Visual Aesthetic Benchmark: Can Frontier Models Judge Beauty?

Yichen Feng[1,2,†*] **Yuetai Li**[2*] **Chunjiang Liu**[1,6,†*]
Yuanyuan Chen[1] **Fengqing Jiang**[2] **Yue Huang**[5] **Hang Hua**[7]
Zhengqing Yuan[5] **Kaiyuan Zheng**[2] **Luyao Niu**[2] **Bhaskar Ramasubramanian**[2,8]
Basel Alomair[9] **Xiangliang Zhang**[5] **Misha Sra**[3]
Zichen Chen[1,3,4‡] **Radha Poovendran**[2] **Zhangchen Xu**[1,2]

[1]Bake AI [2]University of Washington [3]University of California, Santa Barbara
[4]Stanford University [5]University of Notre Dame [6]Carnegie Mellon University
[7]MIT-IBM Watson AI Lab [8]Western Washington University
[9]King Abdulaziz City for Science and Technology

{yfeng42,yuetaili,fqjiang,kaiyuan5,luyaoniu,alomair,rp3,zxu9}@uw.edu,
chunjial@andrew.cmu.edu, {yhuang37,zyuan2,xzhang33}@nd.edu,
hang.hua1@ibm.com, yuanyuan.chen@ntu.edu.sg, ramasub@wwu.edu,
sra@cs.ucsb.edu, zichen_chen@ucsb.edu

Website https://vab.bakelab.ai
Code https://github.com/BakeLab/Visual-Aesthetic-Benchmark
Dataset https://huggingface.co/datasets/BakeLab/Visual-Aesthetic-Benchmark

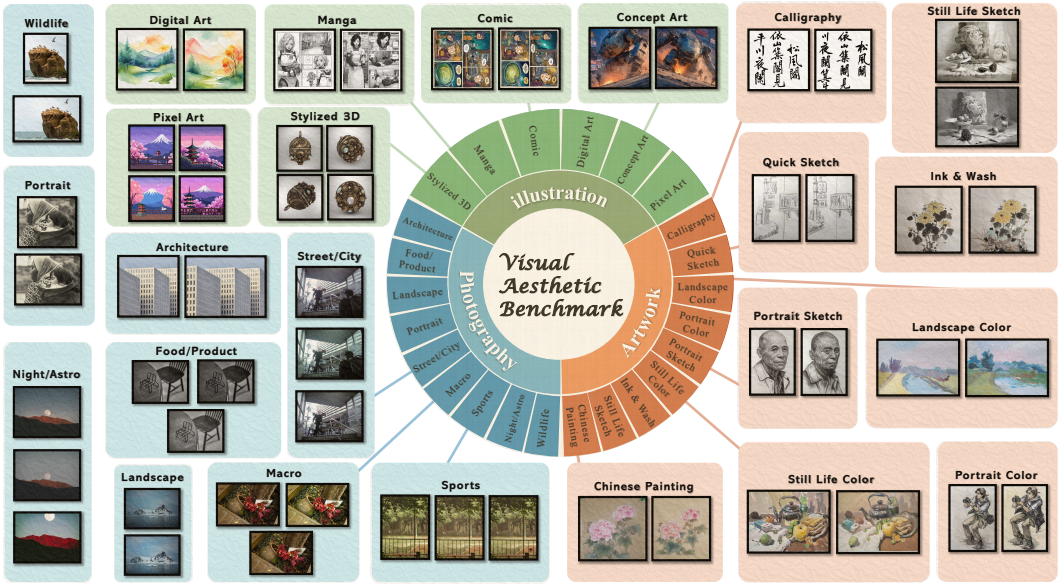


Figure 1: Overview of the VAB benchmark. VAB spans three visual domains — fine art, photography, and illustration — covering 24 topics across 400 evaluation tasks and 1,195 images. Each task presents a set of candidate images sharing a common subject.

[*] Equal contribution.
[†] Work done during internship at Bake AI.
[‡] Corresponding author.

Abstract

Multimodal large language models (MLLMs) are now routinely deployed for visual understanding, generation, and curation. A substantial fraction of these applications require an explicit aesthetic judgment. Most existing solutions reduce this judgment to predicting a scalar score for a single image. We first ask whether such scores faithfully capture comparative preference: in a controlled study with eight expert annotators, score-derived rankings align poorly with the same annotators' direct comparisons, while direct ranking yields substantially higher inter-annotator agreement on best- and worst-image labels. Motivated by this finding, we introduce the *Visual Aesthetic Benchmark* (VAB), an expert-consensus benchmark for ranked aesthetic judgment over matched-subject candidate sets that isolates differences in execution quality from differences in depicted content. VAB contains 400 tasks and 1,195 images across fine art, photography, and illustration, with labels derived from the consensus of 10 independent expert judges per task. Evaluating 20 frontier MLLMs and six dedicated visual-quality reward models, we find that the strongest system identifies both the best and the worst image correctly across three random permutations of the candidate order in only 26.5% of tasks, far below the 68.9% achieved by human experts. Fine-tuning a 35B-parameter model on 2,000 disjoint expert examples yields consistent gains across all four evaluation metrics, providing preliminary evidence that comparative supervision can improve aesthetic judgment within the VAB protocol. Together, these results expose a clear and measurable gap between current multimodal models and expert aesthetic judgment, and establish VAB as a controlled, expert-grounded testbed on which that gap can be tracked and closed.

1 Introduction

Multimodal large language models (MLLMs) have become critical to a widening range of visual tasks such as image generation and curation (Liu et al., 2023; Team et al., 2023; Wang et al., 2024). Many of these applications require aesthetic judgment, asking models to compare visual works and identify differences in execution quality across composition, color harmony, technical execution, and cultural convention. This makes aesthetic judgment a demanding test of higher-order multimodal reasoning that is often omitted by existing visual benchmarks (Yu et al., 2023; Yue et al., 2024).

Existing image-aesthetics datasets and methods typically reduce aesthetic judgment to a mapping from an image to a scalar score (Murray et al., 2012; Kong et al., 2016; He et al., 2022; Lee & Kim, 2019; Hosu et al., 2019; Yi et al., 2023; He et al., 2023), an approach that we argue is poorly suited to the underlying decision being studied. Preference order is inferred from independent scores rather than measured through direct comparison; crowd-averaged labels mix expertise, taste, and annotation noise, so the aesthetic signal is difficult to isolate; and because images in such datasets often depict different subjects, a model can exploit content preferences rather than discriminate execution quality.

We first ask whether scoring faithfully captures comparative preference . Eight expert annotators evaluate the same images under both independent absolute scoring and direct comparative ranking, and the two protocols disagree. Comparative ranking yields 42 percentage points higher inter-annotator agreement on best-image selection than score-derived rankings (Talebi & Milanfar, 2018; Ma et al., 2017; Wu et al., 2023; Zhou et al., 2024a). The two protocols therefore elicit different preference signals from the same annotators, and direct comparison captures the underlying judgment more faithfully than the pointwise scoring used in most existing pipelines.

Building on this finding, we introduce the **Visual Aesthetic Benchmark (VAB)**, which formulates aesthetic judgment as a comparative selection problem over candidate sets with matched subject matter. VAB contains 400 tasks and 1,195 images across fine art,

photography, and illustration, covering 24 topics (Figure 1). Each task presents two to six images sharing a common subject, and the model selects the best image and, under the stricter setting, also the worst image. Three design choices distinguish VAB from prior aesthetic benchmarks. Matched subject matter within each task isolates differences in execution from differences in content; each ground-truth label is the consensus of 10 independent expert judges, filtered through a dual-threshold criterion to remove ambiguous cases; and every task is evaluated under three independent random permutations of the candidate order. We report `ap@1`, the average accuracy across the three evaluations, and `pass^3`, which credits a task only when the model is correct in all three. Appendix A provides a detailed comparison with recent aesthetic benchmarks and datasets.

We benchmark 20 frontier MLLMs and six dedicated visual-quality reward models on VAB. The strongest system, Claude Sonnet 4.6, reaches 26.5% on `TB-1 pass^3`, 42.4 points below the human baseline of 68.9%. Models are also highly sensitive to candidate order, retaining only 34–65% of their `ap@1` accuracy when required to be correct across all three permutations, and they degrade much faster than humans as the candidate set grows: a 7.1× drop from two-image to four-or-more-image tasks, against only 2.0× for human experts. The contributions of this paper are as follows.

- A controlled human study with eight expert annotators establishes that, for selecting the best and worst images within matched-subject candidate sets, direct comparative ranking is substantially more reliable than independent scoring and motivates the design of the benchmark.
- We introduce VAB, an expert-consensus benchmark for ranked aesthetic judgment over matched-subject candidate sets, with permutation-robust evaluation of 400 tasks and 1,195 images across fine art, photography, and illustration.
- A systematic evaluation of 20 frontier MLLMs and six dedicated reward models exposes a persistent 42.4-point gap to human experts on `TB-1 pass^3` together with severe positional instability across permutations of the candidate order.
- Fine-tuning a 35B-parameter model on 2,000 disjoint expert examples yields consistent gains across all four metrics, providing preliminary within-benchmark evidence that comparative supervision can improve aesthetic judgment.

2 Human Annotation Study: Ranking vs. Scoring

Existing aesthetic benchmarks rely on *Scoring* to collect ground truth labels (Murray et al., 2012; Kong et al., 2016; He et al., 2022; Cao et al., 2025). We conduct a controlled human study to explore the following research question regarding the annotation process:

Which protocol yields a more scientifically sound signal for data annotation?

- **Scoring:** expert annotates each image independently on an absolute aesthetic scale.
- **Ranking:** expert compares images and annotates their relative order.

Setup and Metrics. Eight expert annotators evaluate the same images under two protocols: (1) *scoring* assigns each image an independent score $s_{i,a} \in [0, 10]$; (2) *ranking* orders images directly within groups of size $n \in \{2, 3, 4, 5\}$. We construct 119 *homogeneous-content* tasks, in which all candidates share the same underlying content, and 107 *heterogeneous-content* tasks formed by regrouping images across source groups while preserving topic consistency. We compare the two protocols along three dimensions:

- **Fidelity** (Kendall's τ (Kendall, 1938)): we compute $\tau = (C - D)/\binom{n}{2}$, where C and D are the numbers of concordant and discordant pairs between an annotator's score-derived ranking σ_a^{score} and the same annotator's direct comparison. A low τ means the two protocols elicit different preference orderings from the same person. We additionally report the top-1 self-consistency (SC) rate: the fraction of annotator–task pairs for which σ_a^{score} and the direct comparison select the same top image.

Table 1: We report within-annotator agreement between expert's score-derived rankings σ_a^{score} and direct comparison from 8 annotators, measured by Kendall's τ (↑) and top-1 self-consistency rate (SC rate, ↑), defined as the fraction of annotator–task pairs where the score-derived top-1 matches the direct comparison top-1. The results show that nearly 50% of the numeric score annotations do not reliably preserve the order that the annotator would express in direct comparison.

(a) Results on homogeneous-content tasks.

n	N tasks	Mean τ	Std. τ	SC rate
2	28	0.098	0.511	54.9%
3	38	0.274	0.328	49.7%
4	35	0.126	0.243	34.3%
5	18	0.267	0.243	43.6%
All	**119**	**0.188**	**0.354**	**45.5%**

(b) Results on heterogeneous-content tasks.

n	N tasks	Mean τ	Std. τ	SC rate
2	19	0.342	0.427	67.1%
3	37	0.286	0.300	48.0%
4	34	0.441	0.172	51.1%
5	17	0.324	0.256	45.6%
All	**107**	**0.351**	**0.291**	**52.0%**

Table 2: We report inter-annotator consistency on best and worst decisions for 8 annotators, where a majority requires agreement from at least 5 of 8. Each row shows task-level accuracy acc (↑), improvement Δ, and the exact two-sided McNemar p-value (↓) (McNemar, 1947; Fagerland et al., 2013). The Significance level is marked as $^{**}p < 0.001$. Results show that ranking is more likely to yield annotation labels that independent experts can reproduce.

	Homogeneous-content ($N = 119$)				Heterogeneous-content ($N = 107$)			
Metric	Scoring	Ranking	Δ	p	Scoring	Ranking	Δ	p
acc_{best}	52.9%	95.0%	+42.0%	$<0.001^{**}$	65.4%	94.4%	+29.0%	$<0.001^{**}$
$\text{acc}_{\text{worst}}$	55.5%	92.4%	+37.0%	$<0.001^{**}$	69.2%	93.5%	+24.3%	$<0.001^{**}$
$\text{acc}_{\text{best\&worst}}$	39.5%	90.8%	+51.3%	$<0.001^{**}$	49.5%	90.7%	+41.1%	$<0.001^{**}$

- **Reproducibility** ($\text{acc}_{\text{best/worst}}$): acc_{best} is the proportion of tasks where a single image receives ≥ 5 of 8 best-image votes; $\text{acc}_{\text{worst}}$ and $\text{acc}_{\text{best\&worst}}$ are defined analogously. Higher accuracy means the protocol produces ground truth labels that different experts can independently reproduce.
- **Distinguishability** (top-1 entropy (Shannon, 1948)): we measure $H = -\sum_i p_i \log_2 p_i$, where p_i is the fraction of annotators selecting image i as best. Lower entropy means annotators concentrate on a single top choice rather than spreading votes.

Full protocol details, notations, formal definitions, and statistical tests are in Appendix F.

Scoring does not faithfully preserve preference order. Table 1 shows that score-derived rankings often diverge from the direct comparisons given by the same annotator. The mean Kendall's τ is 0.188 for homogeneous-content and 0.351 for heterogeneous-content tasks; the top-1 self-consistency rates are 45.5% and 52.0%, respectively. In other words, nearly half of the numeric score annotations fail to recover the order that the same annotator expresses under direct comparison.

Ranking yields more reproducible labels. Table 2 shows that ranking yields substantially higher inter-annotator agreement on best and worst labels. On homogeneous-content tasks, ranking improves acc_{best} by 42.0 points, $\text{acc}_{\text{worst}}$ by 37.0 points, and $\text{acc}_{\text{best\&worst}}$ by 51.3 points (all $p < 0.001$, McNemar test (McNemar, 1947)); the same pattern holds for heterogeneous-content tasks. Independent experts can therefore reproduce best/worst labels far more reliably under ranking than under scoring.

Ranking makes the top choice stand out more clearly. Table 3 measures how concentrated their top-choice preference is within each task. It shows that ranking produces lower entropy over top-1 selections (Shannon, 1948) (0.729 vs. 1.213) and a much higher rate of unanimous choices (10.1% vs. 0.8%; Wilcoxon $p = 1.7 \times 10^{-15}$ (Wilcoxon, 1945)). More generally,

Table 3: We report top-choice concentration in the homogeneous-content setting (119 tasks, 8 annotators), measured by entropy H (Shannon, 1948) and the rate of unanimous tasks ($H = 0$). The 95% confidence intervals use a normal approximation for mean H and a Wilson interval (Wilson, 1927) for the unanimous-task rate. A Wilcoxon signed-rank test (Wilcoxon, 1945) yields that the difference is statistically significant ($p = 1.7 \times 10^{-15}$).

Metric	Scoring	Ranking
Mean H (↓) [95% CI]	1.213 [1.137, 1.289]	0.729 [0.665, 0.793]
Unanimous rate (↑)	0.8% [0.1%, 4.6%]	10.1% [5.9%, 16.8%]

ranking concentrates votes on a single preferred image rather than spreading them across several plausible winners.

Interpretation. Our findings align with conclusions from two complementary disciplines. **Aesthetically**, modern aesthetics favors a relational view of beauty (Goldman, 1995; Levinson, 2005): people judge preference more reliably when alternatives are juxtaposed than when items are assessed in isolation. **Cognitively**, comparative judgment is a more primitive operation than absolute evaluation (Thurstone, 1927). Absolute scoring requires maintaining an internal reference scale and converting qualitative impressions into numbers, imposing heavier cognitive load and greater instability (Stewart et al., 2005; Parducci, 1965).

We therefore adopt ranking as the annotation protocol for VAB. Extended discussion is provided in Appendix F.2.

3 The VAB Benchmark

Building on the findings of Section 2, we construct the **Visual Aesthetic Benchmark (VAB)** to test whether frontier multimodal models can make expert-aligned aesthetic judgments under a ranking-based protocol. In contrast with prior evaluations that aggregate scalar ratings from crowd annotators (Murray et al., 2012; Kong et al., 2016; He et al., 2022), VAB casts aesthetic evaluation as comparative selection over candidate sets with matched subject matter: each task asks the model to identify the best image and, in the stricter setting, also the worst. The remainder of this section describes the task layout, the data construction pipeline, and the ground-truth formation procedure.

3.1 Overview

Task formulation. Each benchmark task presents a candidate set of images with matched subject matter, so performance depends on differences in execution rather than differences in depicted content. VAB contains 400 released tasks and 1,195 images across fine art, photography, and illustration, covering 24 topics with candidate set sizes from 2 to 6 images. We consider two settings: in `Top-1`, the model selects the best image in the set; in `TB-1`, it identifies both the best and the worst image. Figure 1 provides a visual overview, Figure 2 summarizes the released domain and topic distribution, and Appendix B.1 reports detailed benchmark statistics.

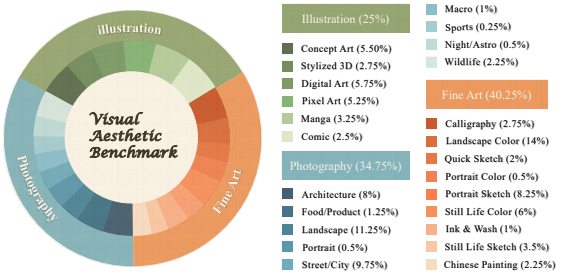


Figure 2: Detailed distribution of VAB.

Evaluation protocol. Labels are derived from direct comparative judgments by independent experts rather than post hoc aggregation of absolute scores. Each task is evaluated three times, each under an independent random permutation of the candidate order. We report `ap@1`, the average accuracy across the three evaluations, and `pass^3`, which counts a task as correct only if the model is correct in all three. Representative topic-level examples are collected in Appendix I.

