HPSv3: Towards Wide-Spectrum Human Preference Score
– Supplementary Materials –

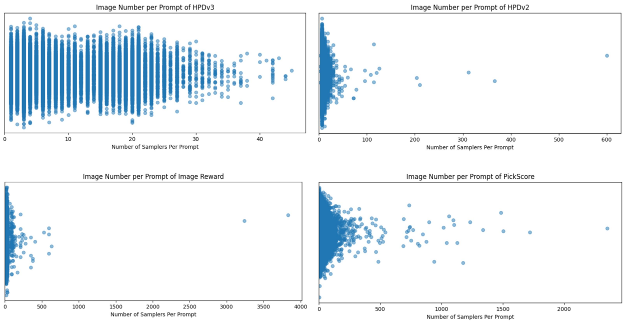


Figure 1. **Image numbers per prompt of each dataset.**

1. Image Sources of HPDv3

Table 1 summarizes the source models and images in HPDv3. Our dataset includes outputs from recent state-of-the-art image generation models, high-quality real-world images, and images generated by Midjourney, resulting in a total of 1.08M text-image pairs. Additionally, we compare the text-image pairs in HPDv3, HPDv2, PickScore, and ImageReward datasets. Figure 1 reveals that HPDv2 [18], PickScore [8], and ImageReward [20] datasets often associate identical prompts with more than 100 images, leading to an uneven distribution with significant outliers. Such imbalances can negatively affect model training. In contrast, HPDv3 maintains a more balanced distribution, with no prompt linked to more than 50 images, ensuring consistent and unbiased training for learning user preferences.

2. Category distribution of HPDv3

To better reflect user preferences for prompt categories, we categorize user prompts in JourneyDB [16] into 12 distinct classes, ensuring that the class proportions in HPDv3 closely match those in JourneyDB.

As shown in Figure 2, we compare the category distributions of HPDv3, HPDv2, ImageReward, and Pick-a-Pic datasets. The result shows that HPDv3 aligns most closely with the category proportions of JourneyDB, indicating that HPDv3 effectively captures user preferences for prompt categories, making it more representative and balanced.

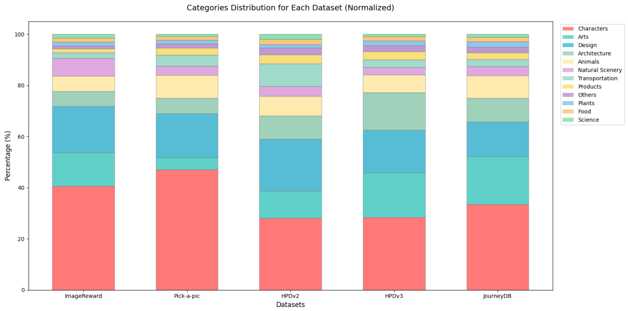


Figure 2. **Distribution of real images across** 12 **categories collected from the Internet.**

3. HPDv3 Dataset Construction

3.1. Real Image Collection

We collect aound 5M high-quality real images from the Internet, covering a wide range of categories such as architecture, people, objects, animals, plants, landscapes, products, and posters. This diverse sampling ensures that the dataset is broad and representative.

The collected images are predominantly authentic photographs. This forms a strong foundation for a high-quality dataset. However, despite our careful collection process, the dataset still contains some noise and irrelevant samples that require further refinement.

3.2. Aesthetic Model Training

To efficiently filter large volumes of images based on their aesthetic quality, we develop a specialized visual assessment model. As our analysis reveals, the model trained by open-source community [15] exhibits a strong preference for oil painting, which may not accurately reflect human aesthetic preferences. This bias could potentially skew our image quality assessments toward a particular visual style rather than capturing more universal aspects of image quality. To address this limitation and better align with diverse human aesthetic judgments, we decide to retrain the model using our newly developed dataset.

Model Training We keep the model architecture similar to the open-source aesthetic model [15] but refine it to address

Image Source	**Type**	**Num Images**	**Prompt Source**	**Split**
High Quality Image (HQI)	Real Image	57759	VLM Caption	Train & Test
Midjourney	-	331955	User Input	Train
CogView4 [23]	DiT	400	HQI+HPDv2+JourneyDB	Test
FLUX.1 dev [1]	DiT	48927	HQI+HPDv2+JourneyDB	Train & Test
Kolors [17]	DiT	49705	HQI+HPDv2+JourneyDB	Train & Test
HunyuanDiT [9]	DiT	46133	HQI+HPDv2+JourneyDB	Train & Test
Stable Diffusion 3 Medium [6]	DiT	49266	HQI+HPDv2+JourneyDB	Train & Test
Stable Diffusion XL [12]	Diffusion	49025	HQI+HPDv2+JourneyDB	Train & Test
PixArt-Σ [3]	Diffusion	400	HQI+HPDv2+JourneyDB	Test
Infinity [7]	Autoregressive	27061	HQI+JourneyDB	Train & Test
Stable Diffusion 2 [14]	Diffusion	19124	HQI+JourneyDB	Train & Test
CogView2 [4]	Autoregressive	3823	HQI+JourneyDB	Train & Test
FuseDream [10]	Diffusion	468	HQI+JourneyDB	Train & Test
VQ-Diffusion [5]	Diffusion	18837	HQI+JourneyDB	Train & Test
Glide [11]	Diffusion	19989	HQI+JourneyDB	Train & Test
Stable Diffusion 1.4 [14]	Diffusion	18596	HQI+JourneyDB	Train & Test
Stable Diffusion 1.1 [14]	Diffusion	19043	HQI+JourneyDB	Train & Test
HPDv2 [18]	/	327763	-	Train
Total		1088274		

Table 1. **Image Sources of HPDv3.** The dataset contains images from both high-quality real photographs and various types of image generation models, including autoregressive models, DiT-based models, and diffusion models.

aesthetic bias. We train our model on a single NVIDIA A100 80GB GPU using our carefully curated $20,000$-image annotation dataset. The training configuration employs a learning rate of 5×10^{-3} and a batch size of 256.

3.3. High-quality Real-image Selection

Using our trained aesthetic model, we evaluate the quality of all collected images. Fisrt, we filter out all images with a quality score below 4.0, as these are consistently lower quality. Then, to ensure category diversity, we apply category-specific evaluation and proportional selection, focusing on the highest-scoring images within each category while maintaining the ratios of each category in HPDv3.

This process results in a final curated dataset comprising 58k high-quality real images. The category distribution of these selected images, shown in Figure 3, closely aligns with the proportions seen in Figure 2, ensuring a well-balanced and high-quality dataset.

3.4. Pairwise Image Generation

We use various generative models with their recommended configurations to create images based on prompts from HPDv2, descriptions of real images, or JourneyDB. For images generated from HPDv2 and JourneyDB prompts, we use square dimensions to maintain consistency. For images based on real image descriptions, we match the aspect ratios of the original images to preserve their structure and visual integrity. assessment of content and aesthetic qualities independent of image proportions.

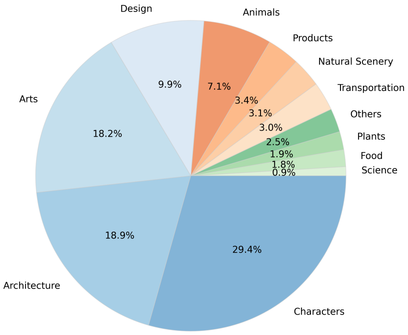


Figure 3. **Category distribution of high-quality real images in HPDv3 dataset.**

After generation, we group images by the same prompt into pairs comparing outputs from different models. These pairwise comparisons allow us to evaluate the relative performance of generative models under identical prompts, providing more detailed insights into their strengths and weaknesses. This pairwise approach forms the foundation for further annotations and model training.

Standard
Based on the text you see two pictures, please combine the degree of detail finesse, chipping, text, artistry, aesthetics and other dimensions of comprehensive consideration, choose the one you prefer. Tips: do not only consider the text and image correlation, do not consider the impact of image size, need to combine multiple dimensions to make a comprehensive judgment!

Table 2. **Details of the annotation guideline.**

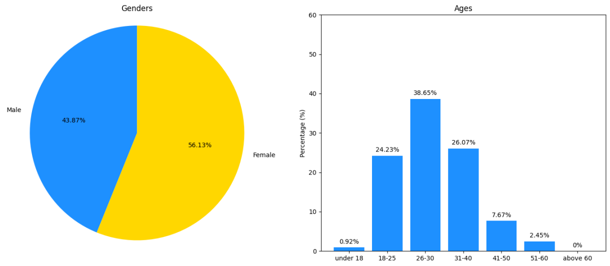
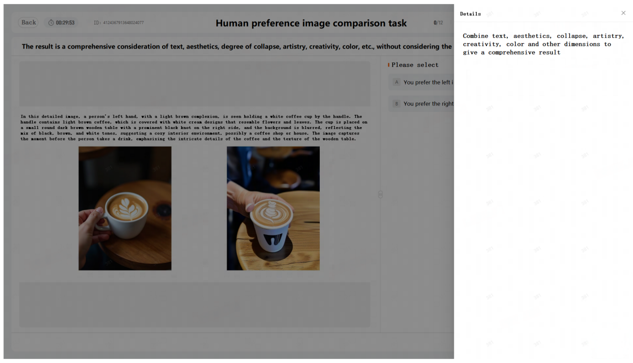


Figure 4. **Demographic Profile of Annotators: Gender Distribution and Age Stratification.**



The result is a comprehensive consideration of text, aesthetics, degree of collapse, artistry, creativity, color, etc., without considering the

In this detailed image, a person's left hand, with a light brown complexion, is seen holding a white coffee cup by the handle. The handle contains light brown coffee, which is covered with white cream designs that resemble flowers and leaves. The cup is placed on a small round dark brown wooden table with a prominent black knot on the right side, and the background is blurred, reflecting the mix of black, brown, and white tones, suggesting a cozy interior environment, possibly a coffee shop or house. The image captures the moment before the person takes a drink, emphasizing the intricate details of the coffee and the texture of the wooden table.

Combine text, aesthetics, collapse, artistry, creativity, color and other dimensions to give a comprehensive result

Figure 5. **Annotation interface of pairwise image comparison.**

4. Annotation Details

4.1. Image Annotation.

To build a reliable training dataset, we subject the pairwise image data to thorough human annotation, following a standardized evaluation protocol.

