Illustrious: an Open Advanced Illustration Model

Sang Hyun Park* **Jun Young Koh*** **Junha Lee** **Joy Song**

Dongha Kim **Hoyeon Moon** **Hyunju Lee** **Min Song**†

Onoma AI Research

Abstract

In this work, we share the insights for achieving state-of-the-art quality in our text-to-image anime image generative model, called Illustrious. To achieve high resolution, dynamic color range images, and high restoration ability, we focus on three critical approaches for model improvement. First, we delve into the significance of the batch size and dropout control, which enables faster learning of controllable token based concept activations. Second, we increase the training resolution of images, affecting the accurate depiction of character anatomy in much higher resolution, extending its generation capability over 20MP with proper methods. Finally, we propose the refined multi-level captions, covering all tags and various natural language captions as a critical factor for model development. Through extensive analysis and experiments, Illustrious demonstrates state-of-the-art performance in terms of animation style, outperforming widely-used models in illustration domains, propelling easier customization and personalization with nature of open source. We plan to publicly release updated Illustrious model series sequentially as well as sustainable plans for improvements on HuggingFace[3] with a license[4].

1 Introduction

Stable Diffusion[1] has brought groundbreaking advancements to the field of image generation. In particular, SDXL[2], which was trained with SD XL architecture with dual CLIP text encoder, based on large-scale datasets, has become even more powerful, offering excelling prompt control over generated images. While photorealistic image generation has benefited from large datasets like ImageNet[3] and OpenImages[4], illustration and animation image generation has comparatively shown slow progress, mainly due to the lack of large-scale finetuned open sourced models and strict dataset requirements.

We introduce a state-of-the-art anime generation model, Illustrious, which surpasses existing various models in various aspects. By leveraging a large dataset and offering detailed prompt guidance, Illustrious can express a wide range of concepts combinations, as depicted in Figure 17 that previous models struggled, with accurate control with prompt guidance, such as CFG[5], and capable for producing high-resolution images with anatomical integrity.

*equal contribution

†corresponding author

[3]https://huggingface.co/OnomaAIResearch/Illustrious-xl-early-release-v0

[4]https://huggingface.co/OnomaAIResearch/Illustrious-xl-early-release-v0/blob/main/README.md

Technical Report.

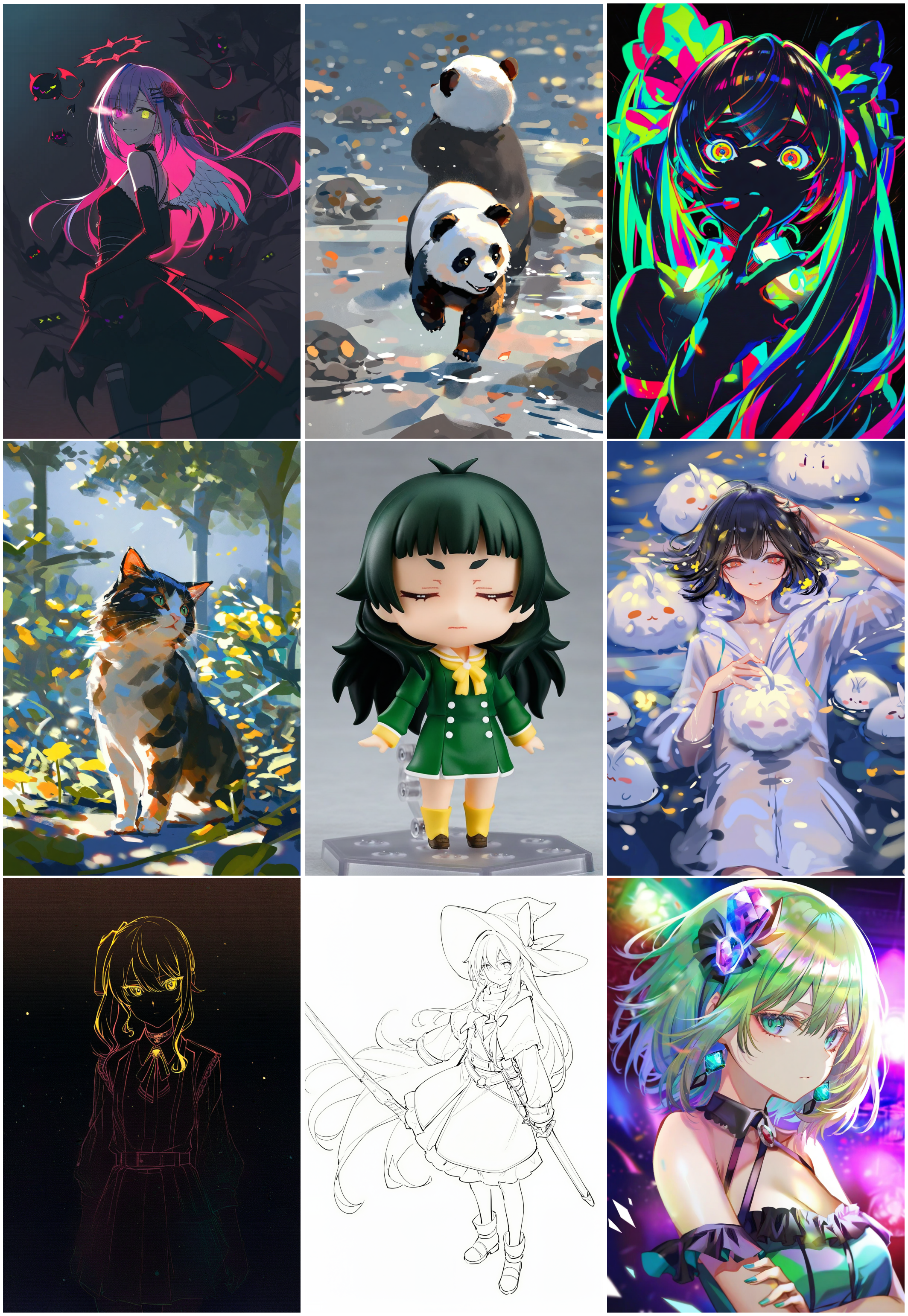


Figure 1: **High-quality samples from Illustrious.** Our model exhibits vibrant color and contrast on a range of image styles.

Illustrious v0.1 Illustrious v1.0 Illustrious v1.1 Illustrious v2.0

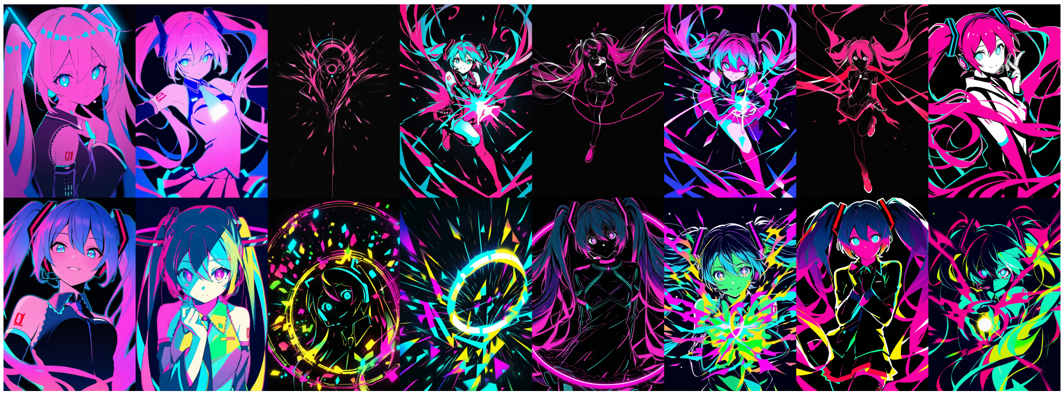


Figure 2: Model Comparison Images

2 Preliminary

2.1 SDXL

Stable Diffusion is a latent Text-to-Image diffusion model used as a foundation model in various image domain fields such as classification[6], controllable image editing[7][8], personalized image generation [9][10][11], and synthetic data generation[12][13]. According to previous studies by Ho et al.[14] and Song et al.[15][16], the diffusion model has arose as a powerful image generation model[17], placing the U-Net[18] backbone as a dominant architecture. In addition to this popular U-Net backbone, SD / SDXL applies improved upscaling layers, and cross-attention for text-to-image synthesis to a Transformer-based architecture. Unlike SD1.5 and SD2.0, which uses CLIP ViT-L, OpenCLIP ViT-H respectively, SDXL employs dual text encoders: CLIP ViT-L and OpenCLIP ViT-bigG. With the addition of a second text encoder, SDXL has significantly improved its understanding of text descriptions for images compared to previous models. The change resulted in the parameter size of the text encoder of 817M and 2.6B parameters in the U-Net.

