This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

K-Sort Arena: Efficient and Reliable Benchmarking for Generative Models via K-wise Human Preferences

Zhikai Li[1,*], Xuewen Liu[1,*], Dongrong Joe Fu[2], Jianquan Li[1], Qingyi Gu[1,†], Kurt Keutzer[2], Zhen Dong[2,†]
[1]Institute of Automation, Chinese Academy of Sciences [2]University of California, Berkeley
lizhikai2020@ia.ac.cn, zhendong@berkeley.edu

Project: https://huggingface.co/spaces/ksort/K-Sort-Arena

Abstract

The rapid advancement of visual generative models necessitates efficient and reliable evaluation methods. Arena platform, which gathers user votes on model comparisons, can rank models with human preferences. However, traditional Arena methods, while established, require an excessive number of comparisons for ranking to converge and are vulnerable to preference noise in voting, suggesting the need for better approaches tailored to contemporary evaluation challenges. In this paper, we introduce K-Sort Arena, an efficient and reliable platform based on a key insight: images and videos possess higher perceptual intuitiveness than texts, enabling rapid evaluation of multiple samples simultaneously. Consequently, K-Sort Arena employs K-wise comparisons, allowing K models to engage in free-for-all competitions, which yield much richer information than pairwise comparisons. To enhance the robustness of the system, we leverage probabilistic modeling and Bayesian updating techniques. We propose an exploration-exploitation-based matchmaking strategy to facilitate more informative comparisons. In our experiments, K-Sort Arena exhibits 16.3× faster convergence compared to the widely used Elo algorithm. To further validate the superiority and obtain a comprehensive leaderboard, we collect human feedback via crowdsourced evaluations of numerous cutting-edge text-to-image and text-to-video models. Thanks to its high efficiency, K-Sort Arena can continuously incorporate emerging models and update the leaderboard with minimal votes. Our project has undergone several months of internal testing and is now available online.

1. Introduction

Visual generation models have made significant advancements, excelling in tasks such as text-to-image [5, 41, 44, 54, 56] and text-to-video [11, 13, 16, 55] generation. Such great progress has attracted more and more researchers, leading to a continuous proliferation of new models. Therefore, an efficient and reliable evaluation of the models' capabilities is urgently desired. However, traditional evaluation metrics, such as IS [45], FID [19], FVD [49], etc., fall short in providing a fair and comprehensive evaluation, especially not reflecting human preferences in the real world.

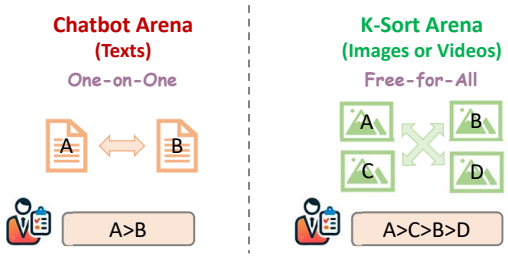


Figure 1. Comparison between K-Sort Arena and Chatbot Arena [10]. K-Sort Arena employs K-wise comparisons (K>2) to get richer information from user votes. Notably, it introduces probabilistic modeling and an effective matchmaking strategy, significantly improving efficiency and reliability.

To this end, Chatbot Arena [10] is proposed as a platform developed for evaluating large language models (LLMs). It constructs randomized, anonymous pairwise comparisons of models and collects user judgments of their outputs, thereby forming an overall ranking of models' capabilities. Despite the great progress, Chatbot Arena still faces challenges regarding efficiency and accuracy: (i) The inefficiency of Chatbot Arena stems primarily from two inherent mechanisms: pairwise comparisons and randomized matching. By allowing only two models to be compared at a time and potentially matching models of vastly different ranks, the system often yields minimal information per comparison. This inefficiency necessitates an excessive number of comparisons to achieve a stable ranking, resulting in a significant waste of valuable human effort in voting. More importantly, as a massive number of new models contin-

