GenAI Arena: An Open Evaluation Platform for Generative Models

Dongfu Jiang* **Max Ku*** **Tianle Li***
Yuansheng Ni **Shizhuo Sun** **Rongqi Fan** **Wenhu Chen**
University of Waterloo
{dongfu.jiang, m3ku, t29li, wenhuchen}@uwaterloo.ca

https://hf.co/spaces/TIGER-Lab/GenAI-Arena

Abstract

Generative AI has made remarkable strides to revolutionize fields such as image and video generation. These advancements are driven by innovative algorithms, architecture, and data. However, the rapid proliferation of generative models has highlighted a critical gap: the absence of trustworthy evaluation metrics. Current automatic assessments such as FID, CLIP, FVD, etc often fail to capture the nuanced quality and user satisfaction associated with generative outputs. This paper proposes an open platform GENAI-ARENA to evaluate different image and video generative models, where users can actively participate in evaluating these models. By leveraging collective user feedback and votes, GENAI-ARENA aims to provide a more democratic and accurate measure of model performance. It covers three arenas for text-to-image generation, text-to-video generation, and image editing respectively. Currently, we cover a total of 35 open-source generative models. GENAI-ARENA has been operating for seven months, amassing over 9000 votes from the community. We describe our platform, analyze the data, and explain the statistical methods for ranking the models. To further promote the research in building model-based evaluation metrics, we release a cleaned version of our preference data for the three tasks, namely GenAI-Bench. We prompt the existing multi-modal models like Gemini, GPT-4o to mimic human voting. We compute the accuracy by comparing the model voting with the human voting to understand their juding abilities. Our results show existing multimodal models are still lagging in assessing the generated visual content, even the best model GPT-4o only achieves an average accuracy of 49.19% across the three generative tasks. Open-source MLLMs perform even worse due to the lack of instruction-following and reasoning ability in the complex vision scenarios.

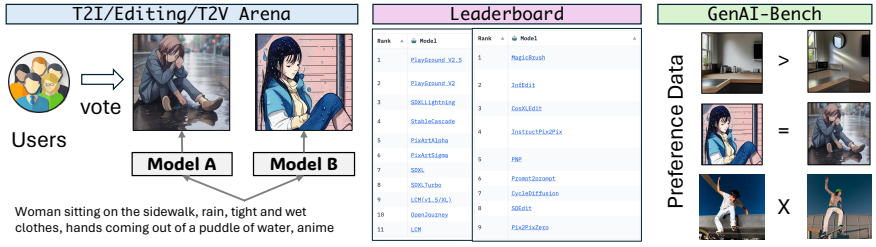

Figure 1: GenAI Arena contains three components: (1) text-to-image, text-to-video and image editing arena, which accept community voting to obtain the preference pairs. (2) The leaderboard utilizes the preference pairs to calculate elo ranking for all the evaluated models. (3) We further release GenAI-Bench to judge different multimodal LLM judges.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-2538

1 Introduction

Image generation and manipulation technologies have seen rapid advancements, leading to their widespread application across various domains such as creating stunning artwork [54, 67, 83, 21], enhancing visual content [6, 44], and aiding in medical imaging [81, 11]. Despite these advancements, navigating through the multitude of available models and assessing their performance remains a challenging task [65]. Traditional evaluation metrics like PSNR, SSIM [76], LPIPS [84], and FID [20], while valuable, offer very specific insights into precise aspects of visual content generation. However, these metrics often fall short in providing a comprehensive assessment of overall model performance, especially when considering subjective qualities like aesthetics and user satisfaction [58].

To address these challenges, we introduce GenAI-Arena—a novel platform designed to enable fair evaluation. Inspired by successful implementations in other domains [86, 53], GenAI-Arena offers a dynamic and interactive platform where users can generate images, compare them side-by-side, and vote for their preferred models. Such a platform not only simplifies the process of comparing different models but also provides a ranking system that reflects human preferences, thereby offering a more holistic evaluation of model capabilities. To our knowledge, GenAI-Arena is the first evaluation platform with comprehensive evaluation capabilities across multiple properties. Unlike other platforms, it supports a wide range of tasks across text-to-image generation, text-guided image editing, and text-to-video generation, along with a public voting process to ensure labeling transparency. The votes are utilized to access the evaluation ability of Multimodal Large Language Model (MLLM) evaluators. Table 1 shows our platform excels in its versatility and transparency.

Since February 11th, 2024, we have collected over 9000 votes for three multimodal generative tasks. We constructed leaderboards for each task with these votes, identifying the state-of-the-art models as PlayGround V2.5, MagicBrush, and StableVideoDiffusion, respectively (until Oct 24th, 2024). Detailed analyses based on the votes are presented. For example, our plotted winning fraction heatmaps reveal that while the Elo rating system is generally effective, it can be biased by imbalances between "easy games" and "hard games". We also performed several case studies for qualitative analysis, demonstrating that users can provide preference votes from multiple evaluation aspects, which help distinguish subtle differences between the outputs and upload high-quality votes for Elo rating computation.

Automatically assessing the quality of generated visual content is a challenging problem for several reasons: (1) images and videos have many different aspects like visual quality, consistency, alignment, artifacts, etc. Such a multi-faceted nature makes the evaluation intrinsically difficult. (2) the supervised data is relatively scarce on the web. In our work, we release the user voting data as GenAI-Bench to enable further development in this field. Specifically, we calculate the accuracy between different image/video auto-raters (i.e. MLLM judges like GPT-4o, Gemini, etc.) with user preference to understand their judging abilities. Our results show that even the best MLLM, GPT-4o achieves at most 49.19% accuracy compared with human preference.

Table 1: Comparison with different evaluation platforms on different properties.

Platform	Text-To-Image Generation	Text-Guided Image Editing	Text-To-Video Generation	Human Label Transparency	Open/Public Voting Process	Judging MLLM judge
T2I-CompBench [24]	✓	✗	✗	✗	✗	✗
HEIM [38]	✓	✗	✗	✓	✗	✗
ImagenHub [32]	✓	✓	✗	✓	✗	✗
VBench [25]	✗	✗	✓	✓	✗	✗
EvalCrafter [50]	✗	✗	✓	✓	✗	✗
GENAI-ARENA	✓	✓	✓	✓	✓	✓

To summarize, our work's contributions include:

- GenAI-Arena, the first open platform to rank multi-modal generative AI based on user preferences.
- Discussion and case studies of collected user votes, showing the reliability of GenAI-Arena.
- GenAI-Bench, a public benchmark for judging MLLM's evaluation ability for generative tasks.

https://doi.org/10.52202/079017-2538

2 Related Work

2.1 Generative AI Evaluation Metrics

Numerous methods have been proposed to evaluate the performance of multi-modal generative models in various aspects. In the context of image generation, CLIPScore [19] is proposed to measure the text-alignment of an image and a text through computing the cosine similarity of the two embeddings from CLIP [64]. IS [68] and FID [20] measure image fidelity by computing a distance function between real and synthesized data distributions. PSNR, SSIM [76] assess the image similarity. LPIPS [84] and the follow-up works [15, 16] measure the perceptual similarity of images. More recent works leverage the Multimodal Large Language Model (MLLM) as a judge. T2I-CompBench [24] proposed the use of miniGPT4 [87] to evaluate compositional text-to-image generation task. TIFA [23] further adapted visual question answering to compute scores for the text-to-image generation task. VIEScore [31] leveraged MLLMs as a unified metric across image generation and editing tasks, reporting that MLLM has great potential in replacing human judges.

