How Can We Achieve Human-Aligned Evaluation for Visual Generation Tasks?

Anonymous authors
Paper under double-blind review

Abstract

The rapid advancement of visual generation models has outpaced traditional evaluation approaches, necessitating the adoption of Vision-Language Models as surrogate judges. In this work, we systematically investigate the reliability of the prevailing absolute pointwise scoring standard, across a wide spectrum of visual generation tasks. Our analysis reveals that this paradigm is limited due to stochastic inconsistency and poor alignment with human perception. To resolve these limitations, we introduce **GenArena**, a unified evaluation framework that leverages a *pairwise comparison* paradigm to ensure stable and human-aligned evaluation. Crucially, our experiments uncover a transformative finding that simply adopting this pairwise protocol enables off-the-shelf open-source models to outperform top-tier proprietary models. Notably, our method boosts evaluation accuracy by over 20% and achieves 78.7% agreement with human annotators. It also achieves a Spearman correlation of 0.86 with the LMArena leaderboard, surpassing the 0.36 correlation of pointwise methods. Based on GenArena, we benchmark state-of-the-art visual generation models across diverse tasks, providing the community with a rigorous and automated evaluation standard for visual generation.

1 Introduction

Recent advancements in diffusion models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Podell et al., 2023; Blattmann et al., 2023; Esser et al., 2024) and unified multimodal models (Team, 2024; Wang et al., 2024b; Pan et al., 2025; Deng et al., 2025; Geng et al., 2025) have shifted the frontier of visual generation from basic text-to-image synthesis to complex tasks requiring multi-input reasoning, such as image editing (Liu et al., 2025b; Team et al., 2025b) and composition (Labs et al., 2025a; Wu et al., 2025a; Labs, 2025). As these systems reach unprecedented levels of capability, establishing a rigorous and reliable evaluation standard becomes increasingly critical yet challenging. To fundamentally assess the efficacy of an evaluation protocol in this evolving landscape, an ideal judge should satisfy three core requirements: (1) *Human Alignment*: exhibiting strong correlation with human perceptual rankings (Zheng et al., 2023); (2) *Self-Consistency*: producing stable judgments for identical samples across independent trials (Li et al., 2024b); and (3) *Discriminability*: effectively distinguishing subtle differences among state-of-the-art models. Satisfying these rigorous standards is critical for visual generation, particularly for *image editing and composition*, where correctness hinges on accurate prompt following and localized visual consistency that are hard to evaluate automatically.

However, satisfying these requirements remains challenging for traditional automated metrics. Widely used metrics such as FID (Heusel et al., 2017; Stein et al., 2023; Jayasumana et al., 2024) and CLIP Score (Hessel et al., 2021) often fail to evaluate fine-grained semantic alignment and aesthetics (Stein et al., 2023; Kynkäänniemi et al., 2023; Jayasumana et al., 2024), struggling to capture the nuances required by high-fidelity generation. While human evaluation serves as the *gold* standard, it incurs prohibitive costs and lacks scalability. Recently, in pursuit of a reliable yet automated alternative, the research community has pivoted towards the *VLM-as-a-judge* paradigm (Ku et al., 2024; Lin et al., 2024; Ye et al., 2025; Liu et al., 2025b). By leveraging the general-purpose visual reasoning capabilities of Vision-Language Models (VLMs) (Google, 2025a; Bai et al., 2025), this paradigm promises a scalable path to automate evaluation while approximating human-like perception.

Within this paradigm, existing evaluation practices typically adopt either *pointwise* scoring (assigning absolute scalar scores) (Ku et al., 2024; Lin et al., 2024; Ye et al., 2025; Liu et al., 2025b) or *pairwise*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

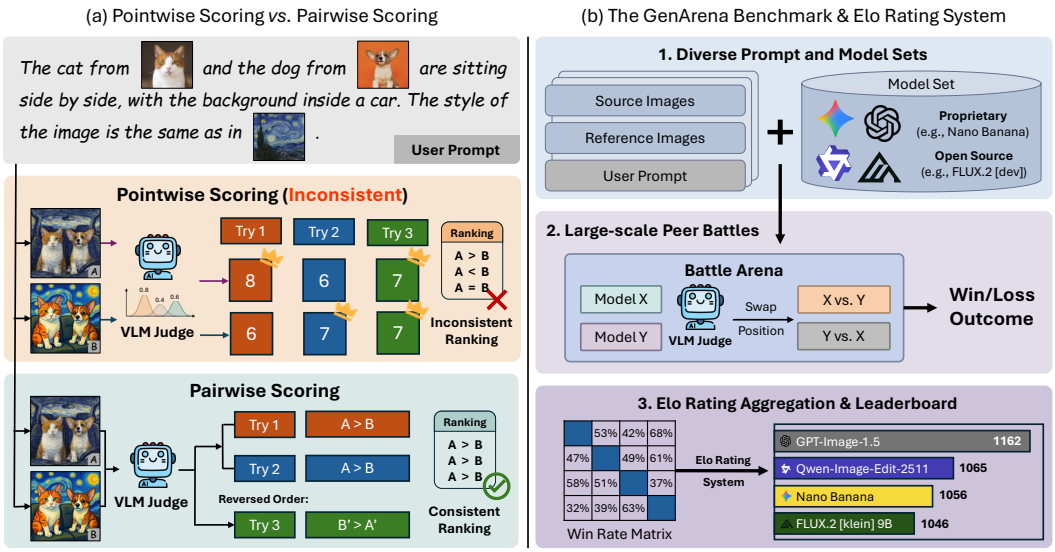


Figure 1: **Schemetric illustration of the comparison between pointwise and pairwise scoring, the GenArena Benchmark, and the Elo Rating System.** **(a)** Current benchmarks rely on absolute pointwise scoring, which suffers from *self-consistency collapse*. As shown, stochastic fluctuations in VLM outputs result in volatile rankings (e.g., $A > B$ in Try 1, but $B > A$ in Try 2) for the same input. In contrast, pairwise comparison yields consistent and robust preferences. **(b)** GenArena builds upon this stable pairwise paradigm. We curate a diverse set of prompts (including multi-reference generation tasks) and conduct large-scale peer battles using VLMs as judges. These pairwise outcomes are aggregated via the Elo Rating System to produce an accurate and reproducible model leaderboard.

comparison (assessing relative quality) (Wang et al., 2025b; Hu et al., 2025). Despite their rapid adoption, the optimal methodology remains underexplored. First, regarding the scoring protocol, a fundamental question arises: *1) Does the prevailing pointwise approach truly fulfill the core criteria of alignment and consistency, or does the pairwise paradigm offer a more robust alternative?* Second, regarding model accessibility, current high-quality evaluation relies predominantly on top-tier proprietary models (Ku et al., 2024; Liu et al., 2025b; Ye et al., 2025) to improve judgmental reliability, while open-source alternatives often lag behind. Bridging this gap typically necessitates fine-tuning on massive and costly human preference datasets (Han et al., 2024; Ma et al., 2025), requiring continuous updates as visual generative models (Wang et al., 2025c; Wu et al., 2025c; Luo et al., 2025a) rapidly evolve. This raises further critical questions: *2) Is reliable evaluation limited to proprietary models, or can open-source VLMs achieve state-of-the-art performance?* Furthermore, *3) is such resource-heavy fine-tuning indispensable, or can off-the-shelf models serve as effective judges directly?*

To answer question 1), we conducted a comprehensive study spanning image generation, editing, and video generation. Our investigation reveals inherent limitations in the pointwise paradigm (Liu et al., 2025b; Ye et al., 2025). As evidenced in Table 1, Table 4, and Figure 4, pointwise scoring suffers from poor human alignment and *self-consistency collapse*, often failing to distinguish subtle differences among models. We attribute this instability to cognitive limitations: maintaining a consistent absolute pointwise grading rubric is inherently prone to variance (Ariely, 1998). In contrast, our analysis identifies the *pairwise* paradigm—a strategy largely undervalued in current visual generation benchmarks—as the superior alternative. By simplifying evaluation into a robust binary choice (Hu et al., 2025), relative comparison effectively mitigates variance and restores alignment. For questions 2) and 3), experiments in Table 1 demonstrate that current open-source VLMs (Bai et al., 2025; Wang et al., 2025a; Team et al., 2025c) possess strong discriminative capabilities that were previously obscured by the pointwise format. We find that adopting the pairwise protocol effectively unlocks this potential, enabling off-the-shelf open-source models (Bai et al., 2025; Team et al., 2025c; Wang et al., 2025a) to achieve state-of-the-art accuracy without any parameter updates, surpassing even top-tier proprietary models using pointwise scoring.

Table 1: **Pairwise scoring universally outperforms pointwise scoring across a wide spectrum of visual generation tasks.** We report the binary classification accuracy with human preferences on GenAI-Bench ("GAI" for short) (Li et al., 2024a), EditScore-Bench (Luo et al., 2025a), and VideoGen-RewardBench (Liu et al., 2025a). The results demonstrate that simply shifting to a pairwise protocol (rows with gray background) enables off-the-shelf open-source VLMs to achieve state-of-the-art accuracy without any parameter updates, consistently surpassing specialized reward models and proprietary systems like GPT-5 that rely on the pointwise paradigm. **Bold** indicates the best result, and underline denotes the second best.

Model	Image Generation	Image Editing		Video Generation	
	GenAI-Bench	GAI	EditScore	GAI	VideoGen-Reward
Proprietary VLMs					
GPT-4.1	–	–	70.5	–	–
GPT-5	–	–	75.5	–	–
Gemini-2.5 Pro	–	–	72.2	–	–
Finetuned Models					
VisionReward	66.4	–	–	73.1	68.2
Qwen2.5-VL-7B	–	–	43.2	–	–
EditScore-7B	–	–	65.9	–	–
Qwen2.5-VL-72B	–	–	62.1	–	–
EditScore-72B	–	–	70.3	–	–
Qwen3-VL 8B Instruct					
EditScore-Qwen3-8B	–	–	69.0	–	–
UnifiedReward-Qwen3-VL-8B	64.2	81.5	75.0	69.1	73.6
+ Pairwise	**67.0**	82.5	73.3	**78.6**	**78.8**
Open Source VLMs					
Qwen3-VL 8B Instruct	49.1	73.4	58.3	62.0	57.0
+ Pairwise	60.5	**83.9**	83.7	73.3	61.5
GLM 4.6V Flash (9B)	48.2	73.2	68.3	63.0	–
+ Pairwise	54.7	81.3	**87.2**	76.6	–
InternVL 3.5 8B	50.7	66.4	53.4	–	–
+ Pairwise	61.9	83.1	75.0	–	–

Building upon these findings, we introduce **GENARENA**, a standardized evaluation framework depicted in Figure 1 (b). Instead of relying on unstable pointwise scores, GENARENA constructs dynamic leaderboards by aggregating the outcomes of massive pairwise comparisons via the Elo Rating System (Elo, 1966). By leveraging the inherent multi-image reasoning capabilities of VLMs (Bai et al., 2025; Team et al., 2025c; Wang et al., 2025a) to adjudicate peer battles, we ensure that model rankings are grounded in consistent comparative criteria. While this framework is universally applicable to visual generation, we specifically target *image editing and composition*, where the disparity between advancing model capabilities (Google, 2025;b; OpenAI, 2025a;b) and lagging evaluation protocols are most pronounced. Specifically, we curate a comprehensive evaluation suite comprising 6,086 high-quality user prompts spanning three categories: basic editing, reasoning-intensive editing, and multi-reference composition. Based on the suite, we conduct an extensive evaluation, as shown in Table 1. It shows that switching to *pairwise* scoring paradigm enables off-the-shelf open-source VLMs (Bai et al., 2025; Team et al., 2025c; Wang et al., 2025a) to achieve a **+20%** accuracy boost, surpassing even top-tier proprietary models. Crucially, our framework achieves **78.7%** agreement with human annotators and yields a Spearman correlation of **0.86** with the LMArena leaderboard, drastically outperforming the 0.36 correlation of pointwise baselines (Table 3).