As shown in Table 2, human evaluators are provided with detailed guidelines that define clear criteria for judgment. This structured approach helps ensure consistent annotations while capturing the multidimensional aspects of human preferences.

Each image is scored by $9-19$ experts, with an inter-annotator agreement threshold set at 0.9 to ensure high reliability. The annotation process follows the same rigorous methodology described in the pairwise data annotation section of the main paper. By maintaining consistent evaluation standards across all stages of dataset creation, we ensure the overall quality and reliability of the annotated data.

Example 1: Question ID: xxxxxxx
Image Type: Hunyuan VS SD3 **Ground Truth:** 1 **Confidence Score:** 0.98 **Average Completion Time:** 9.56 seconds **Response Distribution:** 8/9 users (88.9%) correctly identified the image as synthetic **User Capability:** Only one user had measured capability score (68.0) **Speed Range:** From 1.269s to 22.814s
Example 2: Question ID: xxxxxxx
Image Type: Real image VS SD3 **Ground Truth:** 0 **Confidence Score:** 0.95 **Average Completion Time:** 5.74 seconds **Response Distribution:** 19/19 users (94.7%) incorrectly identified the real image as synthetic **User Capability:** Three users had measured capability scores (68.0, 68.0, 73.5) **Speed Range:** From 0.945s to 22.253s

Table 3. User response analysis for pairwise image analysis tasks

4.2. Demographic of annotators.

As shown in Figure 4, our annotation team has a relatively balanced gender distribution, with 56.13% female and 43.87%male participants. The age demographics show that most annotators fall within the young to middle-aged categories, with 88.95% aged between 18 and 40 years, and the $21-30$ age group being the largest (38.65%). This demographic composition offers several advantages, such as a strong familiarity with modern language patterns and current fashion trends. Additionally, the annotators come from diverse professional backgrounds, including college students, freelancers, artists, teachers, and engineers. This diversity brings a wide range of perspectives to the annotation process. Such a well-balanced and diverse group ensures that the annotations capture the preferences of fashion-conscious consumers effectively, enhancing both the quality and relevance of our annotated dataset.

4.3. Annotation interface and guidelines.

Figure 5 illustrates our user-friendly annotation interface for pairwise image comparison. The interface displays two images generated from the same textual prompt, along with the prompt itself to provide contextual clarity. Annotators are required to select the image they prefer based on clearly defined evaluation guidelines.

The annotation protocol guides evaluators to assess images across three key dimensions:

- **Prompt Alignment**: How well the image matches the given textual description.

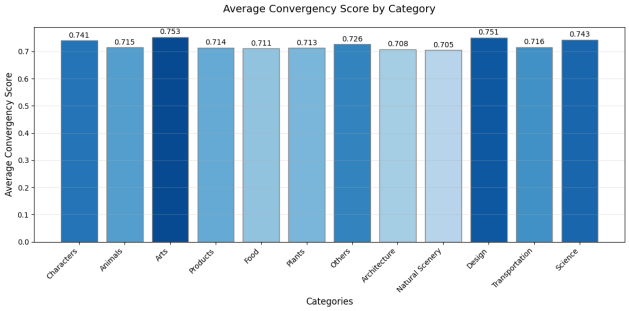


Figure 6. **Average convergency score by categories.**

Data Source	Description	Pair Num
HPDv3	Real images and comparisons Golden trainset filtered from HPDv2	652k 250k
Pick-A-Pic	Randomly selected subset	350k
ImageReward	Randomly selected subset	120k
Midjourney	Real user choice data	150k
Total		**1522k**

Table 4. **Composition of the training dataset used for HPSv3 model training.**

- **Aesthetic Quality**: The visual appeal and technical execution of the image.
- **Overall Coherence**: The logical consistency and naturalness of the scene depicted.

These structured criteria ensure that the annotations capture meaningful differences in quality while minimizing subjective bias. Table 3 provides examples of annotation outcomes, including final preferences, confidence scores, and an assessment of annotator quality. This systematic approach helps maintain the reliability and consistency of the dataset annotations.

4.4. Convergency of Annotations.

We evaluate the annotation convergence for each data category to measure the consistency of annotator decisions. Figure 6 visualizes the level of convergence of annotations in all categories. Convergence is calculated by assessing the agreement among annotators when evaluating image pairs with the same text prompt.

5. HPSv3 Training Details

5.1. Training Dataset

For training our final model, we use data from four sources: HPDv3, subsets of Pick-A-Pic and ImageReward, and real user preference data collected from Midjourney. In total, the training dataset comprises 1.5 million pairwise samples.

Specifically, the HPDv3 dataset is composed of two parts. The first part includes pairwise comparisons of real images and images generated based on prompts from JourneyDB and HPDv2. These comparisons are annotated using the annotation pipeline, and only those with a confidence score of 0.95 or higher are selected for training.

The second part is a manually curated golden training set. This contains high-quality sample data. We use this golden trainset to train a filtering model, which shares the same architecture and training methodology as HPSv3. This filtering model is applied to choose samples from HPDv2. And we randomly select $250,000$ pairs from samples picked from HPDv2. This process enriches the HPDv3 dataset by increasing both model diversity and pairwise data diversity. To further boost the contribution of the golden set, we duplicate some of its samples during the training process.

Additionally, we include $350,000$ pairwise samples from Pick-A-Pic and $120,000$ samples from ImageReward. These datasets provide additional variety to enhance model performance.

Furthermore, we collect $150,000$ pairs of real user-choice data from Midjourney via the Internet. This real-world preference data is crucial for improving HPSv3's ability to handle user selections during CoHP.

In summary, as detailed in Table 4, the training dataset is a diverse and comprehensive mix of high-quality, curated, and real-world user preference data. This careful composition ensures robust and adaptable model performance.

Model	ImageReward	PickScore	HPDv2	HPDv3
HPSv3 (HPDv2)	62.6	64.4	82.3	66.6
HPSv3 (ImageReward)	65.5	64.4	80.5	63.4
HPSv3 (PickScore)	61.0	70.6	80.2	64.9
HPSv3 (Ours)	**66.8**	**72.8**	**85.4**	**76.9**

Table 5. **Dataset ablation**. We train HPSv3 using the training datasets from HPDv2, ImageReward, and PickScore. The results demonstrate that training with HPDv3 training dataset achieves the highest accuracy across all test sets, showcasing its superior performance.

5.2. Training on other datasets

Table 5 presents the results of the dataset ablation study, where HPSv3 is trained on different datasets, including HPDv2, ImageReward, PickScore, and HPDv3. The evaluation is conducted across four metrics: ImageReward, PickScore, HPDv2, and HPDv3. Among the datasets, HPSv3 trained with HPDv3 outperforms others, achieving the best performance across all test sets—66.8% on ImageReward, 72.8% on PickScore, 85.4% on HPDv2, and 76.9% on HPDv3. These results clearly indicate that the HPDv3 dataset provides the most comprehensive and effective supervision for training. It significantly enhances the robustness and generalization of HPSv3, underscoring its superiority over other datasets.

5.3. Clarification on Loss Function

In this section, we clarify that various loss functions mentioned in the literature [8, 18, 19, 22] (including this paper) and the bradley-terry loss [2] share the same underlying optimization objective.

Form 1: Optimizing KL-divergence. In [8, 18, 22], the predicted preference $\hat{y_i}$ is calculated as::

$$\hat{y_i} = \frac{\exp{(r_i)}}{\sum_{j=1}^{2}\exp{(r_j)})}, \tag{1}$$

where r_i denotes the preference score of sample x_i. And the model is optimized by minimizing the KL-divergence between the ground truth y and the predicted distribution. Specifically, $y = [1, 0]$ if sample x_1 is preferred over x_2, and $y = [0, 1]$ otherwise. The loss function is formalized as below to minimize KL-divergence:

$$L_{\text{pref}} = \sum_{j=1}^{2} y_i \left(\log y_i - \log \hat{y_j}\right). \tag{2}$$

To simplify the problem, we assume that the sample x_h is the preferred sample, while x_l is the dispreferred one. According to the order, y_i will always be 1. Substituting them into the loss function, we obtain:

$$\begin{aligned} L_{\text{pref}} &= -\log\left(\frac{\exp(r_h)}{\exp(r_h)+\exp(r_l)}\right) \\ &= -\log\left(\frac{1}{1+\exp(r_l - r_h)}\right) \\ &= \log\left(1+\exp(r_l - r_h)\right). \end{aligned} \tag{3}$$

This shows that the KL-divergence formulation reduces to a logistic loss comparing the preference scores of the two samples.

Form 2: Bradley-Terry Loss. [2, 19] adopt the Bradley-Terry loss to maximize the probability of the winning (higher-ranked) samples over the losing (lower-ranked) samples. The probability of the winning samples in Bradley-Terry model can be defined as:

$$\begin{aligned} P(x_h \succ x_l) &= \frac{\exp(r_h)}{\exp(r_h)+\exp(r_l)} \\ &= \frac{1}{1+\exp(r_l - r_h)} \\ &= \text{sigmoid}(r_l - r_h). \end{aligned} \tag{4}$$

The goal is to maximize this probability. Therefore, the model is optimized by minimizing the negative log-likelihood:

$$\begin{aligned} L_{\text{BT}} &= -\log P(x_h \succ x_l) \\ &= -\log(\text{sigmoid}(r_l - r_h)) \end{aligned} \tag{5}$$

However, we can continue to simplify it.

$$\begin{aligned} L_{\text{BT}} &= -\log(\text{sigmoid}(r_l - r_h)) \\ &= -\log(\frac{1}{1+\exp(r_l - r_h)}) = \log\left(1+\exp(r_l - r_h)\right). \end{aligned} \tag{6}$$

Equivalence Conclusion. We can observe that both Form 1 (KL-divergence) and Form 2 (Bradley-Terry) ultimately converge to the same pair-wise logistic ranking loss:

$$L = \log\left(1+\exp(r_l - r_h)\right), \tag{7}$$

demonstrating their fundamental equivalence in optimization objectives despite different theoretical origins.

6. HPDv3 Dataset Visualization

6.1. Dataset Visualization

Figure 7 showcases examples from the HPDv3 dataset. Each image pair consists of different images generated from the same prompt, with the images sourced from various image generation models as well as real-world photographs. For each prompt, we systematically create pairwise comparisons by pairing all possible image combinations. Human annotators then evaluate these pairs to provide preference judgments, resulting in a detailed collection of pairwise preference data.