Metrics in similar fashions are also proposed for the video domain. For example, FVD [72] measures the coherence shifts and quality in frames. CLIPSIM [64] utilizes an image-text similarity model to assess the similarity between video frames and text. VBench [25] and EvalCrafter [50] also proposed different metrics for evaluating different aspects of the video generation task. However, these automatic metrics still lag compared with human preferences, achieving low correlation and thus giving doubts to their reliability.

2.2 Generative AI Evaluation Platforms

While auto-metric focuses on evaluating a single model's performance, evaluation platforms aim to systematically rank a group of models. Recently, several benchmark suites have been developed to comprehensively assess generative AI models. For image generation, T2ICompBench [24] evaluates compositional text-to-image generation tasks, while HEIM [38] offers a holistic evaluation framework that measures text-to-image tasks across multiple dimensions, including safety and toxicity. Similarly, ImagenHub [32] evaluates text-to-image, image editing, and other prevalent image generation tasks in a unified benchmark suite. For video generation, VBench [25] and EvalCrafter [50] provide structured evaluation approaches ensuring rigorous assessment. Despite their functionality, these benchmarks rely on model-based evaluation metrics, which are less reliable than human evaluation.

To address this issue, variable model arenas have been developed to collect direct human preferences for ranking models. Chatbot Arena by LMsys [13] is the pioneering platform in this regard, setting the standard for evaluation. Subsequent efforts have led to the creation of arenas for vision-language models [78], TTS models [53], and tokenizers [28]. However, there is no existing arena for generative AI models. To fill this gap, we propose GenAI-Arena as a complementary solution in this field.

3 GenAI-Arena: Design and Implementation

3.1 Design

GenAI-Arena is designed to offer an intuitive and comprehensive evaluation platform for generative models, facilitating user interaction and participation. The platform is structured around three primary tasks: text-to-image generation, image edition, and text-to-video generation. Each task is supported by a set of features that include an anonymous and a non-anonyumous battle playground, a direct generation tab, and a leaderboard as shown in Figure 2 . These features are designed to cater to both casual users and researchers, ensuring a democratic and accurate assessment of model performance.

Standardized Inference To ensure a fair comparison between different models, we ported the highly dispersed codebase from the existing works and then standardized them into a unified format. During inference, we fixed the hyper-parameters and the prompt format to prevent per-instance prompt or hyper-parameter tuning, which makes the inference of different models fair and reproducible. Following ImagenHub [32], we build the new library of VideoGenHub (details in subsection A.5), which aims to standardize the inference procedure for different text-to-video and image-to-video models. We find the best hyper-parameters of these models to ensure their highest performance.

https://doi.org/10.52202/079017-2538

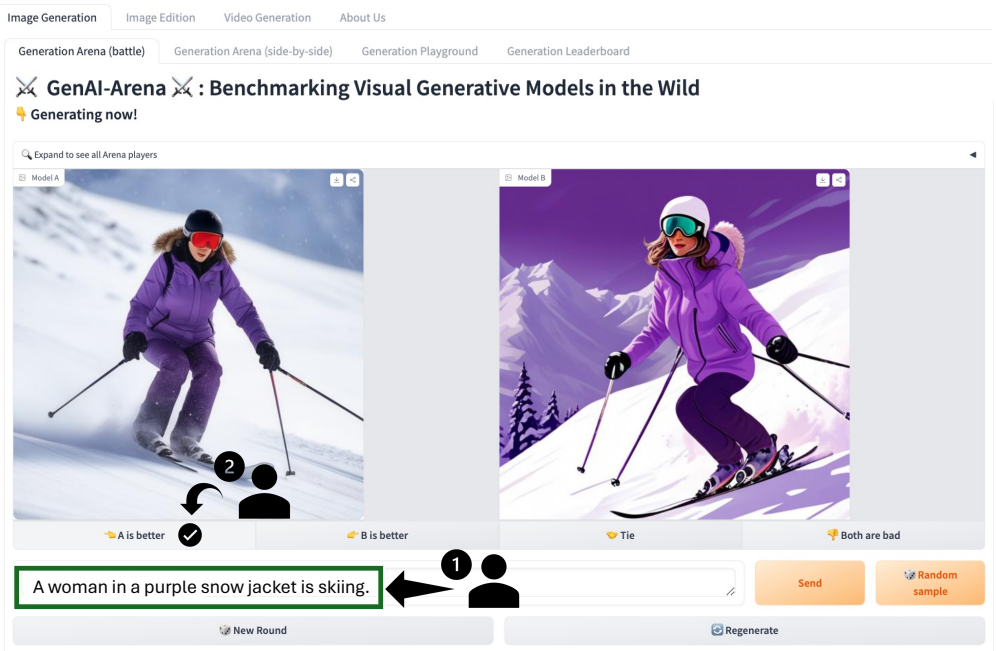

Figure 2: GenAI Arena User Voting Interface.

Voting Rules The anonymous battle section is designed to ensure unbiased voting and accurate evaluation of generative models. The rules for this section are as follows:

1. Users input a prompt, which is then used to generate outputs from two anonymous models within the same category of task.
2. The generated outputs from the two anonymous models are presented side-by-side for comparison.
3. Users can vote based on their preference using the options: 1) left is better; 2) right is better; 3) tie; 4) both are bad. These four options are being used to calculate Elo ranking.
4. Once the user has made their decision, they click the Vote button to submit their vote. It is important to ensure that the identity of the models remains anonymous throughout the process. Votes will not be counted if the model identity is revealed during the interaction.

3.2 Model Integration

In GenAI-Arena, we incorporate a diverse array of state-of-the-art generative models, covering a broad range of generative tasks including text-to-image generation, image edition, and text-to-video generation. To ensure comprehensive evaluations, the platform includes models that employ diverse underlying technologies, such as different types of architectures, training paradigms, training data and acceleration techniques. These variations can offer insights to understand these factors rigorously.