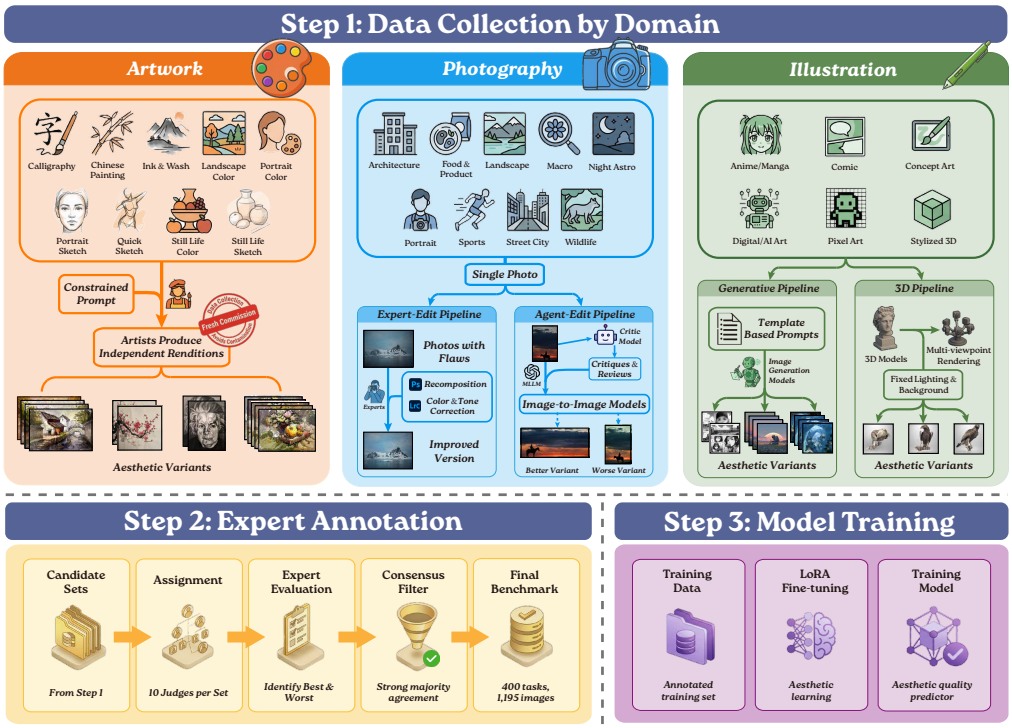


Figure 3: Overview of the VAB data construction pipeline. Images are organized into candidate sets with matched subject matter and varying aesthetic execution, then passed through expert review and consensus filtering before entering the benchmark.

3.2 Data Construction

Because VAB is annotated through direct ranking over candidate sets, data construction is organized around matched image groups rather than standalone images: in every domain, the members of a set share subject matter but vary in execution quality. Figure 3 summarizes the pipeline; Appendix B.3 reports the full raw-pool statistics, and Appendix B.4 contains the domain-level construction details.

Fine Art. We commissioned 426 painting sets comprising 1,142 images. Each set is built from constrained artist prompts that preserve subject matter while allowing variation in composition, value structure, color handling, and finish. All works were commissioned specifically for VAB.

Photography. We constructed 670 photography sets comprising 1,809 images. Each set originates from a single source photograph, many provided directly by professional photographers, and is expanded through two pipelines: an expert-edit pipeline based on professional retouching tools, which can produce either original/improved pairs or larger multi-expert sets, and a scalable agent-edit pipeline that generates improved and degraded variants. The agent-edit prompt sequence and runtime pipeline are recorded in Appendix H.1.

Illustration. We constructed 271 illustration sets comprising 908 images. These sets are formed either by prompt-based generation with text-to-image models or by rendering CC0 3D assets under controlled viewpoint and presentation changes, producing multiple images with matched subject matter but varying execution. Prompt templates are in Appendix H.2.

All collected sets are deduplicated and reviewed for semantic consistency before annotation.

3.3 Expert Annotation and Ground Truth

Following Section 2, VAB uses ranking-based expert consensus for ground truth.

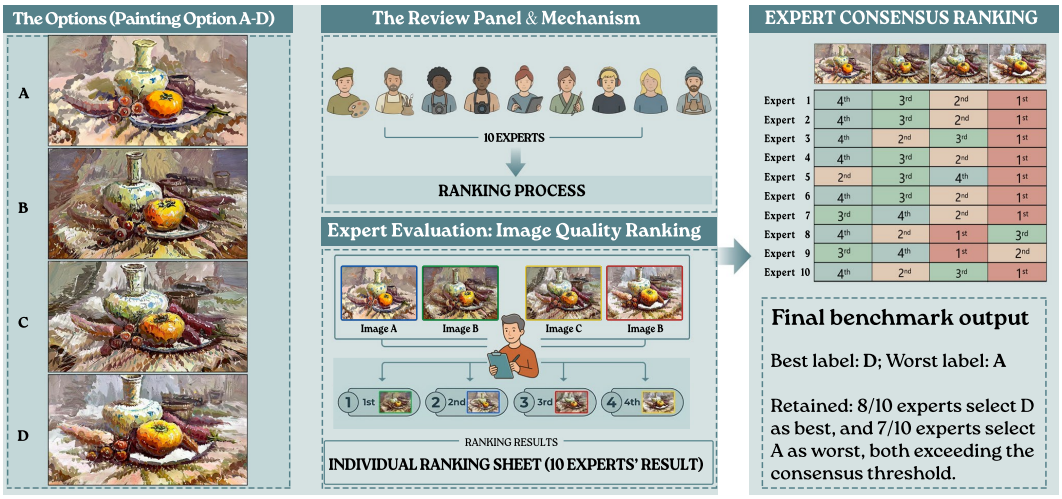


Figure 4: Example of a single VAB annotation task, illustrating the candidate set presented to expert judges and the resulting consensus decision. The consensus threshold is determined through statistical confidence; see Appendix D for details.

Review setup. Creation and evaluation are strictly separated. In the raw pre-filter pool, each collected candidate set is reviewed by 10 independent expert judges drawn from a pool of more than 100 evaluators, yielding over 13,000 assessments and more than 300 hours of review. Judges remain blind to artist identity throughout annotation. Evaluator qualifications, domain-matched assignment, rubrics, and interface examples are provided in Appendix C. Additional topic-level examples are in Appendix I.

Annotation protocol. Each task presents k candidate images side by side. Judges first assess each image against the domain-specific rubric and then make a comparative selection over the full set. For $k = 2$, a judge selects the best image. For $k \geq 3$, a judge selects both the best and the worst image, with distinct choices required. Every task receives exactly 10 judgments. A representative annotation example is shown in Figure 4.

Consensus filter. A task is retained only if expert votes are sufficiently concentrated. For $k = 2$, the best-image votes must meet the threshold; for $k \geq 3$, the best-image and worst-image votes must both meet it, so satisfying only one is insufficient. The thresholds are chosen by statistical confidence under a uniform-random null, and the exact derivation and threshold table are provided in Appendix D. By retaining only these high-confidence cases, VAB targets aesthetic judgments that are reproducible across independent experts while filtering out ambiguous comparisons.

4 Experiments

4.1 Setup

Evaluation scope. We evaluate 20 MLLMs and six reward models on VAB. Unless otherwise noted, each model is evaluated under its default inference configuration.

We compare all systems against two baselines. The *Human Expert* baseline is obtained from a fresh, independent round of annotation in which a separate group of 10 expert judges—none of whom contributed to the released consensus labels—evaluates each released task under the benchmark interface; we report their average accuracy against the released labels. The *Random Guess* baseline is computed analytically for each task size k.

MLLMs. Our MLLM evaluation covers 20 closed-source and open-weight models spanning the Claude family (Anthropic, 2026a; 2025a; 2026b; 2025b; 2026c), the Gemini family (Google DeepMind, 2025b;c; 2026), GPT and o-series models (OpenAI, 2025a;b;c;d;e;f), and additional systems from Doubao, Grok, GLM, Kimi, and Qwen (ByteDance Seed, 2026; xAI,

Table 4: Main VAB results for 20 frontier MLLMs. We report Top-1 and TB-1 accuracy under both ap@1 and pass^3 (%). Best result in each column is shown in **bold**.

	Overall				Fine Art				Illustration				Photography			
	Top-1		TB-1		Top-1		TB-1		Top-1		TB-1		Top-1		TB-1	
Model	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3
Random Guess	37.2	6.6	27.5	5.3	41.0	8.9	36.0	8.5	31.5	4.4	19.2	3.3	36.8	5.6	23.7	3.1
Human Expert	77.7		68.9		80.4		74.7		70.3		54.4		79.9		72.4	
Closed-Source Models																
CLAUDE-HAIKU-4.5	43.3	18.8	32.8	12.5	44.7	19.9	40.4	19.3	32.7	13.0	18.7	5.0	49.4	21.6	34.3	10.1
CLAUDE-SONNET-4.5	47.2	24.2	35.7	14.5	44.5	20.5	40.2	19.3	37.3	17.0	20.7	8.0	57.6	33.8	41.2	13.7
CLAUDE-SONNET-4.6	54.8	**40.2**	40.6	**26.5**	48.4	37.3	44.7	**34.2**	45.3	30.0	**29.3**	**19.0**	**68.8**	**51.1**	43.9	23.0
CLAUDE-OPUS-4.5	51.2	34.2	39.3	20.2	45.3	28.0	40.4	24.2	40.0	27.0	25.3	14.0	65.9	46.8	48.2	20.1
CLAUDE-OPUS-4.6	53.0	35.5	38.8	20.0	48.7	27.3	43.3	23.6	42.3	31.0	25.0	17.0	65.7	48.2	43.6	18.0
DOUBAO-SEED-2.0-PRO	51.3	33.2	**42.8**	23.5	**55.9**	**39.8**	**49.9**	32.9	39.3	18.0	25.0	6.0	54.7	36.7	47.5	25.2
GEMINI-3-FLASH	52.8	28.0	40.7	15.8	51.8	23.0	46.8	20.5	46.0	22.0	26.0	8.0	58.8	38.1	44.1	15.8
GEMINI-3-PRO	**55.4**	35.0	40.0	22.0	52.0	32.9	46.6	29.8	47.3	28.0	28.7	14.0	65.2	42.4	40.5	18.7
GEMINI-3.1-PRO	54.8	35.0	41.8	22.2	51.8	31.1	45.1	26.1	**49.7**	**32.0**	28.3	13.0	62.1	41.7	47.5	24.5
GPT-4.1	49.3	29.5	39.8	21.2	47.2	28.0	43.1	25.5	43.0	24.0	27.7	15.0	56.4	35.3	44.8	20.9
GPT-5	51.0	32.2	40.0	21.8	49.9	30.4	42.7	25.5	36.7	19.0	20.3	9.0	62.6	43.9	51.1	26.6
GPT-5.1	52.3	29.5	40.2	20.0	49.9	30.4	44.5	29.8	45.7	25.0	26.7	11.0	60.0	31.7	44.8	15.1
GPT-5.1-CODEX	50.2	31.8	39.2	22.0	47.4	29.8	42.0	27.3	41.3	22.0	24.3	15.0	60.0	41.0	46.8	20.9
GPT-5.2	46.3	24.0	37.4	15.5	40.8	18.0	35.4	14.9	33.7	13.0	21.3	8.0	61.9	38.8	51.3	21.6
GROK-4.1-FAST	44.6	15.5	35.2	12.5	44.5	18.0	39.5	17.4	36.3	15.0	25.0	11.0	50.6	12.9	37.4	7.9
O4-MINI	48.2	25.0	39.8	21.5	44.7	23.6	41.0	23.0	38.0	16.0	20.7	7.0	59.5	33.1	**52.3**	**30.2**
Open-Weight Models																
GLM-4.6V	43.0	17.5	34.1	11.5	42.7	18.0	39.3	18.6	43.3	19.0	27.0	8.0	43.2	15.8	33.1	5.8
KIMI-K2.5	49.4	26.8	36.8	15.0	43.5	19.9	38.5	18.0	43.0	21.0	29.0	14.0	60.9	38.8	40.5	12.2
QWEN3-VL-235B-A22B	44.9	20.5	35.2	14.0	42.4	22.4	36.9	19.3	38.3	12.0	25.7	8.0	52.5	24.5	40.0	12.2
QWEN3.5-397B-A17B	50.1	29.8	38.9	17.2	42.7	21.1	37.3	17.4	44.3	24.0	26.0	8.0	62.8	43.9	50.1	23.7

2025; Z.ai, 2025; Team et al., 2026; Bai et al., 2023; Qwen Team, 2026). For each task, the model selects either the best image (Top-1) or both the best and worst images (TB-1). We evaluate each task three times under independent random permutations of candidate order and report ap@1 and pass^3 as defined in Section 3. Prompts are provided in Appendix H.3.

Reward Models. We additionally evaluate six reward models: ArtiMuse (Cao et al., 2025), PEAS-aes and PEAS-ava (Yun & Choo, 2024), Q-Align (Wu et al., 2023), Q-Instruct (Wu et al., 2024), and RealQA-ava (Li et al., 2025). Unlike MLLMs, these models assign scalar scores to images independently rather than selecting directly from a set. We therefore follow each model's official scoring interface, score each image three times, average the scores, and then rank images within each task. The main text reports Top-1 and TB-1; model-specific scoring interfaces and full result tables, including Bot-1 and topic-level breakdowns, are deferred to Appendix H.4, Appendix E.4, and Appendix E.7.

4.2 Results

Tables 4 and 5 summarize the main results; Appendices E.1–E.7 provide the full tables, topic-level breakdowns, and auxiliary metrics.

4.2.1 Results for Multimodal Large Language Models

Gap to experts. Current frontier models remain far below expert performance. On the strictest metric, TB-1 pass^3, Claude Sonnet 4.6 attains 26.5%, against 68.9% for human experts; even under the more lenient TB-1 ap@1, the strongest model, Doubao Seed 2.0 Pro, reaches 42.8%, still 26.1 points short of the expert baseline.

Statistical uncertainty. Task-level paired-bootstrap intervals caution against overinterpreting small differences between adjacent leaderboard positions, particularly within the leading cluster on Top-1 ap@1. The main human–model gap is nevertheless robust: on TB-1 pass^3, the best MLLM's 95% confidence interval ends at 30.8%, below the Human Expert interval's lower bound of 66.8%. Appendix E.2 provides the full uncertainty analysis.

Table 5: Main reward-model results on VAB. We report overall and domain-level Top-1 and TB-1 accuracy (%). Best non-baseline result in each column is shown in **bold**.

Model	Overall		Fine Art		Illustration		Photography	
	Top-1	TB-1	Top-1	TB-1	Top-1	TB-1	Top-1	TB-1
Random Guess	37.2	27.5	41.0	36.0	31.5	19.2	36.8	23.7
Human Expert	77.7	68.9	80.4	74.7	70.3	54.4	79.9	72.4
ArtiMuse (Cao et al., 2025)	44.2	34.8	50.9	**46.0**	**37.0**	20.0	41.7	32.4
PEAS-aes (Yun & Choo, 2024)	48.5	37.8	49.7	44.1	34.0	**22.0**	57.6	41.7
PEAS-ava (Yun & Choo, 2024)	49.0	37.0	50.9	44.1	25.0	13.0	64.0	46.0
Q-Align (Wu et al., 2023)	**52.2**	**38.2**	**52.2**	44.7	31.0	16.0	**67.6**	**46.8**
Q-Instruct (Wu et al., 2024)	38.5	31.2	37.9	34.2	32.0	21.0	43.9	35.2
RealQA-ava (Li et al., 2025)	44.2	30.2	40.4	36.6	33.0	15.0	56.8	33.8

Progress across families. Improvements are uneven across model families. Claude Sonnet 4.6 nearly doubles its predecessor on TB-1 pass^3 (14.5% → 26.5%), whereas the GPT-5 series regresses from 21.8% (GPT-5) to 20.0% (GPT-5.1) and 15.5% (GPT-5.2). The closed/open gap persists: the strongest open-weight system, Qwen3.5-397B-A17B, reaches 17.2% on TB-1 pass^3, versus 26.5% for Claude Sonnet 4.6.

Domain difficulty and stability. Accuracy varies strongly across domains and across repeated evaluations. Illustration is the most difficult domain, where Claude Sonnet 4.6 attains only 19.0% on TB-1 pass^3, while photography is comparatively more tractable, with o4-mini reaching 30.2%. The large ap@1-to-pass^3 gap—for example, GLM-4.6V drops from 34.1% to 11.5% on TB-1—indicates that sensitivity to candidate order and response instability remain major failure modes.

4.2.2 *Results for Reward Models*

Overall performance. Reward models also remain well below human experts, although some are competitive with MLLMs on Top-1. Q-Align is the strongest overall, reaching 52.2% on Top-1 and 38.2% on TB-1, against 77.7% and 68.9% for the human baseline.

Domain trends and asymmetric failure modes. Domain trends broadly mirror the MLLM results. Appendix E.4 additionally reveals an asymmetry between best and worst predictions: Q-Instruct reaches 51.5% on overall Bot-1 and 77.0% on photography Bot-1, despite considerably weaker Top-1 scores of 38.5% overall and 43.9% in photography. This pattern suggests that detecting aesthetic failure is in some cases easier than identifying the strongest image in a set.

Comparison with MLLMs. The best reward-model Top-1 result , 52.2%, is numerically close to the best MLLM Top-1 ap@1 result , 55.4%, but the comparison is not strictly like-for-like: MLLMs make a single comparative decision over a candidate set, whereas reward models score images independently and are then ranked within the set.

4.3 Ablations and Additional Analysis

Topic balance. An unweighted macro-average across all 24 topics preserves an 18.6–37.4 point gap between human experts and the best MLLM for each of the four metrics; full results are reported in Appendix E.5.

Construction source. A source-split analysis comparing Agent-Edit and Expert-Edit photography tasks is reported in Appendix E.3.

Temperature ablation. For models with an exposed temperature parameter, greedy decoding improves Top-1 pass^3 by 1.1 points on average and TB-1 pass^3 by 2.3 points. Decoding stochasticity therefore contributes noise, but it does not explain the main performance gap; full model-by-model results are reported in Appendix E.8.

Table 6: Fine-tuning results on VAB for open-weight baselines, the base model, and KALLISTI-35B-A3B. Best result in each column is shown in **bold**.