This rigorous pairwise annotation approach captures nuanced human preferences across different visual representations of the same concept. By including both AI-generated images from diverse models and real photographs, the dataset enables a comprehensive analysis of human preferences across the spectrum of synthetic and authentic visual content. The resulting preference signals serve as a rich foundation for training our HPSv3 model to better align with human judgments.

6.2. Benchmark Visualization

Figure 8, 9 and 10 show sample prompts from the HPDv3 Benchmark. This benchmark is designed as a diverse and standardized evaluation framework for assessing the performance of image generation models. Specifically, the HPDv3 Benchmark includes a carefully curated set of $1,000$ prompts for each of the 12 categories, drawn from three datasets: HPDv3, HPDv2, and JourneyDB. These prompts cover a variety of styles and lengths to ensure comprehensive evaluation.

For prompts sourced from the HPDv3 dataset, we include the corresponding real-world reference images to facilitate comparisons with generated images. On the other hand, prompts from the HPDv2 and JourneyDB datasets are provided as text-only, focusing on evaluating a model's ability to generate images purely from textual input.

This setup enables evaluation from multiple perspectives. The inclusion of real-world reference images provides a way to measure how closely generated images align with actual visuals. At the same time, the text-only prompts test the model's ability to interpret and generate images solely based on textual descriptions. By combining these approaches, the HPDv3 Benchmark offers a comprehensive framework to assess image synthesis quality across various content categories and prompt styles, promoting consistent evaluation and progress in text-to-image generation research.

7. More Result of CoHP

In this section, we present an extensive collection of generation results from CoHP. We showcase diverse outputs produced across multiple iterations. The first row of Figure 11 and 12 shows the best result of each model (Flux, Kolors and Playground v2.5) generated in the Model-wise preference stage. As illustrated in Figure 11 and Figure 12, the Model-wise preference stage plays a critical role in CoHP by selecting the best model that can generate images with strong semantic understanding and well-constructed compositions. Meanwhile, the sample-wise preference stage contributes by refining the images and enhancing their details.

Figure 13 and 14 demonstrate comparative results obtained by implementing various human preference models within our framework. These expanded visualizations provide insights into how different preference modeling approaches influence the quality, diversity, and human-alignment of the generated images.

8. HPSv3 as Reward Model

When using reinforcement learning (RL) to improve the quality of generated images, the design of the reward model is critically important. A well-designed reward model can significantly improve outputs by boosting realism, aesthetic quality, and text-image alignment, or by aligning outputs more closely with human preferences. Leveraging a carefully built wide-spectrum image quality dataset and a backbone based on a Visual Language Model, HPSv3 excels at capturing human preferences more accurately. This reduces reward hacking behaviors in RL, guiding the model to produce content that better matches human expectations.

DanceGRPO We employ DanceGRPO [21] as the reinforcement learning algorithm for image generation and compare the results when using ImageReward, PiscScore, HPSv2 and HPSv3 as reward models, respectively. DanceGRPO performs multiple sampling of diffusion trajectories, scores the final generated image of each trajectory using the reward model, and conducts policy gradient optimization by calculating the advantage value of each trajectory relative to the average reward, thereby improving the model's performance. For all reward models, we use the same default experimental settings in DanceGRPO. We use Stable-Diffusion v1.4[13] as our base model, and performs 300 training iteration.

Experiment Results Figures 15 and 16 presents qualitative results obtained after the same number of training iterations. For convenience, we refer to the image generation models trained with these reward models as $M_{\text{ImageReward}}$, $M_{\text{Pickscore}}$, M_{HPSv2} and M_{HPSv3}, respectively. The results show that all these reward models improve the quality and aesthetic appeal of the generated images. M_{HPSv3} produces images with greater realism—more natural color saturation, smoother lighting and shadows, and fewer artifacts and distortions. Moreover, M_{HPSv3} exhibits less reward hacking. As shown in the first column of the third row, the second column of the fifth row, and the first and second columns of the sixth row, M_{HPSv2} tends to generate many meaningless accessories, objects not mentioned in the prompt, or decorative light effects and spots. This behavior suggests that the model is engaging in reward hacking through these elements, whereas M_{HPSv3} exhibits significantly less of this phenomenon. More results of DanceGRPO using HPSv3 as the reward model are shown in Figure 17.

9. Term of Use of HPDv3

Ownership and Responsibility. The HPDv3 dataset contains some parts of images obtained from the Internet, which are not the property of MizzenAI. MizzenAI is not responsible for the content or the meaning of these images.

Noncommercial Usage. Our funding resources, dataset, and models are strictly limited to noncommercial use. This aligns with the principle of ”fair use” as suggested by the United States Supreme Court for educational and research purposes. Any use of the HPDv3 dataset for commercial purposes is strictly prohibited.

Restrictions on Usage. You agree not to reproduce, duplicate, copy, sell, trade, resell, or exploit, for any commercial purposes, any portion of the images or any portion of derived data from the HPDv3 dataset. You also agree not to further copy, publish, or distribute any portion of the HPDv3 dataset. However, it is permitted to make copies of the dataset for internal use at a single site within the same organization.

Removal of Content. If you wish to have your content or product removed from the HPDv3 dataset, please contact us, and we will address your request promptly.

Acceptance of Terms. By using the HPDv3 dataset, you agree to comply with these Terms of Usage. Any violation of these terms may result in the termination of your access to the dataset and may lead to legal action.

Licensing Policy. To prevent unauthorized commercial usage of our dataset and models, we employ the ”CC BY-NC-SA” license (Creative Commons Attribution-NonCommercial-ShareAlike). This license permits others to freely share and adapt our work for non-commercial purposes, provided proper attribution is given, and derivative works maintain the same licensing terms. This measure ensures ethical distribution while preserving our original intent for non-commercial use.

Image Collection and Licensing Compliance. A significant portion of the real images in our dataset are sourced from Unsplash, a platform offering high-quality images through the CC0 license. The CC0 license permits unrestricted collection, distribution, and use of images, including free utilization for research and educational purposes. By inheriting these licensing terms, we ensure compliance with intellectual property standards while fostering free and open collaboration within the research community.

Commitment to Ethical and Fair Usage. We are committed to maintaining ethical standards in dataset construction and model development. All materials have been vetted to ensure adherence to licensing agreements and proper attribution where applicable. We encourage the broader community to uphold these ethical principles when utilizing our work, thereby fostering responsible research practices and avoiding any misuse of intellectual property.

10. Limitation

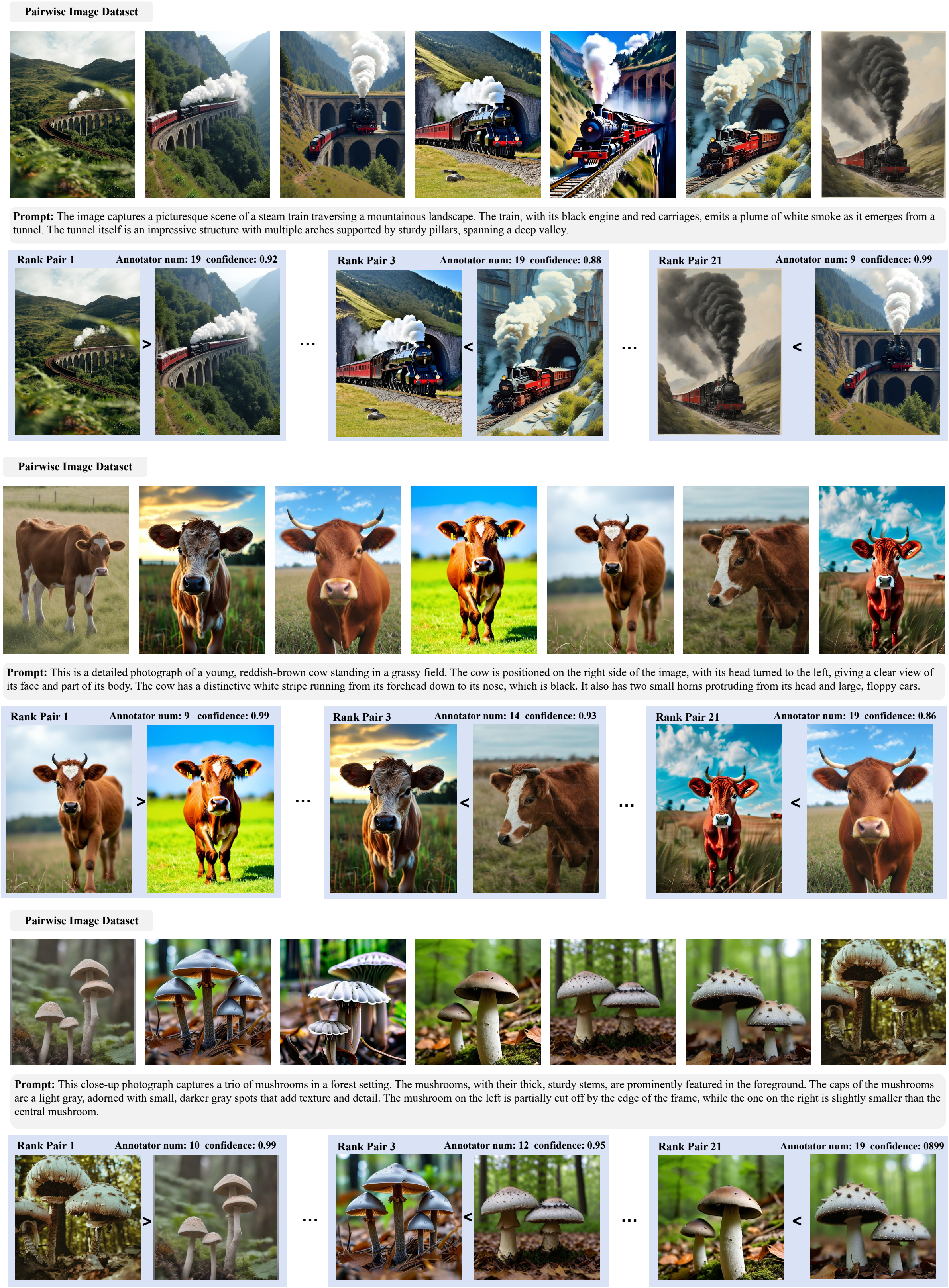
While HPDv3 contains 1.08M text-image pairs and 1.17M pairwise data, aiming to reflect real-world user preferences, it is important to acknowledge its inherent limitations, which may affect its generalizability and applicability in certain contexts.