*Equal Contribution
†Corresponding Author

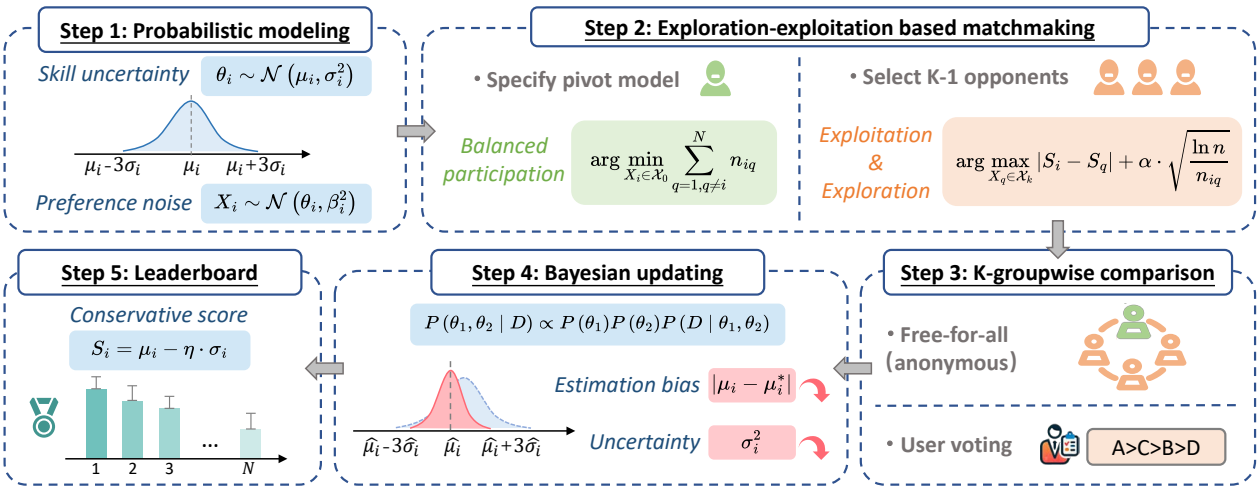


Figure 2. Overview of the proposed K-Sort Arena. K-wise comparisons (K>2) and the advanced matching strategy can significantly accelerate ranking convergence, achieving stable ranking with minimal user votes. Probabilistic modeling and Bayesian updating can enhance the robustness of model capability representation, thus resulting in greater efficiency and reliability.

uously emerge, this inefficiency prevents the rapid evaluation of new models' capabilities and the timely updating of the leaderboard, causing a lagged response to the latest advances. (ii) In user voting, preference noise and subjective bias are inherent, leading to occasional unjustified ratings. Pairwise comparisons are sensitive to this issue, which could introduce bias into the relative rankings. This is especially problematic when the leaderboard is updated frequently and the number of votes is small.

To address the above issues, we propose K-Sort Arena, a novel benchmarking platform for visual generation models. K-Sort Arena offers better efficiency and reliability. Specifically, K-Sort Arena employs K-wise comparisons (K>2), allowing K models to participate in free-for-all battles, which provides greater information benefits than pairwise comparisons, as shown in Figure 1. This approach is based on a practical biological principle: images and videos have higher perceptual intuitiveness compared to texts, enabling rapid evaluation of multiple samples at once. To ensure the robustness of ranking, we introduce probabilistic modeling of the models' capabilities, as well as a Bayesian score updating strategy applied after free-for-all battles among K models, which can dilute the adverse effects of preference noise. Furthermore, we propose an exploration-exploitation-based model matching strategy, which facilitates matchmaking among models with comparable strength while also incorporating under-explored models, thereby maximizing the expected benefit of each comparison. The overview is presented in Figure 2.

To demonstrate the superiority of K-Sort Arena , in Section 4, we design experiments to simulate the scenarios of model comparisons and user voting. Encouragingly, K-Sort Arena shows 16.3× faster ranking convergence than the Elo system in Chatbot Arena and exhibits greater robustness to preference noise. On one hand, it can update the leaderboard accurately and quickly with a minimum number of votes, which can effectively cope with model proliferation; on the other hand, it is more stable and reliable in long-term evaluations, obtaining more trustworthy evaluations with the same number of votes.

K-Sort Arena has served to evaluate dozens of state-of-the-art visual generation models, including both text-to-image and text-to-video models. By statistically organizing user feedback from crowdsourced questions, K-Sort Arena effectively builds comprehensive model leaderboards. To aptly reflect the diverse real-world applications, users are free to choose prompts sampled from open-source datasets or to create fresh prompts. Moreover, K-Sort Arena supports multiple voting modes and user interactions. Users can either select the best output from a free-for-all comparison or rank the K outputs instead. This flexibility ensures a faster, more user-friendly, and versatile evaluation process. Overall, our contributions can be summarized as follows:

- We introduce K-Sort Arena, an efficient and reliable platform for evaluating visual generation models. It can continuously monitor new models and quickly update the leaderboard with minimal votes.
- We propose K-wise comparisons to obtain richer feedback information and save human efforts in evaluation.
- We devise an exploration-exploitation-based matchmaking strategy with probabilistic capability modeling and Bayesian updating mechanisms.
- Ablation study shows that compared to traditional Elo algorithms, K-Sort Arena can achieve 16.3× faster convergence and greater robustness against preference noise.

2. Related Work

2.1. Visual Generation Evaluation

Text-to-Image Benchmarks Various metrics have been proposed to assess the performance of text-to-image models [25, 58]. IS [45], FID [19], sFID [45], and KID [7] calculate the distance between the real and generated data distributions using logits or features from InceptionNet [48]. CLIPScore [18] computes the cosine similarity between two embeddings from CLIP [42], measuring the alignment of texts and images. There are also several variants of CLIPScore, such as AS [47] and HPS [52], which aim to enhance evaluation quality. Similarly, BLIPScore [6] and ImageReward [53] are metrics calculated based on BLIP [26]. Beyond traditional metrics, recent works introduce large multimodal models as judgment assistants. For instance, T2I-CompBench [21] and X-IQE [9] utilize the Chain of Thought to enable MiniGPT-4 [59] to produce self-consistent evaluations. VQAScore [33] employs a visual-question-answering (VQA) model to generate alignment scores by calculating the accuracy of answering simple questions. TIFA [20] also uses VQA to measure the faithfulness of generated images to text inputs.

Text-to-Video Benchmarks Metrics for assessing text-to-video models have also been broadly investigated [31, 50, 51]. FVD [49] is used to measure the discrepancy between the real and synthesized videos. CLIPSIM [43] is extended to evaluate the alignment of texts and videos by measuring the similarity of multiple frames with texts. VBench [22] and FETV [36] decompose the evaluation of video quality into multiple dimensions for fine-grained evaluation. EvalCrafter [35] selects multiple objective metrics, which are expected to summarize real-world situations, to assess the synthesized video quality in various aspects.