Table 1: Finetuned Model

Finetuned Model	**Base Model**	**step**	**batch size**	**Dataset Size**	**Prompt Style**	**Annotation Method**	**Resolution**
Animagine XL V3.1	SDXL 1.0	91,030	96	2.1M	Tag based	Original Prompt	1024 x 1024
Kohaku XL Delta	SDXL 1.0	28,638	128	3.6M	Tag based	Original Prompt	1024 x 1024
Kohaku XL Zeta	SDXL 1.0	16,548	128	8.4M	Tag based	Original Prompt	1024 x 1024
SanaeXL anime V1.0	SDXL 1.0	-	-	7.8M	Tag based	Original Prompt	1024 x 1024
Neta Art XL	SDXL 1.0	-	-	-	Tag based	Original Prompt + CogVLM [19] + WaifuTagger	1024 x 1024
Arti Waifu Diffusion 2.0	SDXL 1.0	-	-	2.5M	Tag based	Original Prompt + Tag Ordering	1024 x 1024
Illustrious v0.1	SDXL 1.0	781,250	192	7.5M	Tag based	Original Prompt + Reorganized / Manual Filtering	1024 x 1024
Illustrious v1.0	SDXL 1.0	625,000	128	10M	Tag based	Original Prompt + Reorganized / Manual Filtering	**1536 x 1536**
Illustrious v1.1	SDXL 1.0	93,750	512	12M	Tag based	Multi-level Captions	**1536 x 1536**
Illustrious v2.0	SDXL 1.0	78,125	512	20M	Tag based	Multi-level Captions	**1536 x 1536**

2.2 Illustration / Animation Domain

Danbooru dataset [20][21] is a public large-scale anime image dataset with over 8 million images contributed and annotated in detail by communities. Annotation of images covers aspects such as characters, scenes, copyrights, and artists. Along with the Danbooru dataset, most available datasets are either processed versions of the Danbooru dataset[22] or face datasets[23][24] used for benchmarking purposes. We note that open sourced Illustrious model variants are being released under a research focused, non-commercial / no-closed source derivative public license, solely for open-source progresses.

2.3 Next-generation Text-to-Image Generative Models

With the advancement of AI technology in recent years, AI-based generative models have attracted a great amount of attention within the illustration field. In particular, next-generation models such as Hunyuan-DiT[25], Kolors[26], Stable Diffusion 3 (SD 3)[27], Flux[28], and AuraFlow[29] utilize additional as well as alternative text encoders to correctly interpret natural language input from users, increasing the sophistication of their ability to generate various, correct compositions of visual content.

2.3.1 Features of next-generation models

Hunyuan-DiT is a text-to-image diffusion transformer with a fine-grained understanding of both English and Chinese. It has redesigned the transformer structure, text encoder, and positional encoding. The model supports multi-turn, multi-modal dialogue with users, allowing it to generate and refine images based on contextual input. Another text-to-image generation model, Kolors, uses GLM[30], instead of T5 to improve comprehension of captions in order to improve the performance of natural language processing. Kolors uses the U-Net architecture and improves performance through a two-stage learning strategy: conceptual learning and learning for quality improvement. SD3 trains a rectified flow model by enhancing existing noise sampling techniques. This approach has demonstrated superior performance compared to traditional diffusion methods in high-resolution text-to-image synthesis. Flux is based on a hybrid architecture of multi-modal and parallel diffusion transformer[31] blocks, scaled to 12B parameters, with various technologies[32][33][34]. AuraFlow replaced the MMDiT block with a large DiT encoder block to improve model performance and computational efficiency. Performance was improved by using zero-shot LR transitions, and all data was re-captioned to reduce noise in the dataset.

2.3.2 Text Encoder

Currently, the text encoder seemly plays a crucial role in text-to-image generative models. A commonly used text encoder in generative models is CLIP[35]. OpenCLIP[36] provides various versions of CLIP. Despite existence of various CLIP model variants, trained in various datasets[37][38][39], the CLIP-only model has not shown significant success on complex compositions and glyph generations. For instance, SD1.5 and DALL-E2[40] use CLIP as their text encoder, however possibly due to limitation of CLIP itself proposed in various researches,[41][42][43], it is unknown whether SD XL architecture is fundamentally limited in complex compositions, and glyph generations.

One valid solution has been proposed by various models such as Imagen[44], PixArt [45], eDiFF-I[46] Hunyuan-DiT[25], Auraflow, and Flux. Through the utilization of the Transformer T5[47], this solution enables delivering more fine-grained local information to their text encoder. Stable Diffusion 3[27] also demonstrated the potential to interpret and generate complex prompts using the T5-XXL model. Remarkably, the CLIP-escaping architectures, like Kolors[26], which use GLM[30] has noted CLIP-dependent architecture as significant cause of limitation.

The Illustrious model is built upon SD XL architecture without changes, may share the noted limitations.

2.4 Data Ethics

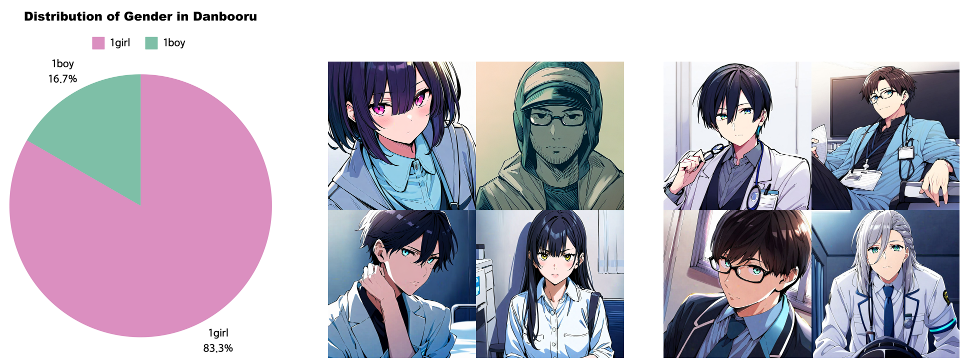
Text-to-image diffusion models are often trained under the pretext of 'aesthetic' considerations. However, this practice sometimes involves unethical data usage, such as obscuring the names of the artists whose works are used in training, thereby enabling the generation of specific styles without crediting the original artists. We believe it is crucial not to exploit or distort the data, even if this leads to a model with a default style that may appear dull or unclear.

To ensure ethical use of data, it is essential to clearly distinguish styles by associating them with the names of the artists and making this information transparent. Moreover, to safeguard artists from potential exploitation, we recommend that any transformative use of data and model to be accompanied by clear specification of training methodologies, modifications, and other relevant details, under fair public AI license terms[48].

Table 2: Baseline Model

Model	Parameter Size	Dataset	Resolution	Domain	Prompt Style	Accessibility
Stability AI Stable Diffusion 1.5	980M	LAION	512×512	Arbitrary	Natural Language	Open Source
Stability AI Stable Cascade	1.4B	-	1024×1024	Arbitrary	Natural Language	Open Source
Stability AI Stable Diffusion XL	2.5B	-	1024×1024	Arbitrary	Natural Language	Open Source
Stability AI Stable Diffusion 3	2B, 8B	ImageNet, CC12M[49]	1024×1024	Arbitrary	Natural Language	Open Source [a]
Midjourney V4	-	COCO[50], Visual Genome[51], Flickr 30K[52]	1024×1024	Arbitrary	Natural Language	Closed Source
OpenAI DALLE-3	-	LAION	1024×1024	Arbitrary	-	Closed Source
Hunyuan DiT	1.5B	-	-	Arbitrary	Natural Language	Open Source
Playground V3.0[53]	24B	-	1024×1024	Arbitrary	Natural Language	Closed Source
Flux	12B	-	2.0MP	Arbitrary	Natural Language	Open Source
Novel AI Image Generator[54]	-	Danbooru	-	Illustrate Picture	Tag based	Closed Source
Illustrious	2.5B	Danbooru, Synthetic datasets*	$\mathbf{2048 \times 2048}$	Illustrate Picture	Tag based [b]	Open Source [c]

[a] SD3 model variants are currently separated source
[b] Illustrious datasets and prompt styles vary by version
[c] Illustrious model variants are currently separated source



(a) **Gender distribution of Danbooru dataset.**

(b) **Example of "doctor" and "boy", showing model's weak understanding.**

(c) **The resolved understanding of "doctor" and "boy" in future epochs.**

Figure 3: Comparison of gender distribution and example generations from the model showing bias and weak understanding of gender-specific terms. The used prompt was **"1boy, doctor, masterpiece, looking at viewer"**.

3 Methodology

3.1 Dataset

3.1.1 Dataset Bias

Danbooru dataset contains a noticeably larger representation of female characters compared to male characters. This imbalance mirrors broader trends in the source material including anime and manga, where female characters are often more prominently featured in the form of images and character designs. Such gender imbalance in anime and manga datasets can lead to biased model performance, with models trained on this dataset potentially performing better on tasks involving female characters while underperforming on tasks related to male characters or other underrepresented categories, as shown in Figure 3a. This disproportionate representation can hinder the model's generalizability and faireness across different character types. We observed significant discrepancies in v0.1 model, which was later solved by **removing unfocused annotations in datasets**.

The dataset presents various issues due to its tag-based structure. Oftentimes, multiple meanings overlap with the same tokens or are used interchangeably, leading to confusion and ambiguity. A prominent example is the token "doctor," which can refer to both a character and a profession. In this case, one concept completely overlaps with the other, as shown in Figure 3b. Despite the fact that some images feature multiple characters, many in the dataset have very few tags or lack detailed annotations. This sparsity can make it difficult for models to learn critical concepts, since important features or attributes of the image may not be captured. The dataset contains extremely high-resolution images that could not be properly downsampled using any existent methods, leading to poor concept comprehension by the model. The Illustrious v0.1 model initially struggled with issues related to absurd aspect ratio, extremely high resolution images, and comic-like datasets. Therefore, careful pruning and refocusing of the dataset is necessary.

Based on insights and analysis on the v0.1 model, we expanded the dataset by including synthetic dataset based on generated images and captions to resolve the issues shown in Figure 3c.