Prompt Distribution Bias The dataset construction is primarily based on the prompt categories frequently observed in the JourneyDB database, which reflects general user input patterns. While this approach captures a broad range of typical generative use cases, it may inadvertently overlook specialized domains such as medicine, biology, physics, and other specialized fields requiring unique data. For example, generative models tailored for medical imaging or scientific diagram generation might not perform accurately when benchmarked with our dataset. This potential bias could limit the dataset’s usefulness for evaluating models designed for these specialized applications.

Unified Scoring Metric Our annotation pipeline employs a unified scoring metric to evaluate text-image pairs holistically across all dimensions. While this approach simplifies the evaluation process, it does not provide insights into more fine-grained dimensions such as color fidelity, artistic style, sharpness, or image clarity. This lack of granularity might hinder more detailed analysis and benchmarking of generative models, especially for applications where specific attributes are critical.

Annotator Demographics The dataset annotation process did not enforce strict demographic controls or categorizations. Information about annotators’ ethnicity, age, professional expertise, and cultural background was not collected or utilized during data annotation. As a result, the annotations may reflect unintended bias based on the subjective perspectives of the annotators. This lack of demographic diversity could reduce the robustness of the dataset in evaluating generative models designed for global or culturally sensitive contexts.

Challenges in Difficult Cases To ensure robust annotations, we adopted a multi-annotator approach for labeling text-image pairs, allowing feedback from multiple individuals to improve the reliability of scores. However, this approach encountered challenges when dealing with difficult or ambiguous cases. For prompts or images that were subjective or had conflicting interpretations, annotators often struggled to converge on a consistent score. These unresolved discrepancies can affect the accuracy of the dataset and limit its ability to serve as a definitive benchmark in such cases. Despite these challenges, the multi-annotator mechanism remains a valuable method for improving dataset reliability overall.



Prompt: The image captures a picturesque scene of a steam train traversing a mountainous landscape. The train, with its black engine and red carriages, emits a plume of white smoke as it emerges from a tunnel. The tunnel itself is an impressive structure with multiple arches supported by sturdy pillars, spanning a deep valley.

Rank Pair 1 Annotator num: 19 confidence: 0.92
>
...
Rank Pair 3 Annotator num: 19 confidence: 0.88
<
...
Rank Pair 21 Annotator num: 9 confidence: 0.99
<
Pairwise Image Dataset

Prompt: This is a detailed photograph of a young, reddish-brown cow standing in a grassy field. The cow is positioned on the right side of the image, with its head turned to the left, giving a clear view of its face and part of its body. The cow has a distinctive white stripe running from its forehead down to its nose, which is black. It also has two small horns protruding from its head and large, floppy ears.

Rank Pair 1 Annotator num: 9 confidence: 0.99
>
...
Rank Pair 3 Annotator num: 14 confidence: 0.93
<
...
Rank Pair 21 Annotator num: 19 confidence: 0.86
<
Pairwise Image Dataset

Prompt: This close-up photograph captures a trio of mushrooms in a forest setting. The mushrooms, with their thick, sturdy stems, are prominently featured in the foreground. The caps of the mushrooms are a light gray, adorned with small, darker gray spots that add texture and detail. The mushroom on the left is partially cut off by the edge of the frame, while the one on the right is slightly smaller than the central mushroom.

Rank Pair 1 Annotator num: 10 confidence: 0.99
>
...
Rank Pair 3 Annotator num: 12 confidence: 0.95
<
...
Rank Pair 21 Annotator num: 19 confidence: 0899
<

Figure 7. **HPDv3 Dataset Visualization.** Our dataset contains a diverse range of images spanning multiple categories including animals, architecture, characters, and other subjects. Each row displays different samples demonstrating the variety and quality of the dataset.

Animals:

- water color Bird similar to alex grey style , opal.
- an armadillo on a bicycle in the rain.
- American opossum in faerieland, cute, in the style of Maurice Sendak.
- The image showcases a magnificent peacock with its tail fully fanned out, displaying its vibrant plumage. The bird's body is a rich, deep blue, contrasting beautifully with the array of colors in its feathers. The head is adorned with a small crest of feathers, and a white marking extends from the beak along the side of the face.The tail feathers are the focal point, featuring iridescent ëyep̈atterns in shades of blue, green, and gold. Each eye is surrounded by a halo of darker hues, creating a mesmerizing effect. The feathers are long and slender, fanning out in a wide semi-circle that fills much of the frame. The overall impression is one of elegance and opulence, highlighting the peacock's role as a symbol of beauty and pride. The background is blurred, allowing the peacock's vibrant colors and intricate patterns to stand out.

Architecture:

- beautiful atmospheric picture of ghosts attacking New York, Y0̆14dkai, visually stunning, highly detailed, 8K,
- fantasy Christmas house, in a field of snow, fairy lights, by Gediminas Pranckevicius
- the palace of the Red Branch Knights
- The image captures a row of Tudor-style buildings under a cloudy sky. The buildings are characterized by their distinctive black and white timber framing, a hallmark of Tudor architecture. The black beams create vertical and diagonal patterns against the white-painted walls, giving the facades a striking contrast and a sense of depth. The roofs are steeply pitched and covered with dark tiles.Windows punctuate the facades, many of which feature small panes of glass and some with projecting bays. A British flag, known as the Union Jack, hangs from one of the buildings, adding a splash of color to the monochromatic scheme.

Arts:

- elf martial artist, meditating, misty background, pink flowers on the ground, fantasy art, oil painting
- dreamy watercolor painting of an angry witch and her ravens in a magical forest
- sketch, Abstract Purple background, illustration, watercolor::2.5,
- The image captures a tourist immersed in art appreciation within a museum setting. The focus is on the back of a young man with light-colored hair, wearing a white t-shirt, dark jeans, and a black backpack. He stands facing a display of classical sculptures.The sculptures include a prominent male nude figure with its arm raised, positioned to the left of a smaller statue. In the center, there is an ornate, large vase topped with a figure. Further to the right stands another statue, possibly of a young boy or man in traditional attire, positioned on a pedestal.

Characters:

- The image features a half cyborg girl character design with a nature meets technology theme, rendered with cinematic lighting and high detail.
- Anime girl in transparent holographic light suit black and yellow, Full body, three quarter length + poster + Guweiz + Cyberpunk, branding, texts, labels, three quarter body
- the most beautiful Andorran woman in the world,
- The image features a medium shot of a man with dark skin, his gaze directed straight at the viewer. He has a serious expression, with a neatly trimmed goatee and dark eyes that carry a piercing intensity. His hair is styled in thin braids, some falling around his face and shoulders, adding to his thoughtful appearance.He is wearing a beige or light-colored t-shirt that appears loose and comfortable, and gray jeans that are visibly worn, suggesting a casual, relaxed setting. On his left wrist, a white smartwatch or fitness tracker is visible.The background is split between a light gray wall on the left, which seems to have some faded graffiti or markings, and a dark void on the right, creating a stark contrast that draws attention to the man's figure. The lighting is soft and diffused, emphasizing the textures of his clothing and the depth of his expression. Overall, the image evokes a sense of introspection and quiet strength.

Design:

- illustration of bluebeard tale, with typographic placement , linocut, by Saul Bass
- A pen and copic marker sketch of the design of an simple robotic fish, white background, no paper background, no words, no pen or marker in photo, high quality
- vector symbol Sci fi Space 70s empty world
- This image shows a stylish and monochromatic interior design. The walls are painted in a deep, matte black, which serves as a dramatic backdrop for the decor. A black upholstered armchair with button detailing and four wooden legs sits prominently in the foreground, adding a touch of luxury. A circular black rug peeks from underneath the chair, enhancing the dark theme.On the wall, there is a gallery-style arrangement of framed black and white art, featuring the word V̈ogueänd fashion-themed images, complementing the sophisticated ambiance.

Figure 8. Representative examples of prompts from the HPDv3 Benchmark. For each category, we include a range of prompts varying from simple descriptions to highly detailed specifications. These prompts are used to generate image pairs from different AI models for calculating HPSv3 Score of each model.

Food:

- a handful of guava fruits drawn by david hockney, detailed, intricate
- sketch round cracker with chocolate Half enrobbed with color
- cartoon wine illustration, vector, simple clean, minimalist, wallpaper, bright, collection, in a set
- The image displays two tall glasses filled with a deep red beverage, likely a type of cocktail or juice, set against a dark background. The glasses are garnished with slices of grapefruit and sprigs of fresh mint, adding a vibrant splash of color and freshness. Ice cubes float within the drink, suggesting a refreshing and chilled experience.The wooden surface beneath the glasses has a weathered, rustic look, enhancing the natural and organic feel of the composition. One glass is positioned closer to the viewer, bringing the details of the drink and garnish into sharp focus, while the other is slightly blurred in the background, creating depth. The overall lighting is dramatic, with the dark backdrop emphasizing the brightness and color of the drinks and their garnishments. The composition is carefully arranged to evoke a sense of relaxation, freshness, and the enjoyment of a well-crafted beverage.

Natural Scenery:

- an epic outdoor sticker of the animas river with the san juan mountains in the background
- water ocean texture shot from bird perspective
- the view from inside a cosmic black hole 8k photorealistic
- The image presents a long, straight stretch of asphalt road leading towards a distant mountain range. The road takes up most of the foreground, its surface a dark gray, with two solid yellow lines running down the center, creating a strong sense of perspective. On either side of the road, there are dark, undefined areas of land.In the distance, snow-capped mountains dominate the horizon. Their jagged peaks and white surfaces contrast sharply with the dark land and road below, adding depth and a sense of grandeur to the scene.The sky above is partially cloudy, with patches of blue peeking through the white clouds. The clouds are scattered across the sky, adding texture and visual interest to the upper part of the image. In the far distance, there is a single car traveling on the road, adding a sense of scale and activity to the otherwise still landscape. The overall composition emphasizes the vastness and isolation of the natural setting.

Plants:

- fotorealistic jungle leaves, repetitive pattern, endless, mid-century modern style, geometric
- dark geisha pink flower with sword blue background ultra hd 8k
- garden full of golden flowers,
- The image presents a vast, golden field of ripe wheat under a clear blue sky. The wheat stalks dominate the foreground, their heads heavy with grain, creating a dense and textured visual tapestry. The color palette is dominated by warm tones, with the golden wheat contrasting nicely with the cool blue above.The wheat field stretches far into the distance, appearing to meet a slightly darker horizon line. A few scattered white clouds add a touch of lightness to the sky. The composition is simple yet striking, emphasizing the expansive nature of the landscape and the abundance of the harvest. The lighting is bright and sunny, casting gentle shadows within the field and highlighting the individual grains of wheat. Overall, the image evokes a sense of warmth, tranquility, and the bounty of nature.