The contributions of this work are summarized as follows:

1. We identify the critical flaws of absolute pointwise scoring, specifically *self-consistency collapse* and *poor human alignment*. In contrast, we demonstrate that shifting to a pairwise paradigm serves as a more robust and accurate alternative for visual evaluation.
2. We propose **GENARENA**, an Elo-based benchmark for visual generation tasks based on the pairwise scoring mechanism. By unifying assessments ranging from basic editing to complex multi-reference generation, GENARENA enables rigorous, consistent, and fair model comparison.

Table 2: **GENARENA Leaderboard.** Elo-based rankings of a variety of SoTA models across Basic, Reasoning, and Multi-Reference editing tasks. Rankings are established via pairwise battles judged by *Qwen3-VL-32B Instruct FP8* and aggregated into Elo scores. The results demonstrate that while basic editing capabilities are converging, significant disparities remain in reasoning and multi-reference tasks, distinguishing frontier proprietary models from current open-source alternatives. The rightmost column shows rankings from the LMArena (Chiang et al., 2024) leaderboard (Jan 16, 2026 version).

Models	Basic		Reasoning		MultiRef		LMArena
	Elo	Rank	Elo	Rank	Elo	Rank	
GPT Image 1.5 [High] (OpenAI, 2025b)	1162	#1	1204	#1	1259	#1	#1
Qwen-Image-Edit-2511 (Wu et al., 2025a)	1065	#2	1005	#4	793	#7	#3
Nano Banana (Google, 2025)	1056	#3	1130	#2	1048	#3	#2
Flux.2 [klein] 9B (Black Forest Labs, 2026)	1046	#4	962	#6	1018	#4	#5
LongCat-Image-Edit (Team et al., 2025b)	1037	#5	944	#8	–	–	–
Qwen-Image-Edit-2509 (Wu et al., 2025a)	1020	#6	962	#7	705	#10	–
GPT Image 1 [High] (OpenAI, 2025a)	1004	#7	1095	#3	1066	#2	#9
Flux.2 [dev] (Labs, 2025)	997	#8	968	#5	948	#6	#4
Flux.2 [klein] 4B (Black Forest Labs, 2026)	987	#9	928	#9	967	#5	#7
Qwen-Image-Edit (Wu et al., 2025a)	979	#10	920	#10	–	–	#6
Flux.1 Kontext [dev] (Labs et al., 2025b)	860	#11	849	#12	713	#9	#8
Bagel (Deng et al., 2025)	773	#12	823	#13	649	#11	#10
Step1X-Edit (Liu et al., 2025b)	739	#13	744	#14	–	–	#11
DreamOmni2 (Xia et al., 2025)	718	#14	858	#11	777	#8	–
Corr. w/ LMArena	0.87		0.80		0.50		–

3. We validate that our benchmark achieves strong direct agreement with human judgments and superior correlation with external human-preference leaderboards compared to existing methods. Our experiments verify that such a pairwise scoring approach unlocks the potential of open-sourced VLMs for visual generation evaluation, establishing a reproducible and highly accurate standard for the research community.

2 RELATED WORK

LLM-as-a-Judge. The capacity of Large Language Models (LLMs) to emulate human reasoning and evaluate specific inputs against predefined criteria has established the *LLM-as-a-Judge* paradigm (Zheng et al., 2023; Chen et al., 2024; Li et al., 2024b; 2025a). This approach has been widely adopted in visual generation (Ku et al., 2024; Ye et al., 2025; Liu et al., 2025b; Han et al., 2025), demonstrating high alignment with human preferences, and has recently been extended to Generative Reward Models to provide precise feedback signals for reinforcement learning (Liu et al., 2025c; Wang et al., 2025c). Nevertheless, these judges remain prone to vulnerabilities that introduce confounding factors into the evaluation, such as position bias (Ye et al., 2024; Shi et al., 2024; Wang et al., 2024a; Thakur et al., 2025), self-enhancement bias (Tan et al., 2024; Ye et al., 2024), overconfidence (Gu et al., 2024), and susceptibility to jailbreaking attacks (Raina et al., 2024; Zheng et al., 2024; Zhao et al., 2025c). In this work, we identify that stochastic inconsistency across repeated evaluations is a critical issue within this paradigm, a problem significantly exacerbated by standard pointwise scoring methods.

Visual Generation Benchmarks. Benchmarks for visual generative models have expanded rapidly across text-to-image generation (Ghosh et al., 2023; Huang et al., 2023; Sun et al., 2025; Niu et al., 2025), text-to-video generation (Liu et al., 2023; 2024; Huang et al., 2024), and image editing tasks (Brooks et al., 2023; Zhang et al., 2023; Sheynin et al., 2024; Liu et al., 2025b; Ye et al., 2025; Wu et al., 2025d; Han et al., 2025). While methodologies have evolved from low-level statistics (e.g., CLIP, SSIM) (Radford et al., 2021; Korhonen & You, 2012; Wang et al., 2004) and detection-based verification (Ghosh et al., 2023; Huang et al., 2023) to recent "VLM-as-a-Judge" paradigms (Ye et al., 2025; Ku et al., 2024), a fundamental limitation persists. Whether assessing compositional constraints (Ghosh et al., 2023; Huang et al., 2023), temporal consistency (Huang et al., 2024; Liu et al., 2024), or instruction following (Liu et al., 2025b), existing protocols predominantly rely on

Table 3: **Elo rating aligns better with LMArena than pointwise scoring.** We compare the ranking alignment of absolute pointwise scores and our pairwise Elo ratings on the widely-used GEdit-Bench-EN (Liu et al., 2025b) prompt set against human preference represented by LMArena (Chiang et al., 2024). For the pointwise baseline, we source scores from Yang et al. (2025) for Nano Banana, Li et al. (2025b) for Flux.1 Kontext [dev], and Wu et al. (2025a) for other models, with Flux.2 [dev] (Labs, 2025) evaluated by us. Compared with pointwise scoring yielding a low correlation of $\rho = 0.36$, our pairwise approach achieves more robust alignment with human judgment, reaching $\rho = 0.86$.

Models	GEdit-EN (Point)		**GEdit-EN (Elo)**		LMArena
	Score	Rank	Score	Rank	
Nano Banana (Google, 2025)	7.10	4	1062	1	1325
Flux.2 [dev] (Labs, 2025)	7.24	3	1053	2	1250
Qwen-Image-Edit (Wu et al., 2025a)	7.56	1	992	4	1231
Flux.1 Kontext [dev] (Labs et al., 2025b)	6.00	7	964	5	1166
GPT Image 1 [High] (OpenAI, 2025a)	7.53	2	1034	3	1155
Bagel (Deng et al., 2025)	6.52	6	890	7	1042
Step1X-Edit (Liu et al., 2025b)	6.70	5	938	6	1014
Corr. w/ LMArena (Chiang et al., 2024)	0.36		**0.86**		

absolute pointwise scoring. As noted in recent critiques (Stein et al., 2023; Kynkäänniemi et al., 2023), these scalar metrics suffer from stochastic inconsistency and the ambiguity of absolute scales, often yielding rankings that diverge from human perception. To address these shortcomings, we propose a scalable *pairwise* comparison framework (Thurstone, 2017) grounded in the Elo rating system (Elo, 1966) to ensure robust and reproducible evaluation.

Elo Rating System in AI Evaluation. The practice of ranking model capabilities via Elo scores derived from pairwise comparisons has become a standard in the evaluation of both LLMs (Team et al., 2025a; xAI, 2025; Google, 2025a) and visual generative models (Labs et al., 2025a; Wu et al., 2025a; Cao et al., 2025). Public benchmarks such as LMArena (Chiang et al., 2024), Artificial Analysis Arena (Artificial Analysis, 2025) and GenAI-Arena (Jiang et al., 2024) leverage crowdsourced voting, where users blindly select the superior response from two models, to calculate win rates and subsequent Elo ratings. Subsequent works (Zhao et al., 2025a; Luo et al., 2025b) further automated this process by utilizing LLMs to generate queries and evaluate responses, thereby reducing the dependency on manual labor. While pairwise Elo aggregation is established, its interaction with automated VLM judges for visual generation remains underexplored. We study this interaction systematically, showing that the evaluation protocol itself can be a stronger lever than judge scale and using Elo to aggregate VLM pairwise judgments into a robust benchmark.

3 Revisiting the VLM-as-a-judge Paradigm

While VLMs are increasingly used as surrogate judges, the standard practice of absolute pointwise scoring remains unexamined. In this section, we conduct a systematic study to explore the reliability of this paradigm. By comparing it with a pairwise approach across diverse tasks, we uncover its significant limitations in both accuracy (Sec. 3.1) and stability (Sec. 3.2). Consequently, our results establish that the pairwise protocol offers a far more robust and accurate evaluation standard, providing the empirical cornerstone for GenArena.

3.1 Pairwise scoring is more accurate than its pointwise alternatives

In this section, we quantitatively validate the superiority of the pairwise paradigm over traditional absolute pointwise scoring across a wide spectrum of visual generation tasks. To isolate the impact of the scoring protocol from model capability, we evaluate a diverse set of open-source VLMs (Team et al., 2025c; Wang et al., 2025a; Bai et al., 2025) under both pointwise and pairwise settings. Our evaluation leverages high-quality human-annotated preference datasets spanning three distinct domains. For image generation, we utilize GenAI-Bench (Li et al., 2024a). For image editing, we employ GenAI-Bench alongside EditScore-Bench (Luo et al., 2025a). For video generation, we use

GenAI-Bench (Li et al., 2024a) and VideoGen-RewardBench (Liu et al., 2025a). For the pointwise baseline, we follow a standard VIEScore-style absolute scoring protocol (Ku et al., 2024) and derive pairwise preferences by comparing the resulting scalar scores (Appendix F). We report the pairwise accuracy, defined as the agreement rate with ground-truth human choices, as our primary metric to assess the reliability of each judge.

The results summarized in Table 1 reveal a systematic performance disparity where pairwise scoring consistently outperforms pointwise scoring across image generation, editing, and video tasks. We observe that the evaluation protocol itself serves as a critical performance lever. Simply switching from a pointwise to a pairwise paradigm triggers an immediate and substantial accuracy boost for identical models without any parameter updates. For instance, Qwen3-VL 8B Instruct achieves a remarkable leap across different modalities, improving accuracy from 49.1% to 60.5% on GenAI-Bench (Li et al., 2024a) of image generation, rising from 58.3% to 83.7% on EditScore-Bench (Luo et al., 2025a) of image editing, and advancing from 57.0% to 61.5% on VideoGen-RewardBench (Liu et al., 2025a) of video generation. Similar consistent gains across different VLM judges confirm that *the pairwise paradigm effectively mitigates calibration noise, thereby unlocking the latent discriminative power of VLMs*.