Model	Top-1		TB-1	
	ap@1	p^3	ap@1	p^3
Open-Weight Models				
GLM-4.6V	43.0	17.5	34.1	11.5
KIMI-K2.5	49.4	26.8	36.8	15.0
QWEN3-VL-235B-A22B	44.9	20.5	35.2	14.0
QWEN3.5-397B-A17B	50.1	**29.8**	**38.9**	17.2
Ours				
QWEN3.5-35B-A3B (base)	47.5	25.5	36.5	15.8
KALLISTI-35B-A3B	**51.3** +3.8	29.5 +4.0	38.8 +2.3	**17.5** +1.7

Candidate set size. Performance degrades sharply as the candidate set grows. Under TB-1 pass^3, the best model falls from 47.3% on two-image tasks to 6.7% on four-image tasks, while human experts decline from 87.1% to 43.6%. Larger candidate sets therefore amplify both comparative difficulty and instability; the full breakdown is reported in Appendix E.9.

5 Fine-Tuning on Expert Aesthetic Judgments

Training Setup. We fine-tune QWEN3.5-35B-A3B using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank 64 and $\alpha = 128$, freezing the vision tower while updating the multimodal projector. Supervision comes from 2,000 additional expert-annotated examples collected with the same annotation pipeline described in Section 3.3. The training examples are disjoint from the benchmark evaluation set at the image, prompt, artist/source, and task-instance levels. We train for three epochs with a learning rate of 1×10^{-4} and cosine scheduling. Additional hyperparameters, separation details, and full fine-tuning results are provided in Appendix G. The fine-tuned model is denoted as KALLISTI-35B-A3B.

Improvement over the base model. Table 6 compares KALLISTI-35B-A3B against its base model and a set of open-weight baselines. Fine-tuning yields consistent gains across all four reported metrics: Top-1 pass^3 improves from 25.5% to 29.5%, TB-1 pass^3 from 15.8% to 17.5%, Top-1 ap@1 from 47.5% to 51.3%, and TB-1 ap@1 from 36.5% to 38.8%. The topic-level breakdown is reported in Appendix G.3.

Scope of evidence. The consistent base-to-fine-tuned gains provide preliminary evidence that expert comparative supervision can improve aesthetic judgment within the VAB protocol. The larger open-weight baselines in Table 6 contextualize the resulting accuracy, but the experiment neither tests external transfer nor supports general model-scaling claims.

6 Conclusion

We introduced VAB, a 400-task, 1,195-image benchmark spanning fine art, photography, and illustration that evaluates matched-subject candidate sets using expert rankings rather than scalar scores. Direct ranking yields 42 percentage points higher inter-annotator agreement than score-derived rankings. Across 20 frontier MLLMs and six reward models, the strongest system achieves 26.5% on TB-1 pass^3 versus 68.9% for human experts, with performance degrading as candidate sets grow and marked positional instability across order permutations. Fine-tuning a 35B-parameter model on 2,000 disjoint expert examples improves all four metrics, providing preliminary within-benchmark evidence for expert comparative supervision. Future work should broaden domain and annotator diversity, improve order robustness, and scale expert supervision. VAB provides a controlled foundation for measuring and advancing expert-aligned multimodal aesthetic judgment.

Ethics Statement

This study involved human participants providing aesthetic judgments through an online evaluation platform. The University of Washington Human Subjects Division (HSD) reviewed the protocol and determined that the research qualifies as exempt human subjects research (Category 2 and 3; minimal-risk behavioral research) under U.S. federal regulations. Participation was voluntary and participants could stop at any time. No sensitive personal information was collected. Only aggregated benchmark results are publicly released, and no identifiable participant data are included in any publications or datasets. The exempt determination letter is on file with the authors.

All fine art works were created as commissioned works for VAB, with participating artists informed of and consenting to the use of their work for research. Professional photographers whose images served as source material for the photography pipeline contributed their work under appropriate licensing terms. Agent-edited variants derived from these source photographs were produced within the scope of those agreements. Illustration sets generated through text-to-image models used structured prompts without reproducing copyrighted material, and 3D-rendered sets were constructed exclusively from CC0-licensed assets. We are committed to crediting and fairly compensating all human creators whose work contributed to this benchmark.

References

Adobe. Adobe Lightroom. Adobe Website, 2026a. URL https://www.adobe.com/products/photoshop-lightroom.html.

Adobe. Adobe Photoshop. Adobe Website, 2026b. URL https://www.adobe.com/products/photoshop.html.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anthropic. Introducing Claude Opus 4.5. Anthropic News, 2025a. URL https://www.anthropic.com/news/claude-opus-4-5.

Anthropic. Introducing Claude Sonnet 4.5. Anthropic News, 2025b. URL https://www.anthropic.com/news/claude-sonnet-4-5.

Anthropic. Introducing Claude Haiku 4.5. Anthropic News, 2026a. URL https://www.anthropic.com/news/claude-haiku-4-5.

Anthropic. Introducing Claude Opus 4.6. Anthropic News, 2026b. URL https://www.anthropic.com/news/claude-opus-4-6.

Anthropic. Introducing Claude Sonnet 4.6. Anthropic News, 2026c. URL https://www.anthropic.com/news/claude-sonnet-4-6.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62, 1936.

ByteDance Seed. SEED2. ByteDance Seed Website, 2026. URL https://seed.bytedance.com/en/seed2.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025.

Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria HUANG, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *Advances in Neural Information Processing Systems*, 38, 2026.

Creative Commons. CC0 1.0 Universal. Creative Commons Website, 2009. URL https://creativecommons.org/publicdomain/zero/1.0/.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Morten W Fagerland, Stian Lydersen, and Petter Laake. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, 13(1):91, 2013.

Alan H. Goldman. *Aesthetic Value*. Westview Press, 1995.

Google DeepMind. Gemini 3 Pro Image (Nano Banana Pro). Google DeepMind Model Card, 2025a. URL https://deepmind.google/models/model-cards/gemini-3-pro-image/.

Google DeepMind. Gemini 3 Flash. Google DeepMind Model Card, 2025b. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Google DeepMind. Gemini 3 Pro. Google DeepMind Model Card, 2025c. URL https://deepmind.google/models/model-cards/gemini-3-pro/.

Google DeepMind. Gemini 3.1 Pro. Google DeepMind Model Card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In *IJCAI*, volume 6, pp. 22, 2022.

Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21838–21847, 2023.

Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9375–9383, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5911–5920, 2024a.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024b.

Ruixiang Jiang and Chang Wen Chen. Multimodal llms can reason about aesthetics in zero-shot. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6634–6643, 2025.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Maurice G Kendall and B Babington Smith. The problem of m rankings. *The annals of mathematical statistics*, 10(3):275–287, 1939.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Keunsoo Ko, Jun-Tae Lee, and Chang-Su Kim. Pac-net: Pairwise aesthetic comparison network for image aesthetic assessment. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2491–2495. IEEE, 2018.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pp. 662–679. Springer, 2016.

Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1191–1200, 2019.

Jerrold Levinson. *The Oxford handbook of aesthetics*. Oxford University Press, 2005.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025.

Boyang Liu, Yifan Hu, Senjie Jin, Shihan Dou, Gonglei Shi, Jie Shao, Tao Gui, and Xuanjing Huang. Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization. *arXiv preprint arXiv:2509.21871*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4535–4544, 2017.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Midjourney. Version. Midjourney Documentation, 2026. URL https://docs.midjourney.com/docs/models.

Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

OpenAI. GPT-4.1. OpenAI API Documentation, 2025a. URL https://developers.openai.com/api/docs/models/gpt-4.1.

OpenAI. Introducing GPT-5. OpenAI Website, August 2025b. URL https://openai.com/index/introducing-gpt-5/.

OpenAI. Introducing GPT-5.1. OpenAI Website, 2025c. URL https://developers.openai.com/api/docs/models/gpt-5.1.

OpenAI. GPT-5.1 Codex. OpenAI API Documentation, 2025d. URL https://developers.openai.com/api/docs/models/gpt-5.1-codex.

OpenAI. Introducing GPT-5.2. OpenAI Website, 2025e. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. o4-mini. OpenAI API Documentation, 2025f. URL https://developers.openai.com/api/docs/models/o4-mini.

Allen Parducci. Category judgment: a range-frequency model. *Psychological review*, 72(6): 407, 1965.

PixelLab. PixelLab. PixelLab Website, 2026. URL https://www.pixellab.ai/.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The Photographer's Eye: Teaching Multimodal Large Language Models to See, and Critique Like Photographers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24807–24816, 2025.

Qwen Team. Qwen3.5. Qwen Blog, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Sketchfab. Sketchfab. Sketchfab Website, 2026. URL https://sketchfab.com/.

Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2. 5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Louis L. Thurstone. A Law of Comparative Judgment. *Psychological Review*, 34(4):273–286, 1927. doi: 10.1037/h0070288.

Guolong Wang, Heng Huang, Zhiqiang Zhang, Wentian Li, Feilong Ma, and Xin Jin. Aestest: Measuring aesthetic intelligence from perception to production. *arXiv preprint arXiv:2511.06360*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenshan Wang, Su Yang, Weishan Zhang, and Jiulong Zhang. Neural aesthetic image reviewer. *IET Computer Vision*, 13(8):749–758, 2019.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25490–25500, 2024.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Tianyi Xu, Rong Shan, Junjie Wu, Jiadeng Huang, Teng Wang, Jiachen Zhu, Wenteng Chen, Minxin Tu, Quantao Dou, Zhaoxiang Wang, et al. Photobench: Beyond visual matching towards personalized intent-driven photo retrieval. *arXiv preprint arXiv:2603.01493*, 2026.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19861–19869, 2022.

Zhichao Yang, Jianjie Wang, Zhixianhe Zhang, Pangu Xie, Xiangfei Sheng, Pengfei Chen, and Leida Li. Fine-grained image aesthetic assessment: Learning discriminative scores from relative ranks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 145–155, 2026.

Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22388–22397, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Jooyeol Yun and Jaegul Choo. Scaling up personalized image aesthetic assessment via task vector customization. In *European Conference on Computer Vision*, pp. 323–339. Springer, 2024.

Z.ai. GLM-4.6V. Z.ai Blog, 2025. URL https://z.ai/blog/glm-4.6v.

Hantao Zhou, Longxiang Tang, Rui Yang, Guanyi Qin, Yan Zhang, Yutao Li, Xiu Li, Runze Hu, and Guangtao Zhai. UniQA: Unified vision-language pre-training for image quality and aesthetic assessment. *arXiv preprint arXiv:2406.01069*, 2024a.

Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024b.

Appendix

A Related Work 17
B Benchmark Details 18
B.1 Final Benchmark Statistics 18
B.2 Post-Filter Construction-Source Composition 19
B.3 Raw Collection Statistics Before Expert Filtering 20
B.4 Data Construction Details 20
C Annotation Details 22
C.1 Evaluator Pool and Assignment 22
C.2 Annotation Rubrics 22
C.3 Annotation Interface 25
D Ground Truth Mathematical Details 25
D.1 Consensus Labels and Complementary Metrics 25
D.2 Selected Thresholds 25
D.3 Derivation of Null Pass Probabilities 26
D.4 Monte Carlo Verification 26
D.5 Special Case for Binary Tasks 26
E Detailed Experimental Results 26
E.1 Comprehensive MLLM Results 27
E.2 Task-Level Bootstrap Uncertainty 27
E.3 Photography Source-Split Analysis 28
E.4 Comprehensive Reward-Model Results 28
E.5 Macro-Topic Robustness 28
E.6 Topic-Level MLLM Results 28
E.7 Topic-Level Reward-Model Results 31
E.8 Temperature Ablation 33
E.9 Results by Candidate Set Size 40
F Human Study Details 40
F.1 Detailed Setup and Metric Definitions 40
F.2 Extended Interpretation 41
F.3 Task Examples 42
F.4 Annotation Protocol 42
F.5 Notation 47
F.6 Statistical Tests 47
F.7 Task Lists 47
G Fine-Tuning Details 53
G.1 Training–Evaluation Separation and Scope 53
G.2 Hyperparameters 54

G.3 Detailed Results 54
H Prompt Templates and Scoring Interfaces 58
H.1 Photography Pipeline Prompts 58
H.2 Illustration Pipeline Prompts 60
H.3 MLLM Evaluation Prompts 61
H.4 Reward-Model Scoring Interfaces 62
I Topic-Level Annotation Examples 64
I.1 Fine Art 64
I.2 Photography 68
I.3 Illustration 73

A Related Work

Multimodal Large Language Models. Recent MLLMs such as Flamingo, BLIP-2, InstructBLIP, LLaVA, Gemini, and Qwen2-VL have rapidly improved multimodal perception, instruction following, and reasoning (Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Team et al., 2023; Wang et al., 2024). Their capabilities are typically assessed on broad benchmarks such as MM-Vet and MMMU (Yu et al., 2023; Yue et al., 2024), where each example usually admits a single correct answer. Aesthetic judgment is different: it is comparative, convention-sensitive, and not reducible to factual correctness alone. Our focus is therefore whether frontier MLLMs can make stable, expert-aligned selections over candidate sets with matched subject matter rather than pointwise judgments on isolated images.

Image Aesthetics Assessment. Image aesthetics assessment (IAA) has traditionally been posed as single-image score prediction. Canonical datasets such as AVA and AADB provide large-scale photographic images with scalar ratings. Later resources extend this paradigm through comments, theme-aware labels, and personalization, including AVA-Reviews, TAD66K, and PARA (Murray et al., 2012; Kong et al., 2016; Wang et al., 2019; He et al., 2022; Yang et al., 2022). These resources support a broad line of regression and ranking models, including NIMA, A-Lamp, theme-aware predictors, and more recent vision-language scorers such as Q-Align and UniQA (Talebi & Milanfar, 2018; Ma et al., 2017; He et al., 2022; Wu et al., 2023; Zhou et al., 2024a). Still, the dominant setup remains pointwise, mapping from one image to a score, with preference order inferred only indirectly from independent estimates.

More recently, MLLM-based efforts have moved toward richer aesthetic perception and reasoning. AesBench evaluates aesthetic perception in frontier MLLMs; AesTest covers perception, appreciation, creation, and photography through a broader multiple-choice task suite; AesExpert and UNIAA develop instruction-tuned systems for critique and assessment; and ArtiMuse and the photography-focused *The Photographer's Eye* move closer to expert-style analysis (Huang et al., 2024b; Wang et al., 2025; Huang et al., 2024a; Zhou et al., 2024b; Cao et al., 2025; Qi et al., 2025). PhotoBench instead studies personalized, intent-driven photo retrieval, rather than set-level aesthetic best/worst selection (Xu et al., 2026). MJ-Bench evaluates multimodal judges for text-to-image feedback across alignment, safety, image quality, and bias, using prompt-conditioned generations (Chen et al., 2026). These benchmarks are complementary, but their retrieval, critique, or text-to-image feedback targets differ from VAB's expert-consensus comparative label.

Relative supervision and comparative resources. Several image-aesthetics methods explicitly use relative supervision. PARN directly models relative photo-aesthetic ranks in its training loss while jointly learning aesthetic ratings and attributes with content-adaptive

branches (Kong et al., 2016); PAC-Net learns a pairwise aesthetic comparator for ranking and classification (Ko et al., 2018); and Pairwise IAA uses pairwise comparisons to learn aesthetic scores for regression, classification, and personalization (Lee & Kim, 2019). More recently, Aes-R1 introduced a training framework whose relative-absolute policy optimization jointly optimizes score regression and relative ranking for aesthetic reasoning (Liu et al., 2025). These methods establish the value of relative supervision, but their primary contribution is a model, loss, comparator, or training procedure rather than controlled benchmark construction.

Comparative datasets and benchmarks provide another related direction. FGAesthetics organizes related image series and calibrates within-series pairwise preferences into ranks and scores for fine-grained assessment (Yang et al., 2026). ImageReward and Pick-a-Pic provide human-preference signals for text-to-image generation, comparing prompt-conditioned generated images for reward learning or preference modeling (Xu et al., 2023; Kirstain et al., 2023). FineArtBench and ArtCoT focus on aesthetic comparison and reasoning for fine art and generative art (Jiang & Chen, 2025). These efforts are complementary to VAB, which differs by combining benchmark-wide matched-subject candidate-set construction with direct set-level judgments from domain experts, consensus best/worst labels, and a common evaluation protocol spanning fine art, photography, and illustration. By controlling depicted content to emphasize execution quality, VAB contributes an expert-consensus benchmark for ranked aesthetic judgment rather than a new relative-supervision method.

B Benchmark Details

B.1 Final Benchmark Statistics

Table 7 reports the final benchmark statistics after expert annotation and consensus filtering.

Table 7: Final benchmark statistics for VAB by domain and topic after expert consensus filtering. T/I denotes the number of tasks and images.

Fine Art (161/458)		**Photography** (139/389)		**Illustration** (100/348)	
Topic	T / I	Topic	T / I	Topic	T / I
Landscape Color	56/243	Landscape	45/130	Digital & AI Art	23/84
Portrait Sketch	33/66	Street & City	39/106	Concept Art	22/88
Still Life Color	24/53	Architecture	32/89	Pixel Art	21/84
Still Life Sketch	14/28	Wildlife	9/24	Anime & Manga	13/26
Calligraphy	11/22	Food & Product	5/14	Stylized 3D	11/44
Chinese Painting	9/18	Macro	4/12	Comic	10/22
Quick Sketch	8/16	Night & Astro	2/6		
Ink & Wash	4/8	Portrait	2/5		
Portrait Color	2/4	Sports	1/3		

Table 8 shows the candidate set size distribution in the final benchmark by domain, and Table 9 provides the corresponding topic-level breakdown.

Table 8: Candidate set size distribution in the final benchmark by domain.

Domain	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Fine Art	109	4	13	34	1
Photography	30	107	2	0	0
Illustration	26	0	74	0	0
Total	165	111	89	34	1

The released scale is the result of the complete control and filtering pipeline rather than a target image count: every retained candidate set must preserve subject matter, pass

Table 9: Topic-level breakdown of the final benchmark after expert consensus filtering.