Products:

- simple knolling, snow removal Snowblower and graders and loaders, white background
- Hand holding megaphone on bright yellow background with plenty of copy space. Magazine collage cut out style
- raw wrapping papers inspired Yeezy 350, hyper-detailed
- The image presents a striking contrast between black and red, with a small, wrapped gift placed on the black portion of the background. The background is divided diagonally, with a textured black area on the left and a smooth, vibrant red area on the right. This division creates a visually compelling composition.The gift itself is small and square, wrapped in shiny red foil. A bright red ribbon is tied around the package in a simple bow, with the ends of the ribbon elegantly trailing onto the black background. The arrangement of the ribbon adds a touch of movement and sophistication to the overall image.The dark background of the left side accentuates the metallic sheen of the foil wrapper and the brightness of the ribbon, making the gift stand out. The smooth red backdrop on the right provides a balance to the dark and textured side, contributing to a clean and modern aesthetic. Overall, the image evokes a sense of gift-giving and celebratory moments, with a clear focus on color contrast and simple elegance.

Figure 9. Representative examples of prompts from the HPDv3 Benchmark. For each category, we include a range of prompts varying from simple descriptions to highly detailed specifications. These prompts are used to generate image pairs from different AI models for calculating HPSv3 Score of each model.

Transportation:

- Sports Bike, race track, realism, 4K, No logo, Ducati, HDR, ar 3:2
- Porsche made from candy, beautiful, editorial photography, color graded, masterpiece
- sunken pirate ship, cinematic lighting, 4k, 8k, unreal engine, octane render
- The image showcases a collection of vintage automobiles, arranged neatly on a paved surface that seems to be a brick-patterned area. The backdrop includes lush greenery of trees and a unique architectural structure, possibly a modern park feature with a curved design, suggesting a blend of historical vehicles in a contemporary setting.Dominating the view are three cars of different makes. The first one, partially visible on the left, is a sleek, dark green, with a closed top. Adjacent to it is a classic black car with an open driver's area, a notable upright grille, and what looks like a soft top. Positioned in the foreground is a lighter-colored vehicle, likely a light grey or off-white, also with a soft top. It features spoked wheels and shiny headlamps. All three vehicles exhibit signs of age and careful preservation. The black and grey cars each have a sign or placard visible in the front window. Overall, the image captures a timeless elegance.

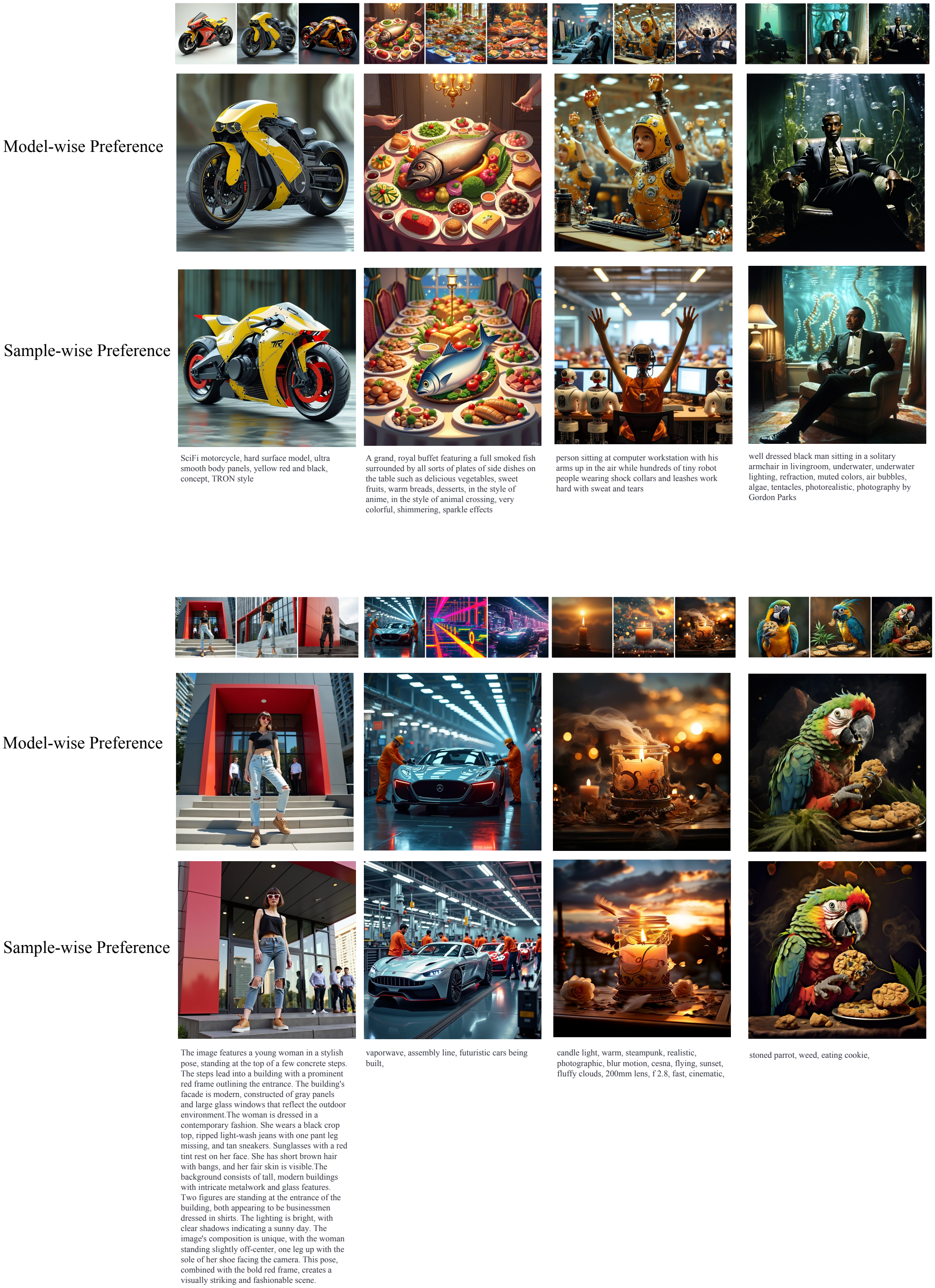
Science:

- a solar system view of a large space battle set in the Star Wars Universe, hyper realistic, 4k resolution
- 3D illustration of employees working in factory, 21st century
- futuristic neuronal network with robotic integration. Black background, 8K, hyperrealistic.
- reactor round underground scifi, hardsurface, HD, cinematography, low viewpoint, photorealistic, epic composition, Cinematic, Color Grading, portrait Photography, Ultra-Wide Angle, hyper-detailed, beautifully color-coded, insane details, intricate details, beautifully color graded, Unreal Engine, Cinematic, Color Grading, Editorial Photography, Photography, Photoshoot, Depth of Field, DOF, Tilt Blur, White Balance, 32k, Super-Resolution, Megapixel, ProPhoto RGB, VR, Halfrear Lighting, Backlight, Natural Lighting, Incandescent, Optical Fiber, Moody Lighting, Cinematic Lighting, Studio Lighting, Soft Lighting, Volumetric, Contre-Jour, Beautiful Lighting, Accent Lighting, Global Illumination, Screen Space Global Illumination, Ray Tracing Global Illumination, Optics, Scattering, Glowing, Shadows, Rough, Shimmering, Ray Tracing Reflections, Lumen Reflections, Screen Space Reflections, Diffraction Grading, Chromatic Aberration, GB Displacement, Scan Lines,

Others:

- realistic eyeball, pupil replaced with heart, up close, red, 1970s documentary photographs 35mm
- A memorial candle is lit on a table in a dark room close up vivid colours
- The image captures the vibrant energy of a live music festival. The foreground is filled with a large crowd, their heads a sea of diverse hairstyles and colors. Many are looking towards a massive stage structure, which dominates the left side of the frame.The stage is a complex scaffolding of metal and screens. The screens are illuminated with bright blue light, suggesting some sort of visual display accompanying the performance. A silhouette of a performer can be seen on stage, adding to the sense of a live event.Above the stage, the sky is a patchwork of white clouds against a bright blue background, creating a sense of expansive space. In the distance, the sun appears to be setting, casting a warm glow over the scene. To the right, trees in the background mark the natural setting of the festival. The overall composition captures the excitement and scale of a large-scale music event.
- The image showcases a single, small, earthen oil lamp, known as a diya, resting on a reflective surface. The diya is dark brown and has a rounded shape, typical of traditional Indian oil lamps. Atop the diya, a small flame flickers, providing the main source of light in the image.The backdrop is a soft, blurred yellow-orange gradient, creating a warm and cozy atmosphere. The surface on which the diya sits is dark, shiny, and reflective, mirroring the light from the flame. This reflection adds depth and dimension to the image, amplifying the glow and creating a sense of tranquility.The image is composed with simplicity and focus, drawing the viewer's attention to the delicate flame and the soft light it casts. The play of light and shadow, combined with the warm color palette, evokes feelings of peace and serenity. The diya symbolizes hope, positivity, and the triumph of light over darkness, often associated with festivals like Diwali.

Figure 10. Representative examples of prompts from the HPDv3 Benchmark. For each category, we include a range of prompts varying from simple descriptions to highly detailed specifications. These prompts are used to generate image pairs from different AI models for calculating HPSv3 Score of each model.



