S}_j, z) \leq 0.5 - \tau \quad (28)$$

This leads to a significant score difference:

$$|\mathcal{Q}(\mathcal{S}_i, z) - \mathcal{Q}(\mathcal{S}_j, z)| \geq |(0.5 + \tau) - (0.5 - \tau)| = 2\tau \quad (29)$$

When $2\tau > \max(\epsilon_k, \epsilon_l)$, this satisfies the inter-group score divergence constraint:

$$\mathbb{E}[|\mathcal{Q}(\mathcal{S}_i, z) - \mathcal{Q}(\mathcal{S}_j, z)|] \geq 2\tau > \max(\epsilon_k, \epsilon_l) \quad (30)$$

Furthermore, for pairwise decisions on an image pair (z_1, z_2) where the groups have opposite preferences, we obtain:

$$D_{u_i}(z_1, z_2) = 1, \quad D_{u_j}(z_1, z_2) = 0 \quad (31)$$

This ensures the inter-group decision divergence requirement is met:

$$\mathbb{P}[D_{u_i}(z_1, z_2) \neq D_{u_j}(z_1, z_2)] = 1 > \beta_{kl} \quad (32)$$

Therefore, by enforcing $\tau > \max(\epsilon_k, \epsilon_l)/2$, our contrastive preference learning method simultaneously satisfies both intra-group consistency and inter-group discrimination constraints specified in Assumption 1.

C MORE EXPERIMENTAL DETAILS

Examples of Training Data. Our dataset, based on Pick-a-Pic v2 dataset (Kirstain et al., 2023), focuses on image pairs annotated with user preferences. To ensure reliability, we filtered entries to include only users with at least 11 unique liked images. Fig. 10 and Fig. 11 present a selection of the training set from the dataset, providing valuable insights into how user-specific preferences. Patterns distinguishing a user’s likes and dislikes are evident.

Training. To conserve memory, each prompt is truncated to a maximum length of 100 tokens, and input images are resized to 512×512 pixels. Following the setup of (Salehi et al., 2024), we set the length of each user’s preference history sequence, N_{ref} , to 8. The learning rate is set to 1×10^{-5} , with a weight decay of 1×10^{-2} . The language model is fine-tuned using QLoRA (Dettmers et al., 2023), while the vision encoder is trained simultaneously. The input tokens template for the MLLM is “<image>The prompt is <prompt>. Score for this image?<label>”. We first train the MLLM using our custom loss function for 5k steps. After this phase, we continue training for 16k steps, during which both the model and the learnable preference tokens are jointly optimized. To prevent the model from overfitting to a fixed input pattern, we randomly shuffle the order of the reference history sequences during training.

User Preference Dimensions and Attribute Space in the Agent Dataset. To simulate diverse and fine-grained user preferences, we construct a dataset using the Claude-3.5-Sonnet agent. We first define a comprehensive taxonomy of aesthetic attributes spanning multiple key dimensions, as shown in Table 7. These dimensions include art styles, color palettes, compositional strategies, skill levels, visual detail, color hues, and artistic mediums. Each agent is assigned a personalized subset of liked and disliked attributes, sampled from the full attribute space. This configuration enables controllable and individualized preference simulation. The richness of the attribute space ensures that agents exhibit highly diverse and nuanced preferences, mimicking the variability observed in real-world users. Some examples of generated agents are illustrated in Fig. 8, with their corresponding attribute configurations listed in Tab. 8.



Figure 8: Some examples in Agent Dataset.