Despite the great advances, the above static metrics still suffer significant flaws in expressing human preferences in the real world. They cannot provide comprehensive evaluations, especially in aspects such as visual aesthetics. Furthermore, with the rapid emergence of diverse tasks such as image editing [2], image captioning [26], video editing [39], video captioning [57], etc., static metrics are increasingly inadequate in capturing the nuanced performance across these varied and evolving domains.

2.2. Arena Evaluation with Human Preferences

To address the limitations of static metrics, DynaBench [24] suggests implementing a live benchmark system that integrates a human-in-the-loop approach, thus allowing for more dynamic and adaptive evaluation. Building on this idea, Chatbot Arena [10] is developed as a platform specifically for LLMs. It constructs model arenas that allow LLMs to make randomized, anonymous pairwise comparisons. Users are required to judge and score the outputs of two models to continuously calibrate the capability scores of each model, resulting in an overall ranking of model capabilities. It also inspires WildVision's efforts to rank vision-language models [37]. However, these Arena algorithms require excessive comparisons to achieve a stable ranking and are susceptible to preference noise in voting. As our concurrent work, GenAI Arena [23] replicates the above workflow to visual generative models and thus has the same issues. Consequently, the coverage of the leaderboard is limited to a few models. In contrast, K-Sort Arena capitalizes on the intuitive advantage of visual information over texts, incorporating more robust modeling methods and more effective matchmaking strategies, which shows great potential in large-scale model evaluations.

3. Methodology

In this section, we describe how to perform robust probabilistic modeling and Bayesian updating of model capabilities in free-for-all comparisons of K models, and how to schedule matches to accelerate ranking convergence.

3.1. K-wise Comparison

The pairwise comparison employed by Chatbot Arena evaluates only two models per round and is inefficient. In contrast, K-Sort Arena evaluates K models (K>2) simultaneously, which naturally provides more information and thus improves the efficiency of the overall ranking. In coordination with K-wise comparisons, the modeling and updating of model capabilities are detailed below.

Probabilistic Capability Modeling Individual numerical modeling, as in the Elo system [12], provides only a certain value of the estimate and thus cannot ensure reliability. Instead, by using probability distributions to represent capabilities, it is possible to capture and quantify the inherent uncertainty and hence become more flexible and adaptive. This idea can be seen in popular ranking systems such as Glicko [14] and TrueSkill [17]. Our approach, while inspired by them, incorporates further improvements to enhance efficiency and reliability. Formally, we represent the capability θ of each model as a normal distribution:

$$\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \tag{1}$$

where μ_i and σ_i denote the i-th model's expected score and uncertainty, respectively. Here, $i = 1, 2, \cdots, N$, and N are the total number of models. As previously mentioned, user voting inevitably has preference noise, which is orthogonal to the uncertainty σ of the model's performance. Therefore, we introduce an additional stochastic variable β over the model's capability θ such that the model's actual performance judged by human evaluation is:

$$X_i \sim \mathcal{N}(\theta_i, \beta_i^2) \tag{2}$$

To build a leaderboard, we use the conservative score [40] to estimate the model's capability, as defined below:

$$S_i^{(n)} = \mu_i^{(n)} - \eta \cdot \sigma_i^{(n)} \tag{3}$$

where η is a coefficient with a typical value of 3.0, $S_i^{(n)}$, $\mu_i^{(n)}$ and $\sigma_i^{(n)}$ are the values after n comparisons and updates. For each update of μ_i^j and σ_i^j, $j = 1, 2, \cdots, N$, we follow Eq. 9 and Eq. 10 specified as follows.

Bayesian Capability Updating Based on probabilistic modeling, we implement the updating process using Bayesian inference with observed match results. We begin by discussing the case of two models and then generalize to the free-for-all comparison of K models. Assuming in the current comparison there are two models M_1 and M_2, the likelihood estimate of observation D that M_1 wins M_2 is:

$$P(D|\theta_1, \theta_2) = P(X_1 > X_2) = \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{\beta_1^2 + \beta_2^2}}\right) \tag{4}$$

Here, $\Phi(x)$ is the cumulative distribution function of standard normal distribution, *i.e.*, $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$, and $\phi(x)$ is the probability density function of standard normal distribution, *i.e.*, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then, based on Bayes' theorem, we can derive the joint posterior density of (θ_1, θ_2) given observation D as follows:

$$\begin{aligned} &P(\theta_1, \theta_2|D) \propto P(\theta_1)P(\theta_2)P(D|\theta_1, \theta_2) \\ &= \phi\left(\frac{\theta_1 - \mu_1}{\sigma_1}\right) \phi\left(\frac{\theta_2 - \mu_2}{\sigma_2}\right) \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{\beta_1^2 + \beta_2^2}}\right) \end{aligned} \tag{5}$$

The marginal posterior density of θ_1 can be subsequently obtained by the following equation:

$$\begin{aligned} P(\theta_1|D) &= \int_{-\infty}^{\infty} P(\theta_1, \theta_2|D)\, d\theta_2 \\ &\propto \phi\left(\frac{\theta_1 - \mu_1}{\sigma_1}\right) \Phi\left(\frac{\theta_1 - \mu_2}{\sqrt{\beta_1^2 + \beta_2^2 + \sigma_2^2}}\right) \end{aligned} \tag{6}$$