3.1.2 Data Preprocessing

We initially adopted the tag ordering approach developed by NovelAI Team[5], which we believe that it functions as an instruction-tuning mechanism. Tags were separated and reordered following a specific schema:

person count ||| character names ||| rating ||| general tags ||| artist ||| score range based rating ||| year modifier

In the v0.1 model training, we split the tags using the "," convention, later occasionally replacing it with spaces based on a certain probability, combined with natural language prompts. Over time, we observed that the score range varied both temporally and across rating categories. To address this, we employed a percentile-based moving window method to determine the score range.

The score criteria and the year modifiers are defined numerically and range-based, depending on post counts:

Table 3: Score Criteria

Score Criteria	Percentage
Worst quality	~8%
Bad quality	~20%
Average quality	~60%
Good quality	~82%
Best quality	~92%
Masterpiece	~100%

Table 4: Year Modifier

Tag	Year
Oldest	~2017
Old	~2019
Modern	~2020
Recent	~2022
Newest	~2023

In subsequent epochs, we slightly modified the shuffling behavior by introducing aesthetic modifiers and filtering based on aesthetic scoring and file compression size metrics. The details of these aesthetic modifiers will be disclosed in future model releases.

For images used in the dataset, when the image size exceeded 4MP, we employed a mixed NEAREST/LANCZOS resizing method to maintain the aspect ratio. Images smaller than 768 $\times$ 768 were pruned from the dataset. Notably, few extremely high resolution images, ~40MP and those with uncommon aspect ratios (>1:10) were also removed.

However, the significant amount of high-resolution data remain problematic during the resize process, regardless of the resize method. Thus, we limited the higher-resolution dataset with minimal resizing, which is used for high-resolution training from v1.0 training. This allows for native high resolution generation, while minimizing down-sampling artifacts in smaller resolution.

Unlike the common practice of removing comics or low-quality images, our approach aimed to prune only a minimum of problematic images. This allowed us to expand the overall knowledge base, enhancing the model's understanding of diverse samples, increasing its ability to handle diverse inputs while maintaining overall control. A broader dataset also enables the model to generate low-quality sparse samples, as depicted in Figure 4.

3.1.3 Resolution

We trained the v0.1 model within 1MP range, as standard resolution. The v1.0 and v1.1 models were further trained at 2.25MP range, enabling native 2MP generation and up to 20MP generation when combined with proper img2img pipelines with reduced artifacts. The v2.0 model was additionally augmented with 0.15MP images, allowing it to generate outputs at a wide range of resolutions. Generated examples are shown in Figure 25.

[5]https://docs.novelai.net/image/tags.html

(a) Intentional low-quality generation, with prompt **1girl, hatsune miku, worst quality, ms paint (medium)**.

(b) Generation of 2-koma typed illustration, with prompt **1girl, happy, smile, crying, 2koma**.

Figure 4: Minimal data pruning strategy has allowed various concept genreation, including extremely rare ms-paint like concepts, without harnessing overall generation quality.

3.1.4 Limited Corpus

We identified several critical limitations in the Danbooru tag vocabulary, making it unsuitable for interpolation tasks. For instance, while the model can accurately generate objects like "stained glass" and "sword", it struggles with more complex concepts like "covering wound with left hand" due to insufficient data for such specific actions. Furthermore, the v0.1 model has difficulty processing natural language-based prompts, especially longer ones, as it was not well adapted to such formats.

3.2 Training Method

Based on the characteristics of the dataset described earlier, we attempted to overcome such problems by conducting the training using the following methods:

Firstly, we implemented a **No Dropout Token approach** to ensure that provocative or specific tokens are never excluded. In conventional training methods, random tokens are dropped during image pairing to prevent overfitting and enhance model generalization. However, this approach led to the occasional generation of provocative images. By ensuring provocative tokens were always retained and training the model to recognize these concepts with 100% accuracy, we found that controlling the sampling the provocative tokens by CFG, or preventing their use entirely effectively prevented the generation of provocative or inappropriate content[55].

Next, we employed **Cosine Annealing scheduler**[56] empirically. Such a schedule enables to achieve a lower learning rate and to gain reasonable converged checkpoints with a focus on improving the quality of image and stability of model training. Therefore, we adopted it into v1.0, v1.1, and v2.0 Illustrious models.

Third, we used **Quasi-Register Tokens** [57] to embed concepts the model doesn't understand into specific tokens for training. Since the dataset cannot contain all metadata, certain image characteristics may not be reflected. We identified these outlier concepts that the model couldn't comprehend and embedded them into register tokens during training. Conversely, when random tokens are included during training, concepts not represented in the text encoder or metadata can be captured by these random tokens. By attaching random alphanumeric strings, the model is allowed to separate 'bad characteristics' into leftover tokens, by separating known concepts from ambiguous ones. However, we observed that padding tokens used for sequence length matching in batching, are also treated as register tokens, a phenomenon we discuss in detail in the appendix.

Fourth, we trained model in **Contrastive Learning by Weak-Probability Dropout Tokens**. Similar to the first method, we prevented certain character names or artist names from being dropped with a set probability during training, which improves the model's ability to understand character names and artist styles, while other tokens are dropped as usual. This approach significantly improved character-wise understanding with fewer mixed features. Additionally, we observed that with this method, character learning accelerated even with smaller batch sizes, allowing more contrastive learning between no-character tokens and character token conditions. However, unlike the tag weighting strategy used by NovelAI, the absence of CFG control over character tags sometimes led to the model generating specific characters inductively, leading to weak dataset leakage, as expected.

Fifth, we implemented a simple **paraphrasing sequence process** to train the model on more diverse texts. Tags like "1girl, 1boy" were paraphrased as "one girl, single women," etc. This process enables the model to understand various inputs, instead of relying strictly on tag-based conditioning.

Finally, we adopted **Multi Level Dropout** by dividing the dropout into 4 stages, ranging from minimal, critical tokens to full tags. This allows the model to adapt to varying levels of caption detail. By 30% chance, we utilize max(30% * total tokens, 10) tags, 20% chance, max (40% * total tokens, 15), 10% chance, min(6, total tokens), 4% chance, min(total tokens, 4) tokens. The no-dropout tokens ignores this rule, for strict controllability.

We applied the eps-prediction loss objective and also utilized Input Perturbation Noise Augmentation with strength

- $0 < \varepsilon < 0.1$

[58], and Debiased Estimation Loss [59]. We observed noise offset [60] to be useful for broader color ranges. However, with lower batch sizes, it was not suitable for the common training procedure.

4 Training Setups

Table 5: Illustrious Training Setups

Model	Dataset Size	Batch Size	LR	TE LR	Epoch	Resolution	Prompt Style	Tag Manipulation		Multi Caption
								Dropout Level	Register Token	
Illustrious v0.1	7.5M	192	3.5e-5	4.5e-6	20	1024 × 1024	Tag	N	N	N
Illustrious v1.0	10M	128	1e-5	6e-6	8	1536 × 1536	Tag	Y	Y	N
Illustrious v1.1	12M	512	3e-5	4e-6	4	1536 × 1536	Tag + Natural Language	Y	Y	N
Illustrious v2.0	20M	512	4e-5	3e-6	2	**1536 × 1536**	Tag + Natural Language	Y	Y	Y

We trained models using different strategies sequentially.

- Illustrious v0.1 was trained on a 7.5M dataset consisting of 1024 × 1024 images with a batch size of 192. The data were tagged using the original Danbooru tags. The learning rate for the U-Net was set to 3.5e-5, and the text encoder learning rate was 4.5e-6, trained over 20 epochs.

- Illustrious v1.0 used a 10M dataset of 1536 × 1536 images with a batch size of 128, also tagged with the original Danbooru tags, with duplicate separated higher-resolution images. The U-Net learning rate was 1e-5, and the text encoder learning rate was 6e-6, trained over 8 epochs. For this dataset, we applied tag manipulation strategies, Dropout-Leveling and Register Tokens.

- Illustrious v1.1 was trained on a 12M dataset of the same 1536 × 1536 resolution images as v1.0. It used a batch size of 512 and was trained for 4 epochs with a U-Net learning rate of 3e-5 and a text encoder learning rate of 4e-6. The dataset for v1.1 was tagged using a combination of natural language descriptions and tags.

- Illustrious v2.0 was trained on a 20M dataset with the same 1536 × 1536 image resolution as v1.1. The model was trained with a batch size of 512 for 2 epochs, using a U-Net learning rate of 4e-5 and a text encoder learning rate of 3e-6. Illustrious v2.0 mainly incorporated the multi-caption method for enhanced text-image correspondence.

5 Evaluation

We conducted evaluations of our models with the well known rating method, Elo Rating and TrueSkill 2, and Character wise similarity, CCIP.

5.1 User Preference with Elo Rating

The ELO Rating system, developed by Arpad Elo, is being widely used to evaluate user's skill levels in competitive survey by adjusting user's ratings based on match outcomes. The rating changes reflect the difference between expected and actual results, providing a dynamic measure of a user's relative strength. The standard Elo rating update formula is given as following:

$$R' = R + K \times (\text{actual} - \text{expected})$$

Where:

- R' is the new rating after the match.
- R is the current rating before the match.
- K is the K-factor, a constant that determines the sensitivity of rating changes.
- actual is the actual result of the match (1 for a win, 0.5 for a draw, 0 for a loss).
- expected is the expected score, calculated using the formula:

$$\text{expected} = \frac{1}{1 + 10^{(R_{\text{opponent}} - R)/400}}$$

Here, R_{opponent} is the rating of the opponent.

Recently, various research studies have been evaluating models based on ELO rating using win rates.[61][62] In the case of images, in particular, traditional metrics[63][64] tend to focus on the similarity of the image itself, such as pixel-level similarity, rather than the meaning of the image. Therefore, human evaluation becomes even more essential in such cases.