Text-to-Image Generation In Table 2, we list all the included text-to-image generation models. For example, SDXL, SDXL-Turbo, and SDXL-Lightning are all derived based on SDXL [63], while SDXL-Turbo [69] and SDXL-Lightning [47] adopt different distillation method. We also include diffusion transformer models [60] like PixArt-α and PixArt-σ. Playground V2 and Playground V2.5 are based on SDXL architecture, but trained by Playground.ai from scratch with an internal dataset. We have also included the latest released HunyuanDiT [45], FLUX.1-dev [35], FLUX.1-schnell [35].

Text-guided Image Editing In Table 3, we list all the image editing models and approaches. Some of them are plug-and-play approaches without requiring any training, like Pix2PixZero [59], InfEdit [79], SDEdit [52], etc. These methods can be applied to a broad range of diffusion models. Some of the models like PnP [71] and Prompt2Prompt [18] require DDIM inversion, which takes

Table 2: The overview of all text-to-image generation models.

Model	Size	Method	Resolution	#Steps
OpenJourney [57]	1B	SD-2.1 + MidJourney Dataset	512x512	50
LCM [51]	1B	SD-2.1 + Consistency Distillation	512x512	4
SDXL [63]	3.5B	Latent Diffusion Model (LDM)	1K×1K	50
SDXL-Turbo [69]	3.5B	LDM + Distillation	1K×1K	1
SDXL-Lightning [47]	3.5B	LDM + Distillation	1K×1K	4
PixArt-α [9]	0.6B	Diffusion Transformer (DiT)	1K×1K	50
PixArt-σ [10]	0.6B	DiT + Weak-to-Strong	4K×4K	50
StableCascade [62]	1.5B + 3.6B	Würstchen	1K×1K	20+10
Playground V2 [42]	3.5B	LDM	1K×1K	50
Playground V2.5 [41]	3.5B	LDM	1K×1K	50
FLUX.1-dev [35]	12B	Guidance-distilled DiT + Flow Matching	1K×1K	20
FLUX.1-schnell [35]	12B	Timestep-distilled DiT + Flow Matching	1K×1K	4
Kolors [33]	2.6B	LDM + ChatGLM3	1K×1K	50
HunyuanDiT [45]	1.5B	DiT + multilingual text encoder	1K×1K	50
Stable Diffusion 3 [70]	8B	Multimodal DiT	1K×1K	50
AuraFlow [14]	6.8B	Flow-based Model	1K×1K	50

Table 3: Overview of all the image editing models.

Model	Trained?	Method	Runtime
Pix2PixZero [59]	Zero-shot	Editing Direction Discovery + Attention Control	21s
SDEdit [52]	Zero-shot	Iteratively Denoising through SDE	13s
CycleDiffusion [77]	Zero-shot	Reconstructable Encoder for Stochastic DPMs	9s
Prompt2Prompt [18]	Zero-shot	Prompt-based Cross-attention Control	120s
PnP [71]	Zero-shot	Feature and Self-attention Injection	120s
InfEdit [79]	Zero-shot	Consistent Model + Uni-Attention Control	5s
InstructPix2Pix [6]	Trained	Instruction-based Fine-tuning with Synthetic Data	12s
MagicBrush [82]	Trained	Instruction-based Fine-tuning with Annotated Data	12s
CosXLEdit [1]	Trained	Cosine-Continuous EDM VPred schedule	50s

Table 4: Overview of all text-to-video generation models.

Model	Base	Len	FPS	Dataset	Resolution	#Steps
AnimateDiff [17]	SD-1.5	2s	8	WebVid10M	512×512	25
AnimateDiff-Turbo [17]	SD-1.5	2s	8	WebVid10M	512×512	4
ModelScope [73]	SD-1.5	2s	8	WebVid10M	256×256	50
LaVie [75]	SD-1.5	2s	8	Vimeo25M	320×512	50
StableVideoDiffusion [4]	SD-2.1	2.5s	10	LVD-500M	576×1024	20
VideoCrafter2 [8]	SD-2.1	2s	16	WebVid10M	320×512	50
T2V-Turbo [43]	VideoCrafter2	2s	8	WebVid10M	320×512	4
OpenSora [55]	Pixart-α	2s	16	WebVid10M	320×512	50
OpenSora v1.2 [55]	Pixart-α	2s	16	WebVid10M	320×512	50
CogVideoX-2B [80]	DiT	2s	8	35M videos + 2B images	480×720	50

much longer time than the other approaches. We also include specialized trained image editing models like InstructP2P [6], MagicBrush [82] and CosXLEdit [1].

Text-to-Video Generation In Table 4, we list all the text-to-video generation models. We include different types of models. For example, AnimateDiff [17], ModelScope [73], Lavie [75] are initialized from SD-1.5 and continue trained by injecting a motion layer to capture the temporal relation between frames. In contrast, StableVideoDiffusion [4] and VideoCrafter2 [7] are iniialized from SD-2.1. Besides these models, we also include OpenSora [55], which utilizes a Sora-like diffusion transformer [60] architecture for joint space-time attention.

3.3 Elo Rating System

Online Elo Rating The Elo rating system models the probability of player i winning against player j, based on their current ratings, R_i and R_j respectively, where $i, j \in N$. We define a binary outcome Y_{ij} for each comparison between player i and player j, where $Y_{ij} = 1$ if player i wins and $Y_{ij} = 0$ otherwise. The logistic probability is formulated as:

$$P(Y_{ij} = 1) = \frac{1}{1 + 10^{(R_j - R_i)/\alpha}} \tag{1}$$

https://doi.org/10.52202/079017-2538

where $\alpha = 400$ for Elo rating computation. After each match, a player's rating is updated using the formula:

$$R'_i = R_i + K \times (S(i,j) - E(i,j)) \tag{2}$$

where $S(i,j)$ is the actual match outcome, $S(i,j) = 1$ for a win $S(i,j) = 0.5$ for a tie, and $S(i,j) = 0$ for a loss, and $E(i,j) = P(Yi, j = 1)$.

For example, given a model's Elo rating as 1200 and the other model's elo rating as 1100, then the estimated probability of the first model winning will be $\frac{1}{1+10^{(1100-1200)/400}} \approx 0.64$. In this way, we can have a direct understanding of the elo rating's meaning. This mapping from absolute number to the pairwise winning rate of two models gives a more straightforward understanding of the meaning of elo rating score.

Another design logic behind the Elo rating is that a higher-rated player should gain fewer points if they win a lower-rated player, but lose more if they lose the game, whereas the lower-rated player experiences the opposite. In this way, the order of a specific set of matches will significantly affect the final computed Elo rating, as the player's Elo rating and the rating gain of each match are both changing dynamically. This online Elo rating system might be good for real-world competitions, where players usually have less than 100 competitions a year. However the arena for AI models usually comes with thousands of votes (competitions), and the quality of votes is not ensured. Thus, it's necessary to acquire an order-consistent and more stable elo rating. To do this, we follow Chatbot Arena [12] to adopt the Bradley–Terry model [5] for a statistically estimated elo rating.