The derivation of the above equation is detailed in Appendix A.1. With the marginal posterior density, the posterior mean of θ_1 can be calculated as follows:

$$\begin{aligned} \hat{\mu}_1 = E[\theta_1|D] &= \frac{\int_{-\infty}^{\infty} \theta_1 P(\theta_1|D) d\theta_1}{\int_{-\infty}^{\infty} P(\theta_1|D) d\theta_1} \\ &= \mu_1 + \frac{\sigma_1^2}{\sqrt{\sum(\beta_i^2 + \sigma_i^2)}} \frac{\phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sum(\beta_i^2 + \sigma_i^2)}}\right)}{\Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sum(\beta_i^2 + \sigma_i^2)}}\right)} \\ &= \mu_1 + \frac{\sigma_1^2}{c_{12}} \cdot \mathcal{V}\left(\frac{\mu_1 - \mu_2}{c_{12}}\right) \end{aligned} \tag{7}$$

where $\mathcal{V}(x) = \phi(x)/\Phi(x)$ and $c_{ij}^2 = \sum(\beta_i^2 + \sigma_i^2)$. The derivation of the above equation is detailed in Appendix A.2. Here, $\hat{\mu}_1$ is the updated mean μ_1 value. Similarly, the updating process of the variance σ_1^2 is given by the following equation:

$$\begin{aligned} \hat{\sigma_1}^2 &= Var[\theta_1|D] = E[\theta_1^2|D] - (E[\theta_1|D])^2 \\ &= \sigma_1^2 \cdot \left(1 - \frac{\sigma_1^2}{\sum(\beta_i^2 + \sigma_i^2)} \cdot \mathcal{W}\left(\frac{\mu_1 - \mu_2}{\sqrt{\sum(\beta_i^2 + \sigma_i^2)}}\right)\right) \\ &= \sigma_1^2 \cdot \left(1 - \frac{\sigma_1^2}{c_{12}^2} \cdot \mathcal{W}\left(\frac{\mu_1 - \mu_2}{c_{12}}\right)\right) \end{aligned} \tag{8}$$

where $\mathcal{W}(x) = \mathcal{V}(x)(\mathcal{V}(x) + x)$.

The above procedure accomplishes Bayesian updating of the two models after comparing them, and as the number of comparisons increases, μ gets closer to the true value and σ tightens up, resulting in a high-confidence capacity estimate [4]. We generalize it to a free-for-all comparison of K models, and the capacity updating formulas for the i-th model are as follows:

$$\begin{aligned} \hat{\mu}_i = \mu_i + \sigma_i^2 \cdot \Bigg(&\sum_{q:r_q > r_q} \frac{1}{c_{iq}} \cdot \mathcal{V}\left(\frac{\mu_i - \mu_q}{c_{iq}}\right) \\ &+ \sum_{q:r_i < r_q} \frac{-1}{c_{iq}} \cdot \mathcal{V}\left(\frac{\mu_q - \mu_i}{c_{iq}}\right) \Bigg) \end{aligned} \tag{9}$$

$$\begin{aligned} \hat{\sigma_i}^2 = \sigma_i^2 \cdot \Bigg(1 - \Bigg(&\sum_{q:r_i > r_q} \frac{\sigma_i^2}{c_{iq}^2} \cdot \mathcal{W}\left(\frac{\mu_i - \mu_q}{c_{iq}}\right) \\ &+ \sum_{q:r_i < r_q} \frac{\sigma_i^2}{c_{iq}^2} \cdot \mathcal{W}\left(\frac{\mu_q - \mu_i}{c_{iq}}\right) \Bigg)\Bigg) \end{aligned} \tag{10}$$

Thanks to the probabilistic modeling and Bayesian updating employed in K-wise comparison, the model's capabilities can be represented with high robustness, thereby facilitating stable and accurate ranking. Additionally, it is important to note that K-wise comparison offers an inherent advantage in terms of efficiency. Generally speaking, a K-wise comparison can be viewed as $C_K^2 = \frac{K(K-1)}{2}$ pairwise comparisons. Assuming that each pairwise comparison provides a certain ranking benefit, and this benefit is additive, we can claim that the total number of comparisons required is significantly less than that for pairwise comparisons.

3.2. Exploration-Exploitation Based Matchmaking

Effective model matchmaking significantly impacts the efficiency of ranking convergence. Here we first examine matchmaking methods used in notable ranking systems. For instance, the Elo system [12], employed by traditional Arena, uses completely random matching. This can result in pairing the lowest-ranked player with the highest-ranked

one, even after numerous comparisons when rankings are nearly stable. Such matchups provide minimal new information, often leading to inefficient use of evaluation resources and slower ranking convergence. To address the above issue, TrueSkill system [17] focuses on matching players whose strengths are as equal as possible. However, it is only effective for assessing the ability of individual players, because each player's opponents are limited to a small, localized group of candidates. This limitation means that it lacks a comprehensive understanding of the overall pool of players, making it less useful for the overall ranking of a large number of players.

