Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation

Daiqing Li **Aleks Kamko** **Ehsan Akhgari** **Ali Sabet** **Linmiao Xu** **Suhail Doshi**

Playground Research

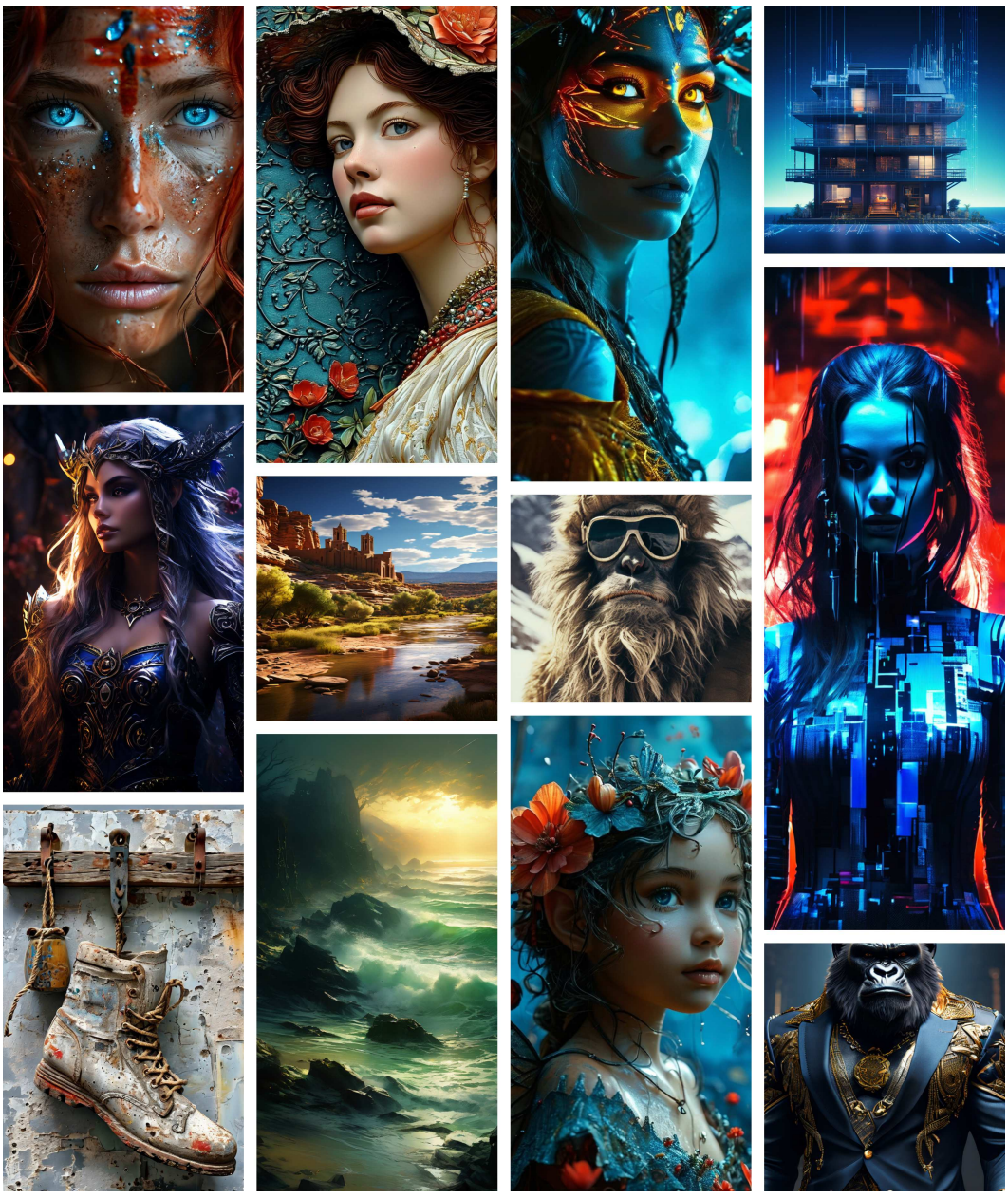


Figure 1: **High-quality samples from Playground v2.5.** Our model exhibits vibrant color and contrast on a range of image styles. Samples generated by the Playground community members.

Abstract

In this work, we share three insights for achieving state-of-the-art aesthetic quality in text-to-image generative models. We focus on three critical aspects for model improvement: enhancing color and contrast, improving generation across multiple aspect ratios, and improving human-centric fine details. First, we delve into the significance of the noise schedule in training a diffusion model, demonstrating its profound impact on realism and visual fidelity. Second, we address the challenge of accommodating various aspect ratios in image generation, emphasizing the importance of preparing a balanced bucketed dataset. Lastly, we investigate the crucial role of aligning model outputs with human preferences, ensuring that generated images resonate with human perceptual expectations. Through extensive analysis and experiments, Playground v2.5 demonstrates state-of-the-art performance in terms of aesthetic quality under various conditions and aspect ratios, outperforming both widely-used open-source models like SDXL [28] and Playground v2 [20], and closed-source commercial systems such as DALL·E 3 [2] and Midjourney v5.2. Our model is open-source, and we hope the development of Playground v2.5 provides valuable guidelines for researchers aiming to elevate the aesthetic quality of diffusion-based image generation models.