Furthermore, this methodological shift *enables general-purpose models to surpass specialized systems that are significantly larger or explicitly finetuned for evaluation*. As evidenced in Table 1, the off-the-shelf Qwen3-VL 8B Instruct utilizing pairwise scoring outperforms the specialized EditScore-72B (Luo et al., 2025a) reward model in editing and surpasses VisionReward (Xu et al., 2024) in video tasks, despite both baselines relying on pointwise scoring. Moreover, this robust protocol allows open-source VLMs to challenge top-tier proprietary models. GLM-4V-Flash with our pairwise protocol attains 87.2% accuracy on EditScore-Bench (Luo et al., 2025a), significantly exceeding GPT-5 at 75.5%. In light of its superior accuracy, efficiency and widespread community adoption, we designate Qwen3-VL 8B Instruct (Bai et al., 2025) as the default judge model for subsequent experiments unless otherwise specified.

3.2 Pairwise scoring is more consistent than its pointwise alternatives

Beyond alignment with human preference, a robust benchmark demands *internal stability*, ensuring that stochastic fluctuations during decoding do not alter rankings. We quantify this self-consistency using *Krippendorff's Alpha* (α) (Krippendorff, 2011), treating independent inference runs with identical inputs as distinct raters. To enable direct comparison, we project pointwise scalar scores into a unified categorical preference space, utilizing a nominal difference metric that only penalizes fluctuations severe enough to flip the decision boundary (see Appendix H for formal definitions).

Table 4: **Judge self-consistency.** Krippendorff's α over 5 independent inference runs using *Qwen3-VL 8B Instruct*. **Bold** denotes the best result.

	Pointwise	**Pairwise**
Reward Model Benchmarks		
GenAI-Bench	0.7256	**0.8628**
EditScore-Bench	0.5753	**0.7087**
Image Editing Benchmarks		
ImgEdit	0.5707	**0.7040**
GEdit-Bench	0.5169	**0.6553**

We evaluate consistency across both human-annotated datasets (Li et al., 2024a; Luo et al., 2025a) and large-scale model matchups (Ye et al., 2025; Liu et al., 2025b). As detailed in Table 4, the absolute pointwise scoring paradigm suffers from *self-consistency collapse*, with α dropping as low as 0.5169 on GEdit-Bench. In stark contrast, our pairwise framework significantly mitigates this noise, maintaining high stability across all benchmarks (e.g., $\alpha = 0.8628$ on GenAI-Bench). These findings confirm that transitioning to comparative reasoning effectively rectifies the inherent volatility of pointwise metrics.

One intuition is that pointwise evaluation requires comparing two absolute scores produced by separate calls, making the final preference sensitive to run-to-run calibration drift. Pairwise evaluation instead presents both candidates in the same context, so the judge can apply a shared comparison criterion within a single decision. This contextual anchoring helps explain the improved stability observed in Table 4.

These findings suggest that reliable automated evaluation benefits more from pairwise preferences rather than pointwise scores. Motivated by these findings, we introduce GENARENA, a scalable framework to aggregate pairwise outcomes into reproducible leaderboards via Elo ratings.

4 GENARENA

4.1 OVERVIEW

In this section, we introduce **GENARENA**, a standardized evaluation framework designed to overcome the "self-consistency collapse" of pointwise scoring by shifting to a robust comparative paradigm. As illustrated in Figure 1, GENARENA orchestrates a scalable tournament among generative models, emulating the dynamics of human preference aggregation in a fully automated, reproducible pipeline:

Stage 1: Competitive Sampling. We first curate a diverse instruction set $\mathcal{I}$ of 6,086 prompts spanning Basic, Reasoning, and MultiRef tasks (see Appendix D) and sample outputs from a pool of candidate models $\mathcal{M} = \{M_1, \ldots, M_K\}$, treating them as active competitors in a tournament rather than isolated subjects.

Stage 2: Robust Pairwise Judging. A VLM judge evaluates model pairs (M_i, M_j). To mitigate the stochasticity and position bias inherent in LLMs, we implement a *bi-directional consistency check* and strictly enforce a *forced-choice* mechanism, only resolving those low-confidence (conflict) judgments into ties (see Sec. 4.2).

Stage 3: Global Elo Aggregation. Pairwise votes are transformed into a continuous leaderboard using the Bradley-Terry statistical model, ensuring a statistically rigorous ranking (see Sec. 4.3).

4.2 AUTOMATIC PAIRWISE SCORING

To mitigate the calibration drift and inconsistency inherent in absolute pointwise scoring, we adopt a *pairwise comparison paradigm*. In detail, a VLM judges a triplet (I, O_A, O_B), comprising an instruction I and two model outputs, i.e., O_A and O_B, to determine the superior candidate. To ensure robustness against position bias and stochasticity (Zheng et al., 2023), we implement a ***Bi-directional Consistency Protocol*** (detailed in Appendix K).

Besides, we enforce a *forced-choice constraint*, prohibiting the judge from declaring ties during a single inference pass. Specifically, we query the judge twice with the image order swapped: (O_A, O_B) and (O_B, O_A). A reliable preference is recorded only if the judge model consistently selects the same image across both permutations (i.e., $A \succ B$ and $B \prec A$). Cases of conflicting preferences (e.g., the judge always selects the first position) are algorithmically resolved as a Tie ($S_{ij} = 0.5$). This rigorous filtering mechanism ensures that only high-confidence, non-stochastic judgments contribute to the final ranking.

4.3 ELO RANKING AGGREGATION

To aggregate discrete pairwise outcomes into a continuous global ranking, we employ the **Elo rating system** (Elo, 1966) modeled via the Bradley-Terry (BT) framework (Bradley & Terry, 1952). We define the probability that model i defeats model j as a logistic function of the difference in their latent skill ratings R_i and R_j:

$$P(i \succ j) = \frac{1}{1 + 10^{(R_j - R_i)/\xi}}, \tag{1}$$

where $\xi = 400$ is a scaling factor standard in Elo systems.

Given a dataset of pairwise comparisons, we estimate the optimal Elo scores $\mathbf{R} = \{R_1, \ldots, R_N\}$ by performing Maximum Likelihood Estimation (MLE). Let W_{ij} denote the total number of wins model i achieves against model j. Following standard practices (Chiang et al., 2024; Zhao et al., 2025a), we incorporate ties generated by our consistency protocol by allocating 0.5 wins to both models. The optimal scores are obtained by maximizing the log-likelihood function:

$$\mathcal{L}(\mathbf{R}) = \sum_{i \neq j} W_{ij} \ln P(i \succ j) = \sum_{i \neq j} W_{ij} \ln \left(\frac{1}{1 + 10^{(R_j - R_i)/400}} \right). \tag{2}$$

Table 5: **Validation of Design Choices.** (a) Confusion matrix under the Explicit Tie option, showing severe *Laziness Bias*: the model defaults to "Tie" in nearly 40% of cases where humans identify a clear winner (highlighted with underline). (b) Judge alignment across Qwen3-VL model scales; *32B-FP8* achieves the best overall agreement with human perception and is adopted as our judge.

(a) Confusion matrix with explicit tie option.

Human Judgment	Model Prediction		
	A > B	**B > A**	**Tie (A=B)**
A > B	55.8%	6.2%	37.9%
B > A	6.8%	54.0%	39.2%
Tie	17.8%	16.8%	65.4%

(b) Judge alignment across model scales.

Size	Basic	Reasoning	MultiRef	Overall
4B	61.5	65.3	56.8	60.9
8B	64.7	66.1	57.6	63.6
32B	**68.2**	**71.5**	63.2	67.6
$32B^{FP8}$	68.1	70.8	**66.3**	**68.0**

This formulation allows us to solve for the global ranking that best explains the observed pairwise preferences, providing a statistically grounded leaderboard resilient to individual judgments.

Although exhaustive all-pair evaluation scales quadratically with the number of models, our subsampling analysis shows that only 5% of battles recovers the full-ranking order with Spearman $\rho > 0.99$ across all three tracks; see Appendix M.

5 Evaluation and Benchmarking

In this section, we validate the effectiveness of GenArena by demonstrating its superior human alignment compared to pointwise baselines (Sec. 5.1). Furthermore, we benchmark various state-of-the-art image editing models (Sec. 5.2) and ablate our design choices (Sec. 5.3).

5.1 Elo ranking aligns more with human preferences

To study the alignment with human preference, we conduct a comparative study using the prompt set from GEdit-Bench (Liu et al., 2025b) across the models listed in Table 3. We assess the candidate models under two protocols: the standard absolute pointwise scoring (Point) utilizing GPT-4.1, and our proposed pairwise Elo rating system (Elo) employing Qwen3-VL 8B Instruct (Bai et al., 2025) as the judge. We then compare the resulting rankings against LMArena (Chiang et al., 2024), an external crowdsourced human-preference leaderboard.

As presented in Table 3, the choice of evaluation paradigm significantly impacts ranking reliability. The absolute pointwise baseline struggles to reflect human consensus, yielding a low Spearman correlation (Spearman, 1961) of 0.36 and notably misranking top-performing models like *Nano Banana* by placing it the 4th in GEdit-EN (Point) compared to 1st in LMArena. In contrast, even when relying on a smaller open-source judge, the pairwise Elo approach effectively mitigates this noise, recovering a hierarchy that closely mirrors the external ranking. This methodological shift drives a substantial improvement in correlation to 0.86, demonstrating that converting the VLM's task from pointwise scoring to comparative ranking is the key factor in aligning automated metrics with human perception.

5.2 Benchmarking existing generation models

To establish a reliable leaderboard, we first determine the optimal evaluator based on the empirical evidence from Section 5.3.2. Our experiments indicate that *Qwen3-VL-32B Instruct FP8* yields the highest comprehensive agreement with human preference across the Basic, Reasoning, and MultiRef evaluation dimensions. Consequently, we adopt it as our designated judge to evaluate various state-of-the-art models. The assessment is conducted across three distinct tracks: *Basic Editing* for standard instruction following and visual quality, *Reasoning Editing* for complex spatial and logical constraints, and *MultiRef* for composition involving multiple reference images. We aggregate the outcomes of pairwise battles into global rankings using the Elo rating system, as presented in Table 2, with visualized results provided in Figure 2 and Figure 3.

The experimental results present the model performance across Basic, Reasoning, and MultiRef categories, revealing that competency in standard visual editing does not necessarily translate to

effectiveness in complex scenarios. While *GPT Image 1.5 [High]* maintains a consistent lead, we observe notable divergences in other models. For instance, *Qwen-Image-Edit-2511* excels in *Basic* tasks but ranks lower in *MultiRef*, whereas *GPT Image 1 [High]* shows the opposite trend by performing significantly better in composition tasks. Quantitatively, our rankings achieve high Spearman correlations with LMArena (Chiang et al., 2024) on Basic (0.87) and Reasoning (0.80). The relatively lower MultiRef correlation mainly reflects a version and task-distribution mismatch in the external LMArena comparison. We analyze this case in Appendix E. Combined with the 85% agreement with human judgments on MultiRef, this supports the reliability of our judge on complex multi-reference generation tasks. Qualitative results in Figure 3 further confirm that our judge remains highly accurate in evaluating these complex generation qualities.