Domain	Topic	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Tasks	Avg. k
Fine Art	Calligraphy	11	0	0	0	0	11	2.00
Fine Art	Chinese Painting	9	0	0	0	0	9	2.00
Fine Art	Ink and Wash	4	0	0	0	0	4	2.00
Fine Art	Landscape Color	6	4	12	33	1	56	4.34
Fine Art	Portrait Color	2	0	0	0	0	2	2.00
Fine Art	Portrait Sketch	33	0	0	0	0	33	2.00
Fine Art	Quick Sketch	8	0	0	0	0	8	2.00
Fine Art	Still Life Color	22	0	1	1	0	24	2.21
Fine Art	Still Life Sketch	14	0	0	0	0	14	2.00
Photography	Architecture	7	25	0	0	0	32	2.78
Photography	Food and Product	1	4	0	0	0	5	2.80
Photography	Landscape	6	38	1	0	0	45	2.89
Photography	Macro	0	4	0	0	0	4	3.00
Photography	Night and Astrophotography	0	2	0	0	0	2	3.00
Photography	Portrait	1	1	0	0	0	2	2.50
Photography	Sports	0	1	0	0	0	1	3.00
Photography	Street and City	11	28	0	0	0	39	2.72
Photography	Wildlife	4	4	1	0	0	9	2.67
Illustration	Anime and Manga	13	0	0	0	0	13	2.00
Illustration	Comic	9	0	1	0	0	10	2.20
Illustration	Concept Art	0	0	22	0	0	22	4.00
Illustration	Digital and AI Art	4	0	19	0	0	23	3.65
Illustration	Pixel Art	0	0	21	0	0	21	4.00
Illustration	Stylized 3D	0	0	11	0	0	11	4.00

Table 10: Construction-source composition of the released benchmark after expert-consensus filtering. Shares are computed within each domain.

Domain	Construction Source	Tasks	Share
Photography	Agent-Edit	69	49.6%
Photography	Expert-Edit	70	50.4%
Illustration	Text-to-Image	89	89.0%
Illustration	3D Rendering	11	11.0%

semantic-consistency review, and meet the 10-expert consensus criterion. The resulting 400 tasks and 1,195 images therefore form a controlled diagnostic benchmark for expert-aligned comparative aesthetic judgment, rather than an exhaustive census of every visual style, subject, or cultural preference.

The uneven topic counts are an outcome of the expert-consensus filter rather than a target distribution imposed in advance. Enforcing uniform counts would require either discarding valid consensus-passing tasks from larger topics or admitting weaker labels into smaller topics. We therefore retain all qualified tasks, treat results from very small topics as diagnostic, and report both the complete topic-level breakdown and an unweighted 24-topic macro-average in Appendix E.5.

B.2 Post-Filter Construction-Source Composition

Table 10 reports the construction-source composition of the released photography and illustration subsets. Construction source is neither used to assign labels nor used as a shortcut for retention. Regardless of provenance, every candidate set is reviewed for semantic consistency, evaluated by 10 independent expert judges, and retained only if it passes the same candidate-size-dependent consensus filter.

The released photography subset is nearly balanced between the Agent-Edit and Expert-Edit pipelines. We maintained this balance after filtering by increasing the pre-filter pool

Table 11: Raw collection statistics by domain before expert filtering.

Domain	Tasks	Images	Avg. k
Fine Art	426	1,142	2.68
Photography	670	1,809	2.70
Illustration	271	908	3.35
Total	1,367	3,859	2.82

Table 12: Topic-level raw collection statistics before expert filtering.

Domain	Topic	Tasks	Images	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Avg. k
Fine Art	Landscape Color	113	502	12	6	20	70	5	4.44
Fine Art	Portrait Sketch	93	186	93	0	0	0	0	2.00
Fine Art	Still Life Color	73	160	68	0	1	4	0	2.19
Fine Art	Still Life Sketch	43	86	43	0	0	0	0	2.00
Fine Art	Calligraphy	44	88	44	0	0	0	0	2.00
Fine Art	Quick Sketch	30	60	30	0	0	0	0	2.00
Fine Art	Chinese Painting	13	26	13	0	0	0	0	2.00
Fine Art	Ink and Wash	10	20	10	0	0	0	0	2.00
Fine Art	Portrait Color	7	14	7	0	0	0	0	2.00
Photography	Landscape	280	761	81	197	2	0	0	2.72
Photography	Street and City	165	450	45	120	0	0	0	2.73
Photography	Architecture	157	420	51	106	0	0	0	2.68
Photography	Wildlife	24	62	11	12	1	0	0	2.58
Photography	Portrait	13	31	8	5	0	0	0	2.38
Photography	Macro	13	36	3	10	0	0	0	2.77
Photography	Night and Astrophotography	7	20	1	6	0	0	0	2.86
Photography	Food and Product	7	18	3	4	0	0	0	2.57
Photography	Sports	4	11	1	3	0	0	0	2.75
Illustration	Digital and AI Art	51	184	10	0	41	0	0	3.61
Illustration	Concept Art	50	200	0	0	50	0	0	4.00
Illustration	Stylized 3D	50	200	0	0	50	0	0	4.00
Illustration	Pixel Art	40	160	0	0	40	0	0	4.00
Illustration	Anime and Manga	40	82	39	0	1	0	0	2.05
Illustration	Comic	40	82	39	0	1	0	0	2.05

available to each pipeline. In illustration, the 11 tasks produced by 3D rendering constitute the Stylized 3D topic; the remaining 89 tasks across the other five illustration topics are produced through text-to-image generation. Because construction source and topic coincide for the 3D-rendered subset, we do not treat this as a like-for-like source comparison; instead, the illustration topic-level results in Appendix E.6 make its behavior directly inspectable.

B.3 Raw Collection Statistics Before Expert Filtering

Before expert filtering, the collected pool contains 1,367 tasks and 3,859 images. Table 11 summarizes the raw collection by domain, Table 12 reports the corresponding topic-level breakdown, and Table 13 further decomposes the photography subset by collection pipeline.

B.4 Data Construction Details

Fine Art. We commissioned 426 painting sets comprising 1,142 images across nine fine-art topics: calligraphy, Chinese painting, ink and wash, landscape color, portrait color, portrait sketch, quick sketch, still-life color, and still-life sketch. For each topic, artists received constrained prompts that fixed the subject and high-level compositional requirements while allowing variation in execution. The resulting sets preserve semantic intent while differing in composition, value structure, color harmony, and paint handling. All works were commissioned specifically for VAB rather than drawn from existing collections, which reduces the risk of contamination from publicly indexed art. Commissioning does not eliminate sampling effects, however, and the resulting collection may still reflect the composition of the creator pool and the design of the prompts. We therefore treat the fine-art subset as a

Table 13: Raw photography collection statistics by topic and collection pipeline before expert filtering. Agent-Edit and Expert-Edit denote task counts from the automated editing pipeline and the expert retouching pipeline, respectively.

Topic	Tasks	Images	$k = 2$	$k = 3$	$k = 4$	Avg. k	Agent-Edit	Expert-Edit
Landscape	280	761	81	197	2	2.72	188	92
Street and City	165	450	45	120	0	2.73	115	50
Architecture	157	420	51	106	0	2.68	104	53
Wildlife	24	62	11	12	1	2.58	10	14
Portrait	13	31	8	5	0	2.38	5	8
Macro	13	36	3	10	0	2.77	10	3
Night and Astrophotography	7	20	1	6	0	2.86	5	2
Food and Product	7	18	3	4	0	2.57	4	3
Sports	4	11	1	3	0	2.75	3	1

controlled expert-evaluation setting rather than a comprehensive census of global artistic traditions. Before annotation, we remove near-duplicates and discard sets with semantic mismatches or trivial cues.

Photography. We constructed 670 photography sets comprising 1,809 images across nine topics: architecture, food and product, landscape, macro, night and astrophotography, portrait, sports, street and city, and wildlife. Each set originates from a single source photograph, many provided directly by professional photographers. We transform these source images into controlled comparison sets through two pipelines.

In the *expert-edit pipeline*, photographers and retouchers improve each source image using tools such as Adobe Photoshop and Lightroom (Adobe, 2026b;a), applying recomposition, color and tone correction, and content-aware expansion. The original photograph and its improved variant form a pair. For a subset of source images, multiple experts independently edit the same photograph, yielding larger candidate sets with matched subject matter and varying aesthetic quality.

In the *agent-edit pipeline*, each source image is first sent to a critic model, ArtiMuse (Cao et al., 2025), which produces a detailed aesthetic review. The review and the original image are then passed to a general-purpose MLLM, GPT-5 (OpenAI, 2025b), to generate two editing prompts: one for improvement and one for degradation. These prompts, together with the original image, are fed to an image-to-image model, Gemini-3-Pro-Image (Nano Banana Pro) (Google DeepMind, 2025a), to generate an improved and a degraded variant. The original, improved, and degraded images form a three-image candidate set. The exact prompt sequence is recorded in Appendix H.1.

All photography sets from both pipelines are deduplicated and reviewed for semantic consistency before annotation.

Illustration. We constructed 271 illustration sets comprising 908 images across six topics: anime and manga, comic, concept art, digital and AI art, pixel art, and stylized 3D. The sets are derived from two distinct sources, prompt-based generation and 3D asset rendering.

In the *generative pipeline*, each set originates from a structured prompt that specifies subject identity, scene context, composition, camera viewpoint, lighting, and stylistic attributes. For each set, the prompt is instantiated using a single text-to-image model, such as Midjourney v7 (Midjourney, 2026), Pixellab.ai (Create M-XL) (PixelLab, 2026), or LlamaGen (Sun et al., 2024), to generate candidate images with matched subject matter and varying aesthetic quality. The prompt templates are provided in Appendix H.2.

In the *3D-rendering pipeline*, each set is constructed from CC0-licensed assets sourced from Sketchfab (Creative Commons, 2009; Sketchfab, 2026). For each asset, we render multiple images by varying camera viewpoint while keeping object identity, scene semantics, lighting, and background conditions fixed. The resulting renders form candidate sets with matched subject matter and controlled variation in composition and visual presentation.

All illustration sets from both pipelines are deduplicated and reviewed for semantic consistency before annotation.

C Annotation Details

C.1 Evaluator Pool and Assignment

The evaluator pool contains more than 100 domain-qualified visual experts rather than general crowdworkers. Every evaluator has at least a relevant undergraduate or graduate degree in fine art, photography, illustration, visual design, or a closely related field, together with practical experience in the corresponding domain. Although domain-qualified, the pool is not constructed as a culturally stratified panel and does not include balanced cultural-background metadata. The resulting labels should therefore be interpreted as expert-consensus judgments under our domain-specific rubrics, rather than as equally representing all cultural communities or aesthetic traditions.

For each candidate set, we randomly sample 10 distinct judges from the qualified pool for that domain: fine-art, photography, and illustration tasks are therefore evaluated by experts with matching backgrounds. Data creation and evaluation are strictly separated, and judges remain blind to artist identity throughout annotation to reduce reputation and halo effects. The domain-specific rubrics below further standardize the criteria used across judges.

Independent human baseline. The released consensus labels and the Human Expert baseline come from two non-overlapping annotation rounds. The labels are formed from the original 10 expert judgments per task, whereas the baseline is collected in a fresh round from a separate group of 10 domain-qualified experts who did not participate in producing those labels. The baseline therefore measures reproducibility against an independent expert group rather than agreement with an annotator's own prior judgment.

C.2 Annotation Rubrics

The rubric serves as a structured comparison guide rather than a separate numerical scoring form. Judges assess each image against domain-specific criteria before making their final comparative selection. The rubrics below describe the criteria used for each domain.

Fine Art Rubric

Given a set of 2 to 6 fine art paintings with the same style, subject matter, and medium, and only subtle differences, determine which work demonstrates greater artistic completeness and more mature aesthetic expression.

Core Comparison Criteria

1. **Composition and Visual Order**
   - Is the primary subject more clearly defined?
   - Does the composition feel more stable or intentionally dynamic?
   - Is the spatial arrangement natural, without unnecessary congestion or emptiness?
2. **Color and Tonal Relationships**
   - Are the color relationships more harmonious or layered?
   - Is there stronger control over warmth/coolness and light/dark contrast?
   - Does the work avoid muddiness, dullness, or scattered color impressions?
3. **Technical Control**
   - Are the brushstrokes more confident, fluid, and rhythmic?
   - Are there signs of hesitation, redundancy, or ineffective marks?
   - Does the technique serve the image rather than draw attention to itself?
4. **Degree of Completion**
   - Are there any areas that appear unfinished or perfunctory?
   - Are edges and transitions handled naturally?
   - Does the work feel resolved, at a point where the artist could reasonably stop?

5. **Artistic Expression and Overall Character**
   - Is the mood or atmosphere more focused and coherent?
   - Does the work convey a stronger overall character or presence?
   - Does it have greater artistic persuasiveness?
6. **Final Selection**
   - Considering the set as a whole, which image is the strongest work?
   - If the set contains three or more images, which image is the weakest work?

Should Not Be Considered

- Personal preference for subject matter.
- Any contextual or background information beyond what is visible in the image itself.

Photography Rubric

Given a set of 2 to 6 photographs taken of similar subjects, in similar scenes, and under closely matched conditions, determine which image is more successful in terms of photographic language and visual expression.

Core Comparison Criteria

1. **Composition and Image Structure**
   - Is the subject more clearly emphasized?
   - Is the frame more balanced or intentionally dynamic?
   - Is the background cleaner and more supportive of the subject?
2. **Quality of Light**
   - Is the light more natural or more expressive?
   - Are highlights and shadows better controlled?
   - Does the image avoid crushed blacks or blown highlights?
3. **Sharpness and Focus**
   - Is focus more accurate where it matters?
   - Is the relationship between sharpness and blur more appropriate?
   - Are there any technical errors that detract from the image?
4. **Color or Tonal Rendering**
   - Is white balance or color temperature more accurate or more intentional?
   - Are colors or tonal values more consistent and unified?
   - Does the image avoid unwanted color casts or muddy tones?
5. **Moment and Emotional Expression**
   - Does the image capture a stronger or more meaningful moment?
   - Is the emotional or narrative content more compelling?
   - Does it carry a stronger sense of photographic presence?
6. **Final Selection**
   - Considering the set as a whole, which photograph is the strongest image?
   - If the set contains three or more images, which photograph is the weakest image?

Should Not Be Considered

- The personal appeal or attractiveness of the subject.
- Assumptions about shooting difficulty.
- Assumptions about post-processing effort.

Digital Illustration Rubric

Given a set of 2 to 6 digital illustrations with similar purpose, style, and level of completion, determine which image demonstrates stronger visual design and higher professional illustration quality.

Core Comparison Criteria

1. **Visual Focus and Information Clarity**
   - Is the visual focal point clearer?
   - Does the visual flow guide the viewer more effectively?
   - Is there less visual noise or competition for attention?
2. **Form and Proportion**
   - Are the structures of figures or objects more accurate?
   - Are poses and silhouettes more aesthetically refined?
   - Does the image avoid stiffness, distortion, or unnatural forms?
3. **Color Design**
   - Is the color palette more unified and thematically coherent?
   - Are value and lighting hierarchies clearer?
   - Is there effective use of contrast and visual rhythm?
4. **Detail Control and Finish**
   - Are details handled appropriately, neither excessive nor careless?
   - Are materials and lighting consistent throughout?
   - Are edges, masks, and gradients clean and well controlled?
5. **Stylistic Consistency and Purpose Fit**
   - Does the illustration better align with its intended stylistic direction?
   - Is quality more consistent within a unified style?
   - Is the image more suitable for its potential use case?
6. **Final Selection**
   - Considering the set as a whole, which illustration is the strongest work?
   - If the set contains three or more images, which illustration is the weakest work?

Should Not Be Considered

- Personal preference for illustration style.
- Software or tools used.
- Hypothetical or imagined client requirements.

Table 14: Consensus thresholds and corresponding null pass probabilities by candidate set size k, with $n = 10$ judges per task.

k	Threshold m	$p_{\text{accept}}(k, m)$
2	8	10.94%
3	7	0.88%
4	6	0.95%
5	6	0.15%
6	6	0.03%

C.3 Annotation Interface

Figure 5 shows an example annotation interface used by expert judges.

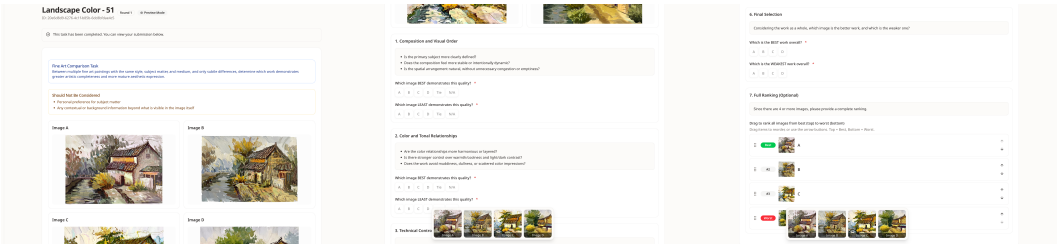


Figure 5: The annotation interface presents all k candidate images side by side. Judges first review each image with respect to the rubric criteria, then make a final comparative selection: selecting the best image for $k = 2$ and selecting both the best and worst images for $k \geq 3$.

D Ground Truth Mathematical Details

This appendix provides additional mathematical details for the consensus filter described in Section 3.3.

D.1 Consensus Labels and Complementary Metrics

VAB uses hard best/worst labels as its primary target because the benchmark asks whether a model can recover a concrete expert-consensus decision over a matched candidate set. These labels are formed only for tasks that pass the 10-judge consensus filter below, so ambiguous subjective variation is not converted into a benchmark target without sufficient agreement. The reported `ap@1` metric gives credit across individual candidate-order permutations, whereas `pass^3` additionally tests whether the decision is robust across all three permutations.

Vote-share-weighted accuracy, top-tier matching, and distributional agreement would provide useful complementary views of minority expert preferences. We leave these metrics to future work because they answer a different question—matching the full vote distribution or an acceptable preference tier—rather than recovering the retained consensus decision that defines the present benchmark.

D.2 Selected Thresholds

Table 14 reports the consensus thresholds used in the benchmark together with their corresponding null pass probabilities.

For $k \geq 3$, all four exact null pass probabilities—0.88%, 0.95%, 0.15%, and 0.03%—are below 1%.