Dimension	Example Attributes
Art Styles	Surrealism, Aboriginal Art, Ukiyo-e, Romanticism, Anime/Manga, Contemporary Abstraction, Ancient Greek Art, Baroque Art, Abstract Expressionism, Art Deco, Cubism, ...
Color Palettes	Oceanic Tones (e.g., Turquoise, Deep Sea Blue), Neon (e.g., Laser Blue, Hot Magenta), Urban Industrial (e.g., Alloy Silver, Iron Black), Pastels (e.g., Mint Green, Peach), Muted Shades (e.g., Dusty Rose), Vibrant Colors, Earthy Palettes, ...
Composition	Invented vs. Real Space, Dynamic/Static Tension, Grid-Based Layouts, Pictorial vs. Installation, Foreground vs. Background Contrast, Rule of Thirds, Negative Space Use, Deep/Shallow Space, Balanced or Fragmented Structures, ...
Skill Level	Rigorous, Intuitive, Spontaneous, Experimental, Polished, Effortless, Graceful, Heavy-handed, Sophisticated, Controlled, Inventive, ...
Detail Level	Tactile, Sharp, Subtle, Elaborate, Vivid, Blurred, Simplified, Defined, Smooth, Intricate, Muted, Textured, ...
Hues	Turquoise, Magenta, Burgundy, Indigo, Crimson, Yellow, Slate Gray, Cerulean, Forest Green, Orange, ...
Artistic Medium	Mixed Media (e.g., Found Object, Assemblage), Printmaking (e.g., Lithography, Woodcut), Digital (e.g., 3D, Virtual Reality), Traditional Painting (e.g., Tempera, Watercolor), Textile Arts (e.g., Weaving, Embroidery), Ceramics, Sculpture, Drawing, ...

Table 7: Overview of user preference attribute space across key aesthetic dimensions.

Evaluation Prompt of Claude-3.5-Sonnet. To provide an additional benchmark for evaluating preference prediction accuracy, we employ Claude-3.5-Sonnet, a powerful large multimodal model, as an automated annotator simulating user-level preference reasoning. For each test case, we supply the Claude agent with a set of reference images representing the user’s preferences (liked and disliked examples), along with two candidate images. The agent is instructed to infer visual preference patterns from the references and select the more preferred candidate based on visual alignment. The exact prompt used for each evaluation instance is as follows: "You are given a set of reference images indicating user preferences: the images in <image>, ..., <image> are liked, and those in <image>, ..., <image> are disliked. Based on the visual

Agent	Dislikes	Likes
1	Vivid Purple, Radiant Red, Social Realism, Romanticism, Contemporary Abstraction, Mesoamerican Art, Charcoal Black, Pink, Foreground, Negative Space, Closed space, Pictorial, Free-flowing, Effortless, Experimental, Polished, Unfocused, Tactile, Smooth, Ethereal, Turquoise, Burgundy, Collage, Metal, Crocheting, Drypoint	Deep Sea Blue, Jungle Green, Oceanic Art, Traditional African Art, Islamic Art, Buttercream, Alloy Silver, Rule of Thirds, Invented space, Rhythmic, Illusion of Depth, Graceful, Intuitive, Powerful, Meticulous, Elaborate, Soft, Sharp, Muted, Slate Gray, Blue, Magenta, Orange, Decoupage, Virtual Reality, Cyanotype, Found Object
2	Yellow, Land Art, Situationist Art, Performance Art, Ukiyo-e, Faded Denim, Mint Green, Turquoise, Closed space, Foreground, Invented space, Centralized, Experimental, Inventive, Graceful, Sophisticated, Smooth, Vivid, Expressive, Magenta, Gold, Emerald, Glass, 3D Modeling, Digital Collage, Ink	Naive Art, Contemporary Abstraction, Color Field Painting, Fauvism, Deep Indigo, Jet, Ocean Green, Electric Lime, Pictorial, Golden Ratio, Symmetry, Fragmented, Effortless, Classic, Sophisticated, Subtle, Fine, Abstracted, Sharp, Teal, Orange, Slate, Polaroid, Spray Paint, Watercolor, Found Object

Table 8: Some examples in Agent Dataset.

patterns and preferences inferred from these references, classify a target image by comparing two candidates: `<image>` and `<image>`. Output only the index (0 or 1) of the image the user would prefer, with explanation."

In each query, `<image>` placeholders are replaced with actual image content using Claude’s multi-modal input interface. The agent’s selection (index 0 or 1) is parsed to compute top-1 accuracy across the test set. The performance of Claude-3.5-Sonnet is reported in Table 3 in main text, achieving a top-1 accuracy of 47.96%. This result offers a meaningful reference point for understanding model performance relative to state-of-the-art language-based multimodal reasoning capabilities.