5.3 Validation of Design Choices

5.3.1 Should we allow tie?

We analyze the impact of the scoring strategy by comparing an "Explicit Tie" option against our "Forced Choice" paradigm. We perform this ablation experiment on GenAI-Bench (Li et al., 2024a), which includes *tie* annotations by human. As detailed in Table 5a, allowing explicit ties introduces severe *Laziness Bias*: the model defaults to a neutral judgment in nearly 40% of cases where humans identify a clear winner ($A > B$ or $B > A$). This high False Neutral Rate suggests the model frequently uses the tie option to evade complex reasoning rather than reflecting true equivalence.

This bias significantly hurts the judge's performance. On the subset of discriminative pairs (where Ground Truth $\neq$ Tie), the Explicit Tie strategy achieves only 54.9% accuracy due to these missed judgments. In contrast, the Forced Choice strategy compels the model to resolve ambiguity, boosting accuracy on the same subset to 83.9%. The resulting **+29.0%** gain demonstrates that enforcing binary decisions is imperative for overcoming model inertia and recovering fine-grained discriminability.

5.3.2 Judge's Alignment with Human Preferences

To evaluate the reliability of automated evaluation, we conduct a scalability analysis using the Qwen3-VL Instruct model family across varying parameter sizes (4B, 8B, and 32B). We assess their alignment with human perception using a subset of preference pairs from (Hu et al., 2025), filtered to correspond with the three core capabilities of our benchmark. Following the pairwise protocol outlined in Section 5.2, we calculate the accuracy of each model variant in predicting the ground-truth human choice.

The results, detailed in Table 5b, demonstrate a strong positive correlation between model scale and alignment accuracy. While smaller models (4B and 8B) show limited capability, the 32B variants deliver a significant performance leap. Most notably, the *Qwen3-VL-32B Instruct FP8* model achieves the highest overall accuracy of 68.0%, exhibiting particular robustness in the challenging MultiRef split (66.3%). Driven by this superior alignment and computationally efficient inference, we designate *Qwen3-VL-32B Instruct FP8* as the default judge for the GenArena framework.

We further sample VLM-judged pairs from GenArena logs and collect human preference annotations to directly assess whether the judge's choices align with human judgments. The judge reaches 78.7% agreement with these human annotations, providing complementary validation beyond public leaderboard correlation; details are provided in Appendix I.

6 Conclusion

In this work, we expose the reliability bottlenecks of absolute pointwise scoring and propose GenArena, a robust pairwise evaluation framework. Our empirical results demonstrate that this methodological shift unlocks state-of-the-art judge accuracy in off-the-shelf open-source VLMs, surprisingly surpassing proprietary counterparts without parameter updates. By achieving superior alignment with human preference, GenArena establishes a rigorous, reproducible, and scalable standard for benchmarking the next generation of visual synthesis models.

AI USE STATEMENT

We primarily used LLMs to polish the writing of this paper. All LLM-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Dan Ariely. Predictably irrational: the hidden forces that shape our decisions. *Ebook, Revised and*, 1998.

Artificial Analysis. Image arena: Text-to-image model leaderboard. https://artificialanalysis.ai/image/arena, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Black Forest Labs. FLUX.2 [klein]: Towards Interactive Visual Intelligence, January 2026. URL https://bfl.ai/models/flux-2-klein.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Arpad E Elo. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation, 1966.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. `https://developers.googleblog.com/introducing-gemini-2-5-flash-image/`, August 2025.

Google. Gemini 3 flash, December 2025a. URL `https://blog.google/products/gemini/gemini-3-flash/`.

Google. Gemini 3 pro image (nano banana pro), November 2025b. URL `https://deepmind.google/models/gemini-image/pro/`.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594, 2024.

Feng Han, Yibin Wang, Chenglin Li, Zheming Liang, Dianyi Wang, Yang Jiao, Zhipeng Wei, Chao Gong, Cheng Jin, Jingjing Chen, et al. Unireditbench: A unified reasoning-based image editing benchmark. *arXiv preprint arXiv:2511.01295*, 2025.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, et al. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. *arXiv preprint arXiv:2512.16899*, 2025.

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.

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*, pp. 21807–21818, 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pp. 37–38. IEEE, 2012.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *ICLR*, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025a. URL https://arxiv.org/abs/2506.15742.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025b.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024a.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025a.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024b.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025b.

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. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

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*, pp. 22139–22149, 2024.

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*, 36:62352–62387, 2023.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025c.

Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, et al. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025a.

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8461–8474, 2025b.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

OpenAI. Introducing 4o image generation. `https://openai.com/index/introducing-4o-image-generation/`, March 2025a.

OpenAI. Gpt image 1.5 model. `https://platform.openai.com/docs/models/gpt-image-1.5`, 2025b.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. *arXiv preprint arXiv:2511.22989*, 2025.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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.

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*, pp. 8748–8763. PmLR, 2021.

Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024.

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*, pp. 10684–10695, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Charles Spearman. The proof and measurement of association between two things. 1961.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025a.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025b.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025c. URL https://arxiv.org/abs/2507.01006.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pp. 404–430, 2025.

Louis L Thurstone. A law of comparative judgment. In *Scaling*, pp. 81–92. Routledge, 2017.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025c.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025d.

xAI. Grok 4.1, November 2025. URL https://x.ai/news/grok-4-1.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

Jian Yang, Dacheng Yin, Xiaoxuan He, Yong Li, Fengyun Rao, Jing Lyu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Wemmu: Enhanced bridging of vision-language models and diffusion models via noisy query tokens, 2025. URL https://arxiv.org/abs/2512.02536.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. Auto-arena: Automating llm evaluations with agent peer battles and committee discussions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4440–4463, 2025a.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025b.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, and Dong Yu. One token to fool llm-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025c.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic llm benchmarks: Null models achieve high win rates. *arXiv preprint arXiv:2410.07137*, 2024.

A LIMITATIONS

GENARENA focuses on image editing and multi-reference composition, where reliable automatic evaluation is especially needed; extending the same protocol to other visual generation domains may require task-specific prompt curation and judge rubrics. Since the framework relies on VLMs as surrogate judges, its judgments may vary with future judge models and prompting choices, although our bi-directional consistency check, forced-choice protocol, and human validation help improve robustness. Pairwise evaluation also requires more judge calls than pointwise scoring, but our battle-sampling analysis shows that reliable rankings can be recovered with a small fraction of all possible comparisons.

B BROADER IMPACTS

This work aims to make visual generation evaluation more reproducible, scalable, and aligned with human preferences, reducing reliance on costly manual evaluation and helping the community compare models under a shared protocol. Potential risks come mainly from over-reliance on automated judges or over-interpreting leaderboard results beyond the benchmark scope, since VLM judges may still reflect imperfect or context-dependent preferences. We mitigate these concerns by making the evaluation protocol explicit, validating judge outputs with human annotations, and positioning GENARENA as an evaluation framework rather than a new image generation model.

C COMPUTE RESOURCES

All experiments, including image synthesis with open-source generation models and VLM-based judging in GENARENA, were conducted on a cluster with 64 NVIDIA Hopper GPUs. Our default judge is Qwen3-VL-32B Instruct FP8, as described in Section 5.2; the FP8 deployment reduces memory usage and inference cost for large-scale pairwise judging. In addition, the battle-sampling analysis in Appendix M shows that stable rankings can be obtained without exhaustive all-pair evaluation.

D DATASET DETAILS

Here, we provide the detailed composition and sources of the 6,086 prompts used in our benchmark.

D.1 BENCHMARK COMPOSITION

To ensure GENARENA serves as a holistic standard for challenging visual generation tasks, particularly image editing and multi-reference image composition, we curate a comprehensive evaluation suite consisting of **6,086** high-quality prompts. We categorize prompts into three distinct capability dimensions to probe the frontiers of model performance:

Basic Instruction Editing (1,948 prompts). This subset evaluates fundamental instruction-following capabilities, including object modification, attribute alteration, and background changes. It aggregates filtered prompts from established datasets (Ye et al., 2025; Liu et al., 2025b; Hu et al., 2025) to ensure broad coverage of standard editing scenarios.

Reasoning-Intensive Editing (1,627 prompts). Modern models should handle queries requiring complex logic, spatial reasoning, and world knowledge. We compile implicit and context-dependent instructions (e.g., "make the image look like a scene from the 1920s") (Zhao et al., 2025b; Wu et al., 2025d) that challenge the model to infer visual edits from abstract textual descriptions.

Multi-Reference Composition (2,511 prompts). To compensate for the lack of multi-reference generation in existing evaluations, this subset targets the integration of multiple visual concepts. We curate challenging prompts (Wu et al., 2025b; Xia et al., 2025; Oshima et al., 2025; Hu et al., 2025) requiring the coherent composition of multiple subject images, reflecting the complexity of real-world workflows like subject-driven generation (Wu et al., 2025a;b; Xia et al., 2025; Labs, 2025).

D.2 Source Breakdown

To ensure the high quality and diversity of our evaluation suite, we aggregated and filtered prompts from the following established benchmarks:

- **Basic Instruction Editing:** Comprises 1,948 prompts sourced from ImgEdit (Ye et al., 2025), GEdit-Bench (Liu et al., 2025b), and MMRB2 (Hu et al., 2025). These prompts cover standard editing tasks such as object addition/removal, attribute modification, and background replacement.
- **Reasoning-Intensive Editing:** Comprises 1,627 prompts sourced from RISEBench (Zhao et al., 2025b) and KRIS-Bench (Wu et al., 2025d). These prompts are selected to test the model's ability to understand implicit instructions, cultural references, and complex spatial relationships.
- **Multi-Reference Composition:** Comprises 2,511 prompts sourced from OmniContext (Wu et al., 2025b), DreamOmni2Bench (Xia et al., 2025), MultiBanana (Oshima et al., 2025), and MMRB2 (Hu et al., 2025). This subset represents the most challenging scenarios, requiring the model to understand complicated multimodal instructions and maintain the identity of multiple reference subjects simultaneously.

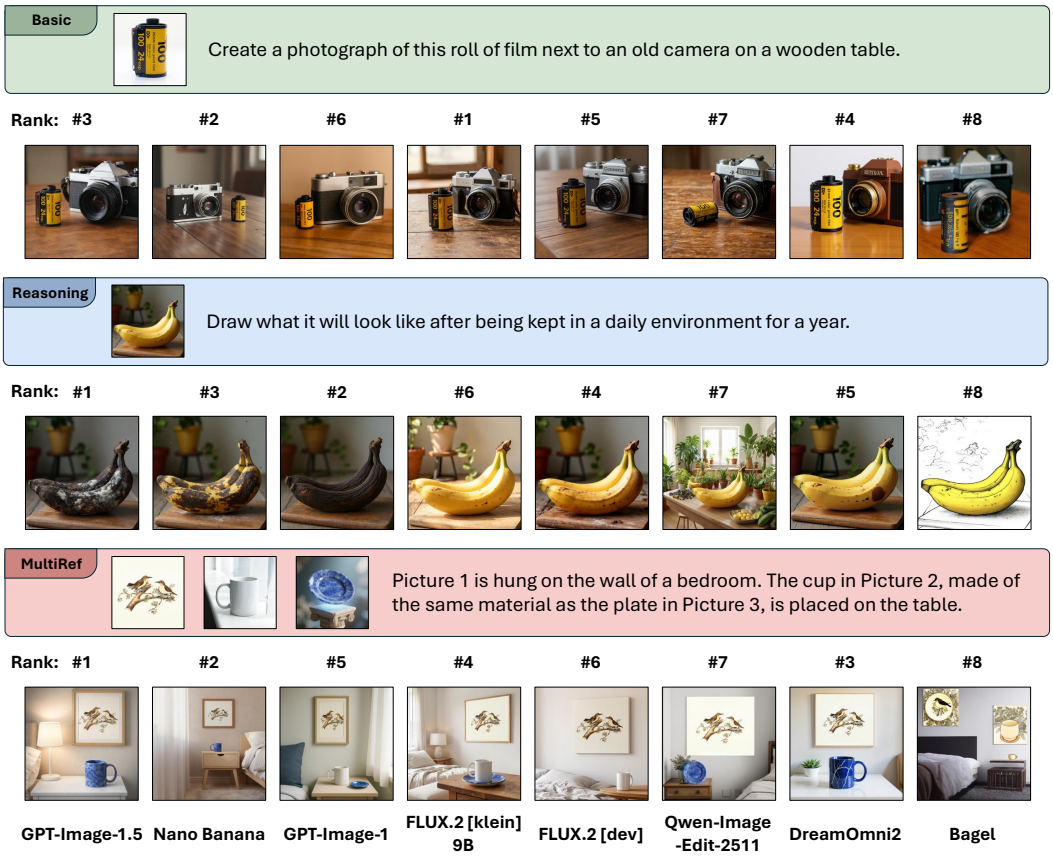


Figure 2: Qualitative comparison of visual generation tasks and generated results from various models in GenArena. The benchmark assesses models on three distinct dimensions: Basic, Reasoning (e.g., predicting environmental effects over time), and MultiRef (composing scenes from multiple image conditions). The bottom row displays sample outputs from leading proprietary and open-source models (e.g., GPT-Image-1.5, FLUX.2, Qwen-Image) on a complex multi-reference composition task, highlighting the variance in adherence to spatial and material constraints.