1 Introduction

Great progress has been made in diffusion-based generative models since the success of better image modeling performance [12, 31, 6] with ImageNet, as compared to performance with the previously dominate framework of generative adversarial networks (GAN) [7, 15, 16]. Open-source models like SDXL [28] have built on top of latent diffusion models (LDM) [29] by scaling up text-to-image pre-training datasets [30] and the latent UNet [6] architecture. PixArt-alpha [3], on the other hand, explores Diffusion Transformer (DiT) [27] as the latent backbone, showing better training efficiency and image quality. Playground v2 [20], an open-source model recently developed by us, focuses on the training recipe and aesthetic quality, achieving $2.5\times$ higher user preference compared to SDXL [28].

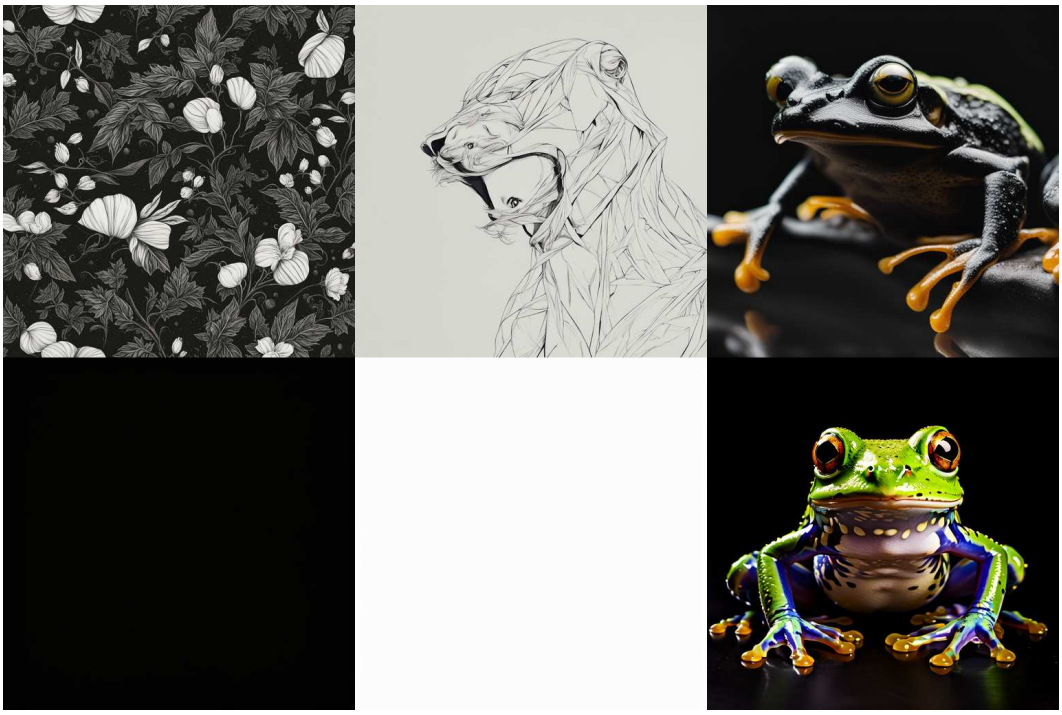
Playground v2 [20] was open-sourced in December 2023 and we were pleased to see the open-source and research community take up our work and reference it. Notably, Playground v2 has amassed over 135,000 downloads in just the last month from HuggingFace, and our work has been cited in recent papers for state-of-the-art image models such as Stable Cascade [1]. Following Playground v2 [20], we chose not to change the underlying model architecture for this work; instead, we focused on analyzing and improving our training recipe and pushing the model's aesthetic quality to a new level.

We focus on three critical issues for image models: enhancing color and contrast (sec. 2.1), improving generation across multiple aspect ratios (sec. 2.2), and improving human-centric fine details (sec. 2.3). More generally, we aim to refine the model's capabilities to produce more realistic and visually compelling outputs. To evaluate the efficacy of our enhancements, we conducted extensive user studies and benchmarked our model against previous state-of-the-art models (sec. 3). We also propose a new automatic-evaluation benchmark *MJHQ-30K* (sec. 3.5) to evaluate the model's performance against 10 unique categories. When evaluating our model's outputs on human preference, we are thrilled to report that Playground v2.5 surpasses state-of-the-art models, including Midjourney 5.2, DALL·E 3 [2], Playground v2 [20], PIXART-α [3], and SDXL [28] (Fig 10). See sec. 3.2 for details. Playground v2.5 endeavors to surpass the performance of its predecessor and establish itself as a leading contender in the space of text-to-image generative models.

We open-source the weights for Playground v2.5, on HuggingFace[1] with a license[2] that makes it easy for research teams to use. We will also provide extensions for using our model in A1111 and ComfyUI, two popular community tools for using diffusion models. Given how much we have benefited from the research and open-source communities, it is important that we make multiple aspects of our work on Playground v2.5 available to the community.