Table 15: Monte Carlo verification of null pass probabilities (10^7 samples).

k	m	Exact	MC Estimate	95% CI
2	8	10.94%	10.9496%	[10.9303%, 10.9690%]
3	7	0.88%	0.8818%	[0.8760%, 0.8876%]
4	6	0.95%	0.9507%	[0.9447%, 0.9567%]
5	6	0.15%	0.1489%	[0.1465%, 0.1513%]
6	6	0.03%	0.0307%	[0.0296%, 0.0318%]

D.3 Derivation of Null Pass Probabilities

For $k \geq 3$ candidates with $n = 10$ judges, each judge's response is modeled as a directed edge $i \to j$ (selecting image i as best and image j as worst, with $i \neq j$) drawn uniformly from $k(k-1)$ possibilities. Let c_{ij} denote the count of judges choosing the pair (i, j). The probability of any configuration C under the random null is:

$$\Pr(C) = \frac{n!}{\prod_{i \neq j} c_{ij}!} \left(\frac{1}{k(k-1)} \right)^n .$$

The exact by-chance pass probability is computed by summing over all configurations satisfying the consensus threshold:

$$p_{\text{accept}}(k, m) = \sum_{\substack{C:\, c_{ii}=0, \\ \sum_{i \neq j} c_{ij} = n}} \frac{n!}{\prod_{i \neq j} c_{ij}!} \left(\frac{1}{k(k-1)} \right)^n \mathbf{1}\left(\max_i B_i \geq m \right) \mathbf{1}\left(\max_j W_j \geq m \right),$$

where $B_i = \sum_{j \neq i} c_{ij}$ (votes for i as best) and $W_j = \sum_{i \neq j} c_{ij}$ (votes for j as worst). The exact probabilities were computed by exhaustive enumeration over all valid count configurations $C = (c_{ij})$ satisfying $c_{ii} = 0$ and $\sum_{i \neq j} c_{ij} = n$, summing the multinomial probability mass over the configurations that satisfy the acceptance condition. For $k \geq 3$, this acceptance condition is conjunctive: a task passes only if the best-vote threshold and the worst-vote threshold are both met, so satisfying only one of the two is insufficient.

D.4 Monte Carlo Verification

We verify the exact probabilities via Monte Carlo simulation with 10^7 samples per (k, m) configuration. Table 15 compares the exact values from Table 14 against Monte Carlo estimates and reports 95% Wald confidence intervals.

D.5 Special Case for Binary Tasks

For binary tasks ($k = 2$), the worst image is implied once the best is selected. The consensus filter reduces to requiring $\max_i B_i \geq m$ where $B \sim \text{Binomial}(n, 1/2)$:

$$p_{\text{accept}}(2, m) = \Pr(B \geq m) + \Pr(B \leq n - m) = 2 \sum_{b=m}^{n} \binom{n}{b} \left(\frac{1}{2} \right)^n .$$

For $n = 10$ and $m = 8$, this gives $p_{\text{accept}}(2, 8) = 2 \times (45 + 10 + 1)/1024 = 112/1024 \approx 10.94\%$. We use this lower binary threshold because $k = 2$ tasks provide only a single comparative decision per annotator; compared with $k \geq 3$, demanding the same level of concentration would remove a disproportionate number of otherwise informative binary comparisons.

E Detailed Experimental Results

This appendix supplements Section 4 with full result tables that are too large for the main text. We retain the same evaluation focus as in the main paper: overall performance relative to expert baselines, cross-domain variation, sensitivity to candidate order, and auxiliary analyses on decoding and task size.

Table 16: Comprehensive VAB results for evaluated MLLMs and reference baselines. We report Top-1, Bot-1, and TB-1 accuracy under both ap@1 and pass^3 (%). Best result in each non-baseline column is shown in **bold**.

	Overall						Fine Art						Illustration						Photography					
	Top-1		Bot-1		TB-1		Top-1		Bot-1		TB-1		Top-1		Bot-1		TB-1		Top-1		Bot-1		TB-1	
Model	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3
Random Guess	37.2	6.6	37.2	6.6	27.5	5.3	41.0	8.9	41.0	8.9	36.0	8.5	31.5	4.4	31.5	4.4	19.2	3.3	36.8	5.6	36.8	5.6	23.7	3.1
Human Expert	77.7		80.1		68.9		80.4		82.7		74.7		70.3		70.4		54.4		79.9		84.0		72.4	
Closed-Source Models																								
Claude-Haiku-4.5	43.3	18.8	50.4	26.5	32.8	12.5	44.7	19.9	49.1	27.3	40.4	19.3	32.7	13.0	41.7	12.0	18.7	5.0	49.4	21.6	58.3	36.0	34.3	10.1
Claude-Sonnet-4.5	47.2	24.2	53.3	28.5	35.7	14.5	44.5	20.5	51.6	26.7	40.2	19.3	37.3	17.0	40.3	17.0	20.7	8.0	57.6	33.8	64.7	38.8	41.2	13.7
Claude-Sonnet-4.6	54.8	**40.2**	54.4	36.8	40.6	**26.5**	48.4	37.3	53.2	34.2	44.7	**34.2**	45.3	30.0	40.3	23.0	**29.3**	**19.0**	**68.8**	**51.1**	65.9	49.6	43.9	23.0
Claude-Opus-4.5	51.2	34.2	55.5	32.8	39.3	20.2	45.3	28.0	51.3	28.0	40.4	24.2	40.0	27.0	44.7	**29.0**	25.3	14.0	65.9	46.8	68.1	41.0	48.2	20.1
Claude-Opus-4.6	53.0	35.5	53.9	31.2	38.8	20.0	48.7	27.3	50.7	27.3	43.3	23.6	42.3	31.0	42.7	26.0	25.0	17.0	65.7	48.2	65.7	39.6	43.6	18.0
Doubao-Seed-2.0-Pro	51.3	33.2	60.8	**44.8**	**42.8**	23.5	**55.9**	**39.8**	60.0	**46.0**	**49.9**	32.9	39.3	18.0	43.3	17.0	25.0	6.0	54.7	36.7	74.1	63.3	47.5	25.2
Gemini-3-Flash	52.8	28.0	60.5	38.2	40.7	15.8	51.8	23.0	**60.7**	40.4	46.8	20.5	46.0	22.0	47.3	15.0	26.0	8.0	58.8	38.1	69.8	52.5	44.1	15.8
Gemini-3-Pro	**55.4**	35.0	55.2	33.8	40.0	22.0	52.0	32.9	55.1	32.9	46.6	29.8	47.3	28.0	44.3	19.0	28.7	14.0	65.2	42.4	63.3	45.3	40.5	18.7
Gemini-3.1-Pro	54.8	35.0	**61.1**	41.5	41.8	22.2	51.8	31.1	59.4	41.6	45.1	26.1	**49.7**	**32.0**	**49.0**	23.0	28.3	13.0	62.1	41.7	71.7	54.7	47.5	24.5
GPT-4.1	49.3	29.5	59.3	43.2	39.8	21.2	47.2	28.0	53.0	37.9	43.1	25.5	43.0	24.0	44.7	21.0	27.7	15.0	56.4	35.3	77.2	**65.5**	44.8	20.9
GPT-5	51.0	32.2	58.1	43.0	40.0	21.8	49.9	30.4	50.5	36.6	42.7	25.5	36.7	19.0	41.7	23.0	20.3	9.0	62.6	43.9	**78.7**	64.7	51.1	26.6
GPT-5.1	52.3	29.5	59.1	43.8	40.2	20.0	49.9	30.4	53.8	41.6	44.5	29.8	45.7	25.0	44.3	26.0	26.7	11.0	60.0	31.7	75.8	59.0	44.8	15.1
GPT-5.1-Codex	50.2	31.8	58.8	42.5	39.2	22.0	47.4	29.8	53.4	39.1	42.0	27.3	41.3	22.0	45.3	19.0	24.3	15.0	60.0	41.0	74.8	63.3	46.8	20.9
GPT-5.2	46.3	24.0	53.8	34.5	37.4	15.5	40.8	18.0	44.9	26.7	35.4	14.9	33.7	13.0	39.7	15.0	21.3	8.0	61.9	38.8	74.3	57.6	51.3	21.6
Grok-4.1-Fast	44.6	15.5	47.8	24.2	35.2	12.5	44.5	18.0	49.5	27.3	39.5	17.4	36.3	15.0	34.3	13.0	25.0	11.0	50.6	12.9	55.6	28.8	37.4	7.9
o4-mini	48.2	25.0	53.7	32.2	39.8	21.5	44.7	23.6	46.4	26.7	41.0	23.0	38.0	16.0	41.7	15.0	20.7	7.0	59.5	33.1	70.7	51.1	**52.3**	**30.2**
Open-Weight Models																								
GLM-4.6V	43.0	17.5	51.4	25.8	34.1	11.5	42.7	18.0	47.4	24.2	39.3	18.6	43.3	19.0	41.7	14.0	27.0	8.0	43.2	15.8	63.1	36.0	33.1	5.8
Kimi-K2.5	49.4	26.8	52.5	28.2	36.8	15.0	43.5	19.9	52.2	31.7	38.5	18.0	43.0	21.0	41.0	15.0	29.0	14.0	60.9	38.8	61.2	33.8	40.5	12.2
Qwen3-VL-235B-A22B	44.9	20.5	50.2	22.8	35.2	14.0	42.4	22.4	46.8	23.0	36.9	19.3	38.3	12.0	42.7	9.0	25.7	8.0	52.5	24.5	59.5	32.4	40.0	12.2
Qwen3.5-397B-A17B	50.1	29.8	54.8	36.2	38.9	17.2	42.7	21.1	46.2	28.6	37.3	17.4	44.3	24.0	41.3	18.0	26.0	8.0	62.8	43.9	74.3	58.3	50.1	23.7

E.1 Comprehensive MLLM Results

Table 16 reports comprehensive MLLM results, including Bot-1, across overall and domain-level evaluations. These results complement Table 4 in the main text by making the asymmetry between identifying the strongest and weakest images explicit. They also make clear that the gap between ap@1 and pass^3 is pervasive rather than concentrated in a small subset of models, consistent with the interpretation in Section 4.2.1 that many MLLM predictions remain sensitive to candidate order and response instability.

The full table also makes several model-level patterns easier to inspect. Within the Claude family, Claude Sonnet 4.6 improves substantially over Claude Sonnet 4.5 on the stricter TB-1 pass^3 metric, whereas the GPT-5 series declines monotonically from GPT-5 to GPT-5.2. The same table also makes the closed/open gap clearer under stability-sensitive metrics: Qwen3.5-397B-A17B is much closer to the strongest closed models under ap@1 than under pass^3, suggesting weaker consistency across repeated evaluations. Finally, the domain columns reinforce the main-text headline that illustration is the most difficult domain, photography is generally more tractable, and fine art lies between them.

E.2 Task-Level Bootstrap Uncertainty

To quantify sensitivity to the finite composition of the benchmark, we resample the 400 tasks with replacement for 10,000 paired bootstrap replicates and recompute all four metrics. Each replicate uses the same resampled task indices for every system, preserving paired comparisons. Table 17 reports the point estimate and 95% bootstrap confidence interval for Human Expert and all 20 MLLMs.

The intervals support a deliberately restrained interpretation of small leaderboard differences. In particular, the leading Top-1 ap@1 intervals overlap substantially, and paired-bootstrap comparisons do not separate adjacent top-ranked systems. They do, however, preserve the broad capability differences: Gemini 3 Pro significantly outperforms 12 of the other 19 MLLMs on Top-1 ap@1, and Claude Sonnet 4.6 significantly outperforms 13 of 19 on TB-1 pass^3 ($p < 0.05$). Most importantly, on the latter metric, Claude Sonnet 4.6 has an upper confidence bound of 30.8%, whereas the Human Expert lower bound is 66.8%. Thus, task composition affects the precision of model-to-model ordering but does not explain the central human–model gap.

E.3 Photography Source-Split Analysis

We test whether photography tasks constructed by the Agent-Edit pipeline exhibit systematically different evaluation behavior from tasks constructed by the Expert-Edit pipeline. This diagnostic is applied to the released post-filter subsets described in Appendix B.2; both subsets therefore use labels obtained through the same expert-consensus procedure. Tables 18 and 19 report the complete results for the 20 MLLMs evaluated in the paper.

Across models and metrics, the source-conditioned differences vary in both sign and magnitude, and neither construction source produces a consistent advantage. The Agent-Edit photography tasks therefore exhibit broadly similar evaluation behavior to the Expert-Edit tasks under the common expert-consensus evaluation, although the model-level variation cautions against treating the two subsets as identical in difficulty.

E.4 Comprehensive Reward-Model Results

Table 20 reports comprehensive reward-model results, including `Bot-1`, across overall and domain-level evaluations. In contrast to the condensed main-text presentation in Table 5, this table shows that some reward models are materially stronger at rejecting poor images than at identifying the best one. In particular, the detailed `Bot-1` columns make the failure mode discussed in Section 4.2.2 directly visible, especially for Q-Instruct in photography.

E.5 Macro-Topic Robustness

The released tasks are unevenly distributed across the 24 topics, so the overall task-weighted results give greater influence to larger topics. As a complementary diagnostic, Table 21 first computes each system's result separately within every topic and then averages the 24 topic values with equal weight.

The expert-model gap remains large under this reweighting. Across the four metrics, the best MLLM for each metric remains 18.6–37.4 percentage points below the Human Expert baseline. The paper's central conclusion therefore does not depend on larger topics receiving more weight in the aggregate evaluation.

E.6 Topic-Level MLLM Results

To match the main-text MLLM table, we report topic-level results under all four metrics: `Top-1 ap@1`, `Top-1 pass^3`, `TB-1 ap@1`, and `TB-1 pass^3`. Together, these tables show that the domain-level trends in Section 4.2.1 are not driven by a single topic or a single evaluation metric. Illustration remains the hardest domain across all four views, while photography is generally more tractable but still highly uneven across topics.

Several topic effects recur across metrics. In fine art, Chinese painting and portrait-color tasks are often relatively tractable, whereas landscape-color and some sketch-based tasks become much less stable under the stricter criteria. In illustration, the remaining signal is concentrated mostly in anime-and-manga and comic tasks, while concept art, digital-and-AI art, and stylized 3D frequently collapse under `pass^3`. In photography, macro and sports tasks are often close to saturation for strong models, but night-and-astrophotography, architecture, and some street-and-city tasks remain much less stable, especially when consistency across repeated evaluations is required.

Cultural and topic sensitivity in fine art. The fine-art subset contains 161 tasks across nine topics (Table 7). Grouping Calligraphy, Chinese Painting, and Ink & Wash as East Asian-specific topics gives 24/161 tasks (14.9%); the remaining 137 tasks cover broader studio-art categories. After excluding those three topics, Doubao Seed 2.0 Pro obtains 54.2% on `Top-1 ap@1` and 47.2% on `TB-1 ap@1`, close to its full fine-art averages of 55.9% and 49.9%, respectively. Its aggregate fine-art result is therefore not driven solely by the East Asian-specific topics. This topic-based sensitivity analysis should not be interpreted as evidence of a model's training-data exposure or as a stratified cross-cultural study; the evaluator-pool and cultural-metadata limitations are described in Appendix C.

Table 17: Task-level bootstrap uncertainty for the overall MLLM results. We use 10,000 paired bootstrap replicates over the 400 tasks. Each cell reports the point estimate followed by its 95% confidence interval in brackets (%).

Model	Top-1 ap@1	Top-1 pass^3	TB-1 ap@1	TB-1 pass^3
Human Expert	77.7 [76.5, 79.0]	77.7 [76.5, 79.0]	68.9 [66.8, 70.8]	68.9 [66.8, 70.8]
Gemini-3-Pro	55.4 [51.5, 59.2]	35.0 [30.2, 39.8]	40.0 [36.2, 43.8]	22.0 [18.0, 26.2]
Gemini-3.1-Pro	54.8 [50.9, 58.8]	35.0 [30.2, 39.8]	41.8 [38.0, 45.6]	22.2 [18.2, 26.2]
Claude-Sonnet-4.6	54.8 [50.5, 59.0]	40.2 [35.5, 45.0]	40.6 [36.5, 44.7]	26.5 [22.2, 30.8]
Claude-Opus-4.6	53.0 [48.9, 57.1]	35.5 [30.8, 40.2]	38.8 [35.1, 42.6]	20.0 [16.2, 24.0]
Gemini-3-Flash	52.8 [49.1, 56.4]	28.0 [23.8, 32.5]	40.7 [37.2, 44.0]	15.8 [12.2, 19.3]
GPT-5.1	52.3 [48.6, 56.2]	29.5 [25.2, 34.0]	40.2 [36.5, 43.9]	20.0 [16.2, 24.0]
Doubao-Seed-2.0-Pro	51.3 [47.3, 55.3]	33.2 [28.7, 37.8]	42.8 [38.8, 46.8]	23.5 [19.5, 27.5]
Claude-Opus-4.5	51.2 [47.1, 55.3]	34.2 [29.8, 38.8]	39.3 [35.5, 43.2]	20.2 [16.5, 24.2]
GPT-5	51.0 [47.1, 55.0]	32.2 [27.8, 37.0]	40.0 [36.2, 43.9]	21.8 [17.8, 25.8]
GPT-5.1-Codex	50.2 [46.2, 54.3]	31.8 [27.3, 36.5]	39.2 [35.3, 43.2]	22.0 [18.0, 26.0]
Qwen3.5-397B-A17B	50.1 [46.2, 53.9]	29.8 [25.2, 34.2]	38.9 [35.3, 42.5]	17.2 [13.8, 21.0]
Kimi-K2.5	49.4 [45.6, 53.2]	26.8 [22.5, 31.2]	36.8 [33.4, 40.3]	15.0 [11.5, 18.5]
GPT-4.1	49.3 [45.5, 53.2]	29.5 [25.2, 34.0]	39.8 [36.1, 43.6]	21.2 [17.2, 25.2]
o4-mini	48.2 [44.3, 51.9]	25.0 [20.8, 29.2]	39.8 [36.0, 43.7]	21.5 [17.5, 25.5]
Claude-Sonnet-4.5	47.2 [43.7, 51.0]	24.2 [20.0, 28.5]	35.7 [32.1, 39.2]	14.5 [11.0, 18.0]
GPT-5.2	46.3 [42.5, 50.1]	24.0 [20.0, 28.2]	37.4 [33.8, 41.0]	15.5 [12.0, 19.2]
Qwen3-VL-235B-A22B	44.9 [41.3, 48.6]	20.5 [16.7, 24.5]	35.2 [31.6, 38.8]	14.0 [10.8, 17.5]
Grok-4.1-Fast	44.6 [41.2, 47.9]	15.5 [12.0, 19.2]	35.2 [31.8, 38.7]	12.5 [9.2, 15.8]
Claude-Haiku-4.5	43.3 [39.8, 46.8]	18.8 [15.0, 22.8]	32.8 [29.5, 36.2]	12.5 [9.2, 15.8]
GLM-4.6V	43.0 [39.6, 46.5]	17.5 [14.0, 21.2]	34.1 [30.8, 37.4]	11.5 [8.5, 14.8]

Table 18: Photography source-split results under pass^3. Each entry reports Agent-Edit / Expert-Edit / Diff, where Diff is Agent-Edit minus Expert-Edit in percentage points. A dash indicates an unavailable Bot-1 result.