To this end, we propose an exploration-exploitation-based matchmaking strategy, which promotes valuable comparisons and thus achieves efficient model ranking with fewer comparisons. Specifically, we model the selection of players as a multi-armed bandit problem, where each pair of players is viewed as an arm. The objective is to maximize the overall benefit after n comparisons, *i.e.*, to provide the most information for the overall ranking after n comparisons. Notably, our approach emphasizes maximizing global gains, offering a broader perspective compared to TrueSkill, which focuses on short-term benefits from an individual player's viewpoint. To solve the multi-armed bandit problem, we apply the Upper Confidence Bound (UCB) algorithm. The UCB algorithm performs exploration with the most optimistic attitude given the current exploitation, which is formulated as follows:

$$U^{(n)}(X_i, X_q) = |S_i^{(n)} - S_q^{(n)}| + \alpha \cdot \sqrt{\frac{\ln n}{n_{iq}}} \tag{11}$$

where $|S_i^{(n)} - S_q^{(n)}|$ indicates the absolute difference in scores between the i-th model and q-th model after n comparisons, n_{iq} denotes the number of comparisons that have been made between the two models, and α is a balancing coefficient with a typical value of 1.0. Eq. 11 realizes the trade-off between exploration and exploitation, where the first part is exploitation and the second part is exploration. In exploitation phase, we prioritize selecting players of similar skill levels to create valuable comparisons, while in the exploration phase, we encourage players who have not been sufficiently evaluated to participate in matches to ensure a comprehensive assessment. We theoretically prove its advantage over random selection (details in Appendix B).

Consequently, for a pre-specified player X_i, designated as the pivot, we can achieve grouping by greedily selecting its K-1 opponents based on their Upper Confidence Bound (UCB) scores, as follows::

$$\{X_{q_1}^*, \cdots, X_{q_{K-1}}^*\} = \bigcup_{k=1}^{K-1} \left\{ \arg\max_{X_q \in \mathcal{X}_k} \{U(X_i, X_q)\} \right\} \tag{12}$$

$$\begin{aligned} \text{where} \quad & \mathcal{X}_k = \mathcal{X}_{k-1} - \{X_{q_{k-1}}^*\}, \\ & \mathcal{X}_0 = \{X_q\}_{q=1}^N, \quad X_{q_0}^* = X_i \end{aligned} \tag{13}$$

The above procedure accomplishes the effective selection of its opponents after specifying the pivot player X_i. In our algorithm, instead of random selection, we specify X_i under the guidance of equalizing the number of comparisons to promote balanced participation in comparisons by each player, which is formulated as follows:

$$X_i = \arg\min_{X_i \in \mathcal{X}_0} \sum_{q=1, q \neq i}^{N} n_{iq} \tag{14}$$

In the following, we present the advantages of the proposed specification policy of the pivot player X_i in a scenario-by-scenario manner.

- ***Scenario 1: Ranking many models from scratch.*** In each round of comparisons, we select the model with the fewest comparisons as the pivot. This promotes balanced participation across all models, preventing insufficient or excessive evaluation of certain models. Such equalization from a global perspective is also an important factor in promoting rapid convergence of the overall ranking.
- ***Scenario 2: Adding new models to an existing ranking.*** Our algorithm facilitates new models to participate in comparisons as pivots frequently in the early rounds so that they can quickly catch up with the number of comparisons of old models. Hence, with an effective matchmaking strategy, we can efficiently evaluate new models' capabilities, allowing us to showcase the latest progress in the leaderboard in real time.

4. Experiments

In this section, we design experiments that simulate user voting and different ranking scenarios, to verify the validity of each proposed component in K-Sort Arena .

Experimental Setup In Sections 4.2, 4.3, and 4.4, we conduct experiments to rank 50 models from scratch. In Section 4.5, we perform experiments by adding a new model to an existing ranking of 50 models. To simulate user voting on model comparisons, we assign a *preset out-of-order label* to each model to indicate its ground-truth ability. The result of a specific comparison depends on the performance of models, which are determined by their ground-truth abilities and the preference noise. Note that this preset label is used solely for evaluating the comparison results and is not involved in any other part of the ranking process, such as model capability modeling and updating. Finally, we calculate the Mean Squared Error (MSE) of the ranked positions against the preset labels to evaluate the convergence speed and accuracy.

Table 1. Number of comparisons required for Elo system and K-Sort Arena to reach convergence.

Method	K	Modeling	Matchmaking	Comparisons
Elo System	2	Numerical	Random	11692
K-Sort Arena	4	Probabilistic	UCB	716 (↓16.3×)

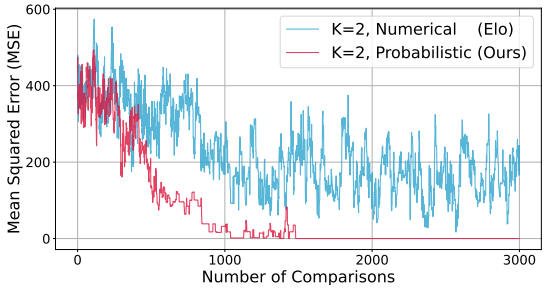
4.1. K-Sort Arena vs. Elo-based Arena

Table 1 shows the number of comparisons required for Elo system and K-Sort Arena to reach convergence, *i.e.*, MSE becomes consistently zero. Encouragingly, with the advanced modeling method and matchmaking strategy, K-Sort Arena is 16.3 times more efficient than Elo system, dramatically reducing the number of user votes required. Below we will verify the advantages of each component.