Bradley–Terry Model Estimation The Bradley–Terry (BT) model [5] estimates Elo ratings using logistic regression and maximum likelihood estimation (MLE). Suppose there are N players and we have a series of pairwise comparisons, where W_{ij} is the number of times player i has won against player j. The log-likelihood function for all pairwise comparisons is written as:

$$\mathcal{L}(\mathbf{R}) = \sum_{i,j \in N, i \neq j} (W_{ij} Y_{ij} \log P(Y_{ij} = 1)) \tag{3}$$

where $\mathbf{R} = \{R_1, \ldots, R_N\}$ represents the Elo ratings of each player. The Bradley–Terry model provides a stable statistical estimation of the players' ratings by consistently incorporating all pairwise comparisons, thus overcoming the limitations of direct Elo computation in online settings.

Since the BT model does not account for ties, we first duplicate all the votes, then allocate half of the "tie" votes to the scenario where model i wins ($Y_{ij} = 1$) and the other half to the scenario where model j wins ($Y_{ij} = 0$) in practice. We model the solver to be a logistic regression model and solve it via the `LogisticRegression` model from `sklearn` for the solving.

Confidence Interval To further investigate the variance of the estimated Elo rating, we use the "sandwich" standard errors described in Huber et al. [26]. That is, for each round, we record the estimated Elo rating based on the same number of battles sampled from the previous round. This process continues for 100 rounds. We select the lowest sampled elo rating as the lower bound of the confidence interval, and the highest sampled elo rating as the upper bound of the elo rating.

Selection of battle pair With a limited number of games, choosing which two players to match up is a crucial issue. The simplest approach, which we currently use, is to randomly select two players. However, this can introduce bias, with some models getting significantly more matches than others. A vote-aware selection system that increases the probability of selecting less-played models and lowers it for more-played ones is needed, and we plan to explore this in future Arena improvements.

3.4 GenAI-Museum

Current GenAI-Arena runs the model on the Hugging Face Zero GPU system [27]. As shown in Table 3, the time for a single generative inference usually ranges from 5 to 120 seconds. Unlike the auto-regression language model, where inference acceleration techniques like VLLM [34], SGLang [85] generate responses in less than a second, diffusion model community does not have such powerful infrastructure. Therefore, pre-computation becomes a necessary way to mitigate computational overhead and streamline user interaction.

To achieve this, we serve GenAI-Museum as a pre-computed data pool comprising various inputs from existing datasets or user collection, along with each model's output. Based on this, a `"Random`

Table 5: GenAI-Arena Leaderboards.
(Last updated on Oct 24th, 2024)

(a) Text-to-Image (Top-10)

Model	Elo	95% CI
PlayGround V2.5	1122	+19/-20
FLUX.1-dev	1114	+45/-42
FLUX.1-schnell	1085	+43/-46
Playground V2	1072	+18/-22
Kolors	1069	+32/-39
StableCascade	1057	+17/-21
HunyuanDiT	1030	+25/-27
PixArt-α	1020	+17/-19
SDXL-Lightning	1020	+17/-15
PixArt-σ	1019	+22/-20

(b) Image Editing

Model	Elo	95% CI
MagicBrush	1108	+32/-28
InfEdit	1075	+26/-32
CosXLEdit	1066	+31/-29
InstructPix2Pix	1038	+32/-24
PNP	998	+35/-34
Prompt2prompt	988	+26/-23
CycleDiffusion	943	+26/-26
SDEdit	924	+24/-23
Pix2PixZero	858	+25/-30

(c) Text-to-Video

Model	Elo	95% CI
StableVideoDiffusion	1148	+31/-28
CogVideoX-2B	1106	+71/-71
T2V-Turbo	1085	+36/-32
VideoCrafter2	1068	+20/-22
AnimateDiff	1068	+25/-21
LaVie	996	+25/-21
OpenSora	912	+23/-23
OpenSora v1.2	894	+54/-71
ModelScope	862	+25/-22
AnimateDiff-Turbo	861	+22/-20

`Sample"` button shown in Figure 2 is additionally implemented to facilitate the random generation of prompts and the immediate retrieval of corresponding images or videos. This functionality operates by sending requests to our deployed GenAI-Museum every time `"Random Sample"` button is hit, receiving input and two random model's pre-computed outputs. In this way, we save the computation time on the GPU, enable users to do instant comparisons and votes on the UI, and balance the votes for each unique input so we gradually collect votes for a full combination of all models. The input prompts were sampled from ImagenHub [32] and VBench [25]. To prevent the bias in the prompt distribution, we also periodically update the input prompts with the lastest collected real-world human votes. We make sure every prompt is filtered via NSFW detector before adding them.

4 Benchmarks and Results Discussion

4.1 Arena Leaderboard

We report our leaderboard at the time of paper publishing in Table 5. For image generation, we collected 6300 votes in total. The currently top-1 model is Playground V2.5, released by Playground.ai, which follows the same architecture as SDXL but is trained with a private dataset. In contrast, SDXL only ranks in the thirteenth position, lagging significantly behind. Such finding highlights the importance of the training dataset. StableCascade is ranked in the sixth place in the leaderboard, which utilizes a highly efficient cascade architecture to lower the training cost. According to Würstchen [62], StableCascade only requires a 10% training cost of SD-2.1, yet it can beat SDXL significantly on our leaderboard. This highlights the importance of the diffusion architecture to achieve strong performance. For image editing, a total of 1154 votes have been collected. MagicBrush, InFEdit, CosXLEdit, and InstructPix2Pix ranked higher as they can perform localized editing on images. PNP preserves the structure with feature injections, thus limiting the edit variety. The older methods such as Prompt-to-Prompt, CycleDiffusion, SDEdit, and Pix2PixZero, frequently result in completely different images during editing despite the high-quality images, which explains the lower ranking of these models. For text-to-video, there is a total of 2024 votes. StableVideoDiffusion leads with the highest Elo score, suggesting it is the most effective model. Close behind, CogVideoX-2B ranks second. The following VideoCrafter2 and AnimateDiff have very close elo scores, showing nearly equivalent capabilities. LaVie, OpenSora, ModelScope, and AnimateDiff-Turbo follow with decreasing scores, indicating progressively lower performance.

4.2 Discussion and Insights

Winning Fraction and Elo Rating We visualize the winning fraction heatmap in Figure 3, where each cell represents the actual winning fraction of Model A over Model B. The models are ordered by their Elo rating in the heatmap. Horizontally across each row, the winning fraction of Model A increases as the Elo rating of Model B decreases, demonstrating the effectiveness of the Elo rating system in ranking different models.