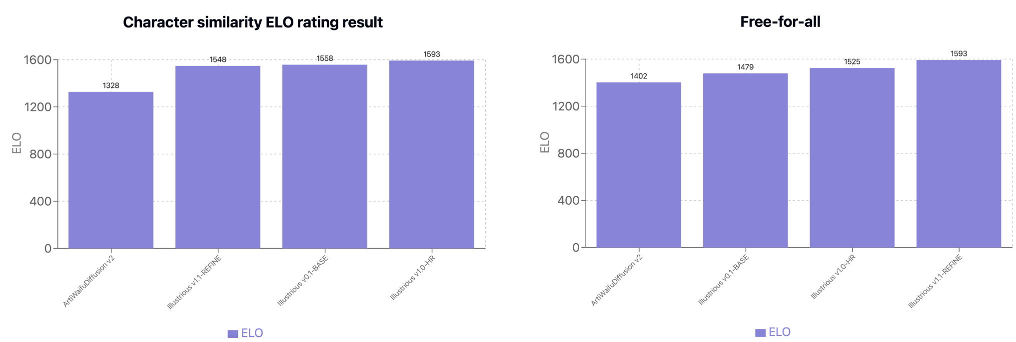


Figure 5: **Character Similarity ELO Ratings Result**, time-weighted average is applied. and **Free-for-all ELO**

Fixed-Characteristics means 2 random images are shown on poll and users select one with fixed prompt generations. This match accepts draw. Free-prompt-duel means 2 random images from free prompt and one is selected. Free-for-All is 1 vs 1 vs 1 vs 1 match.

5.2 CCIP

CCIP [65] is a metric designed to estimate visual differences between given grouped set and given image for character basis, focusing on feature extraction metric based on CLIP. The difference value in CCIP is calculated as average of given formula:

$$D(I_1, I_2) = M(I_1, I_2)$$

Where:

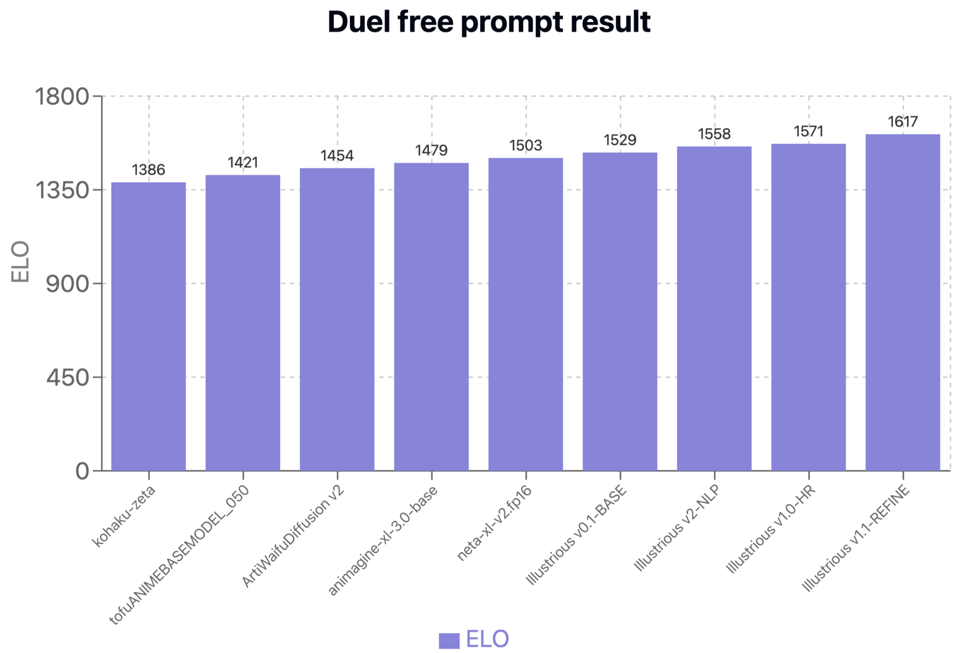


Figure 6: **Duel Free Prompt ELO Result**.

- $D(I_1, I_2)$ represents the difference value between images I_1 and I_2.
- I_1 and I_2 are the two images being compared.
- M is a CCIP model.

CCIP extracts visual features of characters from images and quantifies the differences to assess character similarity. CCIP effectively identifies whether two images contain the same character, focusing on features like facial attributes, clothing, and color schemes.

5.3 TrueSkill Algorithm

TrueSkill is a skill-based ranking system proposed by Microsoft. Unlike the Elo rating system, which was originally developed for chess, TrueSkill requires less trials to estimate users' expected numerical skill scoring, which is more stable for sparse model duels conditions.[66][67] As documented, the update equations are given as following:

$$\mu' = \mu + \frac{\sigma^2}{\sigma^2 + \beta^2} \times (s - \mu)$$

Where:

- μ' is the updated mean skill level of the player.
- μ is the current mean skill level before the update.
- σ^2 is the variance representing the uncertainty in the player's skill estimate.
- β^2 is the variance of the game outcome, reflecting the randomness inherent in game results.
- s is the performance score derived from the game outcome.

The variance σ^2 is also updated every match, to reflect the change in uncertainty after each game.

By integrating these algorithms and metrics into our evaluation framework, we aim to provide a comprehensive assessment that balances quantitative measures with human judgment, which is particularly important in domains like image evaluation where subjective interpretation plays a significant role.

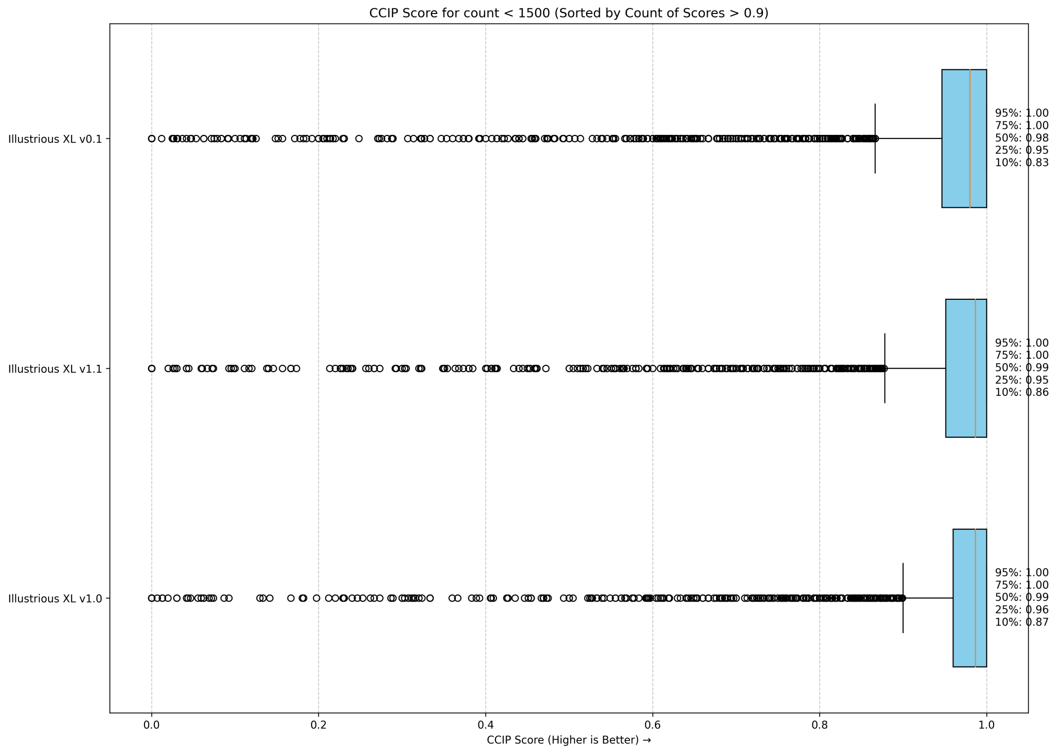


Figure 7: **CCIP Score**.

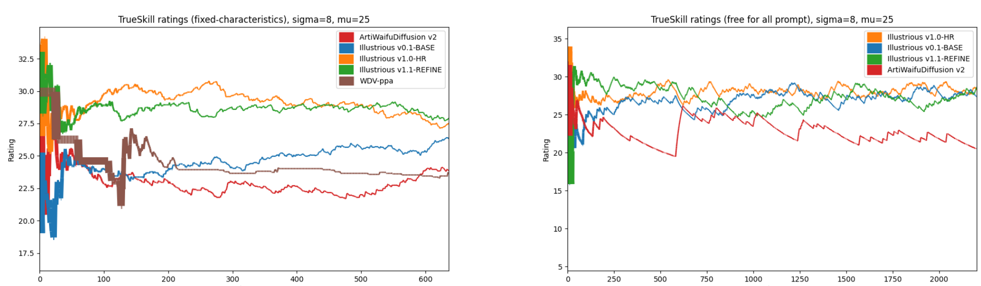


Figure 8: **TrueSkill Ratings (fixed-characteristics)** and **TrueSkill Ratings (Free-for-All Prompt)**

6 Limitations and Future Works

Limitations

Illustrious is a generalized anime image generation model that can create a variety of images through detailed prompts. However, it has the following limitations.

First, the CLIP text encoder's instability in handling character details can lead to less effective performance in embedding similarity calculations. Recently, models such as Flux or Kolors have addressed this issue by using alternatives like T5 and GLM instead of the CLIP text encoder.