[1] https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic

[2] https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic/blob/main/LICENSE.md



a solid black background | a solid white background | frog in front of solid black background

(a) **Generating solid backgrounds.** The top row is sampled from SDXL [28], bottom row is Playground v2.5. SDXL fails to generate pure black or white background while our model can follow the prompt faithfully.

A portrait of an African American woman on a black background at night, photo realistic, ultra realistic | Black UFO in front of a rising sun at the dawn, vibrant colours, dark surroundings, photorealisti | A black obelisk in front of a white background | A town square lit only by torchlight

(b) **Colors and contrast**. The top row is SDXL, bottom row is Playground v2.5. Our model can generate samples with more vibrant colors and contrast.

Figure 2: **Comparing SDXL and Playground v2.5 in generating images with vibrant color and contrast.**

Peter Pan aged 60 years old, with a black background

Bilibin ink detailed masterpiece, National Geographic gracious, filigree acrylic, Slavic folklore, tender face, storybook illustration, art on a cracked wood, young beautiful Red Riding Hood girl, book illustration style, forest, mushrooms, hyperrealism, digital art, cinematic, close portrait, highly detailed expressive glowing eyes, airy, detailed face, shadow play, realistic textures, dynamic pose, unusual, modern. heartwarming, cozy, fairytale, fantasy, detailed textures, artistic dynamic pose, tender, atmospheric, sharp focus, centered composition, complex background, soft haze, masterpiece. animalistic, beautiful, tiny detailed

happy dreamy owl monster sitting on a tree branch, colorful glittering particles, forest background, contoured, surrealism, close up cute, detailed feathers, bioluminescence, leaves, ethereal, ice, looking in camera, sky, sleek, modern, fairytale, fantasy, by Andy Kehoe

Figure 3: **Comparing Playground v2 [20] and v2.5 for color and contrast with more complex prompts.** Top row is Playground v2, bottom row is Playground v2.5. Compared to v2, v2.5 dramatically improves the color and contrast, and the ability to follow style-related prompts.

2 Methods

2.1 Enhanced Color and Contrast

Latent diffusion models have struggled to generate images with high color contrast and vibrant color range since the release of SD1.5. This is a known limitation [21, 4, 13]. For example, SDXL is incapable of generating a pure black image or a pure white image, and fails to place subjects onto solid backgrounds (see Fig 2 (a)).

This issue stems from the noise scheduling of the diffusion process, as pointed out by [21]. The signal-to-noise ratio [17] of Stable Diffusion is too high, even when the discrete noise level reaches its maximum. Several works have attempted to fix this flaw. Guttenberg and CrossLabs [8] propose offset noise. Lin et al. [21] propose Zero Terminal SNR to ensure the last denoising step is pure Gaussian noise. SDXL [28] adopts the strategy of adding offset noise in the last stage of the training, as does Playground v2. However, as can be seen in Fig 2 (b), we still notice SDXL exhibits muted color and contrast.

human-like octopus sitting in a recliner with a human in fish tank on his side table.

8-bit color forest fire visual effect

teampunk vivid illuminated iridescent intricate mechanical racing motorcycle with intricate brown cogwheels sticker, white background, contour, colorful, vector, kawaii, hdr, Watercolor, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowski"

Figure 4: **Qualitative comparison of portrait aspect ratios.** Aspect ratio 3:4 (876×1168), top is SDXL, bottom is Playground v2.5. Our model can generate images with the correct composition to the desired aspect ratio

generate an image of AI-controlled aquatic drones assisting marine biologists in studying the behavior of plankton communities., Miki Asai Macro photography, close-up, hyper detailed, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowsk.

tattoo sketch chicano style, beauty and pit bull, blood, dynamics.

image of a jade green and gold coloured Fabergé egg, 16k resolution, highly detailed, product photography, trending on artstation, sharp focus, studio photo, intricate details, fairly dark background, perfect lighting, perfect composition, sharp features, Miki Asai Macro photography, close-up, hyper detailed, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowski."

Figure 5: **Qualitative comparison of landscape aspect ratios.** Aspect ratio 4:3 (1168×876), the top is SDXL, bottom is Playground v2.5. Our model can generate content following the prompt consistently under extreme aspect ratios, while SDXL sometimes fails to follow the prompt, or generates multiple objects under wide aspect ratios.

For Playground v2.5, we aimed to dramatically improve upon this issue. We wanted to achieve vivid color and contrast in imagery and be able to produce pure-colored backgrounds. To this end, we took a more principled approach and trained our models from scratch using the EDM framework, proposed by Karras et al [14].

EDM brings two distinct advantages: (1) Like Zero Terminal SNR, the EDM noise schedule exhibits a near-zero signal-to-noise ratio for the final "timestep". This removes the need for Offset Noise and fixes muted colors. (2) EDM takes a first-principles approach to designing the training and sampling processes, as well as preconditioning of the UNet. This enables the EDM authors to make clear design choices that lead to better image quality and faster model convergence.

We were also inspired by Hoogeboom et al [13] to skew the noise schedule towards an overall noisier one when training on high-resolution images.