Specific cells in the heatmap reveal notable findings. For instance, although PlayGround 2.5 achieves the state-of-the-art (SOTA) Elo rating in the Text-to-Image task, its winning fraction over PixArt-σ is only 0.58, which is below 60%. The higher Elo rating of T2V-Turbo might be due to our Arena collecting more votes from "easy games" with low-ranked models and fewer from "harder

https://doi.org/10.52202/079017-2538

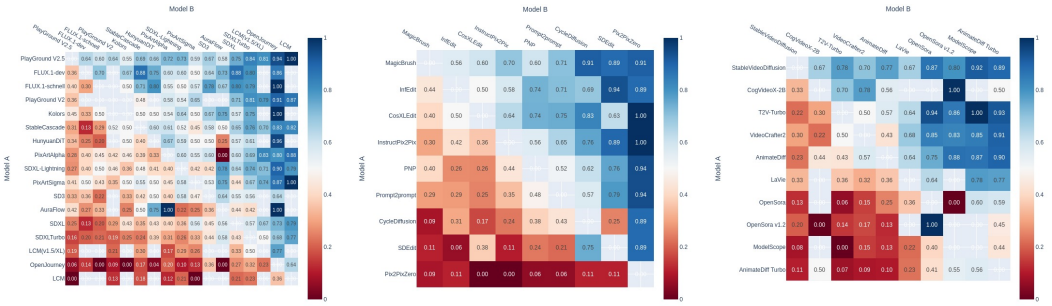

Figure 3: Winning fraction heatmap of different models for the three tasks in GenAI-Arena

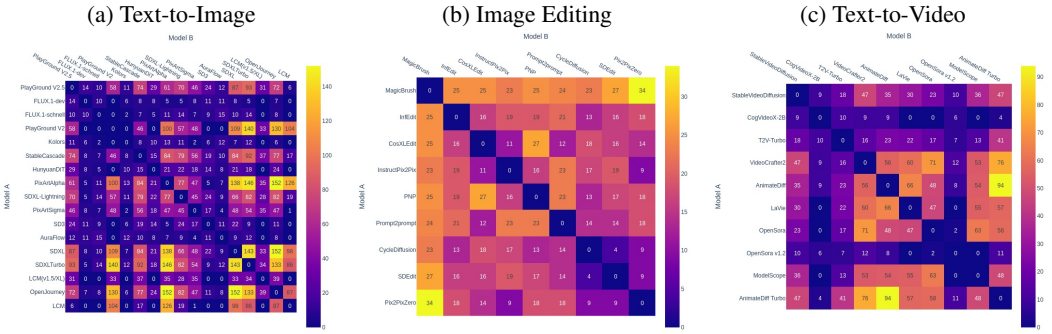

Figure 4: Battle count heatmap of different models for the three tasks in GenAI-Arena (without Ties)

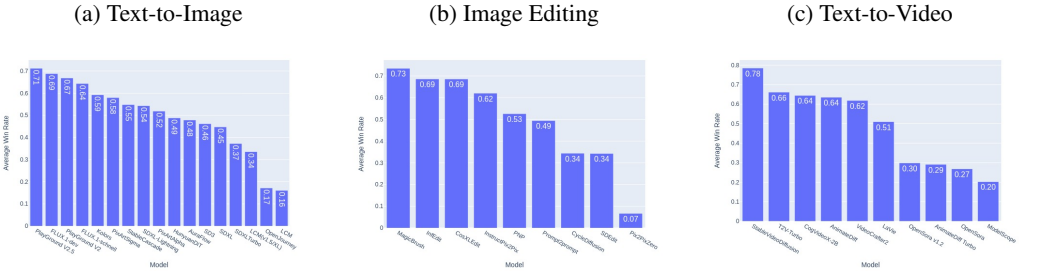

Figure 5: Average Win Rate Against All Other Models (Assuming Uniform Sampling and No Ties)

games" with high-ranked models. For example, the number of battles between PlayGround V2.5 and SDXL-Turbo (93) is way more than PlayGround V2.5 with other models (around 50) in Figure 4.

These anomalies highlight potential drawbacks of the Elo rating system: (1) a reliable and robust Elo rating requires a large amount of voting data, and (2) the estimated Elo rating may be biased by the imbalance between "easy games" and "harder games," as they carry similar weight in the estimation.

As shown in Figure 5, we observe that the average win rates of the top-ranked models are all quite similar, none exceeding 80%. This indicates that there is no dominant, highly powerful model in text-to-image, image editing, or text-to-video generation at this time. The community is still awaiting a "ChatGPT moment"—the release of a breakthrough model with transformative capabilities.

Quality assessment of collected human votes Since our arena users come from different backgrounds and have different preferences, we conduct an expert review on a small set of sampled human vote to ensure there are no severe quality issues of our collected votes. We let different authors review 50 items for each set. A total of 350 items from our GenAI-Bench are evaluated. During the

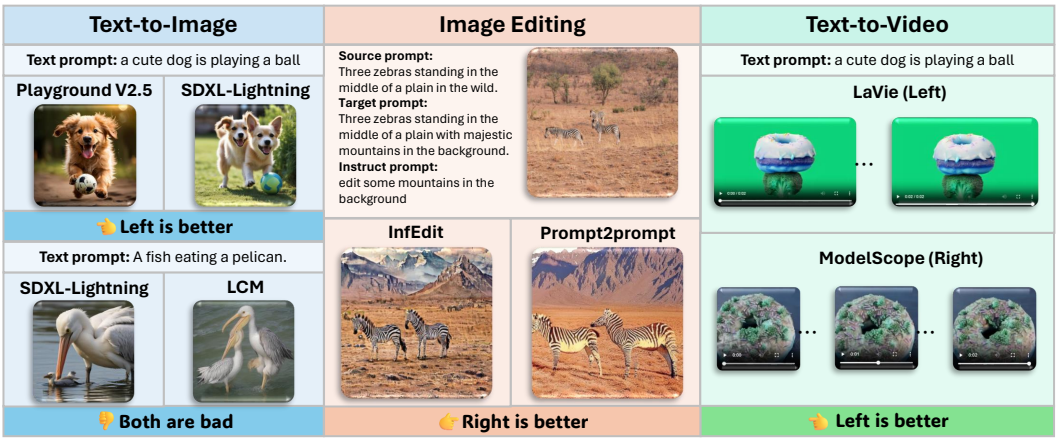

Figure 6: Example of votes from users on the GenAI-Arena for the three generative tasks

annotations, we skipped those bad items due to NSFW or technical issues, and we finally collected 303 valid evaluations. For each vote, 3 available labels for provided for annotating:

- Clearly Reasonable Vote: This vote will be clearly agreed by most of the people.
- Vague Vote: The current vote makes sense. But it's also reasonable if other vote is selected.
- Wrong Vote: This vote will be clearly disagreed by most of the people.