Second, the Danbooru dataset predominantly relies on tag-based metadata, which makes it difficult to describe images across multiple dimensions. This limitation creates challenges in controlling the specific composition and positioning of multiple characters or actions. To fully address this issue, detailed descriptions of each character, their positions, backgrounds, and relationships are necessary—elements often missing in tag-based and other common large-scale datasets.

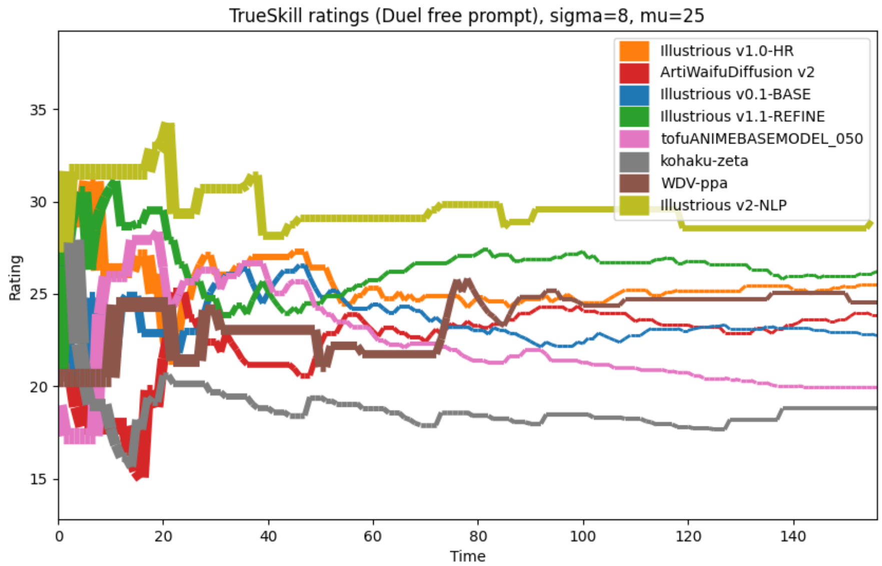


Figure 9: **TrueSkill ratings (Duel Free Prompt)**.

With the enhanced natural language capabilities introduced in v2.0 and a custom-built, sophisticated dataset (to be released in future work), we propose the development of large-scale, refined natural language datasets to overcome these limitations.

Future Works

Below are some possible directions of Illustrious in future work.

One key challenge identified is the task of rendering text within images for anime image generation. While many real-image generation models can partially support the embedding of text in images, open-source anime image generative models struggle with this task. Phrases like "Merry Christmas" or "Happy New Year" can sometimes be rendered correctly due to their frequent appearance in datasets, but generating full sentences or meaningful words within anime images remain a significant challenge.

The Illustrious v2.0 shows notable improvements in generating glyphs, albeit with limited capability, through synthetic captions. Future models could be significantly enhanced by incorporating OCR-based datasets and conditioning as part of the training process.

References

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

[2] 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.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[4] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[6] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion, 2023.

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[10] Jae Wan Park, Sang Hyun Park, Jun Young Koh, Junha Lee, and Min Song. Cat: Contrastive adapter training for personalized image generation, 2024.

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[12] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023.

[13] Jun Young Koh, Sang Hyun Park, and Joy Song. Improving text generation on images with synthetic captions, 2024.

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

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

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

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

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[19] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[20] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2021: A large-scale crowdsourced & tagged anime illustration dataset. `https://gwern.net/danbooru2021`, January 2022. Accessed: DATE.

[21] Anonymous and Danbooru community. Danbooru2023: A large-scale crowdsourced & tagged anime illustration dataset. `https://huggingface.co/datasets/nyanko7/danbooru2023`.

[22] Edwin Arkel Rios, Wen-Huang Cheng, and Bo-Cheng Lai. Daf:re: A challenging, crowd-sourced, large-scale, long-tailed dataset for anime character recognition, 2021.

[23] Kangyeol Kim, Sunghyun Park, Jaeseong Lee, Sunghyo Chung, Junsoo Lee, and Jaegul Choo. Animeceleb: Large-scale animation celebheads dataset for head reenactment, 2022.

[24] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2264–2272, New York, NY, USA, 2020. Association for Computing Machinery.

[25] 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.

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

[27] 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.

[28] black-forest labs. flux, 2024. Available: `https://github.com/black-forest-labs/flux`.

[29] fal. Auraflow, 2024. Available: `https://huggingface.co/fal/AuraFlow?ref=blog.fal.ai`.

[30] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling, 2022.

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

[32] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023.

[33] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[34] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023.

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

[36] Ilharco, Gabriel and Wortsman, Mitchell and Wightman, Ross and Gordon, Cade and Carlini, Nicholas and Taori, Rohan and Dave, Achal and Shankar, Vaishaal and Namkoong, Hongseok and Miller, John and Hajishirzi, Hannaneh and Farhadi, Ali and Schmidt, Ludwig. Openclip, jul 2021.

[37] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*, 2021.

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[39] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[41] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023.

[42] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022.

[43] Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard B W Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation, 2024.

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[45] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[46] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023.

[47] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[48] FreeDev Project. Fair public ai license 1.0-sd, 2024. Retrieved from `https://freedevproject.org/faipl-1.0-sd/`.

[49] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[51] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[52] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[53] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models, 2024.

[54] Juan Ossa, Eren Doğan, Alex Birch, and F. Johnson. Improvements to sdxl in novelai diffusion v3, 2024.

[55] Alphanome.AI. The waluigi effect in ai, 2023. Accessed: 2024-09-28.

[56] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[57] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024.

[58] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models, 2023.

[59] Hu Yu, Li Shen, Jie Huang, Hongsheng Li, and Feng Zhao. Unmasking bias in diffusion model training, 2024.

[60] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5392–5399, 2024.

[61] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[62] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[63] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[64] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[65] Ziyi Dong and narugo1992. Contrastive anime character image pre-training. `https://huggingface.co/deepghs/ccip`, 2024.

[66] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.

[67] Tom Minka, Ryan Cleven, and Yordan Zaykov. Trueskill 2: An improved bayesian skill rating system. *Technical Report*, 2018.

[68] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[69] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023.

[70] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback, 2023.

[71] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[72] Shih-Ying Yeh. Tipo: Text to image with text presampling for prompt optimization, 9 2024. Technical report available at `https://hackmd.io/@KBlueLeaf/BJULOQBR0`. Model available at `https://huggingface.co/KBlueLeaf/TIPO-500M`. Source code available at `https://github.com/KohakuBlueleaf/KGen`.

[73] Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training, 2019.

[74] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George E. Dahl, Christopher J. Shallue, and Roger Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model, 2019.

[75] Aditya Devarakonda, Maxim Naumov, and Michael Garland. Adabatch: Adaptive batch sizes for training deep neural networks, 2018.

[76] Feipeng Ma, Yizhou Zhou, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. Image captioning with multi-context synthetic data, 2023.

[77] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models, 2023.

[78] bdsqlsz. Adapter-based approach to control content safety. `https://huggingface.co/bdsqlsz/filter_nude`, 2024.

[79] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022.

[80] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

[81] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, 2023.

A Appendix / Supplemental Material

A.1 Resolution

Illustrious v1.0+ is capable of generating images in 1536x1536 natively, which can be expanded to 2048x2048 at farmost without any modifcation. In higher resolutions, it allows over 20MP+ generation as depicted in Figure 25, while other models fails to follow.

A.2 Analysis

There are experimental results obtained through various efforts to make Illustrious model. We will describe this in detail as follows.

A.2.1 Limitations regard to aesthetic/biased models

As part of stabilization, careful considerations must be given during the aesthetic tuning stage. Fitting a baseline model into human preferences[68][69][70] can degrade its performance on the true data distribution. This also reduces the diversity of image generation, limiting the model's applicability. Such overfitting makes future fine-tuning significantly more difficult compared to using an unbiased model, as it necessitates re-aligning the model's knowledge with the true data distribution.

(a) LoRA-applied result in Illustrious v0.1. The prompt was **1girl, shinosawa hiro, general, masterpiece**.

(b) LoRA-applied result in Illustrious v2.0.

Figure 10: The LoRA trained on Illustrious v0.1, is widely usable across checkpoints.

For this reason, to ensure broader public usability, we have decided to release non-fine-tuned base models. These models can be adapted for various tasks and concepts. We also demonstrate that model-derived add-ons, such as LoRAs[71], remain compatible with future models and allow for effective model derivation as depicted in Figure 10[6].

A.2.2 Multiple-character generation

We observe that the strict token control approach results in excelling character feature separation in limited budget. The phenomenon is sustained from Illustrious v0.1, toward the cutting edge model, Illustrious v2.0, as depicted in Figure 11.

[6]https://civitai.com/models/794775/llustrious-xl-shinosawa

(a) Multi-character separation result in Illustrious v0.1. The prompt was **2girls, otonose kanade, hatsune miku, side-by-side, masterpiece**.

(b) Character combine result in Illustrious v0.1. The prompt was **1girl, otonose kanade, hatsune miku (cosplay),general, masterpiece, blonde hair**.

(c) Multi-character separation result in Illustrious v2.0. The prompt was **multiple girls, 2girls, nozomi (blue archive), hikari (blue archive),year 2023,dynamic angle,shiny, hat, pointy ears,long hair, shorts, green hair, siblings, pantyhose, thick eyebrows, demon tail, gloves, open mouth, tail, twintails, smile, blush, looking at viewer, sisters, orange eyes, white gloves, skirt parted lips, simple background,white background, masterpiece, absurdres**

Figure 11: The character separation behavior of Illustrious.