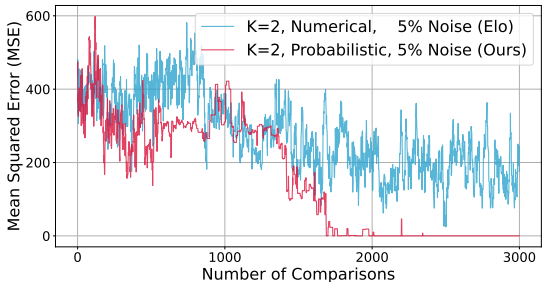
4.2. Probabilistic vs. Numerical Modeling

We begin by verifying the advantages of probabilistic modeling over numerical modeling employed in Elo systems, as shown in Figure 3a. Since the Elo system is designed for pairwise comparisons, we fix K in K-Sort Arena to 2 in this experiment for fairness. Remarkably, numerical modeling exhibits violent oscillation and fails to converge even after 3000 comparisons. This outcome highlights the unreliability of the existing Arena platform, despite the large number of votes that have been collected. On the contrary, our probabilistic modeling provides rapid convergence after about 1500 comparisons.

Figure 3b illustrates the case of voting with preference noise. In Figure 3a, comparison results are directly determined by the preset labels, whereas in Figure 3b we introduce a 5% chance of inconsistency between the comparison results and the labels. As observed, numerical modeling still fails to converge, while probabilistic modeling, despite converging slightly slower due to noise effects, manages to converge after approximately 2000 comparisons. This fully demonstrates the high robustness of probabilistic modeling, offering a strong assurance of the reliability of evaluations.



(a) Case without preference noise.



(b) Case with preference noise.

Figure 3. Comparison of numerical modeling (Elo [12]) and probabilistic modeling (ours) at K=2, separately with and without preference noise. Probabilistic modeling can converge quickly, while numerical modeling stays oscillating and fails to converge.

4.3. K-wise vs. Pairwise Comparison

Next, we verify the effect of different K values (K∈[2,4,6]) on ranking convergence. All three sets of experiments adopt UCB matchmaking strategy, and the experimental results are shown in Figure 4. When K is increased to 4, multiple models engage in free-for-all comparisons in each round, which yields richer information than the case of K=2, resulting in faster convergence (approximately twice as fast). For K=6, while MSE decreases more rapidly in the early stages, small fluctuations occur in the later stages before final convergence, resulting in less pronounced efficiency gains. Therefore, K=4 is considered as a trade-off choice.

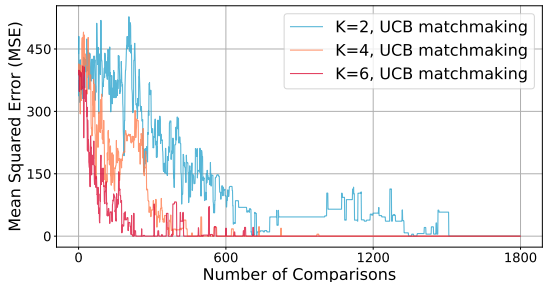


Figure 4. Comparison of different K values when applying UCB matchmaking. K∈[2,4,6]. As K increases, the convergence becomes faster and more stable.

4.4. UCB vs. Traditional Matchmaking

In this section, we demonstrate the advantages of the proposed UCB matchmaking strategy. The comparison methods include random matchmaking in the Elo system and skill-based matchmaking in the TrueSkill system. The experimental results are presented in Figure 5. Since random matching can potentially result in low-information comparisons, such as pairing the highest-ranked player with the lowest-ranked one, it continues to oscillate after 3,000 comparisons. The goal of skill-based matchmaking is that the skills of players in the comparison are as equal as possible. This may promote interesting matches for an individual player, but it ignores exploration and thus fails to en-

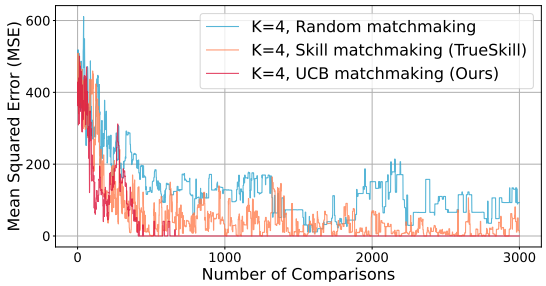


Figure 5. Comparison of different matchmaking strategies at K=4, including random, Skill (TrueSkill [17]), and UCB (ours). The proposed exploitation-exploration based strategy achieves the fastest convergence.

sure convergence and stability of the overall ranking from a global perspective. Fortunately, our UCB matchmaking strategy addresses this issue by balancing exploitation and exploration, achieving ranking convergence with minimal comparisons.

4.5. Specified vs. Random Pivot

Here, we focus on the case of adding a new model to an existing ranking and verify the effectiveness of the proposed pivot specification method. We initialize the new model's ranking at 51 and set its actual label to 31. The experimental results are presented in Figure 6. When both the pivot and its opponents are selected randomly, the new model are less likely to be selected. When the pivot is chosen randomly and UCB is used for matching opponents, the efficiency improves. This improvement is due to the exploration term in Eq. 11, the new model's small n_{iq} increases its probability of being selected as an opponent. Furthermore, when employing our balance-guided specification method, since the new model naturally participates in the minimal number of comparisons, it is always selected as the pivot in the initial period. Notably, only roughly 30 comparisons are needed to determine the new model's ranking, which provides a prerequisite for rapid leaderboard updating.