In Fig 3, we show a qualitative comparison between Playground v2.5 and Playground v2, the latter of which uses offset noise and a DDPM [12] noise schedule. We see in the first column that Playground v2.5 can generate a vivid portrait with enhanced color range, and it exhibits better prompt-image alignment, which enables v2.5 to generate a pure black background.

2.2 Generation Across Multiple Aspect Ratios

The ability to generate images of various different aspect ratios is an important feature in real-world applications of text-to-image models. However, common pre-training procedures [29, 28] for these models start by training only on square images in the early stages, with random or center cropping. This technique is standard practice from conditional generative models trained on ImageNet [6, 23, 14].

Theoretically, this should not pose a problem. A diffusion model like SDXL [29, 28] consisting of mostly convolution layers – mimicking a Convolutional Neural Network (CNN) – should work with any input resolution at inference time, even if it was not trained on that particular resolution. This is due to the transition-invariant [19, 9] property of CNNs. Unfortunately, in practice, diffusion models do not generalize well to other aspect ratios when only trained on square images, as pointed out by NovelAI [24].

To address this challenge, NovelAI proposes doing bucketed sampling, where images with similar aspect ratios are bucketed together in the same forward pass. SDXL [28] adopted this bucketing strategy and also added additional conditioning to specify the source and target image sizes.

SDXL's conditioning strategy forced the model to learn to place the image's subject at the center under different aspect ratios. However, due to an unbalanced distribution of the aspect ratio buckets [28] in SDXL's dataset, i.e. the majority of the dataset's images are square, SDXL also learned the bias of certain aspect ratios in its conditioning. Furthermore, images generated at non-square aspect ratios typically exhibit much lower quality than square images.

In Playground v2.5, one of our explicit goals was to make the model reliably produce high-quality images at multiple aspect ratios, as we had learned from the users that this is a critical element for a high-quality production-grade model. While we followed a bucketing strategy similar to SDXL's, we carefully crafted the data pipeline to ensure a more balanced bucket sampling strategy across various aspect ratios. Our strategy avoids catastrophic forgetting and helps the model not be biased towards one ratio or another.

Fig 4 and Fig 5 show a qualitative comparison between SDXL and Playground v2.5 across portrait and landscape aspect ratios, respectively. Our model can generate high-quality images under various aspect ratios without errors like generating multiple objects or wrong composition.

2.3 Human Preference Alignment

Humans are particularly sensitive to visual errors on human features like hands, faces, and torsos. An image with perfect lighting, composition, and style will likely be voted as low-quality if the face of the human is malformed or the structure of the body is contorted.

Generative models, both language and image, are prone to hallucination. In image models, this can manifest as malformed human features. There are multiple reasons for hallucination, but one evident explanation is a misaligned training objective: generative models are trained to maximize the log-likelihood of the data rather than maximizing human preference. In LLMs, a common strategy to align pre-trained generative models with human preference is called supervised fine-tuning, or SFT. In short [25], SFT fine-tunes a pre-trained base model with a small but very-high-quality dataset. This simple technique often outperforms a more complicated approach like RLHF [33]. However, the question of how to best curate an SFT alignment dataset from different sources to maximize performance on a downstream task remains an ongoing research problem [32].

One of our goals with Playground v2.5 was to reduce the likelihood of visual errors in human features, which is a common critique of open-source diffusion models more broadly, as compared to closed-source models like Midjourney. Emu [5] introduces an alignment strategy similar to SFT for text-to-image generative models. Inspired by Emu, we developed a system that enables us to automatically curate a high-quality dataset from multiple sources via user ratings. Then, we took an iterative, human-in-the-loop [26, 22] training approach to select the best dataset candidates. We monitored the training progress by empirical evaluation, comparing our aligned models by looking at image grids generated from a fixed set of prompts, similar to [33].

Our novel alignment strategy enables us to excel over SDXL in at least four important human-centric categories:

- Facial detail, clarity, and liveliness
- Eye shape and gaze
- Hair texture
- Overall lighting, color, saturation, and depth-of-field

We chose to focus on these categories based on usage patterns in our product and from user feedback.

Seasoned fisherman portrait, weathered skin etched with deep wrinkles, white beard, piercing gaze beneath a fisherman's hat, softly blurred dock background accentuating rugged features, captured under natural light, ultra-realistic, high dynamic range photo

Close-up portrait of a face with big eyes, overflowing with a mysterious, ethereal ambiance, capturing a direct gaze with the viewer amidst the chiaroscuro interplay of light and shadow, featuring soft colors alongside a splash of color from variously hued flowers, all set against a green retro Gothic night scene reminiscent of Mr X's style, face illuminated, reflective surfaces enhancing the rich, vivid textures, and the delicate details enhanced by octane rendering, cinematic portrayal

An old man with a flat cap for his head in a market close up, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowski

Figure 6: **Human aesthetic preference alignment comparison with SDXL.** Top row is SDXL, bottom row is Playground v2.5. Our model can generate better human-centric facial details, overall lighting, color and depth-of-field.