Table 6: Expert Review for 350 sampled human votes

(a) Distribution of Valid Votes

# Valid votes	# NSFW	# Tech issue	Total
303	17	30	350
86.57%	4.86%	8.57%	100%

(b) Distribution of quality labels

# Clearly Reasonable Vote	# Vague Vote	# Wrong Vote	Total
231	51	21	303
76.24%	16.83%	6.93%	100%

We report the distribution of valid votes in Table 6a, and find that 86.57% of the votes are valid without NSFW issues. Among these valid votes, about 76.24% of the votes are clearly reasonable votes and 93.07% of the votes are either clearly reasonable or vaguely reasonable, as shown in Table 6b. We believe this shows the reliability of our preference data.

Case Study We present case studies in Figure 6, showcasing the votes collected for three generative tasks. These cases demonstrate that GenAI-Arena users can provide high-quality votes, even for the most advanced models. For instance, in the text-to-image task, the image generated by PlayGround V2.5 was preferred over that of SDXL-Lightning for the prompt "a cute dog is playing with a ball," as the latter depicted two dogs instead of one. Users can clearly distinguish and vote based on the quality of the outputs, even when both models complete the task. In the image editing task, the edited image from Prompt2Prompt appeared more natural than the one from InfEdit, leading users to make a definitive vote. Similarly, votes collected for the text-to-video task were also of high quality.

5 GenAI-Bench

5.1 Dataset

We applied Llama Guard [29] as an NSFW filter to ensure that the user input prompt is appropriate for a wide range of audiences and protects users of the benchmark from exposure to potentially harmful or offensive content. In the text-to-image generation task, we collect 4.3k anonymous votes in total and there are 1.7k votes left after filtering for the safe content. We observe a large amount of the prompt is filtered out due to sexual content, which takes up 85.6% of the abandoned data. In the text-guided image editing task, we collect 1.1k votes from users before filtering. After applying Llama Guard,

https://doi.org/10.52202/079017-2538

there are 0.9k votes for the image edition being released. In this task, 87.5% of the unsafe inputs contain violent crimes, and the other 12.5% is filtered out resulting from sex-related crimes. For text-to-video generation task, our platform collects 1.2k votes before post-processing. After cleaning it with the NSFW filter, we release the remaining 1.1k votes. All of the unsafe data abandoned in this task is due to the sexual content. We released the current version of GenAI-Bench[1] on the HuggingFace Dataset website, with an MIT license to allow the reuse with or without modification.

5.2 GenAI-Bench Leaderboard

To construct the GenAI-Bench leaderboard, we propmt MLLMs to output preference labels of AI generated contents, where templates are defined in subsection A.6. Specifically, We selected MLLMs including GPT-4o [56], Gemini-1.5-Pro [66], Idefics2 [37], etc., and ask them to output 4 labels: "[[A>B]]", "[[B>A]]", "[[A=B=GOOD]]", and "[[A=B=BAD]]". We then compare them with actual human preference labels collected through the GenAI-Arena using the exact match metric. As shown in Table 7, open-source model still lag behind close-source MLLMs such as GPT-4o and Gemini, indicating a lack of generalization ability in vision reasoning of open-source MLLMs. We also tried models including Fuyu [3], Kosmos-2 [61], Otter [39], Mantis [30], etc., but found that they cannot follow the instruction well to output reasonable labels.

Table 7: GenAI-Bench leaderboard designed to benchmark MLLMs's ability in judging the quality of AI generative contents by comparing with human preferences. Numbers are accuracy (%).

Model	Image Generation	Image Editing	Video Generation	Average
Random	25.36	25.90	25.16	25.47
Idefics1 [36]	0.81	5.66	0.19	2.22
InstructBLIP [6]	3.11	19.80	3.74	8.89
QwenVL [2]	26.63	14.91	2.15	14.56
CogVLM [74]	29.34	0.00	24.60	17.98
VideoLLaVA [46]	37.75	26.66	0.00	21.47
BLIP-2 [40]	26.34	26.01	16.93	23.09
MiniCPM-V-2.5 [22]	37.81	25.24	6.55	23.20
LLaVA-1.6-7B [49]	22.65	25.35	21.70	23.24
Idefics2 [37]	42.25	27.31	16.46	28.67
LLaVA-1.5-7B [48]	37.00	26.12	30.40	31.17
Gemini-1.5-Pro [66]	44.67	**55.93**	46.21	48.94
GPT-4o [56]	**45.59**	53.54	**48.46**	**49.19**

6 Conclusion

In this paper, we introduced GenAI-Arena, an open platform designed to rank generative models across text-to-image, image editing, and text-to-video tasks based on user preference. unlike other platforms, GenAI-Arena is driven by community voting to ensure transparency and sustainable operation. We employed the side-by-side human voting method to evaluate the models and collected over 9000 votes starting from February 11th, 2024. We compiled an Elo leaderboard with the votings and found that PlayGround V2.5, MagicBrush, and StableVideoDiffusion are the current state-of-the-art models in the three tasks (until Oct 24th, 2024). Analysis based on the collected votes shows that while the Elo rating is generally functional, but can biased by the imbalance of the "easy games" and "hard games". Our expert review of 350 sampled human votes confirmed that 93.07% of the votes can be viewed as either clearly reasonable or vaguely reasonable, demonstrating the high quality of our collected votes What's more, we also released the human preference voting as GenAI-Bench. We prompt the existing MLLMs to evaluate the generated images and videos on GenAI-Bench and compute the accuracy with human voting. The experiment showed that the open-source MLLMs achieve very low performance, even the best model GPT-4o can only achieve 49.19% accuracy. This is mostly because their lack of instruction-following and reasoning ability in complex vision scenarios. In the future, we will continue collecting human votes to update the leaderboard, helping the community to keep track of the research progress. We also plan to develop a more robust MLLM to better approximate human ratings in GenAI-Bench.

[1] https://huggingface.co/datasets/TIGER-Lab/GenAI-Bench

References

[1] S. AI. CosXL. https://huggingface.co/stabilityai/cosxl, 2024. Accessed on: 2024-04-13.

[2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023. URL https://api.semanticscholar.org/CorpusID:263875678.

[3] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

[4] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, and D. Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023. URL https://api.semanticscholar.org/CorpusID:265312551.

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

[6] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[7] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[8] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C.-L. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *ArXiv*, abs/2401.09047, 2024. URL https://api.semanticscholar.org/CorpusID:267028095.

[9] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. T. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ArXiv*, abs/2310.00426, 2023. URL https://api.semanticscholar.org/CorpusID:263334265.

[10] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *ArXiv*, abs/2403.04692, 2024. URL https://api.semanticscholar.org/CorpusID:268264262.

[11] Q. Chen, X. Chen, H. Song, Z. Xiong, A. Yuille, C. Wei, and Z. Zhou. Towards generalizable tumor synthesis, 2024.

[12] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference. *ArXiv*, abs/2403.04132, 2024. URL https://api.semanticscholar.org/CorpusID:268264163.