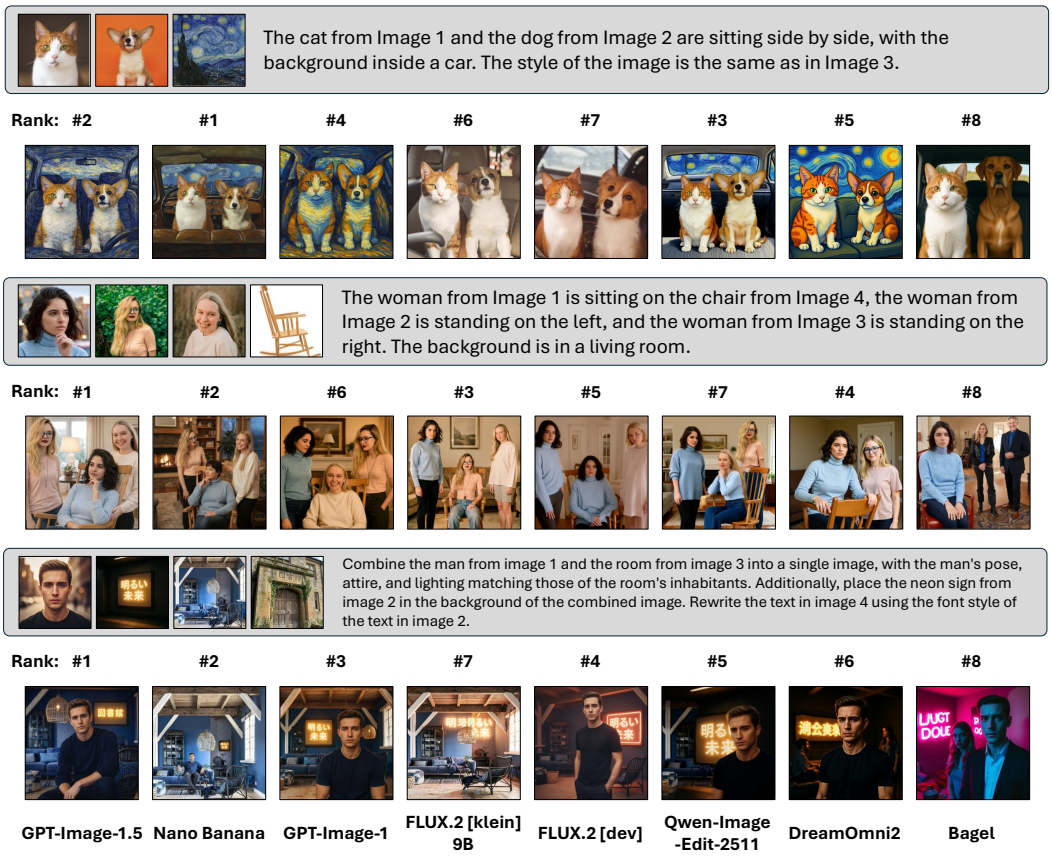


Figure 3: Qualitative comparison of multi-reference generation tasks and generated results from various models in GENARENA.

D.3 QUALITATIVE RESULTS

We provide samples of our benchmark in Figure 2 and Figure 3.

E MULTIREF CORRELATION ANALYSIS

The MultiRef split shows a lower correlation with LMArena than Basic and Reasoning in Table 2. We find that this difference is primarily associated with *GPT Image 1 [High]*, which ranks #2 in GENARENA MultiRef but #9 in the LMArena overall leaderboard. This discrepancy is consistent with a data-vintage mismatch: LMArena aggregates votes collected over earlier model versions, while GENARENA evaluates the model outputs generated for our benchmark. Since Spearman's ρ is sensitive to a single large displacement when the number of compared models is small, we report a sensitivity analysis with and without this model in Table 6.

Table 6: **MultiRef correlation sensitivity.** We report rank correlations with external LMArena references before and after accounting for the GPT Image 1 version mismatch.

Reference ranking	ρ **(all models)**	ρ **(excluding GPT Image 1)**
LMArena overall	0.50	0.79
LMArena multi-image	0.57	0.89

F Pointwise Baseline Details

For the pointwise baseline, we follow a standard VIEScore-style absolute scoring protocol (Ku et al., 2024). Given the source image, edited image, source prompt, target prompt, and editing instruction, the judge assigns scalar scores to the edited image and returns a JSON-formatted rationale and score list. We aggregate the returned scores into a single scalar score for each candidate image, and derive the final preference by independently scoring both candidates and comparing their aggregate scores. Equal aggregate scores are treated as ties.

G Pointwise Scoring Causes "False Tie"

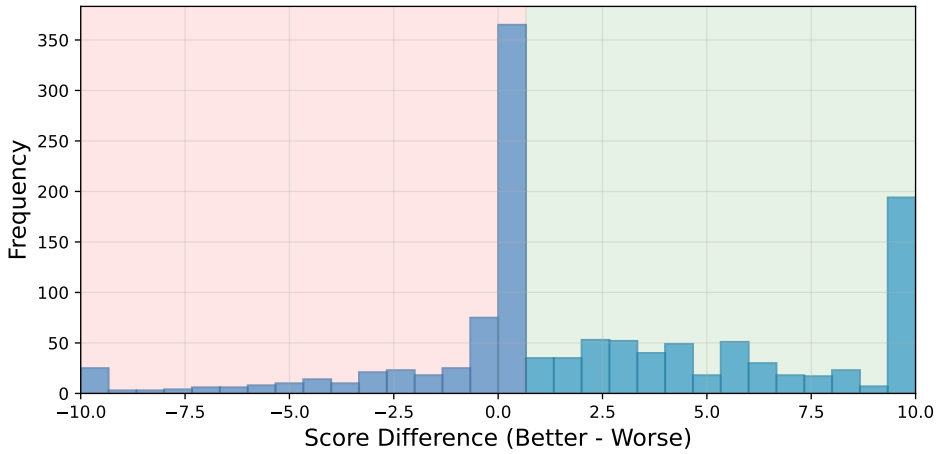


Figure 4: **Distribution of Score Differences in Pointwise Evaluation.** We visualize the score difference ($\Delta S = S_{\text{better}} - S_{\text{worse}}$) assigned by *Qwen3-VL 8B Instruct* on EditScore-Bench (Luo et al., 2025a) under the pointwise paradigm. The green region ($\Delta S > 0$) denotes correct alignment with human preference, while the red region ($\Delta S \leq 0$) indicates contradictory rankings. Notably, the distribution reveals a significant limitation in discriminative power: only 58.3% of cases are correctly ranked. A substantial portion of comparisons result in ties (23.5%, spike at $\Delta S = 0$) or direct errors (18.2%), confirming that absolute pointwise scoring struggles to resolve fine-grained visual differences.

We analyze the distribution of score differences generated by VLM judges. Specifically, we employ *Qwen3-VL 8B Instruct* to evaluate sample pairs from EditScore-Bench (Luo et al., 2025a) using its pointwise scoring protocol (0-10 scale). For each pair with a known ground-truth human preference, we calculate the score difference $\Delta S = S_{\text{better}} - S_{\text{worse}}$. Ideally, a robust judge should consistently yield $\Delta S > 0$.

However, as illustrated in Figure 4, the empirical distribution reveals severe deficiencies in the pointwise scoring mechanism. The model correctly identifies the superior image in only **58.3%** of cases (Green Zone). A critical failure mode is the high prevalence of **ties** (23.5%), represented by the prominent spike at $\Delta S = 0$. In the context of preference evaluation, a tie is effectively a failure to distinguish quality, rendering the metric useless for ranking closely matched models. Furthermore, **18.2%** of the evaluations fall into the negative spectrum (Red Zone), where the judge explicitly favors the inferior image.

The combined error rate (Ties + Errors) of 41.7% underscores a fundamental inability of the pointwise paradigm to capture subtle visual nuances. The VLM tends to map distinct visual outputs to the same coarse scalar values, lacking the sensitivity required for high-fidelity image editing assessment. This empirical evidence strongly motivates our transition to the pairwise comparison framework, which forces explicit binary decision-making to resolve these ambiguities.

G.1 Random Tie-Breaking Control

For a pointwise judge with strict accuracy a and tie rate t, random tie-breaking yields expected accuracy $a + t/2$. Applying this adjustment to the original predictions in Figure 4 gives approximately 70.1%, compared with the pairwise reference of 83.7%.

We also apply the same adjustment to the low-temperature EditScore-Bench control in Appendix H.3, using Qwen3-VL-8B-Instruct. Its strict accuracy is 59.2%, with 15.0% ties, yielding 66.7% expected accuracy after random tie-breaking.

Table 7: **Tie-adjusted pointwise accuracy on EditScore-Bench.** All entries are percentages. The pairwise reference from Table 1 is 83.7%.

Pointwise setting	Strict accuracy	Tie rate	Random tie-break
Original score distribution	58.3	23.5	70.1
Low-temperature control ($T = 0$)	59.2	15.0	66.7

Tie scoring explains part of the pointwise–pairwise gap, but both adjusted pointwise accuracies remain below the pairwise reference.