SciFi motorcycle, hard surface model, ultra smooth body panels, yellow red and black, concept, TRON style

A grand, royal buffet featuring a full smoked fish surrounded by all sorts of plates of side dishes on the table such as delicious vegetables, sweet fruits, warm breads, desserts, in the style of anime, in the style of animal crossing, very colorful, shimmering, sparkle effects

person sitting at computer workstation with his arms up in the air while hundreds of tiny robot people wearing shock collars and leashes work hard with sweat and tears

well dressed black man sitting in a solitary armchair in livingroom, underwater, underwater lighting, refraction, muted colors, air bubbles, algae, tentacles, photorealistic, photography by Gordon Parks

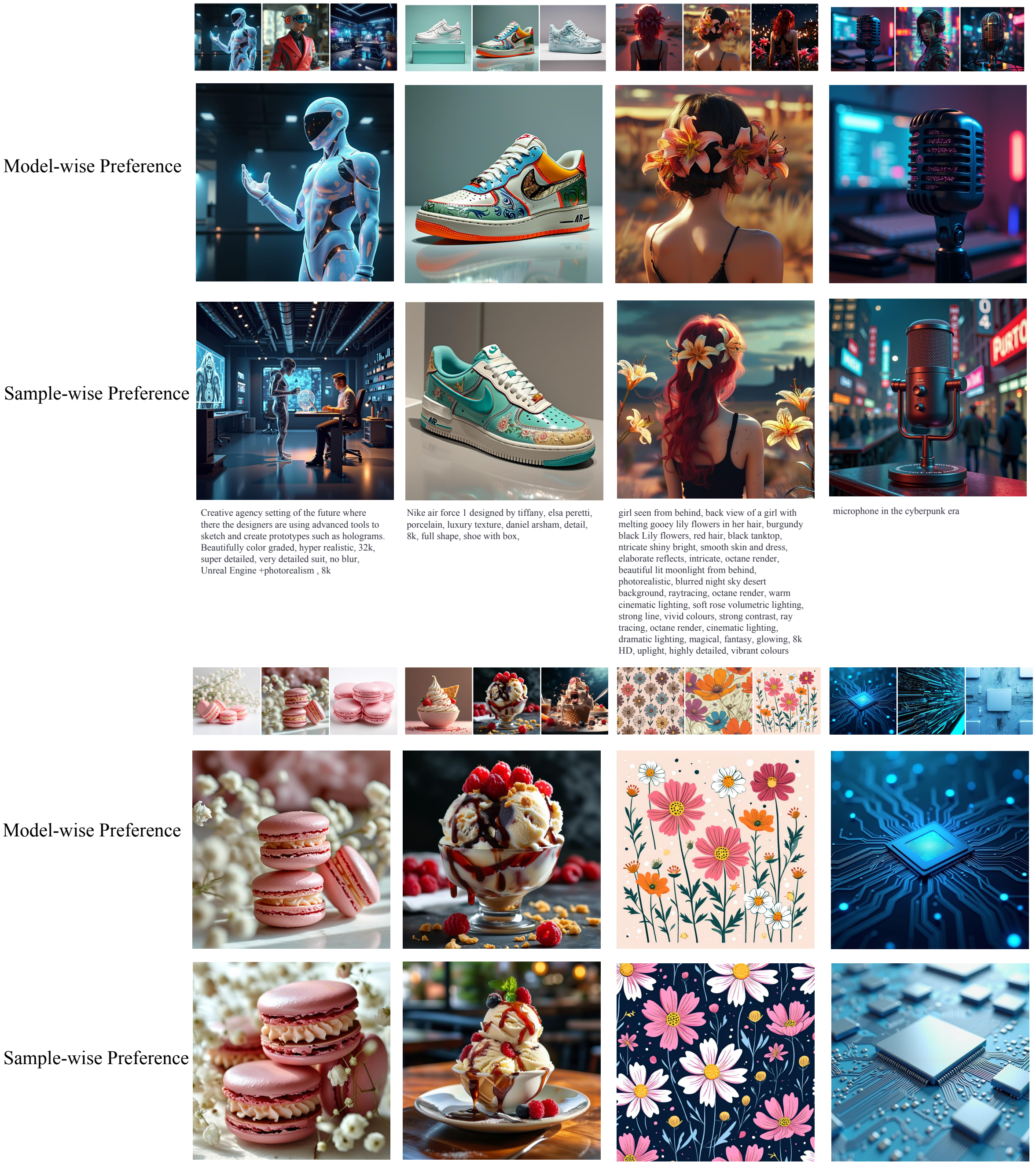
The image features a young woman in a stylish pose, standing at the top of a few concrete steps. The steps lead into a building with a prominent red frame outlining the entrance. The building's facade is modern, constructed of gray panels and large glass windows that reflect the outdoor environment.The woman is dressed in a contemporary fashion. She wears a black crop top, ripped light-wash jeans with one pant leg missing, and tan sneakers. Sunglasses with a red tint rest on her face. She has short brown hair with bangs, and her fair skin is visible.The background consists of tall, modern buildings with intricate metalwork and glass features. Two figures are standing at the entrance of the building, both appearing to be businessmen dressed in shirts. The lighting is bright, with clear shadows indicating a sunny day. The image's composition is unique, with the woman standing slightly off-center, one leg up with the sole of her shoe facing the camera. This pose, combined with the bold red frame, creates a visually striking and fashionable scene.

vaporwave, assembly line, futuristic cars being built,

candle light, warm, steampunk, realistic, photographic, blur motion, cesna, flying, sunset, fluffy clouds, 200mm lens, f 2.8, fast, cinematic,

stoned parrot, weed, eating cookie,

Figure 11. **More Result of CoHP**



Creative agency setting of the future where there the designers are using advanced tools to sketch and create prototypes such as holograms. Beautifully color graded, hyper realistic, 32k, super detailed, very detailed suit, no blur, Unreal Engine +photorealism , 8k

Nike air force 1 designed by tiffany, elsa peretti, porcelain, luxury texture, daniel arsham, detail, 8k, full shape, shoe with box,

girl seen from behind, back view of a girl with melting gooey lily flowers in her hair, burgundy black Lily flowers, red hair, black tanktop, ntricate shiny bright, smooth skin and dress, elaborate reflects, intricate, octane render, beautiful lit moonlight from behind, photorealistic, blurred night sky desert background, raytracing, octane render, warm cinematic lighting, soft rose volumetric lighting, strong line, vivid colours, strong contrast, ray tracing, octane render, cinematic lighting, dramatic lighting, magical, fantasy, glowing, 8k HD, uplight, highly detailed, vibrant colours

microphone in the cyberpunk era

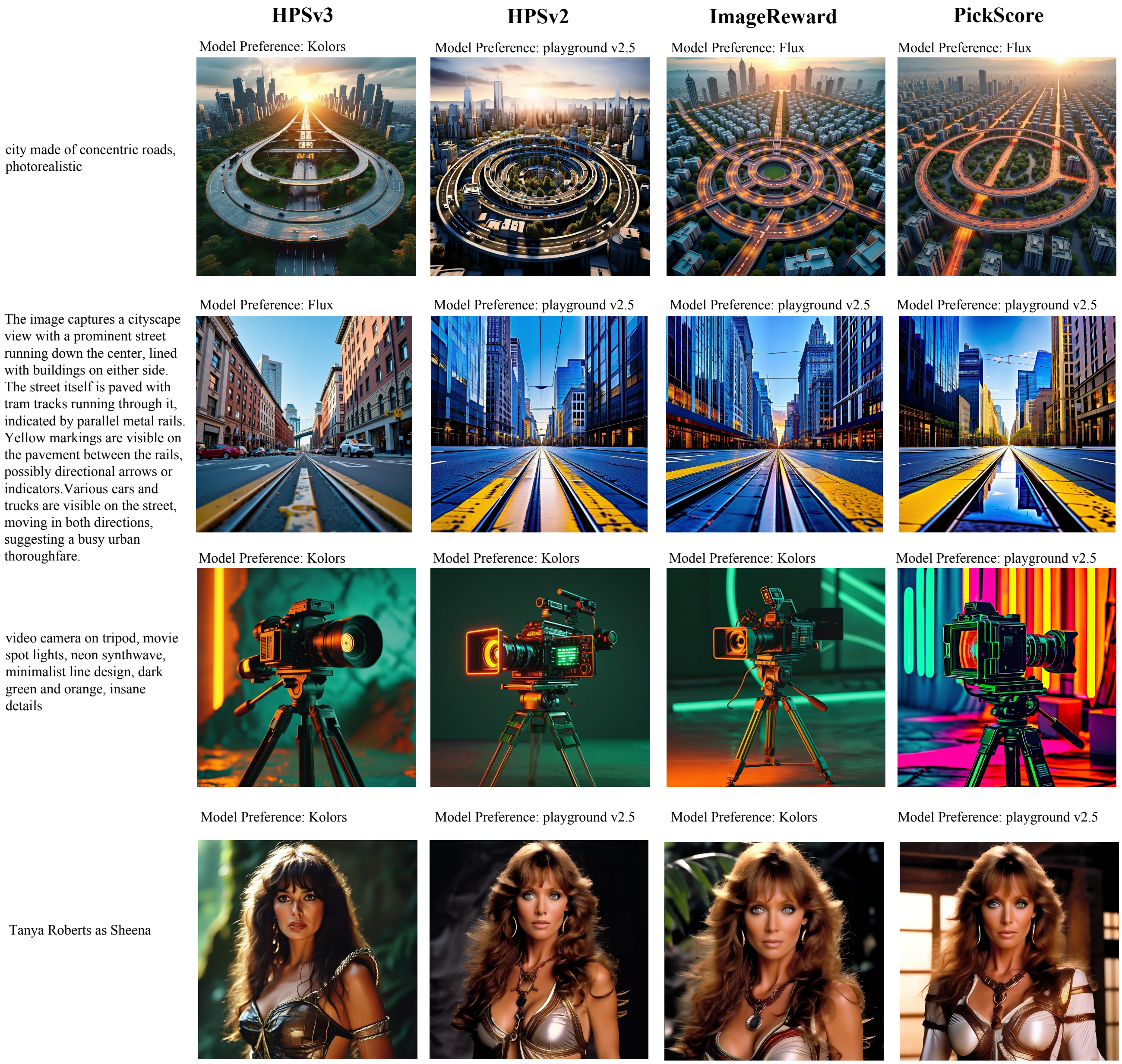
The image showcases a close-up of four pink macarons artfully arranged on a white surface. The macarons, the focal point of the composition, have a smooth, delicate shell with a slightly textured edge, revealing a creamy filling in the center. The color of the macarons is a soft, inviting pink, adding a touch of elegance to the image.In the background, out-of-focus white flowers, possibly baby's breath, create a soft and airy ambiance. The blurred backdrop adds depth to the image, drawing attention to the macarons while providing a gentle, romantic setting.The lighting is soft and diffused, casting gentle shadows that emphasize the texture of the macarons and flowers. The overall composition is clean and balanced, with a focus on color and texture. The image evokes a sense of delicacy and indulgence, making it visually appealing.