Playground v2.5 | Midjourney 5.2 | DALL·E 3 [2] | Playground v2 [20] | PixArt-alpha [3] | SDXL-1.0 [28]

Figure 7: **Qualitative comparison between methods.** Prompts for the top row: "a person with a feeling of dryness in the mouth.", the bottom row: "a jeweler's hands, each holding a tiny gemstone, aligning them perfectly for a custom ring.". Our model can generate lively expressions, fine-details of a person's teeth, eyes, and expression, and the correct hands.

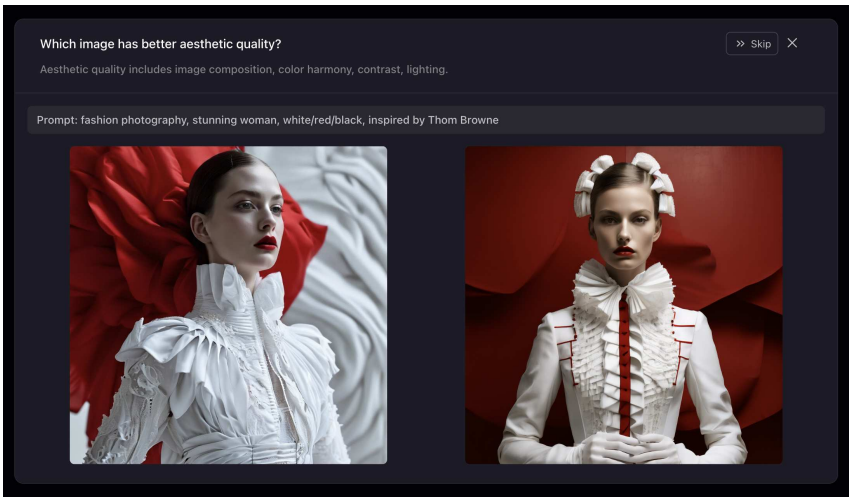


Figure 8: An example of an image pair shown to a user in our product.

In Fig 6, we showcase some examples of the difference in fine details between images generated using Playground v2.5 and those using SDXL. In Fig 7, we compare our model with other SoTA methods in generating human-centric images.

3 Evaluations

3.1 User Study Interface

Since we ultimately build our models to be used by our hundreds of thousands of users, it is critical for us to understand their preferences on model outputs. To this end, we conduct our user studies directly within our product (see Fig 8). We believe this is the best context to gather preference metrics and provides the harshest test of whether a model actually delivers on making something valuable for an end user.

For a given user study, we choose a fixed set of prompts and sample images from two models. We then show a pair of images with the prompt to the user (without showing them which model corresponds to which image) and ask them to pick the best one according to some attribute, e.g. aesthetic preference. Because a single user's rating is prone to bias, we show each image pair to at least 7 unique users. To further reduce bias, an image pair only "wins" for a given model if its output is preferred by at least a

Playground v2.5 | Midjourney 5.2 | DALL·E 3 [2] | Playground v2 [20] | PixArt-alpha [3] | SDXL-1.0 [28]

Figure 9: **More qualitative comparison between methods.** Prompts for the top row: "blurred landscape, close-up photo of man, 1800s, dressed in t-shirt", the bottom row: "human with pig head wearing streetwear fashion clothes, skateboarding on a skatepark".

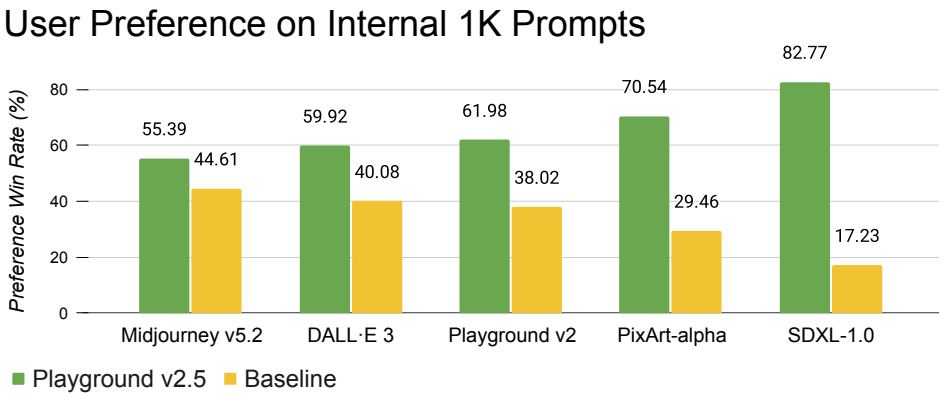


Figure 10: **User study against SoTA Methods.** We report human aesthetic preference metrics of Playground v2.5 against various publicly available text-to-image models. Playground v2.5 outperforms Midjourney 5.2, DALL·E 3 [2], Playground v2 [20], PIXART-α [3], and SDXL [28].

2-vote margin. A 1-vote margin is considered a tie. Lastly, we involve thousands of unique users on each user study. All user studies mentioned in this report are conducted through this interface.

We conducted studies to measure overall aesthetic preference, as well as for the specific areas we aimed to improve with Playground v2.5, namely generation across multiple aspect ratios and human preference alignment.