Model	Top-1	Bot-1	TB-1
Claude-Haiku-4.5	18.8/24.3/−5.4	39.1/32.9/+6.3	13.0/7.1/+5.9
Claude-Sonnet-4.5	34.8/32.9/+1.9	34.8/42.9/−8.1	7.2/20.0/−12.8
Claude-Sonnet-4.6	50.7/51.4/−0.7	43.5/55.7/−12.2	17.4/28.6/−11.2
Claude-Opus-4.5	46.4/47.1/−0.8	34.8/47.1/−12.4	18.8/21.4/−2.6
Claude-Opus-4.6	50.7/45.7/+5.0	34.8/44.3/−9.5	20.3/15.7/+4.6
Doubao-Seed-2.0-Pro	40.6/32.9/+7.7	62.3/64.3/−2.0	23.2/27.1/−4.0
Gemini-3-Flash	31.9/44.3/−12.4	53.6/51.4/+2.2	11.6/20.0/−8.4
Gemini-3-Pro	40.6/44.3/−3.7	43.5/47.1/−3.7	20.3/17.1/+3.1
Gemini-3.1-Pro	39.1/44.3/−5.2	–	27.5/21.4/+6.1
GPT-4.1	31.9/38.6/−6.7	66.7/64.3/+2.4	21.7/20.0/+1.7
GPT-5	43.5/44.3/−0.8	63.8/65.7/−1.9	31.9/21.4/+10.5
GPT-5.1	33.3/30.0/+3.3	56.5/61.4/−4.9	14.5/15.7/−1.2
GPT-5.1-Codex	40.6/41.4/−0.8	68.1/58.6/+9.5	21.7/20.0/+1.7
GPT-5.2	39.1/38.6/+0.6	49.3/65.7/−16.4	21.7/21.4/+0.3
Grok-4.1-Fast	15.9/10.0/+5.9	24.6/32.9/−8.2	7.2/8.6/−1.3
o4-mini	27.5/38.6/−11.0	43.5/58.6/−15.1	20.3/40.0/−19.7
GLM-4.6V	13.0/18.6/−5.5	37.7/34.3/+3.4	5.8/5.7/+0.1
Kimi-K2.5	37.7/40.0/−2.3	30.4/37.1/−6.7	11.6/12.9/−1.3
Qwen3-VL-235B-A22B	20.3/28.6/−8.3	29.0/35.7/−6.7	7.2/17.1/−9.9
Qwen3.5-397B-A17B	42.0/45.7/−3.7	–	18.8/28.6/−9.7

Table 19: Photography source-split results under ap@1. Each entry reports Agent-Edit / Expert-Edit / Diff, where Diff is Agent-Edit minus Expert-Edit in percentage points. A dash indicates an unavailable Bot-1 result.

Model	Top-1	Bot-1	TB-1
CLAUDE-HAIKU-4.5	46.4/52.4/−6.0	58.5/58.1/+0.4	35.3/33.3/+1.9
CLAUDE-SONNET-4.5	55.6/59.5/−4.0	60.4/69.0/−8.7	35.7/46.7/−10.9
CLAUDE-SONNET-4.6	64.7/72.9/−8.1	60.9/71.0/−10.1	35.7/51.9/−16.2
CLAUDE-OPUS-4.5	62.8/69.0/−6.2	62.8/73.3/−10.5	43.5/52.9/−9.4
CLAUDE-OPUS-4.6	65.2/66.2/−1.0	61.8/69.5/−7.7	42.5/44.8/−2.2
DOUBAO-SEED-2.0-PRO	56.0/53.3/+2.7	72.0/76.2/−4.2	46.9/48.1/−1.2
GEMINI-3-FLASH	55.1/62.4/−7.3	70.0/69.5/+0.5	39.6/48.6/−9.0
GEMINI-3-PRO	65.2/65.2/−0.0	64.3/62.4/+1.9	41.1/40.0/+1.1
GEMINI-3.1-PRO	60.4/63.8/−3.4	–	48.3/46.7/+1.6
GPT-4.1	55.6/57.1/−1.6	78.7/75.7/+3.0	45.9/43.8/+2.1
GPT-5	62.3/62.9/−0.5	77.3/80.0/−2.7	51.2/51.0/+0.3
GPT-5.1	59.9/60.0/−0.1	75.4/76.2/−0.8	44.4/45.2/−0.8
GPT-5.1-CODEX	58.9/61.0/−2.0	76.3/73.3/+3.0	46.4/47.1/−0.8
GPT-5.2	63.8/60.0/+3.8	70.0/78.6/−8.5	50.2/52.4/−2.1
GROK-4.1-FAST	51.7/49.5/+2.2	52.2/59.0/−6.9	37.2/37.6/−0.4
O4-MINI	56.0/62.9/−6.8	66.7/74.8/−8.1	41.5/62.9/−21.3
GLM-4.6V	42.0/44.3/−2.3	63.3/62.9/+0.4	31.9/34.3/−2.4
KIMI-K2.5	59.4/62.4/−3.0	58.5/63.8/−5.4	40.1/41.0/−0.9
QWEN3-VL-235B-A22B	51.7/53.3/−1.6	56.0/62.9/−6.8	32.9/47.1/−14.3
QWEN3.5-397B-A17B	63.3/62.4/+0.9	–	48.8/51.4/−2.6

Table 20: Comprehensive reward-model results on VAB. Each image is scored independently three times and the scores are averaged before within-task ranking. We report `Top-1`, `Bot-1`, and `TB-1` accuracy across overall and domain-level evaluations (%). Best result in each non-baseline column is shown in **bold**.

	Overall			Fine Art			Illustration			Photography		
Model	Top-1	Bot-1	TB-1	Top-1	Bot-1	TB-1	Top-1	Bot-1	TB-1	Top-1	Bot-1	TB-1
Random Guess	37.2	37.2	27.5	41.0	41.0	36.0	31.5	31.5	19.2	36.8	36.8	23.7
Human Expert	77.7	80.1	68.9	80.4	82.7	74.7	70.3	70.4	54.4	79.9	84.0	72.4
ArtiMuse	44.2	51.2	34.8	50.9	**54.7**	**46.0**	**37.0**	33.0	20.0	41.7	60.4	32.4
PEAS-aes	48.5	49.5	37.8	49.7	51.6	44.1	34.0	**36.0**	**22.0**	57.6	56.8	41.7
PEAS-ava	49.0	48.2	37.0	50.9	49.1	44.1	25.0	31.0	13.0	64.0	59.7	46.0
Q-Align	**52.2**	47.8	**38.2**	**52.2**	49.7	44.7	31.0	28.0	16.0	**67.6**	59.7	**46.8**
Q-Instruct	38.5	**51.5**	31.2	37.9	39.1	34.2	32.0	**36.0**	21.0	43.9	**77.0**	35.2
RealQA-ava	44.2	42.2	30.2	40.4	43.5	36.6	33.0	27.0	15.0	56.8	51.8	33.8

Table 21: Unweighted macro-topic averages across all 24 topics. For each metric, Best MLLM reports the highest macro-average among the 20 evaluated models; Gap is Human Expert minus Best MLLM in percentage points.

Metric	Random	Human Expert	Best MLLM	Gap
`Top-1 ap@1`	39.7	80.7	62.1 (Gemini-3-Pro)	18.6
`Top-1 pass^3`	7.6	80.7	48.2 (Claude-Sonnet-4.6)	32.5
`TB-1 ap@1`	31.3	72.5	51.4 (Gemini-3.1-Pro)	21.1
`TB-1 pass^3`	6.4	72.5	35.1 (Claude-Sonnet-4.6)	37.4

`Top-1 ap@1`.

Tables 22, 23, and 24 report topic-level `Top-1 ap@1` results for fine art, illustration, and photography, respectively.

`Top-1 pass^3`.

Tables 25, 26, and 27 report topic-level `Top-1 pass^3` results for fine art, illustration, and photography, respectively.

`TB-1 ap@1`.

Tables 28, 29, and 30 report topic-level `TB-1 ap@1` results for fine art, illustration, and photography, respectively.

`TB-1 pass^3`.

Tables 31, 32, and 33 report topic-level `TB-1 pass^3` results for fine art, illustration, and photography, respectively.

E.7 Topic-Level Reward-Model Results

To match the main-text reward-model table, we report topic-level results for both `Top-1` and `TB-1`. As in the MLLM case, the domain-level averages in the main text mask substantial variation across topics. The topic-level reward-model tables show that relatively stronger domain-level results in photography are driven by a subset of topics, while illustration remains difficult across most topics for all six reward models.

The topic breakdown also clarifies where reward models do and do not transfer. Fine-art performance is mixed but not uniformly weak, with several models reaching competitive results on Chinese painting and portrait-color tasks. Illustration remains systematically difficult: outside anime-and-manga and, in a few cases, comic, most models stay close to

Table 22: Topic-level MLLM Top-1 ap@1 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Random Guess	50.0	50.0	50.0	25.2	50.0	50.0	50.0	47.7	50.0	41.0
Human Expert	89.1	94.4	90.0	66.6	85.0	90.3	88.8	81.7	86.5	80.4
Closed-Source Models										
CLAUDE-HAIKU-4.5	33.3	81.5	66.7	29.8	50.0	53.5	54.2	52.8	42.9	44.7
CLAUDE-SONNET-4.5	42.4	85.2	66.7	27.4	50.0	54.5	62.5	48.6	40.5	44.5
CLAUDE-SONNET-4.6	54.5	74.1	50.0	39.9	83.3	48.5	62.5	48.6	47.6	48.4
CLAUDE-OPUS-4.5	42.4	81.5	50.0	32.7	83.3	46.5	58.3	51.4	47.6	45.3
CLAUDE-OPUS-4.6	48.5	81.5	50.0	37.5	66.7	50.5	62.5	47.2	59.5	48.7
DOUBAO-SEED-2.0-PRO	42.4	96.3	58.3	44.6	66.7	64.6	70.8	61.1	45.2	55.9
GEMINI-3-FLASH	36.4	85.2	66.7	42.9	66.7	59.6	66.7	48.6	50.0	51.8
GEMINI-3-PRO	27.3	88.9	58.3	41.7	100.0	58.6	66.7	58.3	45.2	52.0
GEMINI-3.1-PRO	24.2	81.5	41.7	44.0	100.0	60.6	58.3	50.0	59.5	51.8
GPT-4.1	33.3	88.9	41.7	32.7	100.0	44.4	66.7	52.8	69.0	47.2
GPT-5	27.3	96.3	58.3	37.5	100.0	44.4	70.8	59.7	61.9	49.9
GPT-5.1	33.3	92.6	50.0	35.7	83.3	49.5	70.8	54.2	69.0	49.9
GPT-5.1-CODEX	30.3	92.6	50.0	35.1	100.0	39.4	62.5	56.9	66.7	47.4
GPT-5.2	21.2	70.4	50.0	31.5	50.0	40.4	54.2	47.2	52.4	40.8
GROK-4.1-FAST	36.4	55.6	58.3	28.0	33.3	57.6	54.2	59.7	45.2	44.5
O4-MINI	18.2	81.5	58.3	26.2	83.3	43.4	62.5	72.2	52.4	44.7
Open-Weight Models										
GLM-4.6V	21.2	63.0	33.3	26.8	66.7	52.5	54.2	58.3	52.4	42.7
KIMI-K2.5	30.3	85.2	50.0	27.4	100.0	46.5	62.5	50.0	52.4	43.5
QWEN3-VL-235B-A22B	15.2	85.2	50.0	26.8	66.7	48.5	66.7	47.2	57.1	42.4
QWEN3.5-397B-A17B	24.2	96.3	66.7	33.9	100.0	37.4	58.3	43.1	45.2	42.7

Table 23: Topic-level MLLM Top-1 ap@1 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Random Guess	50.0	47.5	25.0	29.3	25.0	25.0	31.5
Human Expert	87.7	82.0	66.8	65.2	62.4	71.8	70.3
Closed-Source Models							
CLAUDE-HAIKU-4.5	56.4	26.7	40.9	20.3	25.4	33.3	32.7
CLAUDE-SONNET-4.5	61.5	26.7	33.3	33.3	41.3	27.3	37.3
CLAUDE-SONNET-4.6	71.8	63.3	28.8	37.7	46.0	45.5	45.3
CLAUDE-OPUS-4.5	64.1	46.7	27.3	31.9	42.9	42.4	40.0
CLAUDE-OPUS-4.6	61.5	50.0	31.8	36.2	46.0	39.4	42.3
DOUBAO-SEED-2.0-PRO	53.8	50.0	24.2	34.8	47.6	36.4	39.3
GEMINI-3-FLASH	59.0	56.7	34.8	44.9	52.4	33.3	46.0
GEMINI-3-PRO	69.2	60.0	36.4	43.5	41.3	51.5	47.3
GEMINI-3.1-PRO	71.8	60.0	37.9	42.0	55.6	42.4	49.7
GPT-4.1	79.5	36.7	28.8	42.0	41.3	39.4	43.0
GPT-5	61.5	30.0	37.9	26.1	33.3	39.4	36.7
GPT-5.1	69.2	40.0	43.9	39.1	46.0	39.4	45.7
GPT-5.1-CODEX	59.0	26.7	37.9	42.0	39.7	42.4	41.3
GPT-5.2	48.7	30.0	30.3	24.6	38.1	36.4	33.7
GROK-4.1-FAST	61.5	36.7	34.8	39.1	28.6	18.2	36.3
O4-MINI	56.4	33.3	36.4	30.4	42.9	30.3	38.0
Open-Weight Models							
GLM-4.6V	51.3	53.3	31.8	37.7	60.3	27.3	43.3
KIMI-K2.5	71.8	56.7	36.4	36.2	39.7	30.3	43.0
QWEN3-VL-235B-A22B	69.2	33.3	30.3	30.4	39.7	36.4	38.3
QWEN3.5-397B-A17B	71.8	53.3	33.3	27.5	50.8	48.5	44.3

random on concept art, digital-and-AI art, pixel art, and stylized 3D, especially under TB-1. Photography remains the strongest domain, but the gains are concentrated in a subset of topics and do not transfer uniformly to the harder best-and-worst selection setting.

Top-1.

Tables 34, 35, and 36 report topic-level Top-1 results for fine art, illustration, and photography, respectively.

TB-1.

Tables 37, 38, and 41 report topic-level TB-1 results for fine art, illustration, and photography, respectively.

Table 24: Topic-level MLLM Top-1 ap@1 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Random Guess	37.0	36.7	35.4	33.3	33.3	41.7	33.3	38.0	39.8	36.8
Human Expert	81.6	78.0	78.9	75.0	80.0	85.0	90.0	80.3	78.9	79.9
Closed-Source Models										
Claude-Haiku-4.5	40.6	60.0	49.6	83.3	0.0	66.7	100.0	47.9	66.7	49.4
Claude-Sonnet-4.5	43.8	40.0	61.5	83.3	66.7	33.3	100.0	58.1	81.5	57.6
Claude-Sonnet-4.6	58.3	46.7	75.6	91.7	50.0	100.0	100.0	70.1	63.0	68.8
Claude-Opus-4.5	58.3	46.7	69.6	83.3	50.0	83.3	100.0	66.7	70.4	65.9
Claude-Opus-4.6	59.4	66.7	65.9	83.3	50.0	100.0	100.0	65.0	74.1	65.7
Doubao-Seed-2.0-Pro	39.6	73.3	59.3	83.3	50.0	33.3	100.0	49.6	85.2	54.7
Gemini-3-Flash	52.1	86.7	60.0	83.3	33.3	0.0	100.0	58.1	66.7	58.8
Gemini-3-Pro	58.3	73.3	65.9	83.3	66.7	66.7	100.0	66.7	63.0	65.2
Gemini-3.1-Pro	47.9	80.0	68.1	100.0	66.7	0.0	100.0	62.4	63.0	62.1
GPT-4.1	36.5	46.7	60.7	100.0	33.3	50.0	100.0	64.1	59.3	56.4
GPT-5	49.0	66.7	65.2	100.0	50.0	33.3	100.0	67.5	63.0	62.6
GPT-5.1	47.9	60.0	62.2	75.0	50.0	33.3	100.0	62.4	77.8	60.0
GPT-5.1-Codex	44.8	66.7	64.4	100.0	50.0	66.7	100.0	62.4	55.6	60.0
GPT-5.2	49.0	53.3	68.9	83.3	50.0	50.0	100.0	62.4	66.7	61.9
Grok-4.1-Fast	45.8	60.0	50.4	58.3	50.0	50.0	100.0	47.9	66.7	50.6
o4-mini	52.1	60.0	63.7	91.7	50.0	0.0	100.0	57.3	70.4	59.5
Open-Weight Models										
GLM-4.6V	35.4	53.3	39.3	58.3	16.7	0.0	100.0	48.7	63.0	43.2
Kimi-K2.5	49.0	80.0	63.7	75.0	50.0	50.0	100.0	61.5	70.4	60.9
Qwen3-VL-235B-A22B	37.5	66.7	54.1	58.3	50.0	50.0	66.7	57.3	66.7	52.5
Qwen3.5-397B-A17B	44.8	66.7	64.4	83.3	50.0	50.0	100.0	72.6	66.7	62.8

Table 25: Topic-level MLLM Top-1 pass^3 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Random Guess	12.5	12.5	12.5	2.4	12.5	12.5	12.5	11.6	12.5	8.9
Human Expert	89.1	94.4	90.0	66.6	85.0	90.3	88.8	81.7	86.5	80.4
Closed-Source Models										
Claude-Haiku-4.5	9.1	66.7	50.0	5.4	50.0	24.2	50.0	16.7	21.4	19.9
Claude-Sonnet-4.5	9.1	77.8	25.0	7.1	0.0	33.3	37.5	20.8	7.1	20.5
Claude-Sonnet-4.6	36.4	66.7	50.0	28.6	50.0	36.4	62.5	37.5	35.7	37.3
Claude-Opus-4.5	36.4	77.8	50.0	16.1	50.0	21.2	37.5	37.5	21.4	28.0
Claude-Opus-4.6	36.4	66.7	50.0	19.6	50.0	21.2	37.5	25.0	28.6	27.3
Doubao-Seed-2.0-Pro	18.2	88.9	50.0	30.4	50.0	42.4	50.0	45.8	35.7	39.8
Gemini-3-Flash	0.0	66.7	50.0	17.9	50.0	24.2	37.5	20.8	14.3	23.0
Gemini-3-Pro	9.1	77.8	50.0	17.9	100.0	36.4	50.0	45.8	28.6	32.9
Gemini-3.1-Pro	0.0	77.8	25.0	23.2	100.0	36.4	37.5	33.3	28.6	31.1
GPT-4.1	9.1	77.8	25.0	12.5	100.0	30.3	50.0	25.0	50.0	28.0
GPT-5	0.0	88.9	50.0	19.6	100.0	21.2	50.0	37.5	42.9	30.4
GPT-5.1	9.1	88.9	50.0	12.5	50.0	30.3	62.5	33.3	50.0	30.4
GPT-5.1-Codex	0.0	88.9	25.0	12.5	100.0	27.3	62.5	37.5	50.0	29.8
GPT-5.2	0.0	55.6	25.0	12.5	50.0	15.2	37.5	12.5	28.6	18.0
Grok-4.1-Fast	9.1	22.2	50.0	5.4	0.0	30.3	25.0	33.3	7.1	18.0
o4-mini	0.0	66.7	50.0	3.6	50.0	21.2	50.0	41.7	42.9	23.6
Open-Weight Models										
GLM-4.6V	9.1	55.6	0.0	5.4	50.0	21.2	25.0	33.3	14.3	18.0
Kimi-K2.5	9.1	77.8	25.0	5.4	100.0	18.2	50.0	16.7	28.6	19.9
Qwen3-VL-235B-A22B	9.1	77.8	50.0	8.9	50.0	24.2	50.0	20.8	21.4	22.4
Qwen3.5-397B-A17B	18.2	88.9	50.0	16.1	100.0	6.1	37.5	16.7	14.3	21.1

E.8 Temperature Ablation

Table 39 reports the full model-by-model temperature ablation referenced in Section 4.3. The overall trend is that greedy decoding modestly improves stability, but the effect size varies substantially across models and does not alter the main conclusion that the dominant limitation is comparative judgment rather than decoding noise alone.