H Consistency Analysis Details

H.1 Mathematical Formulation

To compare the internal stability of pointwise and pairwise judges, we unify their outputs into a categorical preference space $\mathcal{L} = \{A \succ B, B \succ A, \text{Tie}\}$. For the pointwise baseline, given scalar scores $S_A^{(k)}$ and $S_B^{(k)}$ from the k-th inference run, the derived preference $l_{ij}^{(k)}$ is determined by the sign of the score differential:

$$l_{ij}^{(k)} = \begin{cases} A \succ B & \text{if } S_A^{(k)} > S_B^{(k)} \\ B \succ A & \text{if } S_A^{(k)} < S_B^{(k)} \\ \text{Tie} & \text{if } S_A^{(k)} = S_B^{(k)} \end{cases}$$

We calculate Krippendorff's Alpha (α) based on the observed (D_o) and expected (D_e) disagreement across m runs:

$$\alpha = 1 - \frac{D_o}{D_e}, \quad \text{where } D_o = \frac{1}{n} \sum_c \sum_k o_{ck} \delta(c, k). \tag{3}$$

Here, $\delta(c, k)$ is the nominal difference metric, where $\delta(c, k) = 0$ if $c = k$ and 1 otherwise. This metric strictly measures the stability of the decision boundary; numerical fluctuations in pointwise scores are only penalized if they alter the win/loss outcome. In this context, $\alpha = 1$ signifies perfect determinism, while $\alpha \approx 0$ implies agreement indistinguishable from random chance.

H.2 Experimental Setup

To ensure a rigorous evaluation, we curated a testbed from two sources:

- **Human-Annotated Datasets:** We utilize the full evaluation sets of GenAI-Bench (Li et al., 2024a) and EditScore-Bench (Luo et al., 2025a).
- **Model Matchups:** We use the full prompt sets from ImgEdit (Ye et al., 2025) and GEdit-Bench (Liu et al., 2025b) to generate outputs using the models listed in Table 3. From these outputs, we randomly sample 500 pairs to constitute the evaluation set.

For each pair, we conduct 5 independent inference runs using *Qwen3-VL 8B Instruct* as the judge. To ensure a strictly controlled comparison, all system prompts and generation hyperparameters (e.g., temperature) are kept identical to those used in the main experiments described in Section 3.1.

H.3 POINTWISE ROBUSTNESS CONTROLS

We evaluate Qwen3-VL-8B-Instruct on the GenAI-Bench image-editing and EditScore-Bench sets, changing one pointwise setting at a time: temperature $T = 0$, criterion-level reasoning before scoring (CoT), or a rubric matched to the dimensions in Appendix J. Candidates are scored independently in all controls.

Table 8: **Pointwise robustness controls.** Accuracy and tie rate are percentages. For the new controls, they are computed from the most frequent preference over five runs, with equal vote counts resolved as ties; α uses all five per-run preferences. Default and pairwise rows reproduce Tables 1 and 4. Dashes indicate quantities not reported for those reference rows.

Dataset	Setting	T	Accuracy	α	Tie rate
GenAI-Bench editing	Default pointwise	0.7	73.4	0.7256	–
	Low temperature	0.0	71.6	0.8375	11.3
	CoT	0.7	70.2	0.6093	17.3
	GENARENA rubric	0.7	76.9	0.6965	8.0
	Pairwise reference	–	83.9	0.8628	–
EditScore-Bench	Default pointwise	0.7	58.3	0.5753	–
	Low temperature	0.0	59.2	0.7763	15.0
	CoT	0.7	55.1	0.5433	27.4
	GENARENA rubric	0.7	58.7	0.6316	27.9
	Pairwise reference	–	83.7	0.7087	–

Lower temperature improves repeatability; on EditScore-Bench, its α exceeds the pairwise reference, while accuracy remains substantially lower. CoT improves neither metric, and the matched rubric gives mixed results. None of the tested pointwise controls reaches the pairwise reference accuracy.

I HUMAN ANNOTATION STUDY

I.1 ORIGINAL STUDY AND ANNOTATION PROTOCOL

To validate GENARENA independently of external Elo-style leaderboards, we conduct a human annotation study over 300 battles randomly sampled from GENARENA logs. We use stratified sampling with 100 battles per subset: Basic, Reasoning, and MultiRef. The sampled pairs are distributed to 10 annotators with diverse backgrounds. Annotators are calibrated with example pairs before annotation and then provide binary pairwise judgments following the same evaluation criteria as the VLM judge prompt, including instruction faithfulness, preservation of input/reference image content, and overall visual quality.

We define agreement as the fraction of battles where the VLM judge selects the same winner as the human judgment. Each sampled battle is labeled by one annotator, so the metric measures individual-level agreement rather than majority-vote agreement. GENARENA achieves 78.7% overall agreement, with 79% on Basic, 72% on Reasoning, and 85% on MultiRef. A binomial confidence interval gives a 95% CI of [73.7%, 83.1%], well above the 50% random baseline ($p < 10^{-25}$).

The human study is used only as an external validation of judge quality. Disagreement cases are not filtered or down-weighted in the GENARENA pipeline; all pairwise outcomes are aggregated by the same Elo procedure described in Section 4.3. The lower agreement on Reasoning reflects the greater ambiguity of physical commonsense and causal edits, while MultiRef achieves the highest agreement among the three tracks.

I.2 EXTENDED ANNOTATION AND INTER-ANNOTATOR AGREEMENT

We collect two additional independent, blind annotations for a randomly sampled, track-balanced subset of 100 battles. This yields 500 forced-choice labels over the original 300 battles, with three labels on each overlap battle. Annotators are not shown model identities or VLM decisions.

On the triply annotated subset, human–human agreement is 80.0%, with Fleiss' $\kappa = 0.5822$. Table 9 reports agreement with individual labels and majority votes. These results place the original 78.7% judge–human agreement close to the observed human–human agreement.

Table 9: **Extended human validation.** The 500-vote result weights triply annotated battles three times; the majority-vote comparisons use only the 100-battle overlap subset.

Comparison	Evaluation set	Agreement (%)
Judge vs. original human label	300 battles	78.7
Judge vs. all individual human labels	500 labels	79.2
Human vs. human	100 triply annotated battles	80.0
Original label vs. human majority	100 triply annotated battles	88.0
Judge vs. human majority	100 triply annotated battles	85.0

I.3 Same-Judge Pointwise Baseline

We evaluate the same 300 battles with the same Qwen3-VL-32B-Instruct-FP8 judge under the pointwise protocol in Appendix F. As shown in Table 10, pointwise agreement remains below pairwise agreement after assigning half credit to ties.

Table 10: **Protocol comparison on the human-study battles.** All methods use Qwen3-VL-32B-Instruct-FP8 and the original 300 human labels.

Protocol	Agreement (%)
Pointwise, strict	63.7
Pointwise, random tie-breaking (expected)	69.7
Pairwise (GenArena)	78.7

I.4 Human-Derived Leaderboard Audit

We aggregate human votes into track-wise BT/Elo rankings using the same fitting procedure as GenArena. The audit uses the actual benchmark prompts and generated outputs, with 100 battles per track covering all models in Table 2. We also give each battle equal weight by replacing the three overlap labels with their majority.

Table 11: **Human-derived ranking agreement with Table 2.** "All votes" uses all 500 labels. "Equal battle weight" replaces the three labels on each overlap battle with their majority, retaining the original labels elsewhere. All-vote Spearman intervals are 95% cluster-bootstrap intervals over unique battles within each track.

	Votes	All votes		Equal battle weight	
Track		ρ [95% CI]	τ	ρ	τ
Basic	168	0.464 [−0.090, 0.644]	0.341	0.459	0.319
Reasoning	166	0.776 [0.204, 0.859]	0.626	0.684	0.516
MultiRef	166	0.964 [0.436, 0.964]	0.891	0.945	0.855

Human-derived rankings align more strongly with GenArena on Reasoning and MultiRef than on Basic; equal battle weighting preserves this pattern. GPT Image 1.5 [High] ranks second on Basic and Reasoning and first on MultiRef under both aggregations. The Basic correlation interval includes zero, so this limited audit does not establish the exact model ordering in every track.

J Evaluation Criteria and Prompts for VLM Judge

In our pairwise scoring mechanism (Sec. 4.2), the VLM judge is instructed to evaluate image pairs based on four primary standards. These standards ensure that the judgment captures both semantic precision and perceptual fidelity:

1. **Text Faithfulness**: This metric measures the strictness with which the response follows the text editing prompts. It examines the model's ability to accurately implement specific instructions, such as adding distinct objects or altering stylistic attributes.
2. **Image Faithfulness**: This criterion focuses on the preservation of the source image's intrinsic characteristics. The edited response should maintain the original's composition, lighting, and background elements, ensuring that modifications do not disrupt the visual continuity.
3. **Overall Image Quality**: This standard assesses the output's aesthetic and technical fidelity. Preference is given to responses that minimize visual flaws, such as artifacts or unnatural distortions, thereby ensuring high-quality generation.
4. **Text Rendering**: This metric examines the correctness of embedded text. When present, the text is judged on spelling accuracy, readability, and its harmony within the scene; otherwise, it is marked as "Not Applicable."

We use the following system prompt for the VLM-as-a judge, following (Hu et al., 2025).

PROMPT

```
   You are an expert in image editing quality analysis and AI
       evaluation. Your role is to act as an objective judge for
       comparing two AI-generated image editing responses to the
       same prompt. You will evaluate which response is better based
        on a comprehensive rubric specifically designed for image
       editing tasks.

**Important Guidelines:**
- Be completely impartial and avoid any position biases
- Ensure that the order in which the responses were presented does
    not influence your decision
- Do not allow the length of the responses to influence your
    evaluation
- Do not favor certain model names or types
- Be as objective as possible in your assessment
- Focus on image editing specific factors: faithfulness to editing
    instructions, preservation of input image elements, and overall
    editing quality

**Understanding the Content Structure:**
- **[ORIGINAL PROMPT TO MODEL:]**: This is the image editing
    instruction given to both AI models
- **[INPUT IMAGE FROM PROMPT:]**: This is the source image provided
    to both models for editing
- **[RESPONSE A:]**: The first model's edited image response
- **[RESPONSE B:]**: The second model's edited image response

Your evaluation must be based on a fine-grained rubric that covers
    the following criteria. For each criterion, you must provide
    detailed step-by-step reasoning comparing both responses. You
    will use a 1-6 scoring scale.

**Evaluation Criteria:**
1. **text_faithfulness:** Which response better adheres to the text
    editing instruction? Consider how well each response follows the
     specific editing instructions (e.g., adding objects, changing
    colors, modifying scenes).
```

```
2. **image_faithfulness:** Which response better respects and
    incorporates the key elements of the input image? Consider how
    well each response preserves important aspects of the original
    image (composition, lighting, style, background elements) while
    making the requested changes.

3. **overall_image_quality:** Which response has better general
    technical and aesthetic quality, with fewer visual artifacts,
    distortions, or inconsistencies introduced during the editing
    process?

4. **text_rendering:** If either response contains rendered text,
    which one has better text quality (spelling, legibility,
    integration with the image)? If no text is rendered, state "Not
    Applicable."

**Scoring Rubric:**
- Score 6 (A is significantly better): Response A is significantly
    superior across most criteria
- Score 5 (A is marginally better): Response A is noticeably better
    across several criteria
- Score 4 (Unsure or A is negligibly better): Response A is slightly
     better or roughly equivalent
- Score 3 (Unsure or B is negligibly better): Response B is slightly
     better or roughly equivalent
- Score 2 (B is marginally better): Response B is noticeably better
    across several criteria
- Score 1 (B is significantly better): Response B is significantly
    superior across most criteria

**Confidence Assessment:**
After your evaluation, assess your confidence in this judgment on a
    scale of 0.0 to 1.0:

**CRITICAL**: Be EXTREMELY conservative with confidence scores. Most
     comparisons should be in the 0.2-0.5 range.

- **Very High Confidence (0.8-1.0)**: ONLY for absolutely obvious
    cases where one response is dramatically better across ALL
    criteria with zero ambiguity. Use this extremely rarely (less
    than 10% of cases).
- **High Confidence (0.6-0.7)**: Clear differences but some
    uncertainty remains. Use sparingly (less than 20% of cases).
- **Medium Confidence (0.4-0.5)**: Noticeable differences but
    significant uncertainty. This should be your DEFAULT range.
- **Low Confidence (0.2-0.3)**: Very close comparison, difficult to
    distinguish. Responses are roughly equivalent or have
    conflicting strengths.
- **Very Low Confidence (0.0-0.1)**: Essentially indistinguishable
    responses or major conflicting strengths.

**IMPORTANT GUIDELINES**:
- DEFAULT to 0.3-0.5 range for most comparisons
- Only use 0.6+ when you are absolutely certain
- Consider: Could reasonable people disagree on this comparison?
- Consider: Are there any strengths in the "worse" response?
- Consider: How obvious would this be to a human evaluator?
- Remember: Quality assessment is inherently subjective

After your reasoning, you will provide a final numerical score,
    indicate which response is better, and assess your confidence.
```

```
    You must always output your response in the following structured
     JSON format:

{
   "reasoning": {
      "text_faithfulness": "YOUR REASONING HERE",
      "image_faithfulness": "YOUR REASONING HERE",
      "overall_image_quality": "YOUR REASONING HERE",
      "text_rendering": "YOUR REASONING HERE",
      "comparison_summary": "YOUR OVERALL COMPARISON SUMMARY HERE"
   },
   "score": <int 1-6>,
   "better_response": "A" or "B",
   "confidence": <float 0.0-1.0>,
   "confidence_rationale": "YOUR CONFIDENCE ASSESSMENT REASONING
       HERE"
}
```