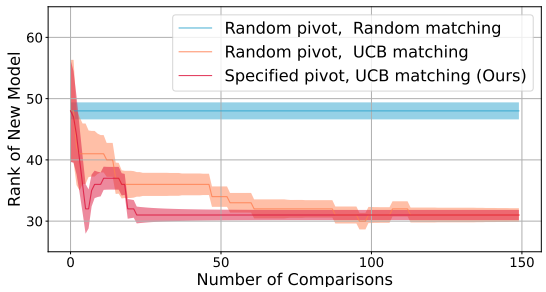


Figure 6. Comparison of different matchmaking strategies when adding a new model. The pivot specification method, coupled with UCB opponent matching, enables the fastest convergence.

5. K-Sort Arena Platform

In this section, we build an open and live evaluation platform with human-computer interactions in Huggingface Space, which integrates the proposed algorithms to improve efficiency and reliability. On this platform, users can input a prompt and receive outputs from K anonymous generative models. Users then cast a ranked vote for these models based on their preferred responses, and these votes are saved for updating the leaderboard. K-Sort Arena platform has the following highlights:

- ***Open-source platform:*** K-Sort Arena platform is open-source, open-access, and non-profit, fostering collaboration and sharing in the community.
- ***Extensive model coverage:*** It covers a comprehensive range of models, including numerous open-source and closed-source models across various types and versions.
- ***Real-time update:*** It continuously adds new models, completes its evaluation with minimal votes, and updates the leaderboard in real-time.
- ***Robust evaluation:*** Bayesian modeling and anonymous comparisons reduce preference noise and model prejudice, making the leaderboard reliable and authoritative.
- ***User-friendly interaction:*** It supports various prompt input modes, voting modes, and user interaction styles, offering users a high degree of flexibility.

5.1. Covered Tasks and Models

K-Sort Arena is dedicated to evaluating visual generation tasks with human preferences, with a particular focus on text-to-image and text-to-video tasks. To ensure a comprehensive and thorough evaluation, we strive to cover as many mainstream models as possible, including both open-source and closed-source models, as well as multiple versions of a single model, if available. Currently, K-Sort Arena has served to evaluate dozens of state-of-the-art models. A detailed list of models is presented in Appendix C.

5.2. Platform Construction

K-Sort Arena platform is designed using Gradio and hosted in Huggface Space. Model inference is performed on ZeroGPU Cloud or Replicate API calls.

Interface Overview The interface features two main functionalities: leaderboard display and user voting for model battles. When participating in voting, after the user enters the prompt, the interface can display 4 generated images or videos from the anonymous models, *i.e.*, $K = 4$ is taken as default. The interface layout is illustrated in Appendix E.

Prompt Input Mode To aptly reflect diverse real-world applications, K-Sort Arena supports two prompt input modes.

- ***Ready-made prompts:*** Users have the option to randomly extract pre-designed prompts from our extensive data pool for input into the models. This feature eliminates the need for users to spend time creating their own

prompts, thereby significantly improving the efficiency of their interactions. At present, the data pool contains 5000 representative prompts, which are sampled from popular datasets such as MS COCO [32] and WebVid [3].

- ***Custom prompts:*** Users are also free to create fresh input prompts, allowing them to tailor and customize the generated content to meet their specific needs.

Voting Mode K-Sort Arena supports two voting modes for K-wise free-for-all comparisons, called Best Mode and Rank Mode. In Best Mode, users compare the outputs of K models and vote for the most preferred answer. For users who are unsure, a tie option is also available. In Rank Mode, users can rank the outputs of K models, providing a more fine-grained comparison (tie is also available).

- ***Best Mode:*** In this mode, the user only needs to select the best model, making one K-wise comparison theoretically equivalent to $K-1$ pairwise comparisons. Since it requires only one mouse click, as in pairwise comparisons, it is $K-1$ times more efficient.
- ***Rank Mode:*** In this mode, the user provides feedback by ranking the K models. One K-wise comparison is theoretically equivalent to $\frac{K(K-1)}{2}$ pairwise comparisons. Since it requires clicking on the rank of each model, *i.e.*, K clicks, it is $\frac{K-1}{2}$ times more efficient.

5.3. Leaderboard Building

Crowdsourced Voting Our project has been undergoing internal testing for several months, during which we have collected over 1,000 high-quality votes. All voters are professors and graduate students in the field of visual generation. To ensure high quality and mitigate preference noise, we organize pre-voting training and provide evaluation guidelines. Specifically, for text-to-image models, the evaluation criteria consist of Alignment (50%) and Aesthetics (50%). Alignment encompasses entity, style, and other matching aspects, while Aesthetics includes photorealism, light and shadow rendering, and the absence of artifacts. Text-to-video models are similarly evaluated based on Alignment (50%) and Aesthetics (50%). Alignment is broken down into video content matching, movement matching, and inter-frame consistency. Aesthetics comprises photorealism, physical correctness, and the absence of artifacts.

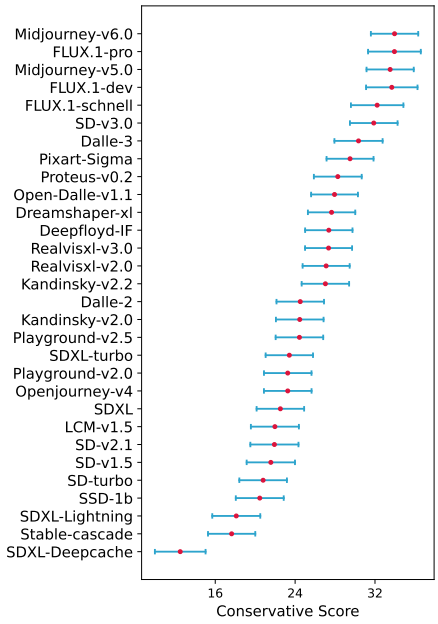


Figure 7. Leaderboard of text-to-image models, which are ranked by conservative score S. We also show μ and σ for each model. The data is as of Aug 2024.