Image Generation Guided by Our Models. Following the method outlined in (Eyring et al., 2024), we assign the weight 0.75 to our model. The initial image is optimized over 30 steps. For our model, we replace non-differentiable components of the vision preprocessor such as numpy-based resizing and similar operations with PyTorch operations. The preprocessed image is then integrated into the model’s input for optimization, ensuring that gradients flow seamlessly from the output score (in Eq. (6)) back to the initial image.

Visualization of Attention Scores. After applying the softmax operation in the self-attention mechanism, we extract attention weights, which are used to compute the weighted average within the self-attention heads. For visualization, we use the attention scores from head No. 28.

D THE ROLE OF LARGE LANGUAGE MODELS (LLMs)

We employed large language models (LLMs) exclusively as writing aids during the preparation of this manuscript. Their use was limited to enhancing readability through language refinement, grammar correction, and stylistic polishing. The models were not involved in formulating research questions, developing methods, conducting experiments, or interpreting findings. All conceptual contributions, technical innovations, and conclusions presented in this paper are entirely attributable to the authors.

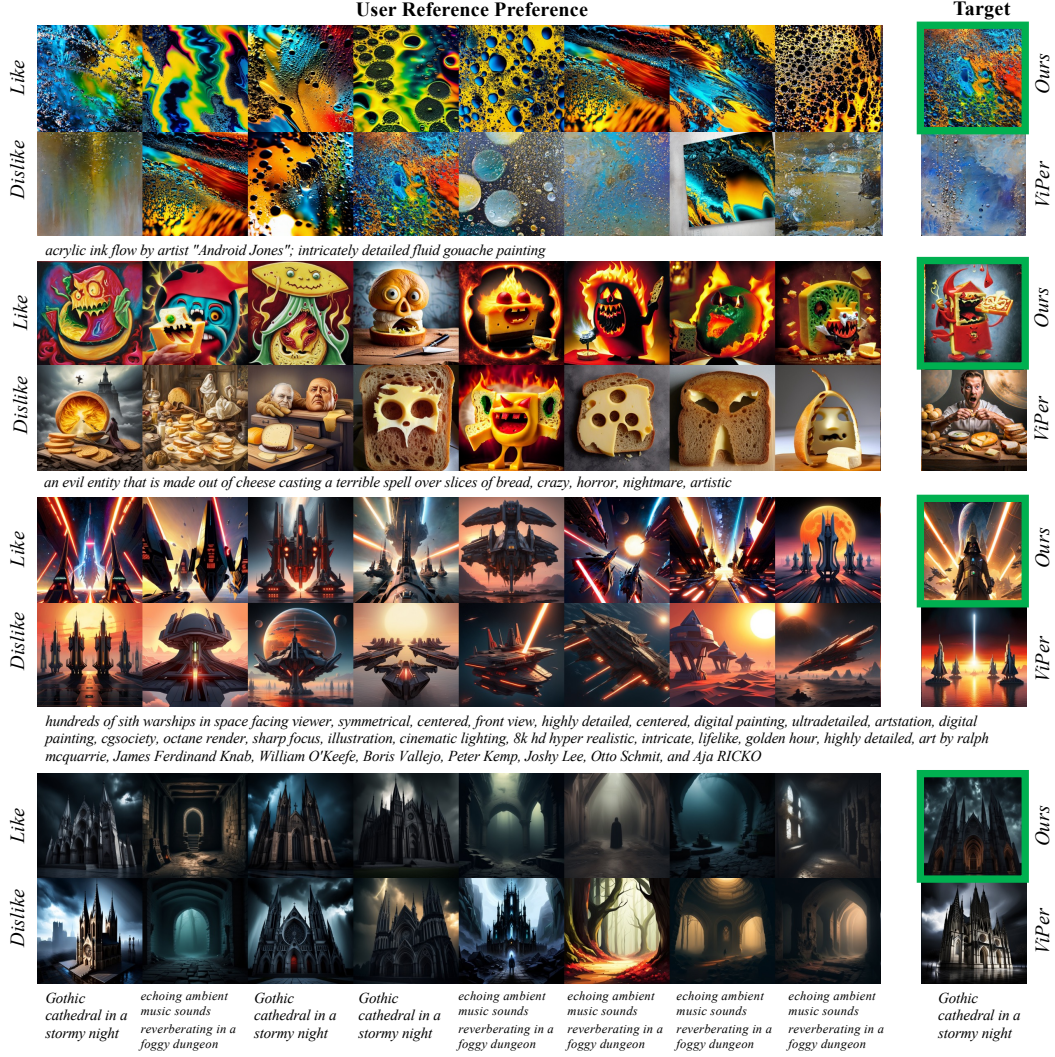


Figure 9: Visual comparison of user-specific preference alignment between our model and ViPer (Salehi et al., 2024) across varying preferences. Target images with green borders indicate preferences aligned with the user. Our method demonstrates effective capture of user-specific personalized results.