K Pairwise Scoring Implementation Details

In Section 4.2, we introduced the Bi-Directional Consistency Protocol. Here, we elaborate on the specific logic used to resolve pairwise comparisons:

1. Forced-Choice Constraint: Unlike prior methods that allow models to output "Tie" directly, we force the VLM to select a winner. This effectively mitigates the *laziness bias*, where models default to a tie to avoid complex reasoning. The judge is prompted to evaluate based on specific multi-dimensional criteria (e.g., semantic alignment, visual fidelity) provided in the system prompt.

2. Consistency Check & Tie Resolution: For every pair of models A and B, we conduct two inference runs:

- **Run 1:** Input order (O_A, O_B). Output: $w_1 \in \{A, B\}$.
- **Run 2:** Input order (O_B, O_A). Output: $w_2 \in \{A, B\}$.

The final outcome S_{AB} is determined as follows:

$$S_{AB} = \begin{cases} 1(\text{Win for } A) & \text{if } w_1 = A \text{ and } w_2 = A \\ 0(\text{Win for } B) & \text{if } w_1 = B \text{ and } w_2 = B \\ 0.5(\text{Tie}) & \text{if } w_1 \neq w_2 \end{cases} \tag{4}$$

A tie (0.5) implies that the visual difference is either negligible or that the judge is heavily influenced by position bias (e.g., always preferring the first image). By filtering these instances, we ensure high precision in the final rankings.

K.1 Order Disagreement and Tie-Handling Sensitivity

We measure final tie rates in the logs used for Table 2. These ties predominantly arise from disagreement between the two order-swapped calls (Table 12).

Table 12: **Tie-handling sensitivity.** Final tie rates are percentages. Maximum rank shift compares the original ranking with the ranking recomputed after removing all ties.

Track	Final tie rate (%)	Max. rank shift
Basic	23.86	0
Reasoning	21.41	0
MultiRef	15.47	0

After removing all tied battles and refitting BT/Elo, every model retains its original rank in all three tracks. Thus, ties affect rating values but do not change the observed model ordering.

L ELO RATING OPTIMIZATION

While Section 4.3 outlines the objective function, this section details the optimization procedure.

Data Construction: We construct a win-matrix where each entry W_{ij} aggregates the results from the bi-directional protocol. If a comparison results in a tie ($S_{ij} = 0.5$), we increment both W_{ij} and W_{ji} by 0.5.

Optimization: Since the negative log-likelihood is convex, we use Logistic Regression to solve for the ratings $\mathbf{R}$. We effectively treat the problem as training a classifier without a bias term, where the input features are one-hot encoded vectors representing the opposing models. We employ the L-BFGS algorithm or SGD to minimize $-\mathcal{L}(\mathbf{R})$.

M BATTLE SAMPLING EFFICIENCY

To quantify the practical cost of pairwise evaluation, we perform a battle sampling sensitivity analysis. Starting from the complete battle logs, we randomly subsample different ratios of battles, recompute BT-Elo rankings from each subsample, and measure Spearman correlation against the full-data ranking. We repeat this process with 50 random seeds for each ratio and track. The full logs contain approximately 174K battles for Basic, 147K for Reasoning, and 136K for MultiRef.

Table 13: **Battle sampling sensitivity.** Spearman correlation between subsampled BT-Elo rankings and full-data rankings, averaged over 50 random seeds.

Sampling ratio	ρ **(Basic)**	ρ **(Reasoning)**	ρ **(MultiRef)**
5%	0.993 ± 0.006	0.991 ± 0.007	0.996 ± 0.006
10%	0.997 ± 0.004	0.994 ± 0.005	0.997 ± 0.005
20%	0.999 ± 0.002	0.996 ± 0.004	0.998 ± 0.004
30%	1.000 ± 0.001	0.997 ± 0.004	1.000 ± 0.002

Even 5% sampling recovers rankings that are nearly identical to the full-data rankings across all three tracks. Therefore, GENARENA does not require exhaustive all-pair evaluation in practical deployments. The default benchmark uses fixed-seed uniform sampling for determinism and cross-run comparability. Fisher-information-guided sampling can be used as an optional acceleration strategy after an initial uniform round by allocating more battles to uncertain, near-even matchups, but it is not required by the standard pipeline.

N ADDITIONAL JUDGE COMPARISONS

N.1 CROSS-MODAL POINTWISE–PAIRWISE COMPARISONS

We extend the pointwise–pairwise comparison to GPT-5 and Qwen3-VL-32B-Instruct-FP8, the judge used for the GENARENA leaderboard. Table 14 reports results across image generation, image editing, and video generation, with Qwen3-VL-8B results included for reference.

Table 14: **Additional judge comparisons.** Accuracy is reported as pointwise $\rightarrow$ pairwise (%). GenAI-Bench is abbreviated as GAI, and EditScore-Bench as EditScore. Qwen3-VL-8B entries reproduce Table 1. Dashes indicate results not included in this comparison.

	Image gen.	Image editing		Video generation	
Judge	GAI	GAI	EditScore	GAI	VideoGen-Reward
GPT-5	$57.0 \rightarrow 67.4$	$84.5 \rightarrow 86.3$	$77.3 \rightarrow 80.7$	$65.9 \rightarrow 76.6$	–
Qwen3-VL-8B	$49.1 \rightarrow 60.5$	$73.4 \rightarrow 83.9$	$58.3 \rightarrow 83.7$	$62.0 \rightarrow 73.3$	$57.0 \rightarrow 61.5$
Qwen3-VL-32B-FP8	$51.5 \rightarrow 60.2$	$80.1 \rightarrow 87.5$	$69.5 \rightarrow 85.9$	$65.6 \rightarrow 72.1$	$52.4 \rightarrow 59.6$

Accuracies for the new runs use successfully judged examples. GPT-5 has slightly different judged subsets across protocols due to a small number of unsuccessful calls; the 32B-FP8 runs have complete coverage.

Pairwise accuracy is higher in every comparison shown, including those using the leaderboard judge. Under pairwise scoring, 32B-FP8 exceeds GPT-5 on both editing evaluations, whereas GPT-5 performs better on GenAI-Bench image and video generation. The advantage of either judge therefore depends on the task.

N.2 Full-Precision versus FP8 Judges

We repeat the judge-alignment experiment in Table 5b five times for full-precision Qwen3-VL-32B-Instruct and its FP8 variant, using the same MMRB2 editing pairs and accuracy convention.

Table 15: **Repeated-run comparison of full-precision and FP8 judges.** Accuracies are mean $\pm$ sample standard deviation (%) over five paired repeats. Differences and their 95% paired t-intervals are in percentage points (pp).

Track	32B	32B-FP8	FP8 − 32B (pp)	95% CI (pp)
Basic	66.90 ± 0.75	66.42 ± 0.78	-0.47	$[-1.07, 0.12]$
Reasoning	72.50 ± 1.54	71.07 ± 2.87	-1.43	$[-5.88, 3.02]$
MultiRef	64.66 ± 0.87	65.39 ± 1.37	$+0.73$	$[-1.12, 2.59]$
Overall	66.97 ± 0.62	66.63 ± 0.57	-0.34	$[-1.03, 0.35]$

All paired difference intervals include zero. The original $+0.4$ pp FP8 advantage in Table 5b is consistent with run-to-run variation, supporting FP8 as an efficiency-oriented deployment with similar observed accuracy.

O Controlled Comparison with LMArena

We evaluate pointwise and pairwise protocols with the same Qwen3-VL-8B-Instruct judge on the same 407 GEdit-Bench-EN prompts and generated images from seven models. Pointwise scores are averaged across prompts, while pairwise judgments are aggregated with BT/Elo. Both rankings are compared with the same LMArena reference.

Table 16: **Controlled ranking comparison.** Both protocols use Qwen3-VL-8B-Instruct on identical prompts and generated images. Parentheses indicate ranks; LMArena provides the external human-preference reference.

Generation model	Pointwise score	Pairwise Elo	LMArena Elo
Nano Banana	7.0985 (3)	1061.73 (1)	1325 (1)
Flux.2 [dev]	7.2899 (2)	1053.81 (2)	1250 (2)
Qwen-Image-Edit	6.5502 (6)	992.79 (4)	1231 (3)
Flux.1 Kontext [dev]	6.4230 (7)	964.49 (5)	1166 (4)
GPT Image 1 [High]	7.5290 (1)	1033.42 (3)	1155 (5)
Bagel	6.9635 (4)	885.16 (7)	1042 (6)
Step1X-Edit	6.7564 (5)	939.84 (6)	1014 (7)
Spearman ρ	0.1786	0.8571	–
95% CI for ρ	$[-0.2143, 0.3929]$	$[0.7500, 0.8571]$	–
Kendall τ	0.1429	0.7143	–

Confidence intervals are obtained by bootstrapping prompts and recomputing both rankings with the models and LMArena reference held fixed. Pairwise evaluation achieves higher alignment under this controlled setup (Table 16). LMArena provides a model-level human-preference reference rather than prompt-matched ground truth.

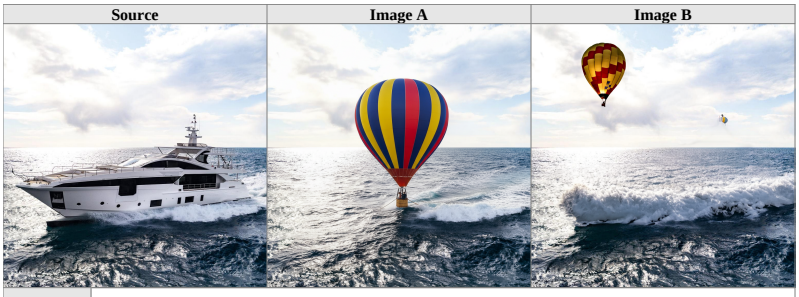
P CLOSE-MATCH DISCRIMINABILITY ANALYSIS

We retain all mutual battles among the top five models in each track of Table 2 and refit BT/Elo with the same tie-handling rule. Table 17 reports decisive outcomes and rank correlations with the full leaderboard restricted to these models.