A.2.3 the Effect of "Long Prompts"

It is commonly known that long prompts or detailed tags are capable of generating sophisticated images. Illustrious also benefits from upsampled / detailed captions, especially when controlled by sophisticated models. While simple prompts directly exposes model's creativity, we recommend sophisticated captions to further utilize the prompts and models' capability.

(a) Simple prompt generation with Illustrious v0.1, prompt **1girl, hatsune miku, angel, masterpiece, general**.

(b) Complex prompt generation upsampled by TIPO with Illustrious v0.1, prompt **1girl, hatsune miku. An illustration of a girl with long white hair and wings. she is wearing a school uniform with a red bow on her head and a pair of headphones on her ears. the wings are spread out behind her, creating a sense of movement and energy. the overall style of the illustration is anime-inspired. solo, skirt, feathered wings, necktie, smile, very long hair, collared shirt, long hair, headset, blue eyes, aqua necktie, looking at viewer, black footwear, black skirt, twintails, grey shirt, bare shoulders, detached sleeves, full body, zettai ryouiki, closed mouth, miniskirt, sleeveless, boots, thighhighs, shirt, standing, wing collar, aqua hair, sleeveless shirt, pleated skirt, angel wings, absurdly long hair, wings, black thighhighs,masterpiece, general**.

Figure 12: The upsampling prompt can escape trivial solutions by providing details.

For this, we utilize TIPO library[72], to show the drastic sample differences across the models, in Figure 12.

A.2.4 Batch Size and Learning Rates

We found that large batch sizes can effectively help sparse tags to learn, making the model more stable against parameter updates.[73][74] In contrast, small batch sizes lead to more frequent attention binding, which benefits general / broader concept handling. This suggests that when training on large datasets which are focused on few new concepts, using small batch sizes can accelerate the learning process. However, for sparse concepts, larger batch sizes promote more stable training.

Additionally, if the learning rate falls below a specific threshold, the model may struggle to learn new concepts, favoring convergence toward stable attention splits rather than forming new attention bindings. Based on these observations, we propose that using adaptive batch sizes[75], combined with learning rate scheduling, could offer a more effective alternative for model training.

A.3 Inpainting

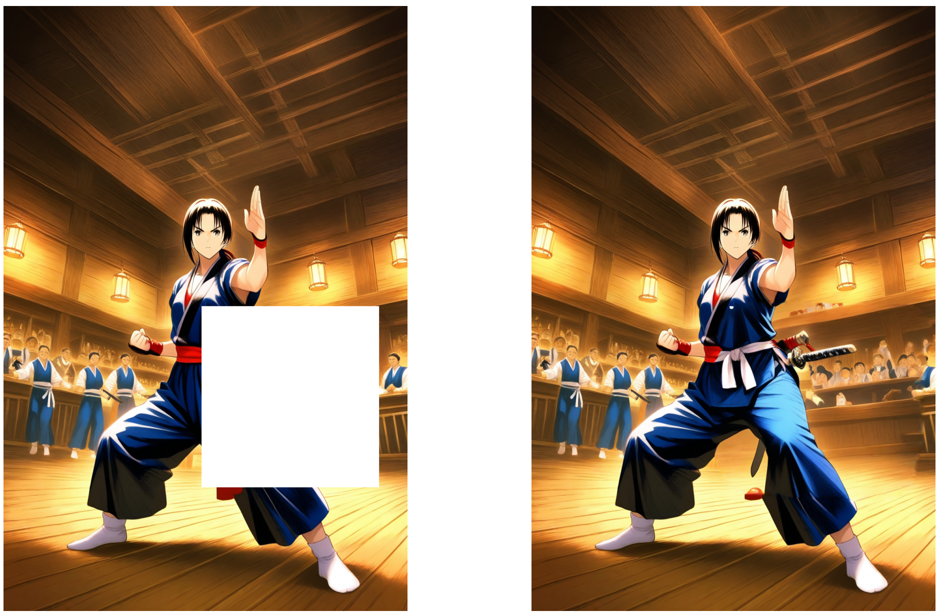


Figure 13: **Enhanced Inpainting** As the model's generation capability and prompt control improve, we can also observe significant advancements in its inpainting functionality.

As Illustrious's prompt control capabilities have improved, it has become capable of supporting powerful image generation. Based on this, we conducted various experiments not only on text-to-image generation but also on image-to-image generation. One of the most interesting findings was that as the model's image generation abilities improved, so did its inpainting capabilities. To demonstrate the improvements in inpainting, we partially cropped and corrupted images, masked the damaged areas, and then applied inpainting using Illustrious. Unlike previous models, which struggled with color or saturation mismatches in inpainting, Illustrious successfully generates images that harmonized seamlessly with the original content. The example image is shown in figure 13.

A.4 Dynamic Color Range

Illustrious has significantly improved its understanding of color, allowing control over color and brightness through prompts. In particular, its understanding of brightness has significantly improved. It successfully generates images with colors that are present even at very low brightness levels. We generated images with low brightness and then increased the brightness to demonstrate that the subject could clearly form a silhouette. The example images is shown in figure 14.

Figure 14: **Dynamic Color Range** Our model can adjust brightness through silhouette generation and similar techniques. Left is the original image generated from our model. Right is the same Image but upper the brightness 0 to 230.

Caption : The image depicts a character named Remilia Scarlet, an iconic figure from the Touhou series, created by the artist Yutazou. Remilia is portrayed with her signature features: light purple hair, red eyes, and bat-like wings. She is dressed in an elaborate alternate costume, consisting of a black and red ribbon hair accessory, a black and red floral-patterned top, and a vibrant pink and black striped ruffled skirt. She also wears black pantyhose and black high-heeled shoes. The character is looking directly at the viewer, with a simple white background that highlights her striking appearance. The overall composition is detailed and vibrant, showcasing Yutazou's distinct artistic style.

Tag : Tyutazou,touhou,remilia_scarlet,bad_id,bad_pixiv_id,1girl,alternate_costume,bat_wings,black_pantyhose, hair_ribbon,light_purple_hair,looking_at_viewer,pantyhose,red_eyes,ribbon,short_hair,simple_background,solo, white_background,wings

Figure 15: **Example of Multi Caption**.

A.5 Multi Level Captions

Starting from Illustrious v2.0, Multi Level Captions has been introduced. We realized that it is difficult to control multiple objects simultaneously through prompts using tagging alone.[76] Even when grouping the sequence of tags or the subcomponents of objects, expressing context solely through tagging proved to be quite challenging. Therefore, it is crucial to tag in a way that makes the context easily understandable through natural language. However, having humans manually caption large amounts of data is labor-intensive and has its limitations. At the same time, we could not abandon the advantages of tagging, so we implemented Multi-Captioning for images. Multi-Captioning involves assigning multiple captions to a single image likes natural language and tags. In the future, we plan to increase the number of captions to not only provide detailed descriptions of the image but also include context and narrative elements. The example of multi caption is shown in figure 15.

A.6 Padding token wise analysis

We find that allowing padding tokens to be trained can cause multiple problems. During training, text encoder outputs must be padded to be packed in batch. This makes padding token usage in CFG setups problematic with imbalanced token lengths, as it retains significant composition knowledge unlike different models. We recommend masked loss to overcome this problem in future training. We show the example in Figure 16.

Figure 16: The intensive padding token being used in CFG, causes problem since padding token was not utilized via masked loss. Left, with 2 tokens + 75 tokens padding, right, no CFG. The phenomenon is reduced when minimal padding token is used.

A.7 Further finetuning recommendations

We found that the Illustrious XL Text Encoders are stably converged - the text encoders are interchangable without major issues, despite of current tradition of not tuning text encoders for knowledge conservation and memory requirements. Despite of our method's empirical success, we do **not** recommend to finetune text encoder, unless datasets are sufficiently large enough to counter possible catastrophic forgetting issues.

As noted previously, we found that character learning trend fluctuates with lower batch sizes, whilst higher batch size stabilizes its forgetting phenomenon. Even larger batch size may be required for sparse concepts.

A.8 Safety control and Red-Teaming

Image dataset domain is abstract and not well researched, publicly available solutions and systems, and its detail lacks, which makes user uncontrollable from unwanted content generation. Following waluigi dillema, we instead finetune with strict control condition to make model understand the concepts separately, then utilize LECO[77] method-based approach, allowing safety control over provocative generations, as released in GUIDED variants, suggested as reference [78]. However, we also note here that simple control can be achieved by rating tokens, inputting "general" in prompt conditioning.

B Model Compare

B.1 Illustrious Qualitative Images

B.1.1 Illustrious v0.1

Figure 17: **Hatsune miku, cosplaying hakurei reimu, in 90s animation style, with glowing eyes**, generated in Illustrious v2.0 with 840×1216 resolution.

Illustrious v0.1's sample image is depicted as Figure 18.