a dessert

simple vector pattern Cosmos Flower simple boho color

sleek futuristic textured desktop wallpaper, computer chip print design using a light blue colour pallet

Figure 12. **More Result of CoHP**



city made of concentric roads, photorealistic

The image captures a cityscape view with a prominent street running down the center, lined with buildings on either side. The street itself is paved with tram tracks running through it, indicated by parallel metal rails. Yellow markings are visible on the pavement between the rails, possibly directional arrows or indicators.Various cars and trucks are visible on the street, moving in both directions, suggesting a busy urban thoroughfare.

video camera on tripod, movie spot lights, neon synthwave, minimalist line design, dark green and orange, insane details

Tanya Roberts as Sheena

Figure 13. **More Result of CoHP with different human preference models**

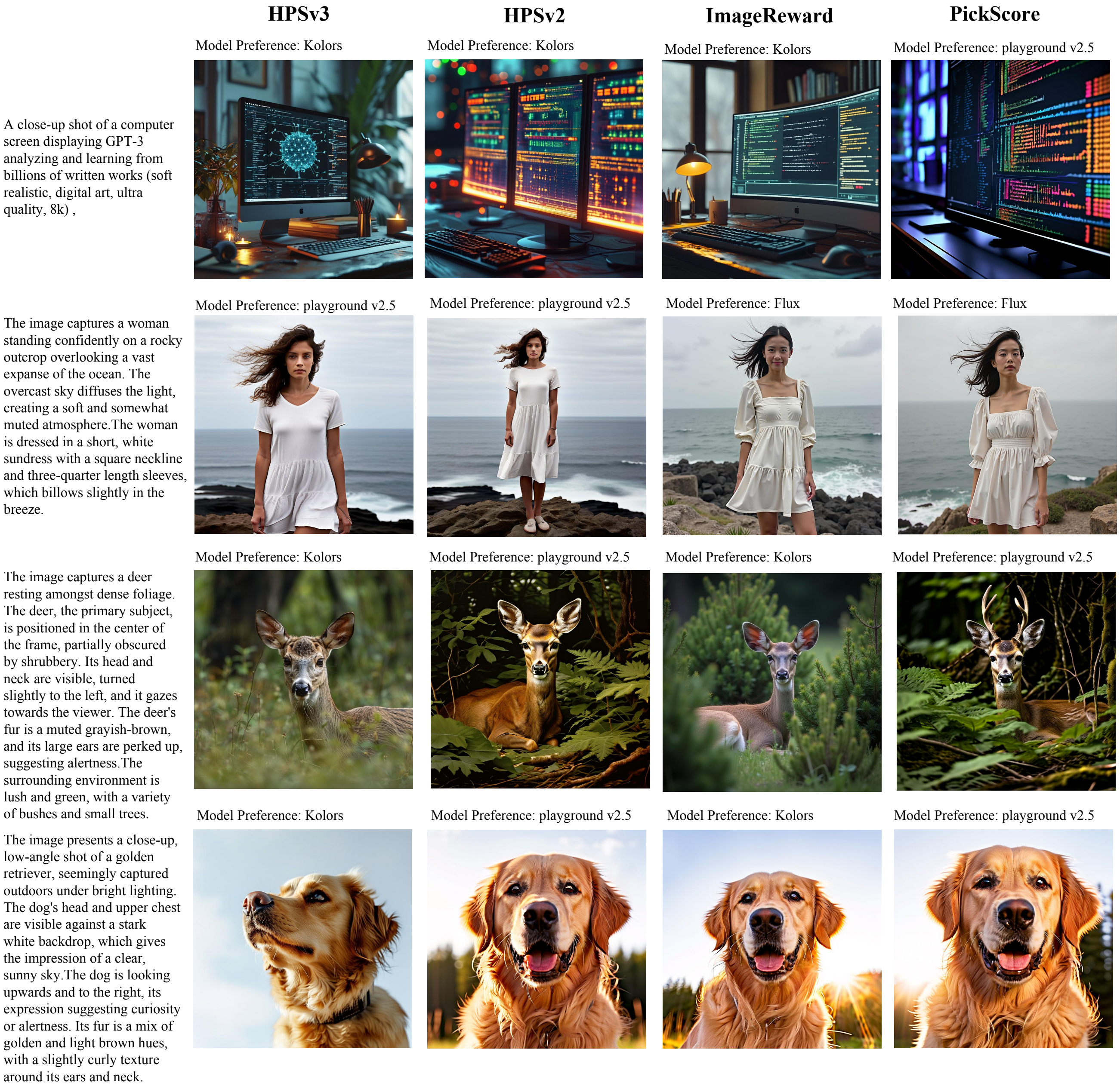
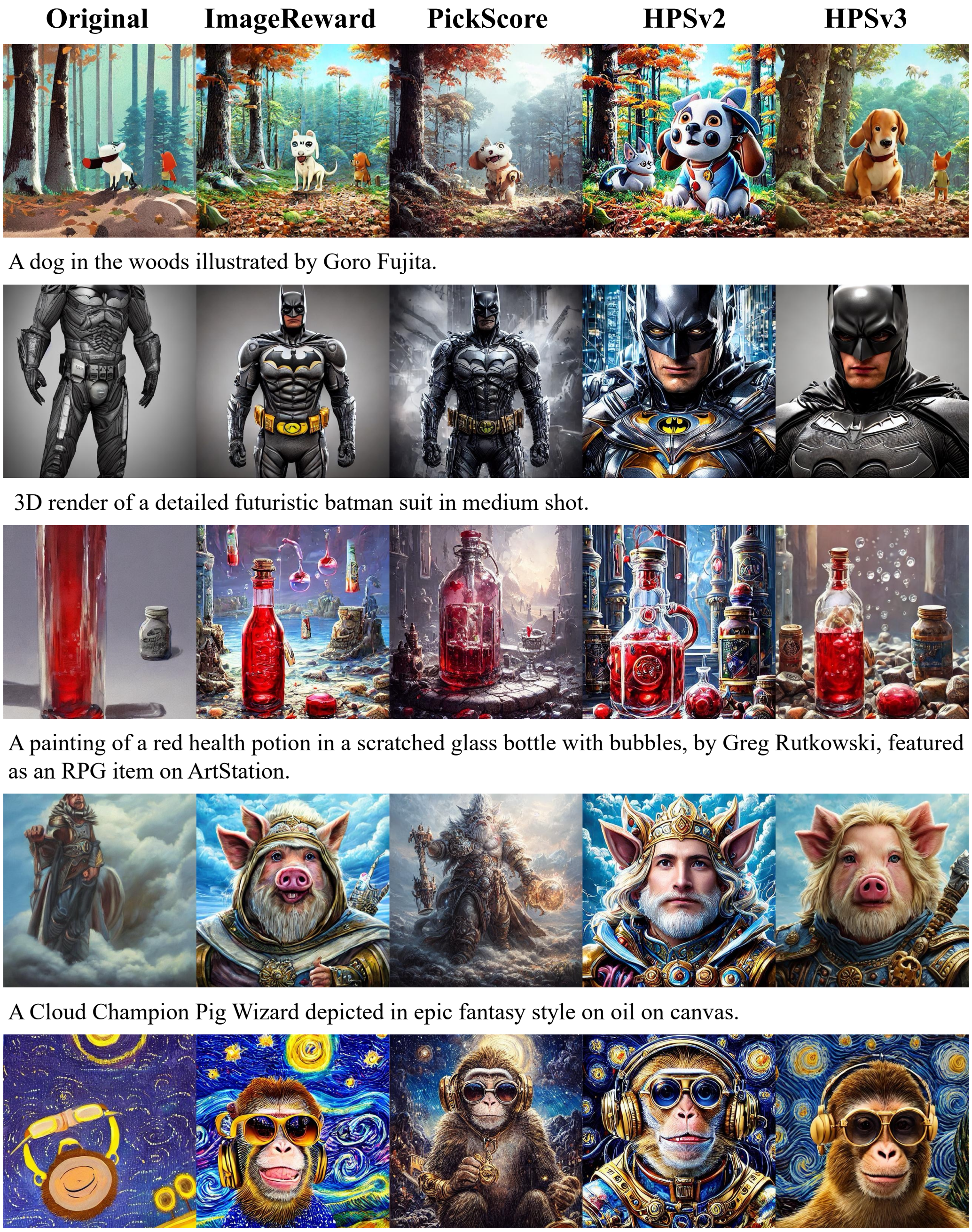

HPSv3
HPSv2
ImageReward
PickScore
A close-up shot of a computer screen displaying GPT-3 analyzing and learning from billions of written works (soft realistic, digital art, ultra quality, 8k) ,
Model Preference: Kolors
Model Preference: Kolors
Model Preference: Kolors
Model Preference: playground v2.5
The image captures a woman standing confidently on a rocky outcrop overlooking a vast expanse of the ocean. The overcast sky diffuses the light, creating a soft and somewhat muted atmosphere.The woman is dressed in a short, white sundress with a square neckline and three-quarter length sleeves, which billows slightly in the breeze.
Model Preference: playground v2.5
Model Preference: playground v2.5
Model Preference: Flux
Model Preference: Flux
The image captures a deer resting amongst dense foliage. The deer, the primary subject, is positioned in the center of the frame, partially obscured by shrubbery. Its head and neck are visible, turned slightly to the left, and it gazes towards the viewer. The deer's fur is a muted grayish-brown, and its large ears are perked up, suggesting alertness.The surrounding environment is lush and green, with a variety of bushes and small trees.
Model Preference: Kolors
Model Preference: playground v2.5
Model Preference: Kolors
Model Preference: playground v2.5
The image presents a close-up, low-angle shot of a golden retriever, seemingly captured outdoors under bright lighting. The dog's head and upper chest are visible against a stark white backdrop, which gives the impression of a clear, sunny sky.The dog is looking upwards and to the right, its expression suggesting curiosity or alertness. Its fur is a mix of golden and light brown hues, with a slightly curly texture around its ears and neck.
Model Preference: Kolors
Model Preference: playground v2.5
Model Preference: Kolors
Model Preference: playground v2.5

Figure 14. **More Result of CoHP with different human preference models**



A dog in the woods illustrated by Goro Fujita.