3.2 Overall Aesthetic Preference against other SoTA models

We used a prompt set called Internal-1K to compare Playground v2.5 against other state-of-the-art models. Internal-1K is a prompt set collected from real user prompts on Playground.com, making it representative of real users' prompting styles. We showed image pairs to thousands of users, specifically focusing on aesthetic preference for this study. This is the same study setup as our previous release of Playground v2 [20]. For reference, our previous studies demonstrated that images produced from Playground v2 were favored 2.5x more than those produced by SDXL. We aimed to surpass this for Playground v2.5 and succeeded: v2.5 is favored 4.8x over SDXL.

Fig. 10 shows our results against various publicly available text-to-image models. Across the board, Playground v2.5 dramatically outperforms the current state-of-the-art open source models of SDXL [28] and PIXART-α [3], as well as Playground v2 [20]. Because the performance differential between Playground v2.5 and SDXL was so large, we also tested against state-of-the-art closed-source

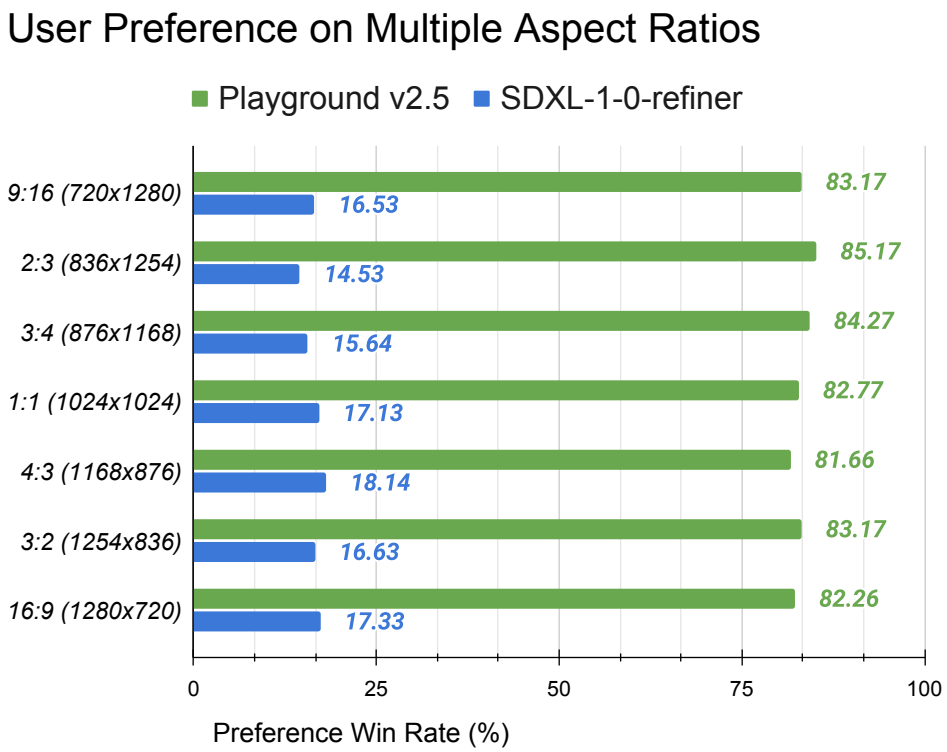


Figure 11: **User study against SDXL on multiple aspect ratios.** We conduct user studies for images generated in various commonly-used aspect ratios, height-to-width ratios ranging from 9:16 to 16:9. Our model outperforms SDXL in all aspect ratios by a large margin.

models like DALL·E 3 [2] and Midjourney 5.2, and found that Playground v2.5 still outperforms these models in aesthetic quality.

3.3 Evaluation of Generation Across Multiple Aspect Ratios

We report user preference study metrics on commonly-used aspect ratios using the Internal-1K prompt set. We conducted a separate user study for each aspect ratio, ranging from 9:16 to 16:9. For a given study, we used the same aspect ratio conditioning for both models on all images. Fig 11 shows our results. Our model outperforms SDXL in all aspect ratios by a large margin.

3.4 Evaluation on People-centric Prompts

As discussed in Section 2.3 about improving human preference alignment, people-related prompts are an important practical use-case for commercial text-to-image models. Indeed, they are quite popular in our product. To assess our model's ability to generate people-related images, we curated 200 high-quality people-related prompts from real user prompts in our product. We call this the *People-200* prompt set. We will release this prompt set to the community for benchmarking purposes.

We conducted our user study using portrait aspect ratio 3:2, since this is the most popular choice in the community for images showing people. We compared Playground v2.5 against two commonly-used baseline models: SDXL and RealStock v2, a community fine-tune of SDXL that was trained on a realistic people dataset.

Fig 12 shows that Playground v2.5 outperforms both baselines by a large margin.