The gains are not uniform. Gemini 3.1 Pro and Claude Sonnet 4.5 benefit the most from greedy decoding, with especially visible improvements on the stricter TB-1 pass^3 metric. A small set of models, including Claude Sonnet 4.6, Doubao Seed 2.0 Pro, GLM-4.6V, and Qwen3.5-397B-A17B, are flat or slightly worse on at least one metric, reinforcing the point that decoding choice matters but does not dominate the benchmark outcome.

Table 26: Topic-level MLLM Top-1 pass^3 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Random Guess	12.5	11.4	1.6	3.5	1.6	1.6	4.4
Human Expert	87.7	82.0	66.8	65.2	62.4	71.8	70.3
Closed-Source Models							
Claude-Haiku-4.5	30.8	0.0	18.2	4.3	9.5	18.2	13.0
Claude-Sonnet-4.5	30.8	0.0	9.1	17.4	23.8	18.2	17.0
Claude-Sonnet-4.6	61.5	60.0	4.5	26.1	28.6	27.3	30.0
Claude-Opus-4.5	53.8	30.0	4.5	21.7	33.3	36.4	27.0
Claude-Opus-4.6	53.8	30.0	9.1	34.8	38.1	27.3	31.0
Doubao-Seed-2.0-Pro	7.7	10.0	4.5	26.1	28.6	27.3	18.0
Gemini-3-Flash	23.1	20.0	9.1	21.7	38.1	18.2	22.0
Gemini-3-Pro	46.2	40.0	13.6	26.1	28.6	27.3	28.0
Gemini-3.1-Pro	53.8	30.0	22.7	21.7	38.1	36.4	32.0
GPT-4.1	61.5	20.0	13.6	21.7	19.0	18.2	24.0
GPT-5	46.2	10.0	13.6	17.4	14.3	18.2	19.0
GPT-5.1	46.2	20.0	18.2	21.7	23.8	27.3	25.0
GPT-5.1-Codex	38.5	10.0	9.1	30.4	19.0	27.3	22.0
GPT-5.2	30.8	10.0	4.5	8.7	14.3	18.2	13.0
Grok-4.1-Fast	38.5	20.0	9.1	17.4	9.5	0.0	15.0
o4-mini	23.1	10.0	13.6	13.0	23.8	9.1	16.0
Open-Weight Models							
GLM-4.6V	23.1	20.0	9.1	17.4	38.1	0.0	19.0
Kimi-K2.5	53.8	30.0	13.6	21.7	9.5	9.1	21.0
Qwen3-VL-235B-A22B	46.2	0.0	4.5	4.3	9.5	18.2	12.0
Qwen3.5-397B-A17B	53.8	10.0	13.6	4.3	33.3	45.5	24.0

Table 27: Topic-level MLLM Top-1 pass^3 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Random Guess	5.6	5.5	4.8	3.7	3.7	8.1	3.7	6.2	7.4	5.6
Human Expert	81.6	78.0	78.9	75.0	80.0	85.0	90.0	80.3	78.9	79.9
Closed-Source Models										
Claude-Haiku-4.5	15.6	20.0	24.4	75.0	0.0	0.0	100.0	15.4	33.3	21.6
Claude-Sonnet-4.5	15.6	0.0	37.8	75.0	50.0	0.0	100.0	35.9	66.7	33.8
Claude-Sonnet-4.6	28.1	20.0	60.0	75.0	50.0	100.0	100.0	56.4	55.6	51.1
Claude-Opus-4.5	31.2	0.0	53.3	75.0	50.0	50.0	100.0	48.7	66.7	46.8
Claude-Opus-4.6	25.0	60.0	51.1	75.0	50.0	100.0	100.0	53.8	55.6	48.2
Doubao-Seed-2.0-Pro	18.8	40.0	44.4	75.0	50.0	0.0	100.0	30.8	66.7	36.7
Gemini-3-Flash	28.1	60.0	42.2	75.0	0.0	0.0	100.0	33.3	55.6	38.1
Gemini-3-Pro	31.2	40.0	44.4	75.0	50.0	50.0	100.0	43.6	44.4	42.4
Gemini-3.1-Pro	25.0	60.0	46.7	100.0	50.0	0.0	100.0	41.0	44.4	41.7
GPT-4.1	15.6	20.0	40.0	100.0	0.0	50.0	100.0	41.0	33.3	35.3
GPT-5	28.1	40.0	48.9	100.0	50.0	0.0	100.0	46.2	44.4	43.9
GPT-5.1	15.6	20.0	35.6	50.0	50.0	0.0	100.0	35.9	44.4	31.7
GPT-5.1-Codex	25.0	40.0	48.9	100.0	50.0	50.0	100.0	35.9	44.4	41.0
GPT-5.2	25.0	20.0	48.9	75.0	50.0	0.0	100.0	35.9	44.4	38.8
Grok-4.1-Fast	6.2	20.0	8.9	25.0	50.0	0.0	100.0	12.8	33.3	12.9
o4-mini	21.9	40.0	33.3	75.0	50.0	0.0	100.0	33.3	44.4	33.1
Open-Weight Models										
GLM-4.6V	6.2	20.0	15.6	50.0	0.0	0.0	100.0	15.4	33.3	15.8
Kimi-K2.5	18.8	60.0	44.4	75.0	50.0	50.0	100.0	35.9	55.6	38.8
Qwen3-VL-235B-A22B	6.2	60.0	26.7	50.0	50.0	0.0	0.0	28.2	33.3	24.5
Qwen3.5-397B-A17B	31.2	40.0	42.2	75.0	50.0	0.0	100.0	51.3	55.6	43.9

Table 28: Topic-level MLLM TB-1 ap@1 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Random Guess	50.0	50.0	50.0	11.3	50.0	50.0	50.0	46.4	50.0	36.0
Human Expert	89.1	94.4	90.0	50.5	85.0	90.3	88.8	80.8	86.5	74.7
Closed-Source Models										
CLAUDE-HAIKU-4.5	33.3	81.5	66.7	17.9	50.0	53.5	54.2	51.4	42.9	40.4
CLAUDE-SONNET-4.5	42.4	85.2	66.7	15.5	50.0	54.5	62.5	47.2	40.5	40.2
CLAUDE-SONNET-4.6	54.5	74.1	50.0	28.6	83.3	48.5	62.5	50.0	47.6	44.7
CLAUDE-OPUS-4.5	42.4	81.5	50.0	19.0	83.3	46.5	58.3	50.0	47.6	40.4
CLAUDE-OPUS-4.6	48.5	81.5	50.0	22.0	66.7	50.5	62.5	47.2	59.5	43.3
DOUBAO-SEED-2.0-PRO	42.4	96.3	58.3	27.4	66.7	64.6	70.8	61.1	45.2	49.9
GEMINI-3-FLASH	36.4	85.2	66.7	28.6	66.7	59.6	66.7	48.6	50.0	46.8
GEMINI-3-PRO	27.3	88.9	58.3	26.2	100.0	58.6	66.7	58.3	45.2	46.6
GEMINI-3.1-PRO	24.2	81.5	41.7	25.0	100.0	60.6	58.3	50.0	59.5	45.1
GPT-4.1	33.3	88.9	41.7	22.0	100.0	44.4	66.7	50.0	69.0	43.1
GPT-5	27.3	96.3	58.3	17.9	100.0	44.4	70.8	56.9	61.9	42.7
GPT-5.1	33.3	92.6	50.0	21.4	83.3	49.5	70.8	51.4	69.0	44.5
GPT-5.1-CODEX	30.3	92.6	50.0	19.6	100.0	39.4	62.5	56.9	66.7	42.0
GPT-5.2	21.2	70.4	50.0	17.3	50.0	40.4	54.2	44.4	52.4	35.4
GROK-4.1-FAST	36.4	55.6	58.3	13.1	33.3	57.6	54.2	61.1	45.2	39.5
O4-MINI	18.2	81.5	58.3	16.7	83.3	43.4	62.5	69.4	52.4	41.0
Open-Weight Models										
GLM-4.6V	21.2	63.0	33.3	17.3	66.7	52.5	54.2	58.3	52.4	39.3
KIMI-K2.5	30.3	85.2	50.0	13.1	100.0	46.5	62.5	50.0	52.4	38.5
QWEN3-VL-235B-A22B	15.2	85.2	50.0	10.7	66.7	48.5	66.7	47.2	57.1	36.9
QWEN3.5-397B-A17B	24.2	96.3	66.7	18.5	100.0	37.4	58.3	43.1	45.2	37.3

Table 29: Topic-level MLLM TB-1 ap@1 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Random Guess	50.0	45.8	8.3	15.6	8.3	8.3	19.2
Human Expert	87.7	79.0	42.3	49.1	43.3	49.1	54.4
Closed-Source Models							
CLAUDE-HAIKU-4.5	56.4	26.7	13.6	15.9	9.5	0.0	18.7
CLAUDE-SONNET-4.5	61.5	23.3	10.6	17.4	12.7	12.1	20.7
CLAUDE-SONNET-4.6	71.8	66.7	16.7	26.1	15.9	3.0	29.3
CLAUDE-OPUS-4.5	64.1	43.3	13.6	20.3	17.5	12.1	25.3
CLAUDE-OPUS-4.6	61.5	43.3	12.1	24.6	19.0	3.0	25.0
DOUBAO-SEED-2.0-PRO	53.8	50.0	16.7	21.7	17.5	6.1	25.0
GEMINI-3-FLASH	59.0	53.3	21.2	18.8	17.5	3.0	26.0
GEMINI-3-PRO	69.2	60.0	12.1	21.7	20.6	15.2	28.7
GEMINI-3.1-PRO	71.8	60.0	7.6	21.7	25.4	9.1	28.3
GPT-4.1	79.5	33.3	9.1	30.4	12.7	21.2	27.7
GPT-5	61.5	30.0	12.1	20.3	7.9	3.0	20.3
GPT-5.1	69.2	36.7	18.2	26.1	14.3	9.1	26.7
GPT-5.1-CODEX	59.0	33.3	9.1	36.2	12.7	3.0	24.3
GPT-5.2	48.7	30.0	15.2	14.5	17.5	15.2	21.3
GROK-4.1-FAST	61.5	40.0	19.7	18.8	17.5	6.1	25.0
O4-MINI	56.4	33.3	13.6	21.7	7.9	3.0	20.7
Open-Weight Models							
GLM-4.6V	51.3	50.0	13.6	20.3	28.6	15.2	27.0
KIMI-K2.5	71.8	53.3	18.2	20.3	17.5	18.2	29.0
QWEN3-VL-235B-A22B	69.2	33.3	22.7	14.5	22.2	3.0	25.7
QWEN3.5-397B-A17B	71.8	46.7	18.2	15.9	20.6	0.0	26.0

Table 30: Topic-level MLLM TB-1 ap@1 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Random Guess	24.0	23.3	20.9	16.7	16.7	33.3	16.7	26.1	30.6	23.7
Human Expert	73.8	72.0	71.3	60.0	60.0	80.0	70.0	74.4	72.2	72.4
Closed-Source Models										
CLAUDE-HAIKU-4.5	21.9	60.0	32.6	41.7	0.0	66.7	33.3	35.0	66.7	34.3
CLAUDE-SONNET-4.5	33.3	33.3	40.7	41.7	33.3	16.7	33.3	47.9	55.6	41.2
CLAUDE-SONNET-4.6	34.4	26.7	41.5	58.3	50.0	100.0	0.0	50.4	55.6	43.9
CLAUDE-OPUS-4.5	42.7	46.7	44.4	66.7	50.0	83.3	66.7	50.4	59.3	48.2
CLAUDE-OPUS-4.6	37.5	40.0	40.0	58.3	50.0	66.7	66.7	47.9	51.9	43.6
DOUBAO-SEED-2.0-PRO	36.5	53.3	43.7	66.7	50.0	33.3	66.7	49.6	85.2	47.5
GEMINI-3-FLASH	36.5	80.0	36.3	66.7	50.0	33.3	66.7	47.9	63.0	44.1
GEMINI-3-PRO	36.5	73.3	27.4	58.3	66.7	66.7	33.3	44.4	66.7	40.5
GEMINI-3.1-PRO	32.3	66.7	45.2	91.7	50.0	33.3	100.0	50.4	66.7	47.5
GPT-4.1	30.2	20.0	40.7	100.0	50.0	50.0	33.3	57.3	51.9	44.8
GPT-5	40.6	73.3	45.2	91.7	50.0	66.7	66.7	57.3	55.6	51.1
GPT-5.1	33.3	73.3	37.8	75.0	66.7	33.3	100.0	50.4	59.3	44.8
GPT-5.1-CODEX	30.2	66.7	43.7	100.0	33.3	66.7	33.3	53.0	59.3	46.8
GPT-5.2	39.6	26.7	51.1	83.3	33.3	50.0	66.7	56.4	74.1	51.3
GROK-4.1-FAST	27.1	46.7	38.5	41.7	33.3	16.7	0.0	39.3	63.0	37.4
O4-MINI	43.8	53.3	52.6	91.7	50.0	16.7	100.0	52.1	66.7	52.3
Open-Weight Models										
GLM-4.6V	28.1	40.0	28.1	33.3	33.3	0.0	66.7	39.3	48.1	33.1
KIMI-K2.5	29.2	80.0	39.3	33.3	16.7	0.0	66.7	45.3	59.3	40.5
QWEN3-VL-235B-A22B	28.1	53.3	36.3	91.7	33.3	66.7	66.7	42.7	51.9	40.0
QWEN3.5-397B-A17B	31.2	66.7	49.6	83.3	50.0	66.7	0.0	57.3	66.7	50.1

Table 31: Topic-level TB-1 pass^3 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Random Guess	12.5	12.5	12.5	1.4	12.5	12.5	12.5	11.5	12.5	8.5
Human Expert	89.1	94.4	90.0	50.5	85.0	90.3	88.8	80.8	86.5	74.7
Closed-Source Models										
CLAUDE-HAIKU-4.5	9.1	66.7	50.0	3.6	50.0	24.2	50.0	16.7	21.4	19.3
CLAUDE-SONNET-4.5	9.1	77.8	25.0	3.6	0.0	33.3	37.5	20.8	7.1	19.3
CLAUDE-SONNET-4.6	36.4	66.7	50.0	19.6	50.0	36.4	62.5	37.5	35.7	34.2
CLAUDE-OPUS-4.5	36.4	77.8	50.0	5.4	50.0	21.2	37.5	37.5	21.4	24.2
CLAUDE-OPUS-4.6	36.4	66.7	50.0	8.9	50.0	21.2	37.5	25.0	28.6	23.6
DOUBAO-SEED-2.0-PRO	18.2	88.9	50.0	10.7	50.0	42.4	50.0	45.8	35.7	32.9
GEMINI-3-FLASH	0.0	66.7	50.0	10.7	50.0	24.2	37.5	20.8	14.3	20.5
GEMINI-3-PRO	9.1	77.8	50.0	8.9	100.0	36.4	50.0	45.8	28.6	29.8
GEMINI-3.1-PRO	0.0	77.8	25.0	8.9	100.0	36.4	37.5	33.3	28.6	26.1
GPT-4.1	9.1	77.8	25.0	5.4	100.0	30.3	50.0	25.0	50.0	25.5
GPT-5	0.0	88.9	50.0	5.4	100.0	21.2	50.0	37.5	42.9	25.5
GPT-5.1	9.1	88.9	50.0	10.7	50.0	30.3	62.5	33.3	50.0	29.8
GPT-5.1-CODEX	0.0	88.9	25.0	5.4	100.0	27.3	62.5	37.5	50.0	27.3
GPT-5.2	0.0	55.6	25.0	3.6	50.0	15.2	37.5	12.5	28.6	14.9
GROK-4.1-FAST	9.1	22.2	50.0	3.6	0.0	30.3	25.0	33.3	7.1	17.4
O4-MINI	0.0	66.7	50.0	1.8	50.0	21.2	50.0	41.7	42.9	23.0
Open-Weight Models										
GLM-4.6V	9.1	55.6	0.0	7.1	50.0	21.2	25.0	33.3	14.3	18.6
KIMI-K2.5	9.1	77.8	25.0	0.0	100.0	18.2	50.0	16.7	28.6	18.0
QWEN3-VL-235B-A22B	9.1	77.8	50.0	0.0	50.0	24.2	50.0	20.8	21.4	19.3
QWEN3.5-397B-A17B	18.2	88.9	50.0	5.4	100.0	6.1	37.5	16.7	14.3	17.4