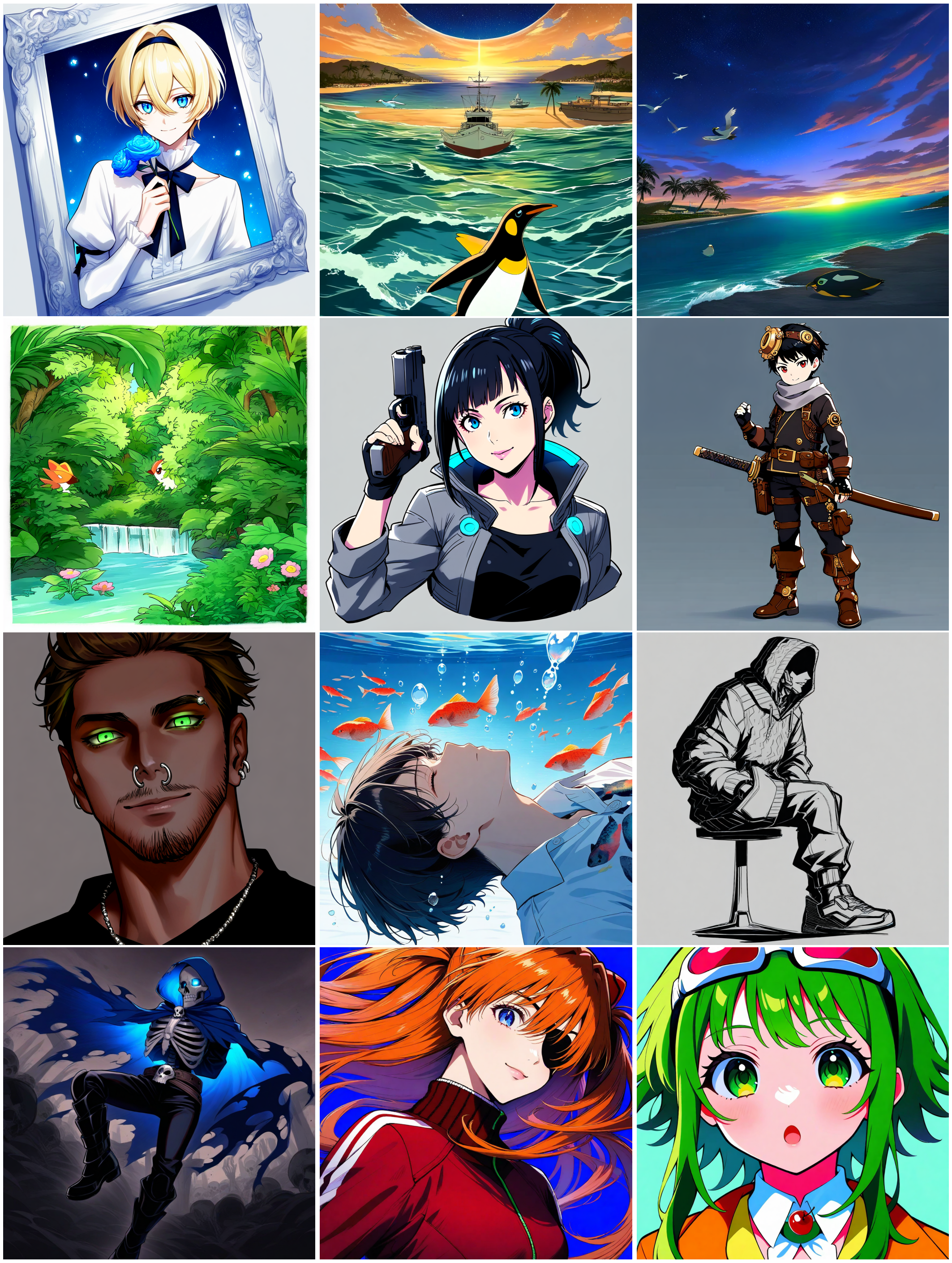


Figure 18: **High-quality samples from Illustrious v0.1.** Illustrious v0.1 can generate creative pictures.

Figure 19: **High resolution samples from Illustrious v1.0.** Illustrious v1.0 can generate the high resolution images. This image is 2048 $\times$ 2048 pixels with no upscale.

B.1.2 Illustrious v1.0

Illustrious v1.0's sample image is depicted as Figure 19 and 20.

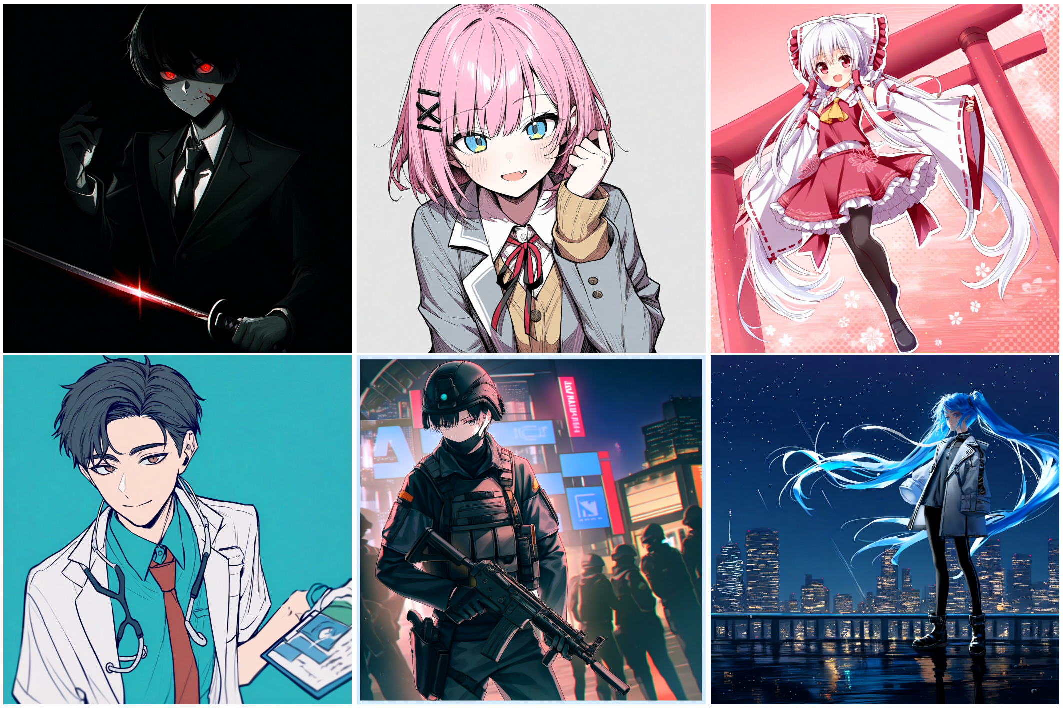


Figure 20: **High-quality samples from Illustrious v1.0.** Illustrious v1.0 can generate various styles. These images are all 1536×1536 pixels.

B.1.3 Illustrious v1.1

Illustrious v1.1's sample image is depicted as Figure 21.

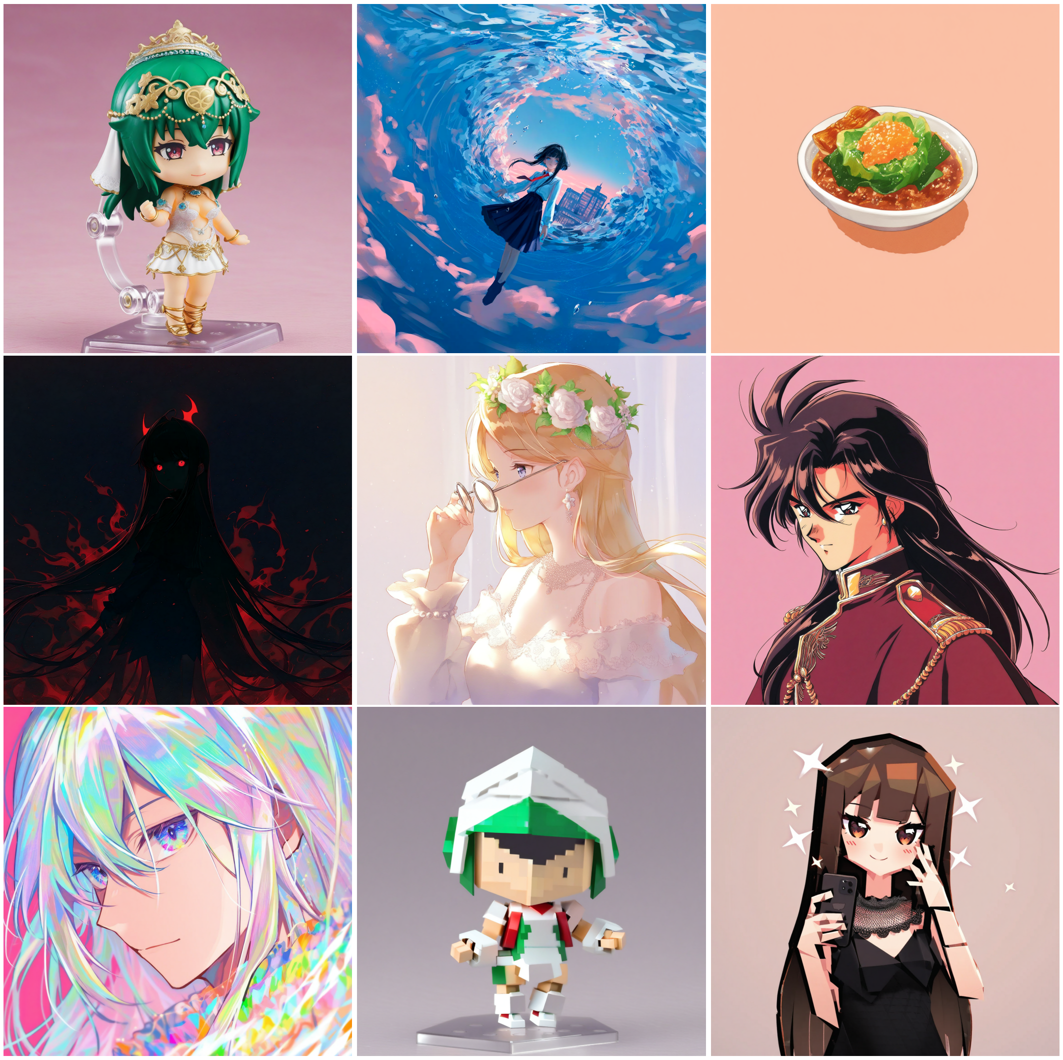


Figure 21: **High-quality samples from Illustrious v1.1.** Illustrious v1.1 can generate various styles. These images are all 1536 × 1536 pixels.