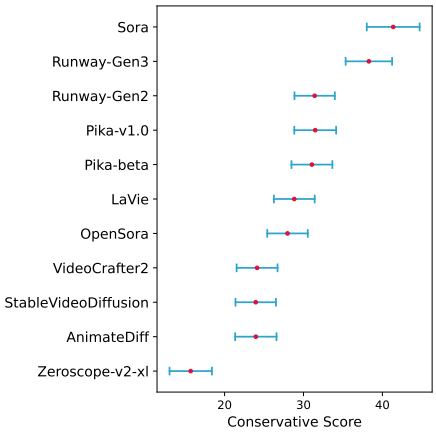


Figure 8. Leaderboard of text-to-video models, which are ranked by conservative score S. We also show μ and σ for each model. The data is as of Aug 2024. Note that the comparisons of Sora only take Sora's official samples due to the lack of available API.

Leaderboard Showcase The leaderboard of text-to-image models is illustrated in Figure 7. We can observe that proprietary models like MidJourney and Dalle dominate the top of the charts. Among open-source models, FLUX.1 and SD-v3.0 stand out with impressive performance. The leaderboard of text-to-video models is in Figure 8.

6. Conclusion

In this paper, we introduce K-Sort Arena, a benchmarking platform for visual generation models. K-Sort Arena employs K-wise comparisons (K>2), allowing K models to play free-for-all games, along with probabilistic modeling and Bayesian updating to improve efficiency and robustness. Furthermore, an exploration-exploitation based matchmaking strategy is proposed to facilitate valuable comparisons, which further accelerates convergence. We validate the superiority of the proposed algorithms via multiple simulated experiments. To date, K-Sort Arena has collected extensive high-quality votes to build comprehensive leaderboards for image and video generation.

Acknowledgement

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0119402; in part by the National Natural Science Foundation of China under Grant 62276255.

Moreover, we would like to express our gratitude to all those who contributed their time, expertise, and insights during the internal testing phase of K-Sort Arena. Listed in no particular order: Collov Labs; Daquan Zhou from NUS; Yang Zhou from CMU; Vijay Anand from Texas A&M University; Ying Li, Chun-Kai Fan, Menghang Dong and Aosong Cheng from Peking University HMI Lab; Yinglong from Meituan; Yinsheng Li from Shao's Lab at McGill; Mingfei Guo from Stanford; and Chenyue Cai from Princeton. We are profoundly grateful for their commitment and the unique perspectives they brought to this project.

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002. 2

[2] Jinbin Bai, Zhen Dong, Aosong Feng, Xiao Zhang, Tian Ye, Kaicheng Zhou, and Mike Zheng Shou. Integrating view conditions for image synthesis. *arXiv preprint arXiv:2310.16002*, 2023. 3

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 8

[4] James L Beck and Lambros S Katafygiotis. Updating models and their uncertainties. i: Bayesian statistical framework. *Journal of Engineering Mechanics*, 124(4):455–461, 1998. 4

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[6] Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. Improving image captioning descriptiveness by ranking and llm-based fusion. *arXiv preprint arXiv:2306.11593*, 2023. 3

[7] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3

[8] Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2024. 2

[9] Yixiong Chen, Li Liu, and Chris Ding. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*, 2023. 3

[10] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024. 1, 3

[11] Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024. 1

[12] Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967. 3, 4, 6

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 1

[14] Mark E Glickman. The glicko system. *Boston University*, 16 (8):9, 1995. 3

[15] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 2

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1

[17] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006. 3, 5, 7

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 3

[20] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah 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. 3

[21] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 3

[22] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 3

[23] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*, 2024. 3

[24] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021. 3

[25] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[27] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545, 2023. 2

[28] Zhikai Li and Qingyi Gu. I-vit: Integer-only quantization for efficient vision transformer inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17065–17075, 2023.

[29] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, pages 154–170. Springer, 2022.

[30] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17227–17236, 2023. 2

[31] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094*, 2024. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 8

[33] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 3

[34] Xuewen Liu, Zhikai Li, and Qingyi Gu. Cachequant: Comprehensively accelerated diffusion models. *arXiv preprint arXiv:2503.01323*, 2025. 2

[35] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 3

[36] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[37] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 3

[38] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[39] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024. 3

[40] Lawrence D Phillips and Ward Edwards. Conservatism in a simple probability inference task. *Journal of experimental psychology*, 72(3):346, 1966. 4

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1, 3

[46] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 2

[47] Christoph Schuhmann. CLIP+MLP Aesthetic Score Predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. 3

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[49] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1, 3

[50] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023. 3

[51] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024. 3

[52] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 3

[53] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[54] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 1

[55] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1

[56] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Gate opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. *arXiv preprint arXiv:2411.18499*, 2024. 1

[57] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 3

[58] Yutong Zhou and Nobutaka Shimada. Vision+ language applications: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 826–842, 2023. 3

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3