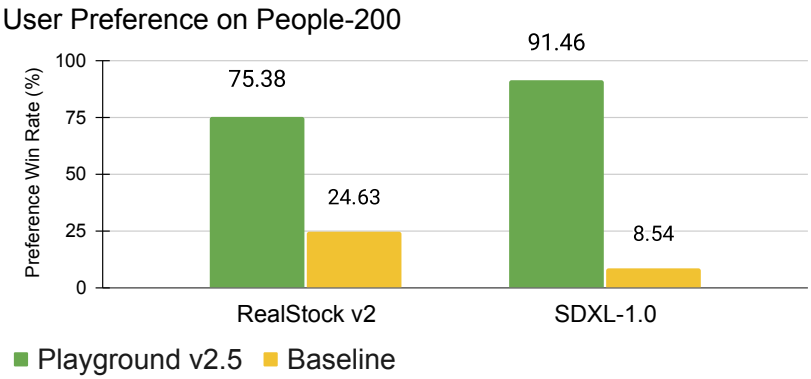


Figure 12: **People-200 benchmark.** We conduct a user study using the People-200 prompt set, which focuses on generating people. We compare Playground v2 against two baseline models: SDXL[28] and RealStock v2, a popular community fine-tune trained on a realistic people dataset. All images were generated in 3:2 aspect ratios with resolution 1254x836.

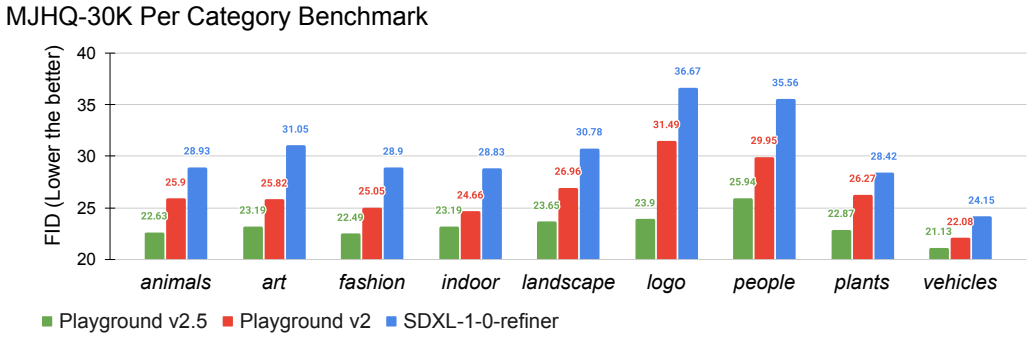


Figure 13: **MJHQ-30K benchmark.** We report FID of Playground v2.5, Playground v2 [20], and SDXL[28] on 10 common categories. Playground v2.5 outperforms Playground v2 in all categories, and most significantly on *logo* and *people* categories.

3.5 Automatic Evaluation Benchmark

Lastly, we introduce a new benchmark, *MJHQ-30K*, for automatic evaluation of a model's aesthetic quality. The benchmark computes Fréchet Inception Distance (FID) [11] on a high-quality dataset to gauge aesthetic quality. By spot-checking the FID and ensuring it was trending lower, we were able to quickly gauge progress throughout the different stages of pre-training and alignment.

We curated a high-quality dataset from images made on Midjourney 5.2. The dataset covers 10 common categories, and each category has 3K samples. Following common practice, we used aesthetic score [18] and CLIP score [10] to ensure high image quality and high text-to-image alignment. Furthermore, we took extra care to make the images and prompts well-varied within each category.

We report both the overall FID (Table 1) and per category FID (Fig 13). All FID metrics are computed at resolution 1024x1024. Our results show that Playground v2.5 outperforms both Playground v2 and SDXL in overall FID and all category FIDs, especially in the people and fashion categories. This is in line with the results of the user study, which indicates a correlation between human preferences and the FID score of the MJHQ30K benchmark.

Method	Overall FID
SDXL 1.0 + refiner[28]	9.55
Playground v2 [20]	7.07
Playground v2.5	**4.48**

Table 1: **MJHQ-30K overall FID.**

We release this benchmark to the public on HuggingFace[3] and encourage the community to adopt it for benchmarking their models' aesthetic quality during pre-training and alignment.

4 Conclusion

In this work, we share three insights for achieving state-of-the-art aesthetic quality in text-to-image generative models, and we analyze and empirically evaluate Playground v2.5 against SoTA models in various conditions and setups. Playground v2.5 demonstrates: (1) superior performance in enhancing image color and contrast, (2) ability to generate high-quality images under various aspect ratios, and (3) alignment to human preference for aesthetic quality in generated images, especially for fine details in images of humans.

We are excited to release Playground v2.5 to the public. The model is available today to use at our product website[4] for all users, and we have open-sourced the weights on HuggingFace[5]. Furthermore, we will soon provide extensions for using Playground v2.5 in A1111 and ComfyUI, two popular community tools for using diffusion models.

For future works, we hope to tackle improving text-to-image alignment, enhancing the model's variation capabilities, and exploring new architectures.

At Playground, our goal is to build a unified general-purpose vision system that deeply understands pixels and enables humans of all skill levels to masterfully generate and edit pixels. We see Playground v2.5 as a stepping stone towards this vision, and we encourage the community to build with us.