B.1.4 Illustrious v2.0

Illustrious v2.0's sample image is depicted as Figure 22 and 23.

Figure 22: **High-quality samples from Illustrious v2.0.** Illustrious v2.0 can understand the natural language prompts.

B.2 Recommended generation configuration

We used Euler A Discrete sampler with step count >20, with CFG 5~7.5 for generation examples, however it may depend on styles, setups. For instance, we found that generating with DPM-based schedulers[80][81], then piping through img2img pipeline with Euler discrete, works well for aesthetic / detailed image setups. Illustrious v0.1 supports 1MP resolutions. Illustrious v1.0+ supports native 1MP~2.25MP resolutions, up to 4MP with some loss. All images, which exceeds 1:10 ratio, was not targetted and included in training.

C Thanks To

Kohaku (KBlueLeaf), with massive supports and initiatives to train the large scale base models,

WDV team and community, with initial thoughts and benchmarks over various prompts,

DeepGHS team, with open-minded datasets and tools, massive contributions fostering open source research,

and **OnomaAI**, supporting the research and training, allowing the model to exist.

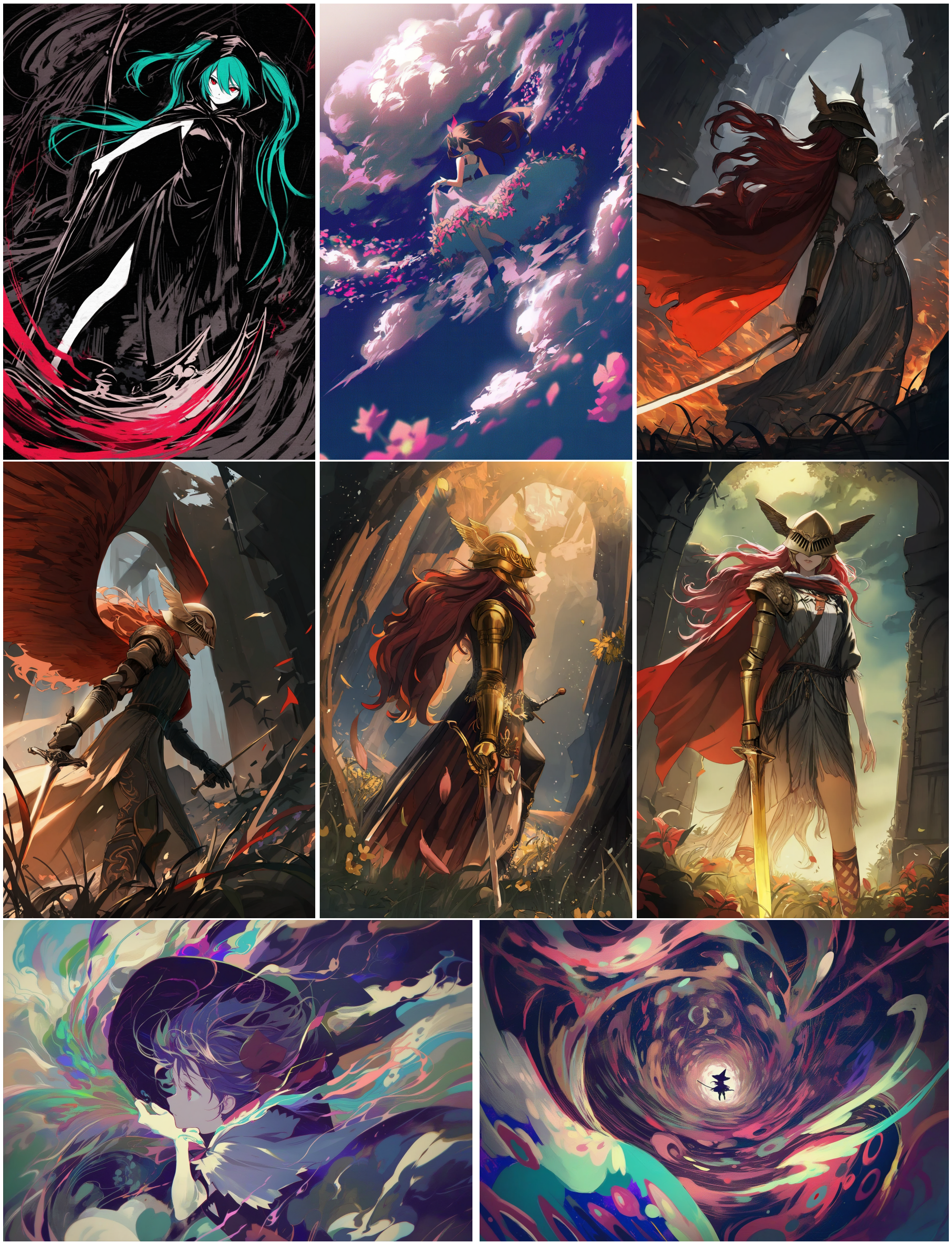


Figure 23: **Horizontal and Vertical High-quality samples from Illustrious v2.0.** Illustrious v2.0 can understand the natural language prompts.

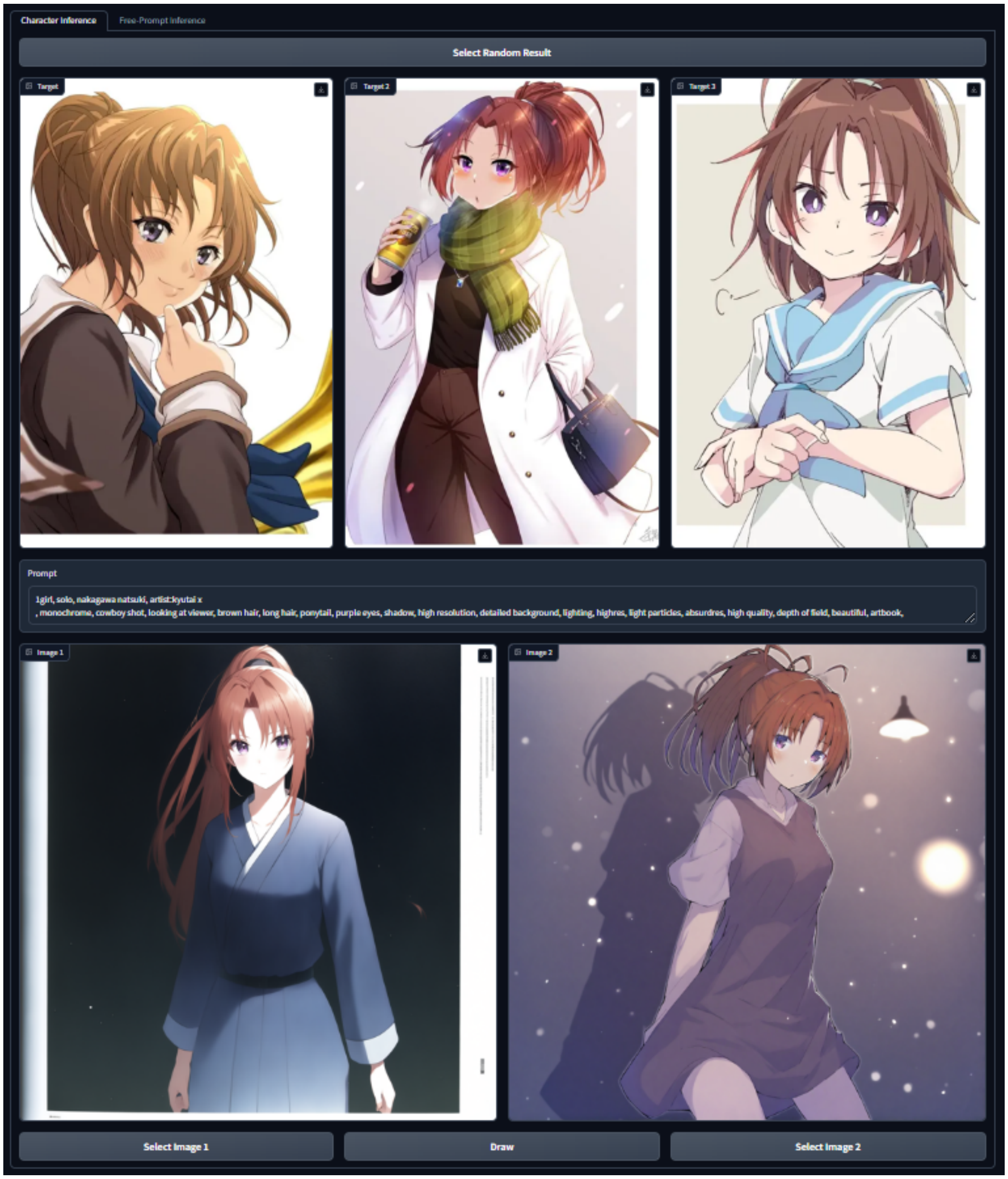


Figure 24: Model Compare Site Image

Figure 25: Illustrious v1.0+ can create the high resolution images. This image is 3744x5472 resolution by v1.0, firstly generated in 1248x1824, then upscaled toward 3744x5472 as same method using SDEdit[79]

.