[13] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[14] fal. Auraflow, 2024. URL https://huggingface.co/fal/AuraFlow. Hugging Face repository.

[15] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] S. Ghazanfari, A. Araujo, P. Krishnamurthy, F. Khorrami, and S. Garg. Lipsim: A provably robust perceptual similarity metric. In *The Twelfth International Conference on Learning Representations*, 2023.

 https://doi.org/10.52202/079017-2538

[17] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

[18] A. Hertz, R. Mokady, J. M. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ArXiv*, abs/2208.01626, 2022. URL https://api.semanticscholar.org/CorpusID:251252882.

[19] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[21] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[22] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao, X. Zhang, Z. L. Thai, K. Zhang, C. Wang, Y. Yao, C. Zhao, J. Zhou, J. Cai, Z. Zhai, N. Ding, C. Jia, G. Zeng, D. Li, Z. Liu, and M. Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. *ArXiv*, abs/2404.06395, 2024. URL https://api.semanticscholar.org/CorpusID:269009975.

[23] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[24] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[25] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[26] P. J. Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA: University of California Press, 1967.

[27] Hugging Face. Zerogpu. https://huggingface.co/zero-gpu-explorers, 2024. Accessed: 2024-06-02.

[28] Hugging Face Spaces. Tokenizer arena. https://huggingface.co/spaces/eson/tokenizer-arena, 2024. Accessed: 2024-06-05.

[29] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *ArXiv*, abs/2312.06674, 2023. URL https://api.semanticscholar.org/CorpusID:266174345.

[30] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

[31] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2024.

[32] M. Ku, T. Li, K. Zhang, Y. Lu, X. Fu, W. Zhuang, and W. Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OuV9ZrkQlc.

[33] Kwai-Kolors. Kolors, 2024. URL https://github.com/Kwai-Kolors/Kolors. GitHub repository.

[34] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[35] B. F. Labs. Flux, 2024. URL https://github.com/black-forest-labs/flux. GitHub repository.

[36] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[37] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models?, 2024.

[38] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024.

[39] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726, 2023. URL https://api.semanticscholar.org/CorpusID:258547300.

[40] D. Li, J. Li, and S. C. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023.

[41] D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *ArXiv*, abs/2402.17245, 2024. URL https://api.semanticscholar.org/CorpusID:268033039.

[42] D. Li, A. Kamko, A. Sabet, E. Akhgari, L. Xu, and S. Doshi. Playground v2, 2024. URL https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic.

[43] J. Li, W. Feng, T.-J. Fu, X. Wang, S. Basu, W. Chen, and W. Y. Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *ArXiv*, 2024. URL https://api.semanticscholar.org/CorpusID:270094742.

[44] T. Li, M. Ku, C. Wei, and W. Chen. Dreamedit: Subject-driven image editing. *Transactions on Machine Learning Research*, 2023.

[45] Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, X. Deng, Y. Zhang, X. Liu, M. Huang, Z. Xiao, D. Chen, J. He, J. Li, W. Li, C. Zhang, R. Quan, J. Lu, J. Huang, X. Yuan, X. Zheng, Y. Li, J. Zhang, C. Zhang, M. Chen, J. Liu, Z. Fang, W. Wang, J. Xue, Y. Tao, J. Zhu, K. Liu, S. Lin, Y. Sun, Y. Li, D. Wang, M. Chen, Z. Hu, X. Xiao, Y. Chen, Y. Liu, W. Liu, D. Wang, Y. Yang, J. Jiang, and Q. Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[46] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122, 2023. URL https://api.semanticscholar.org/CorpusID:265281544.

[47] S. Lin, A. Wang, and X. Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *ArXiv*, abs/2402.13929, 2024. URL https://api.semanticscholar.org/CorpusID:267770548.

[48] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023. URL https://api.semanticscholar.org/CorpusID:263672058.

[49] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

 https://doi.org/10.52202/079017-2538

[50] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[51] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *ArXiv*, abs/2310.04378, 2023. URL https://api.semanticscholar.org/CorpusID:263831037.

[52] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[53] mrfakename, V. Srivastav, C. Fourrier, L. Pouget, Y. Lacombe, main, and S. Gandhi. Text to speech arena. https://huggingface.co/spaces/TTS-AGI/TTS-Arena, 2024.

[54] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[55] N. U. of Singapore. Open-Sora: Democratizing Efficient Video Production for All. https://github.com/hpcaitech/Open-Sora/blob/main/docs/report_01.md, 2024. Accessed on: 2024-05-24.

[56] OpenAI. Gpt-4 technical report, 2023.

[57] openjourney.ai. Openjourney is an open source stable diffusion fine tuned model on midjourney images, 2023. URL https://huggingface.co/prompthero/openjourney.

[58] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.

[59] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[60] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[61] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824, 2023. URL https://api.semanticscholar.org/CorpusID:259262263.

[62] P. Pernias, D. Rampas, M. L. Richter, C. Pal, and M. Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[63] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Muller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

[64] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[65] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. URL https://api.semanticscholar.org/CorpusID:248097655.

[66] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[67] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

[68] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

[69] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. Adversarial diffusion distillation. *ArXiv*, abs/2311.17042, 2023. URL https://api.semanticscholar.org/CorpusID:265466173.

[70] Stability AI. Stable diffusion 3 release, 2024. URL https://stability.ai/news/stable-diffusion-3. News release.

[71] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[72] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[73] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang. Modelscope text-to-video technical report. *ArXiv*, abs/2308.06571, 2023. URL https://api.semanticscholar.org/CorpusID:260887737.

[74] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023. URL https://api.semanticscholar.org/CorpusID:265034288.

[75] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[76] Z. 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.

[77] C. H. Wu and F. D. la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023.

[78] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

[79] S. Xu, Y. Huang, J. Pan, Z. Ma, and J. Chai. Inversion-free image editing with natural language. In *Conference on Computer Vision and Pattern Recognition 2024*, 2024.

[80] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[81] H. Zhang, J. Yang, S. Wan, and P. Fua. Lefusion: Synthesizing myocardial pathology on cardiac mri via lesion-focus diffusion models, 2024.

[82] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS dataset and benchmark track*, 2023.

[83] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

 https://doi.org/10.52202/079017-2538

[84] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[85] L. Zheng, L. Yin, Z. Xie, J. Huang, C. Sun, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett, and Y. Sheng. Efficiently programming large language models using sglang, 2023.

[86] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[87] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

A Appendix

A.1 Broader Society Impacts

The establishment of GENAI-ARENA and the release of GenAI-Bench have broader societal implications. By democratizing the evaluation of generative models, GENAI-ARENA encourages transparency and community engagement in AI development. This can lead to more trust in AI technologies as the public can gain insights into how models perform according to peer evaluations. Moreover, involving the community in such evaluations can accelerate the identification of potentially harmful biases or unethical uses of AI technologies. However, there are potential risks associated with the widespread use of generative AI technologies that GENAI-ARENA evaluates. For instance, advancements in text-to-image and text-to-video generation can be misused for creating misleading or harmful content, such as those filtered by NSFW Filter.