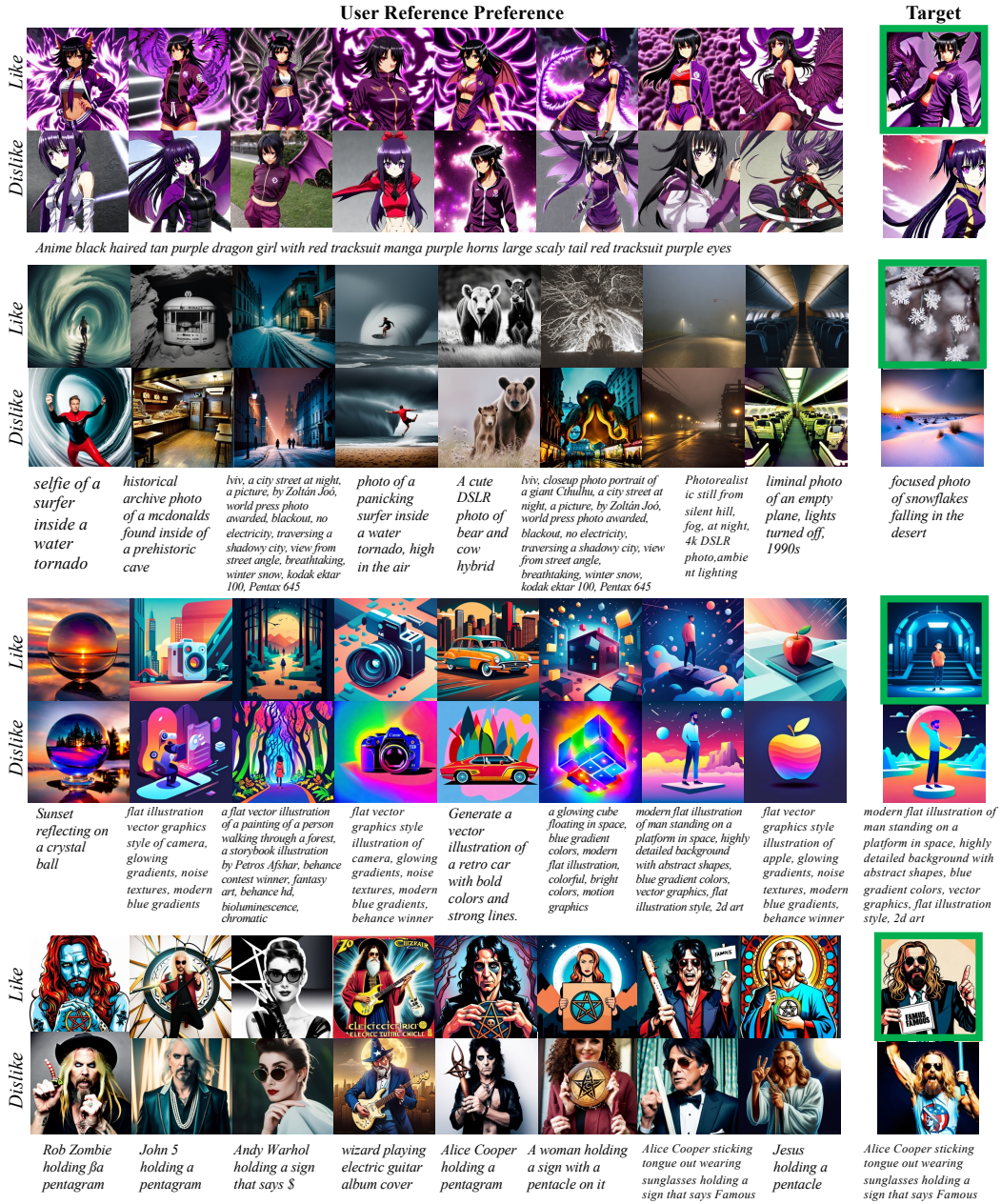


Figure 10: Some examples of the training data.