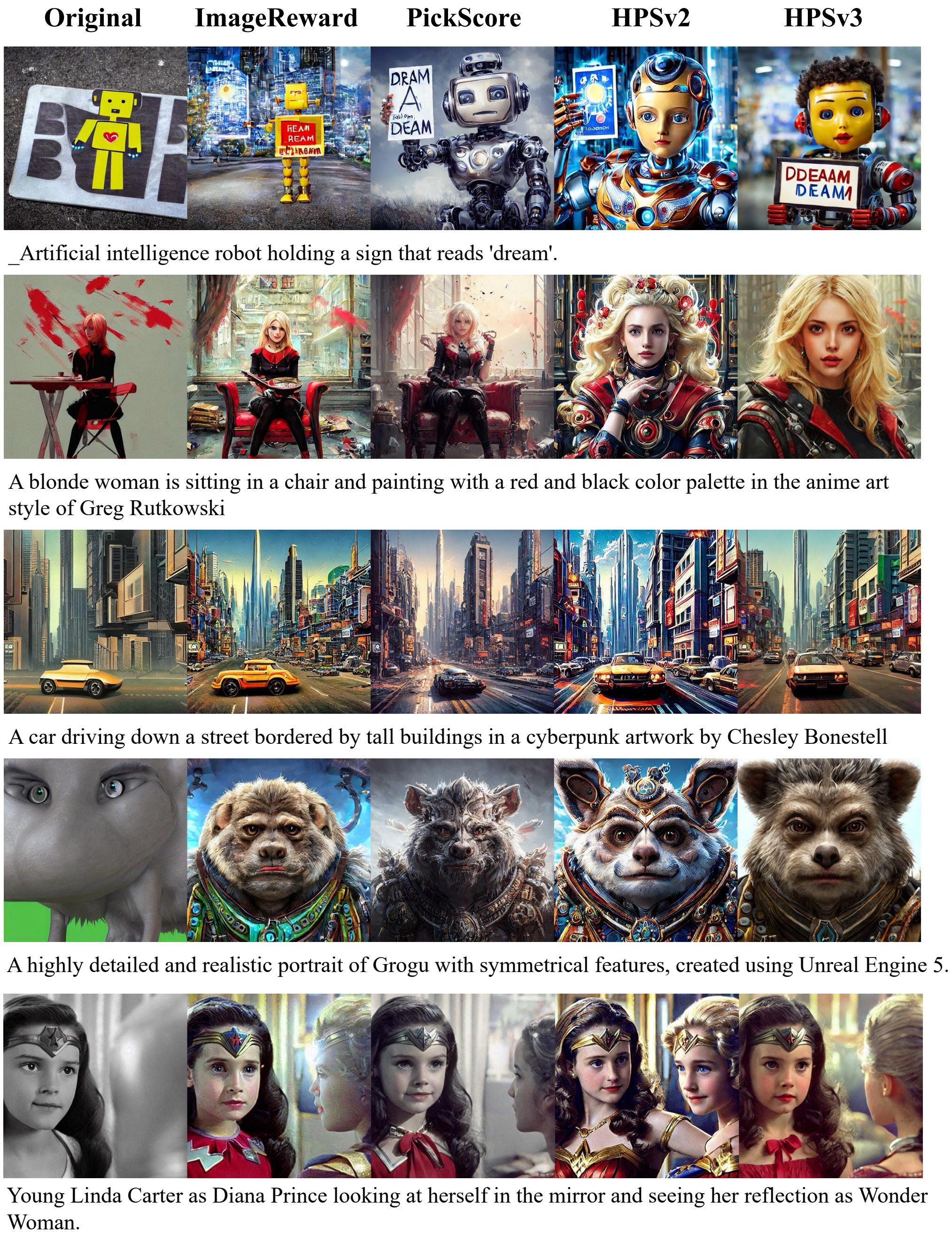
3D render of a detailed futuristic batman suit in medium shot.

A painting of a red health potion in a scratched glass bottle with bubbles, by Greg Rutkowski, featured as an RPG item on ArtStation.

A Cloud Champion Pig Wizard depicted in epic fantasy style on oil on canvas.

A painting of a monkey wearing gold headphones and sunglasses looking up at a starry night sky.

Figure 15. **Results of DanceGRPO Using Different human preference models**



_Artificial intelligence robot holding a sign that reads 'dream'.

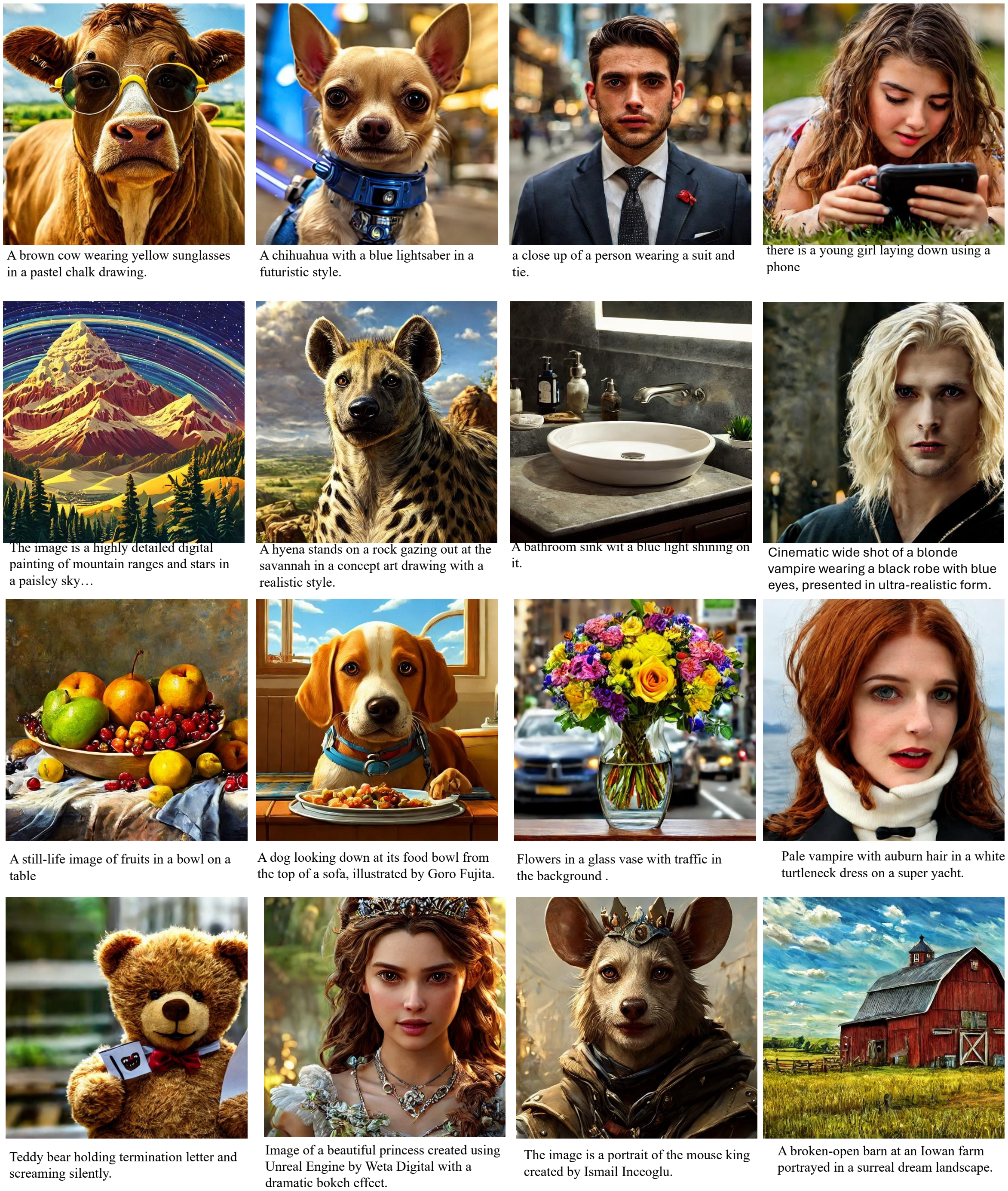
A blonde woman is sitting in a chair and painting with a red and black color palette in the anime art style of Greg Rutkowski

A car driving down a street bordered by tall buildings in a cyberpunk artwork by Chesley Bonestell

A highly detailed and realistic portrait of Grogu with symmetrical features, created using Unreal Engine 5.

Young Linda Carter as Diana Prince looking at herself in the mirror and seeing her reflection as Wonder Woman.

Figure 16. **Results of DanceGRPO Using Different human preference models**



A brown cow wearing yellow sunglasses in a pastel chalk drawing.

A chihuahua with a blue lightsaber in a futuristic style.

a close up of a person wearing a suit and tie.

there is a young girl laying down using a phone

The image is a highly detailed digital painting of mountain ranges and stars in a paisley sky…

A hyena stands on a rock gazing out at the savannah in a concept art drawing with a realistic style.

A bathroom sink wit a blue light shining on it.

Cinematic wide shot of a blonde vampire wearing a black robe with blue eyes, presented in ultra-realistic form.

A still-life image of fruits in a bowl on a table

A dog looking down at its food bowl from the top of a sofa, illustrated by Goro Fujita.

Flowers in a glass vase with traffic in the background .

Pale vampire with auburn hair in a white turtleneck dress on a super yacht.

Teddy bear holding termination letter and screaming silently.

Image of a beautiful princess created using Unreal Engine by Weta Digital with a dramatic bokeh effect.

The image is a portrait of the mouse king created by Ismail Inceoglu.

A broken-open barn at an Iowan farm portrayed in a surreal dream landscape.

Figure 17. **More Results of DanceGRPO Using HPSv3 as Reward Model**

References

[1] Black Forest Lab. Flux.1 [dev]: 12b-parameter open-source text-to-image diffusion model, 2024. Architecture: 12B-parameter diffusion model with T5-XXL text encoder; Features: Anatomically precise human generation, native multilingual support, commercial-grade text rendering; License: Non-commercial research use (Apache 2.0 with additional restrictions); Model Size: 23.8GB FP16 weights.

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[3] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.

[4] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *NeurIPS*, 35:16890–16902, 2022.

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

[7] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

[8] 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. *arXiv preprint arXiv:2305.01569*, 2023.

[9] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[10] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021.

[11] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 2021.

[12] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[13] 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 (CVPR)*, pages 10684–10695, 2022.

[14] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*, pages 10674–10685, 2022.

[15] Christoph Schuhmann. Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor.

[16] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023.

[17] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

[18] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023.

[19] 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, 2023.

[20] 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, 2024.

[21] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[22] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation, 2024.

[23] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion, 2024.