A.2 Limitation

While the release of GENAI-ARENA can enable a more reasonable evaluation of the generative models, there are several limitations in its development. First, the diversity and representativeness of the user base participating in GENAI-ARENA may not fully encapsulate the broader population's preferences, which will potentially bias the evaluation results. Despite efforts to attract voters with diverse backgrounds, there is an inherent challenge in ensuring a balanced representation across different cultures or professional backgrounds. In addition, the reliance on user feedback and votes introduces subjectivity into the evaluation process. While this is partially mitigated by the volume of data collected, individual biases and varying levels of expertise among users can skew the results.

A.3 Data Collection

We stated in the GENAI-ARENA UI that the input and votes will be collected for research purposes only. By using this GENAI-ARENA tool, the users agree to the collection of their input and votes for research purposes. The users are acknowledged that their data will be anonymized and will not be used for commercial purposes.

A.4 Extra Visualization on GenAI-Arena

We included more analysis in Figure 7 and 5 to show the reliability of GenAI-Arena. Specifically, Figure 7 shows the error bar of the Elo rating to prove the reliability. For Figure 5, it predicts the average win rate if the model is played against other models.

(a) Text-to-Image (b) Image Editing (c) Text-to-Video

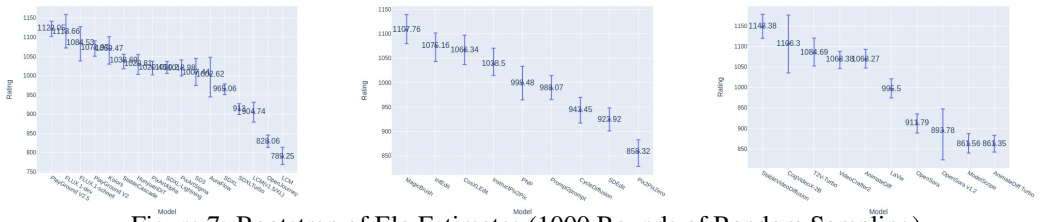

Figure 7: Bootstrap of Elo Estimates (1000 Rounds of Random Sampling)

A.5 VideoGenHub

VideoGenHub is an open-source library to standardize the inference and evaluation of all the conditional video generation models, similar to ImagenHub [32] in the image domain. In the library, all models are implemented with the literature standard, and the seeds are set as 42 for a fair comparison, which is the same standard as ImagenHub [32] implementation.

https://doi.org/10.52202/079017-2538

A.6 Prompt Templates for GenAI-Bench

We provide the prompt templates used to prompt MLLM to output their preferences for the genai bench data in the followings. MLLMs are required to output 4 labels including [[A>B]], [[B>A]], [[A=B=GOOD]], and [[A=B=BAD]]. Videos are extracted into image frames and fed into them as an image sequence, or directly fed into the model if the model have a specific video processing unit. We then compare their output labels with the real-world users preferences collected from out GenAI-Arena to judge a MLLM's ability in judging the quality of AI generative contents.

For text-to-image generation task, the prompt is as follows:

> Please act as an impartial judge and a professional digital artist to evaluate the quality of the responses provided by two AI image generation models to the user inputs displayed below. You will be given model A's generated image and model B's generated image. Your job is to evaluate which assistant's generated image is better.
>
> Text prompt: <prompt>
> Model A Generated Image: <left_image>
> Model B Generated Image: <right_image>
>
> When evaluating the quality of the generated images, you must identify the any inappropriateness in the edited images by considering the following criteria:
> 1. Whether the text prompt has been followed successfully in the generated image.
> 2. Whether the generated image looks natural, such as the sense of distance, shadow, and lighting.
> 3. Whether the generated image contains any artifacts, such as distortion, watermark, scratches, blurred faces, unusual body parts, or subjects not harmonized.
> 4. Whether the generated image is visually appealing and esthetically pleasing.
>
> After providing your explanation, you must output only one of the following choices as your final verdict with a label:
>
> 1. Model A is better: [[A>B]]
> 2. Model B is better: [[B>A]]
> 3. Tie, relatively the same acceptable quality: [[A=B=Good]]
> 4. Both are bad: [[A=B=Bad]]

For image-edition task, the prompt is as follows:

Please act as an impartial judge and a professional digital artist to evaluate the quality of the responses provided by two AI image edition models to the user inputs displayed below. You will be given model A's edited image and model B's edited image. Your job is to evaluate which assistant's edited image is better.

Source Image prompt: <source_prompt>
Target Image prompt after editing: <target_prompt>
Editing instruction: <instruct_prompt>
Source Image: <source_image>

Model A Edited Image: <left_output_image>
Model B Edited Image: <right_output_image>

When evaluating the quality of the edited images, you must identify the any inappropriateness in the edited images by considering the following criteria:
1. Whether the editing instruction has been followed successfully in the edited image.
2. Whether the edited image is overedited, such as the scene in the edited image is completely different from the original.
3. Whether the edited image looks natural, such as the sense of distance, shadow, and lighting.
4. Whether the edited image contains any artifacts, such as distortion, watermark, scratches, blurred faces, unusual body parts, or subjects not harmonized.
5. Whether the edited image is visually appealing and esthetically pleasing.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Model A is better: [[A>B]]
2. Model B is better: [[B>A]]
3. Tie, relatively the same acceptable quality: [[A=B=Good]]
4. Both are bad: [[A=B=Bad]]

https://doi.org/10.52202/079017-2538

For video-generation tasks, the prompt is as follows:

Please act as an impartial judge and a professional digital artist to evaluate the quality of the responses provided by two AI video generation models to the user inputs displayed below. You will be given model A's generated video and model B's generated video. Your job is to evaluate which assistant's generated video is better.

Text prompt: <prompt>
Model A Generated Video: <left_video>
Model B Generated Video: <right_video>

When evaluating the quality of the generated videos, you must identify the any inappropriateness in the edited videos by considering the following criteria:
1. Whether the text prompt has been followed successfully in the generated video.
2. Whether the generated video looks natural, such as the sense of distance, shadow, and lighting.
3. Whether the generated video is good visual quality, such as clearness, resolution, brightness, and color.
4. Whether the generated video is consistent and coherent in terms of the scene, objects, and characters.
5. Whether the generated video is dynamic and not static like a single image.
6. Whether the generated video is visually appealing and esthetically pleasing.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Model A is better: [[A>B]]
2. Model B is better: [[B>A]]
3. Tie, relatively the same acceptable quality: [[A=B=Good]]
4. Both are bad: [[A=B=Bad]]