[3] https://huggingface.co/datasets/playgroundai/MJHQ-30K

[4] https://playground.com

[5] https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic

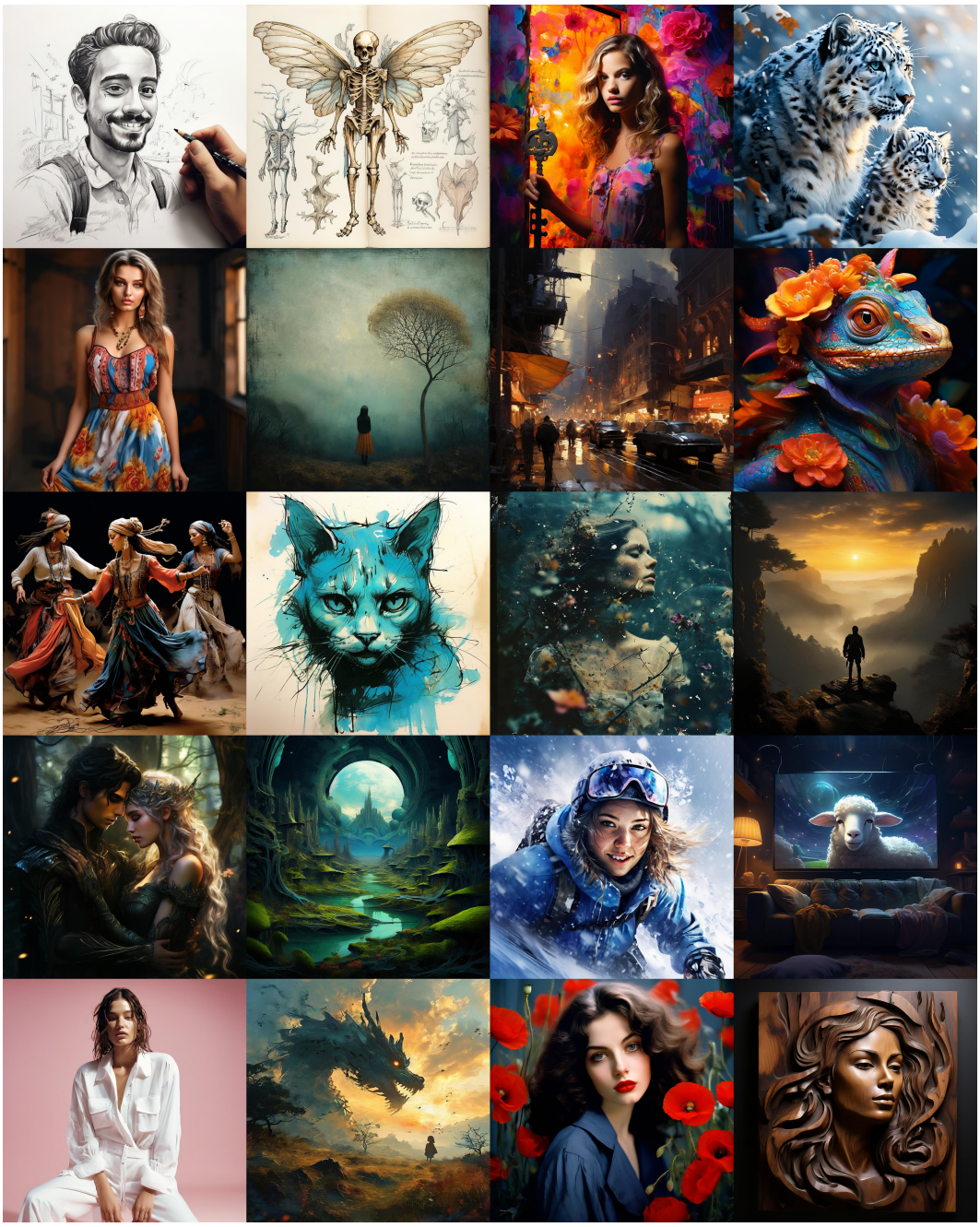


Figure 14: Playground v2.5 random samples with popular user prompts.

References

[1] Stability AI. Introducing stable cascade. https://stability.ai/news/introducing-stable-cascade, 2024. Accessed: 2024-02-20.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[4] Ting Chen. On the importance of noise scheduling for diffusion models, 2023.

[5] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack, 2023.

[6] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[8] Nicholas Guttenberg. Diffusion with offset noise. https://www.crosslabs.org/blog/diffusion-with-offset-noise, 2023. Accessed: 2024-02-20.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

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

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[13] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images, 2023.

[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.

[17] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023.

[18] 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, 2023.

[19] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19(143-155):18, 1989.

[20] Daiqing Li, Aleks Kamko, Ali Sabet, Ehsan Akhgari, Linmiao Xu, and Suhail Doshi. Playground v2.

[21] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed, 2024.

[22] Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, and Marc T. Law. How much more data do i need? estimating requirements for downstream tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 275–284, June 2022.

[23] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[24] NovelAI. Novelai improvements on stable diffusion. https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac, 2022. Accessed: 2024-02-20.

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[26] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pages 354–368. Springer, 2012.

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[28] 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, 2023.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

[32] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning, 2024.

[33] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.