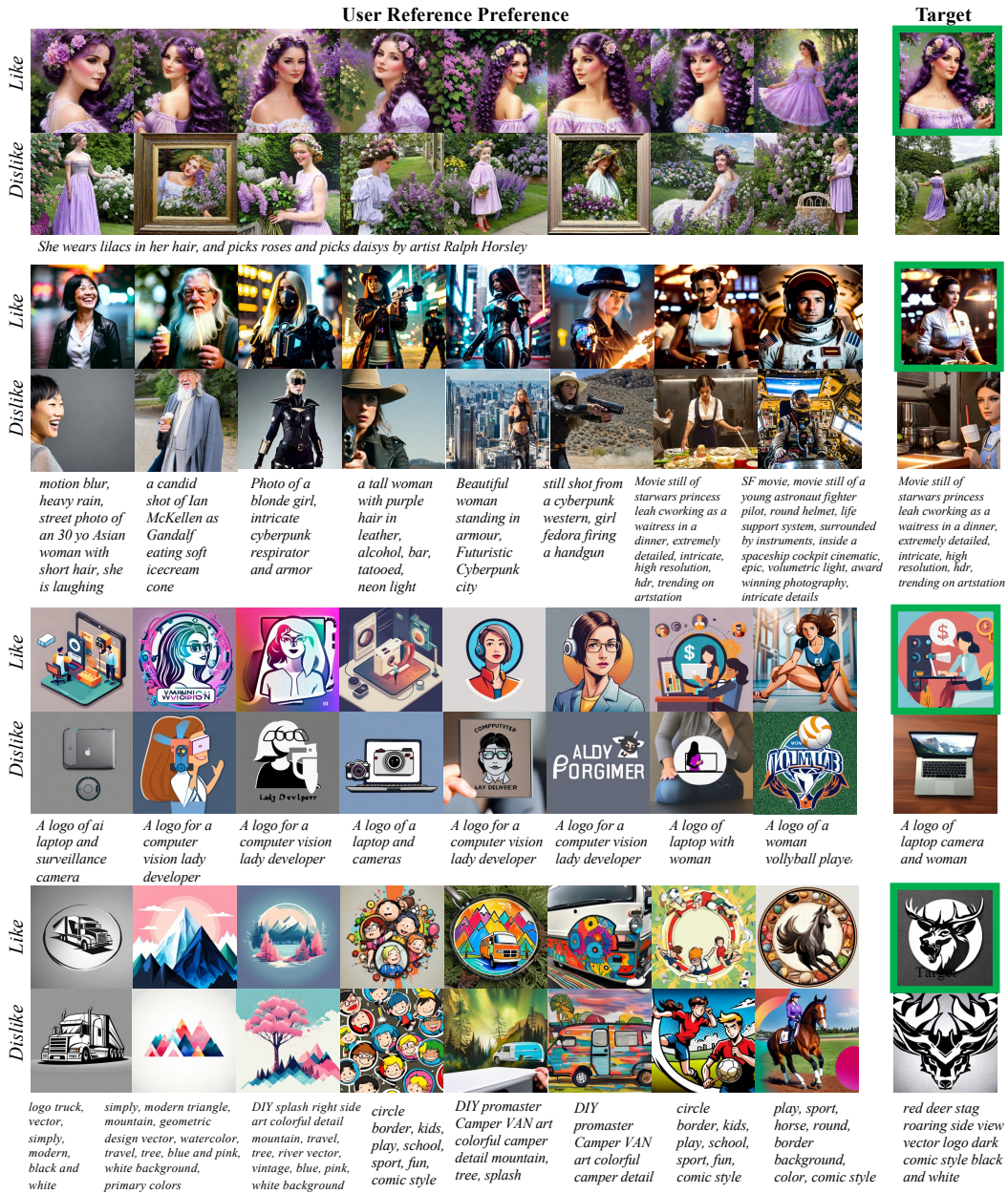


Figure 11: Some examples of the training data.