Table 17: **Top-five model comparisons.** Decisive and tie rates are percentages. Rank correlations compare the restricted ranking with the full leaderboard ordering of the same five models.

Track	Decisive rate	Tie rate	ρ	τ
Basic	68.79	31.21	0.90	0.80
Reasoning	74.65	25.35	1.00	1.00
MultiRef	75.57	24.43	1.00	1.00

The judge yields decisive outcomes on most top-model battles and largely preserves the full ordering. Higher tie rates than in Table 12 indicate greater ambiguity. This is an internal ranking consistency check; human validation is reported in Appendix I.4.

<table>
<tr><th>Source</th><th>Image A</th><th>Image B</th></tr>
<tr><td colspan="3"></td></tr>
<tr><td>Prompt</td><td colspan="2">Replace the yacht in the image with a hot air balloon floating just above the ocean surface.</td></tr>
<tr><td>Ground Truth</td><td colspan="2">A > B</td></tr>
<tr><td rowspan="2">Pointwise Scoring</td><td>Score: 4.47

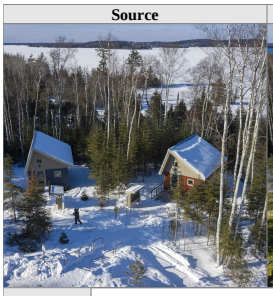
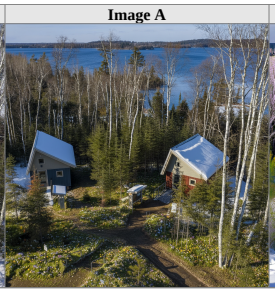
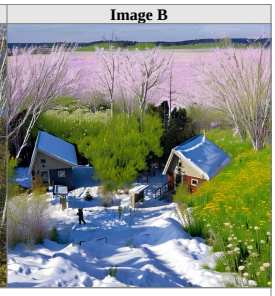
SC: The hot air balloon is correctly placed above the ocean surface, replacing the yacht as instructed. The edit is minimal and effective, preserving the original composition and lighting while only changing the central subject. No over-editing is present.

PQ: The image shows a hot air balloon floating over the ocean, which is physically impossible as hot air balloons cannot fly over water without landing. This violates basic physics and creates an unnatural scene. The balloon's colors are vibrant and well-defined, and there are no visible artifacts like distortion or watermarks. However, the core issue is the implausible scenario, which severely impacts naturalness.</td><td>Score: 5.37

SC: The editing successfully replaces the yacht with a hot air balloon floating above the ocean, fulfilling the instruction. The addition of a second, smaller balloon adds visual interest but is not part of the instruction, making it a minor over-edit. The original yacht is completely removed, which is a minimal edit in terms of composition, though the scene is significantly altered.

PQ: The image depicts a hot air balloon over the ocean, which is physically implausible since hot air balloons cannot fly over open water without land nearby. The lighting and shadows are inconsistent with the scene, and the second balloon in the distance appears distorted and unrealistic. These elements give the image an unnatural, AI-generated feel with visible artifacts.</td></tr>
<tr><td colspan="2">Result: A < B</td></tr>
<tr><td>Pairwise Scoring</td><td colspan="2"><pre>{
 "reasoning": {
 "text_faithfulness": "The instruction is to replace the yacht with a hot air balloon floating just above the ocean surface. Model A places the balloon directly above the ocean surface, visually fulfilling the 'just above' requirement. Model B places a balloon in the sky and adds a second, smaller balloon, deviating from the singular replacement instruction and misrepresenting the 'just above ocean surface' placement.",
 "image_faithfulness": "Model A preserves the ocean surface, sky, and wave patterns from the original image while replacing the yacht with a hot air balloon. Model B retains the ocean and sky but adds an unintended second balloon and alters the wave dynamics, suggesting a more complex, less faithful edit to the original composition.",
 "overall_image_quality": "Model A has better technical quality with fewer artifacts; the balloon is well-integrated, and the lighting and shadows match the original. Model B has visual inconsistencies — the second balloon and altered wave textures — which reduce its overall quality and realism.",
 "text_rendering": "Not Applicable",
 "comparison_summary": "Model A better fulfills the editing instruction by accurately replacing the yacht with a hot air balloon positioned just above the ocean surface while preserving the original image's elements and quality. Model B deviates by adding an extra balloon and altering the scene's dynamics, making it less faithful and lower quality."
 },
 "score": 6,
 "better_response": "A",
 "confidence": 0.95,
 "confidence_rationale": "The evaluation is clear-cut: Model A precisely follows the instruction with no deviations, while Model B introduces multiple inaccuracies (extra balloon, altered wave dynamics) that directly contradict the editing task. The visual quality of Model A is also superior, with no artifacts or inconsistencies. This judgment is highly confident due to the unambiguous nature of the instruction and the clear failure of Model B to meet it."
}</pre></td></tr>
<tr><td colspan="2">Result: A > B</td></tr>
</table>

Figure 5: Qualitative comparison of pointwise judgment and pairwise judgment.

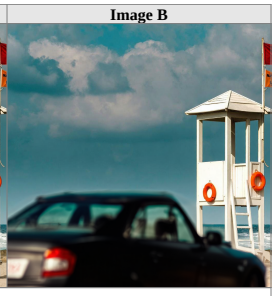
Source	Image A	Image B
		

Prompt	Change the snowy forest environment to a springtime forest with budding trees and wildflowers.	
Ground Truth	**A > B**	
Pointwise Scoring	**Score: 8.49** **SC:** The edit successfully replaces the snowy ground with spring grass and wildflowers, and changes the frozen lake to open water, aligning with the instruction. However, the snow on the roofs and some trees remains, indicating over-editing as the winter elements are not fully removed. **PQ:** The image depicts a serene lakeside forest scene with two cabins, birch trees, and wildflowers. The lighting and shadows are consistent, and the composition feels natural with realistic textures and proportions. There are no visible artifacts such as distortions, watermarks, or unnatural body parts. The scene is harmonious and visually coherent, suggesting high-quality AI generation.	**Score: 7.20** **SC:** The edited image successfully replaces the snowy ground with spring grass and wildflowers, and adds blossoming trees with pink flowers, fulfilling the instruction. However, the snow remains on the roofs and ground, which is a significant inconsistency with the 'springtime' request, indicating overediting. The edit is visually striking but not fully coherent with the instruction. **PQ:** The image has strong visual appeal with vibrant colors and detailed textures, but contains several unnatural elements: the snow appears inconsistent with the blooming flowers and greenery, the trees have odd, almost painterly blossoms, and the overall lighting and depth feel slightly off. There are no major artifacts like blurring or distortions, but the scene's logic is compromised by the conflicting seasonal elements.
	Result: A > B	

Pairwise Scoring	{
	"reasoning": {
	"text_faithfulness": "The instruction is to change the snowy forest to a springtime forest with budding trees and wildflowers. Model A partially fulfills this by removing snow and adding wildflowers, but leaves the trees bare. Model B fully fulfills the instruction by adding blossoming trees and wildflowers, though it inconsistently retains snow. Model B is more faithful to the instruction's intent.",
	"image_faithfulness": "Model A preserves the original composition and lighting well, but fails to show budding trees. Model B preserves the original houses and layout but introduces unnatural blossoms and removes snow inconsistently. Model A is more faithful to the original image's structure and elements.",
	"overall_image_quality": "Model A has a more naturalistic and coherent transition, with believable wildflowers and consistent lighting. Model B has more vibrant colors but appears overly stylized and inconsistent, with some visual artifacts and unnatural blending. Model A has better overall technical quality.",
	"text_rendering": "Not Applicable",
	"comparison_summary": "Model A is more faithful to the instruction's intent and preserves the original image's structure better, while Model B has more visually striking but less coherent results. Model A is the better edit overall."
	},
	"score": 5,
	"better_response": "A",
	"confidence": 0.5,
	"confidence_rationale": "The comparison is nuanced; while Model B is more visually dramatic, Model A better executes the editing instruction with greater consistency and realism. There is reasonable disagreement possible on which edit is 'better,' so confidence is medium."
	}	

	Result: A > B	

Figure 6: Qualitative comparison of pointwise judgment and pairwise judgment.

Source	Image A	Image B
		

Prompt	Add a car in the foreground to the right side of the image.	
Ground Truth	**A > B**	
Pointwise Scoring	**Score: 9.59** **SC:** The edited image successfully adds a white car to the foreground on the right side, as instructed. The car is well-integrated into the scene, maintaining consistent lighting and perspective. There is no overediting; the addition is minimal and effective without altering other elements significantly. **PQ:** The image depicts a vintage VW Beetle parked beside a lifeguard tower on a beach, under a cloudy sky. The lighting and shadows are consistent, the car and tower appear structurally plausible, and the ocean and sky have natural textures. There are no visible artifacts like distortions, watermarks, or unusual body parts. The scene feels cohesive and realistic, though the composition is slightly staged. The naturalness score is high due to realistic rendering, and the artifact score is high due to absence of flaws.	**Score: 10.00** **SC:** The edited image successfully adds a car in the foreground on the right side, as instructed. The car is blurred, suggesting depth of field, which makes it appear in the foreground without overpowering the scene. The addition is subtle and contextually appropriate for a beach setting. There is no over-editing; the original composition is largely preserved with only the car added. **PQ:** The image depicts a lifeguard tower and a car with a shallow depth of field, creating a natural-looking scene. Lighting and shadows are consistent, and the composition feels realistic. No visible artifacts like distortions, watermarks, or unusual anatomy are present. The image appears clean and well-rendered.
	Result: A < B	

Pairwise Scoring	{
	"reasoning": {
	"text_faithfulness": "The instruction is 'Add a car in the foreground to the right side of the image.' Model A places a white car clearly in the foreground on the right side, directly fulfilling the instruction. Model B places a car in the foreground on the right side, but it is heavily blurred and partially obstructs the view of the lifeguard tower, which may not be the intended effect. However, both models satisfy the positional requirement.",
	"image_faithfulness": "Model A preserves the original composition and lighting of the beach scene while adding a car that integrates well with the environment. Model B also preserves the original scene but uses a blurred effect that might not be intended by the instruction, potentially reducing the clarity of the edit. The lifeguard tower and background elements remain intact in both.",
	"overall_image_quality": "Model A has higher overall image quality. The car is rendered with clear details, appropriate lighting, and natural shadows, blending well with the scene. Model B's car is blurred, which introduces visual artifacts and reduces the quality of the edit, making it less convincing.",
	"text_rendering": "Not Applicable.",
	"comparison_summary": "Model A better fulfills the instruction by adding a clear, well-rendered car in the foreground on the right side. Model B adds a car but with a blurred effect that may not be intended, reducing the quality of the edit. Model A is superior in faithfulness to the instruction and overall image quality."
	},
	"score": 5,
	"better_response": "A",
	"confidence": 0.8,
	"confidence_rationale": "The difference is clear: Model A adds a well-integrated, clear car as instructed, while Model B adds a blurred car that may not be the intended effect. The superiority of Model A is evident in both faithfulness and quality, with minimal ambiguity."
	}	

	Result: A > B	

Figure 7: Qualitative comparison of pointwise judgment and pairwise judgment.