Table 32: Topic-level TB-1 pass^3 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Random Guess	12.5	11.3	0.1	2.2	0.1	0.1	3.3
Human Expert	87.7	79.0	42.3	49.1	43.3	49.1	54.4
Closed-Source Models							
CLAUDE-HAIKU-4.5	30.8	0.0	0.0	4.3	0.0	0.0	5.0
CLAUDE-SONNET-4.5	30.8	0.0	4.5	8.7	4.8	0.0	8.0
CLAUDE-SONNET-4.6	61.5	60.0	9.1	13.0	0.0	0.0	19.0
CLAUDE-OPUS-4.5	53.8	30.0	4.5	13.0	0.0	0.0	14.0
CLAUDE-OPUS-4.6	53.8	30.0	0.0	21.7	9.5	0.0	17.0
DOUBAO-SEED-2.0-PRO	7.7	10.0	0.0	13.0	4.8	0.0	6.0
GEMINI-3-FLASH	23.1	20.0	0.0	8.7	4.8	0.0	8.0
GEMINI-3-PRO	46.2	40.0	0.0	13.0	4.8	0.0	14.0
GEMINI-3.1-PRO	53.8	30.0	0.0	13.0	0.0	0.0	13.0
GPT-4.1	61.5	20.0	0.0	21.7	0.0	0.0	15.0
GPT-5	46.2	10.0	0.0	8.7	0.0	0.0	9.0
GPT-5.1	46.2	20.0	0.0	8.7	4.8	0.0	11.0
GPT-5.1-CODEX	38.5	20.0	0.0	26.1	9.5	0.0	15.0
GPT-5.2	30.8	10.0	4.5	4.3	4.8	0.0	8.0
GROK-4.1-FAST	38.5	20.0	9.1	8.7	0.0	0.0	11.0
O4-MINI	23.1	10.0	4.5	8.7	0.0	0.0	7.0
Open-Weight Models							
GLM-4.6V	23.1	20.0	4.5	8.7	0.0	0.0	8.0
KIMI-K2.5	53.8	30.0	4.5	8.7	4.8	0.0	14.0
QWEN3-VL-235B-A22B	46.2	0.0	4.5	0.0	4.8	0.0	8.0
QWEN3.5-397B-A17B	53.8	10.0	0.0	0.0	0.0	0.0	8.0

Table 33: Topic-level TB-1 pass^3 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Random Guess	3.1	2.9	2.1	0.5	0.5	6.5	0.5	3.9	5.8	3.1
Human Expert	73.8	72.0	71.3	60.0	60.0	80.0	70.0	74.4	72.2	72.4
Closed-Source Models										
CLAUDE-HAIKU-4.5	3.1	40.0	4.4	25.0	0.0	0.0	0.0	12.8	33.3	10.1
CLAUDE-SONNET-4.5	9.4	0.0	13.3	25.0	0.0	0.0	0.0	15.4	33.3	13.7
CLAUDE-SONNET-4.6	12.5	0.0	17.8	50.0	50.0	100.0	0.0	28.2	44.4	23.0
CLAUDE-OPUS-4.5	18.8	20.0	15.6	25.0	50.0	50.0	0.0	17.9	44.4	20.1
CLAUDE-OPUS-4.6	12.5	0.0	13.3	25.0	50.0	50.0	0.0	23.1	33.3	18.0
DOUBAO-SEED-2.0-PRO	25.0	20.0	17.8	50.0	50.0	0.0	0.0	25.6	55.6	25.2
GEMINI-3-FLASH	9.4	40.0	8.9	25.0	50.0	0.0	0.0	17.9	44.4	15.8
GEMINI-3-PRO	9.4	60.0	8.9	50.0	50.0	50.0	0.0	20.5	44.4	18.7
GEMINI-3.1-PRO	9.4	60.0	17.8	75.0	50.0	0.0	100.0	28.2	44.4	24.5
GPT-4.1	9.4	0.0	8.9	100.0	50.0	50.0	0.0	35.9	22.2	20.9
GPT-5	9.4	40.0	22.2	75.0	50.0	0.0	0.0	35.9	44.4	26.6
GPT-5.1	12.5	20.0	8.9	25.0	50.0	0.0	100.0	12.8	44.4	15.1
GPT-5.1-CODEX	3.1	40.0	13.3	100.0	0.0	50.0	0.0	30.8	33.3	20.9
GPT-5.2	9.4	0.0	17.8	75.0	0.0	0.0	0.0	25.6	66.7	21.6
GROK-4.1-FAST	3.1	0.0	6.7	0.0	0.0	0.0	0.0	10.3	33.3	7.9
O4-MINI	9.4	40.0	37.8	75.0	50.0	0.0	100.0	28.2	44.4	30.2
Open-Weight Models										
GLM-4.6V	0.0	20.0	2.2	0.0	0.0	0.0	0.0	10.3	22.2	5.8
KIMI-K2.5	3.1	60.0	6.7	25.0	0.0	0.0	0.0	15.4	33.3	12.2
QWEN3-VL-235B-A22B	3.1	20.0	11.1	75.0	0.0	0.0	0.0	12.8	22.2	12.2
QWEN3.5-397B-A17B	15.6	20.0	17.8	50.0	50.0	0.0	0.0	30.8	44.4	23.7

Table 34: Topic-level reward-model Top-1 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Random Guess	50.0	50.0	50.0	25.2	50.0	50.0	50.0	47.7	50.0	41.0
Human Expert	89.1	94.4	90.0	66.6	85.0	90.3	88.8	81.7	86.5	80.5
Reward Models										
ARTIMUSE	54.5	88.9	50.0	39.3	50.0	54.5	50.0	50.0	64.3	50.9
PEAS-AES	54.5	44.4	50.0	30.4	100.0	72.7	50.0	62.5	42.9	49.7
PEAS-AVA	36.4	66.7	50.0	41.1	100.0	54.5	62.5	58.3	57.1	50.9
Q-ALIGN	63.6	66.7	50.0	39.3	100.0	48.5	50.0	66.7	64.3	52.2
Q-INSTRUCT	9.1	44.4	75.0	21.4	100.0	57.6	50.0	41.7	42.9	37.9
REALQA-AVA	36.4	66.7	50.0	21.4	100.0	48.5	50.0	45.8	57.1	40.4

Table 35: Topic-level reward-model Top-1 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Random Guess	50.0	47.5	25.0	29.3	25.0	25.0	31.5
Human Expert	87.7	82.0	66.8	65.2	62.4	71.8	70.3
Reward Models							
ARTIMUSE	53.8	60.0	22.7	39.1	33.3	27.3	37.0
PEAS-AES	30.8	20.0	22.7	43.5	47.6	27.3	34.0
PEAS-AVA	30.8	20.0	9.1	30.4	42.9	9.1	25.0
Q-ALIGN	38.5	10.0	36.4	21.7	38.1	36.4	31.0
Q-INSTRUCT	46.2	30.0	18.2	34.8	42.9	18.2	32.0
REALQA-AVA	38.5	50.0	45.5	13.0	38.1	18.2	33.0

Table 36: Topic-level reward-model Top-1 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Random Guess	37.0	36.7	35.4	33.3	33.3	41.7	33.3	38.0	39.8	36.8
Human Expert	81.6	78.0	78.9	75.0	80.0	85.0	90.0	80.3	78.9	79.9
Reward Models										
ARTIMUSE	25.0	60.0	46.7	50.0	100.0	50.0	0.0	38.5	66.7	41.7
PEAS-AES	50.0	80.0	57.8	100.0	50.0	0.0	0.0	61.5	55.6	57.6
PEAS-AVA	56.2	80.0	68.9	100.0	50.0	0.0	0.0	66.7	55.6	64.0
Q-ALIGN	62.5	80.0	77.8	75.0	50.0	100.0	100.0	59.0	55.6	67.6
Q-INSTRUCT	37.5	40.0	42.2	25.0	50.0	0.0	0.0	51.3	66.7	43.9
REALQA-AVA	46.9	60.0	57.8	100.0	50.0	50.0	100.0	59.0	55.6	56.8

Table 37: Topic-level reward-model TB-1 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Random Guess	50.0	50.0	50.0	11.3	50.0	50.0	50.0	46.4	50.0	36.0
Human Expert	89.1	94.4	90.0	50.5	85.0	90.3	88.8	80.8	86.5	74.7
Reward Models										
ARTIMUSE	54.5	88.9	50.0	25.0	50.0	54.5	50.0	50.0	64.3	46.0
PEAS-AES	54.5	44.4	50.0	16.1	100.0	72.7	50.0	58.3	42.9	44.1
PEAS-AVA	36.4	66.7	50.0	23.2	100.0	54.5	62.5	54.2	57.1	44.1
Q-ALIGN	63.6	66.7	50.0	19.6	100.0	48.5	50.0	62.5	64.3	44.7
Q-INSTRUCT	9.1	44.4	75.0	10.7	100.0	57.6	50.0	41.7	42.9	34.2
REALQA-AVA	36.4	66.7	50.0	12.5	100.0	48.5	50.0	41.7	57.1	36.6

Table 38: Topic-level reward-model TB-1 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Random Guess	50.0	45.8	8.3	15.6	8.3	8.3	19.2
Human Expert	87.7	79.0	42.3	49.1	43.3	49.1	54.4
Reward Models							
ARTIMUSE	53.8	60.0	4.5	13.0	4.8	18.2	20.0
PEAS-AES	30.8	20.0	9.1	30.4	28.6	9.1	22.0
PEAS-AVA	30.8	20.0	4.5	17.4	9.5	0.0	13.0
Q-ALIGN	38.5	10.0	13.6	13.0	19.1	0.0	16.0
Q-INSTRUCT	46.2	30.0	9.1	21.7	19.1	9.1	21.0
REALQA-AVA	38.5	50.0	13.6	4.3	4.8	0.0	15.0

Table 39: Effect of decoding temperature on MLLM performance. We compare default decoding (t=1) with greedy decoding (t=0); Δ denotes the change from t=1 to t=0 (%).

Model	Top-1 p^3			TB-1 p^3		
	t=1	t=0	Δ	t=1	t=0	Δ
Closed-Source Models						
Claude-Haiku-4.5	18.8	21.2	+2.4	12.5	17.2	+4.7
Claude-Sonnet-4.5	24.2	27.3	+3.1	14.5	20.5	+6.0
Claude-Sonnet-4.6	40.2	39.2	−1.0	26.5	26.8	+0.3
Claude-Opus-4.5	34.2	35.2	+1.0	20.2	22.5	+2.3
Claude-Opus-4.6	35.5	36.5	+1.0	20.0	21.5	+1.5
Doubao-Seed-2.0-Pro	33.2	32.8	−0.4	23.5	22.5	−1.0
Gemini-3-Flash	28.0	29.5	+1.5	15.8	21.2	+5.4
Gemini-3-Pro	35.0	36.2	+1.2	22.0	24.8	+2.8
Gemini-3.1-Pro	35.0	40.0	+5.0	22.2	27.8	+5.6
GPT-4.1	29.5	30.8	+1.3	21.2	22.5	+1.3
Grok-4.1-Fast	15.5	16.8	+1.3	12.5	13.8	+1.3
Open-Weight Models						
GLM-4.6V	17.5	16.5	−1.0	11.5	12.2	+0.7
Kimi-K2.5	26.8	28.2	+1.4	15.0	15.5	+0.5
Qwen3-VL-235B-A22B	20.5	20.8	+0.3	14.0	15.8	+1.8
Qwen3.5-397B-A17B	29.8	29.2	−0.6	17.2	18.5	+1.3
Average	28.2	29.4	+1.1	17.9	20.2	+2.3

Table 40: TB-1 results by candidate set size k for the evaluated MLLMs and reference baselines. For each set size, we report pass^3 and ap@1 (%).

Model	$k=2$		$k=3$		$k=4$		$k=5$		$k=6$	
	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1	p^3	ap@1
Random Guess	12.5	50.0	0.5	16.7	0.1	8.3	0.0	5.0	0.0	3.3
Human Expert	87.1		69.3		43.6		45.9		40.0	
Claude Haiku 4.5	24.8	49.7	8.1	32.1	0.0	12.7	0.0	6.9	0.0	0.0
Claude Sonnet 4.5	27.3	53.1	9.9	37.5	2.2	12.0	0.0	7.8	0.0	0.0
Claude Sonnet 4.6	47.3	58.2	17.1	39.0	6.7	17.2	8.8	22.5	0.0	0.0
Claude Opus 4.5	37.0	55.2	15.3	45.0	3.4	13.5	0.0	12.7	0.0	0.0
Claude Opus 4.6	37.6	58.0	9.9	38.4	6.7	13.9	2.9	13.7	0.0	0.0
Doubao Seed 2.0 Pro	38.8	59.8	20.7	43.8	5.6	17.6	5.9	24.5	0.0	0.0
Gemini 3 Flash	28.5	57.6	11.7	42.0	3.4	16.9	0.0	17.6	0.0	0.0
Gemini 3 Pro	43.6	61.4	12.6	34.8	2.2	16.1	0.0	16.7	0.0	0.0
Gemini 3.1 Pro	37.6	58.8	22.5	45.6	1.1	16.1	2.9	13.7	0.0	33.3
GLM-4.6V	24.2	50.7	1.8	29.1	3.4	19.9	2.9	7.8	0.0	0.0
GPT-4.1	37.0	55.6	18.9	43.2	2.2	15.7	2.9	16.7	0.0	0.0
GPT-5	33.9	54.1	27.0	51.1	0.0	11.6	2.9	10.8	0.0	0.0
GPT-5.1	37.0	57.4	13.5	42.3	1.1	15.4	8.8	15.7	0.0	0.0
GPT-5.1 Codex	35.8	52.5	19.8	45.9	6.7	16.1	2.9	14.7	0.0	0.0
GPT-5.2	23.0	47.1	18.9	48.6	3.4	14.6	0.0	14.7	0.0	0.0
Grok 4.1 Fast	26.1	53.1	4.5	33.6	2.2	14.2	0.0	8.8	0.0	0.0
Kimi K2.5	29.7	53.3	8.1	36.3	2.2	15.4	0.0	15.7	0.0	0.0
o4-mini	30.9	53.1	30.6	53.2	1.1	9.7	0.0	11.8	0.0	0.0
Qwen3-VL-235B	26.7	50.3	9.0	37.2	2.2	14.6	0.0	9.8	0.0	0.0
Qwen3.5-397B	28.5	52.5	18.0	45.6	1.1	14.2	2.9	15.7	0.0	33.3

Table 41: Topic-level reward-model TB-1 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Random Guess	24.0	23.3	20.9	16.7	16.7	33.3	16.7	26.1	30.6	23.7
Human Expert	73.8	72.0	71.3	60.0	60.0	80.0	70.0	74.4	72.2	72.5
Reward Models										
ArtiMuse	15.6	40.0	28.9	50.0	100.0	50.0	0.0	35.9	66.7	32.4
PEAS-aes	37.5	60.0	42.2	75.0	0.0	0.0	0.0	46.2	33.3	41.7
PEAS-ava	40.6	40.0	46.7	100.0	0.0	0.0	0.0	51.3	44.4	46.0
Q-Align	46.9	40.0	44.4	75.0	50.0	100.0	0.0	46.2	44.4	46.8
Q-Instruct	31.2	20.0	31.1	0.0	50.0	0.0	0.0	43.6	66.7	35.2
RealQA-ava	25.0	60.0	24.4	50.0	50.0	50.0	0.0	41.0	55.6	33.8

E.9 Results by Candidate Set Size

Table 40 provides the full breakdown by candidate set size. Consistent with the summary in the main text, performance drops sharply as k increases, and the decline is especially severe under the stricter TB-1 pass^3 criterion. The table also shows that the gap between models and human experts widens noticeably on larger candidate sets.

F Human Study Details

F.1 Detailed Setup and Metric Definitions

This subsection provides the full metric definitions summarized in Section 2.

Scope of inference. This study is designed to justify the annotation protocol for VAB's target label: a comparative best/worst decision within a matched-subject candidate set. It is a paired protocol comparison rather than a population survey of aesthetic taste: the same eight experts evaluate the same image pool under both scoring and ranking, holding stable individual differences in aesthetic preference fixed, and statistical inference uses paired task-level outcomes rather than an unpaired sample of annotators. This protocol study is also separate from the construction of the released ground truth, for which each task is evaluated by 10 independent expert judges drawn from a pool of more than 100 evaluators and retained through the consensus filter described in Appendix C.

The study does not claim that every possible scoring protocol is invalid or unsuitable for other purposes. Rather, in this controlled setting, independent absolute scoring introduces a scale-mapping step before the desired ordering can be inferred, whereas direct ranking collects the target comparative decision itself.

Under scoring, each annotator independently assigns an absolute score $s_{i,a} \in [0, 10]$ to each image. Under ranking, annotators provide a complete within-task ordering when presented with groups of size $n \in \{2, 3, 4, 5\}$. We operationalize "more scientifically sound" along three dimensions: fidelity to human comparative preference, reproducibility across annotators, and sharpness of the resulting label signal.

The study comprises 119 *homogeneous-content* tasks drawn directly from the curated benchmark, where all candidates share the same underlying content, and 107 *heterogeneous-content* tasks formed by regrouping 370 images from distinct homogeneous task groups while preserving topic consistency. Each image appears in at most one regrouped task. Examples of both settings are given in Appendix F.3.

Fidelity. We ask whether the ranking induced by absolute scores matches the direct ranking produced by the same annotator. Let ρ_a^{same} and ρ_a^{diff} denote annotator a's direct rankings for homogeneous-content and heterogeneous-content tasks, and let σ_a^{score} denote the ranking induced by the absolute scores. We report Kendall's τ (Kendall, 1938),

$$\tau = \frac{C - D}{\binom{n}{2}},$$

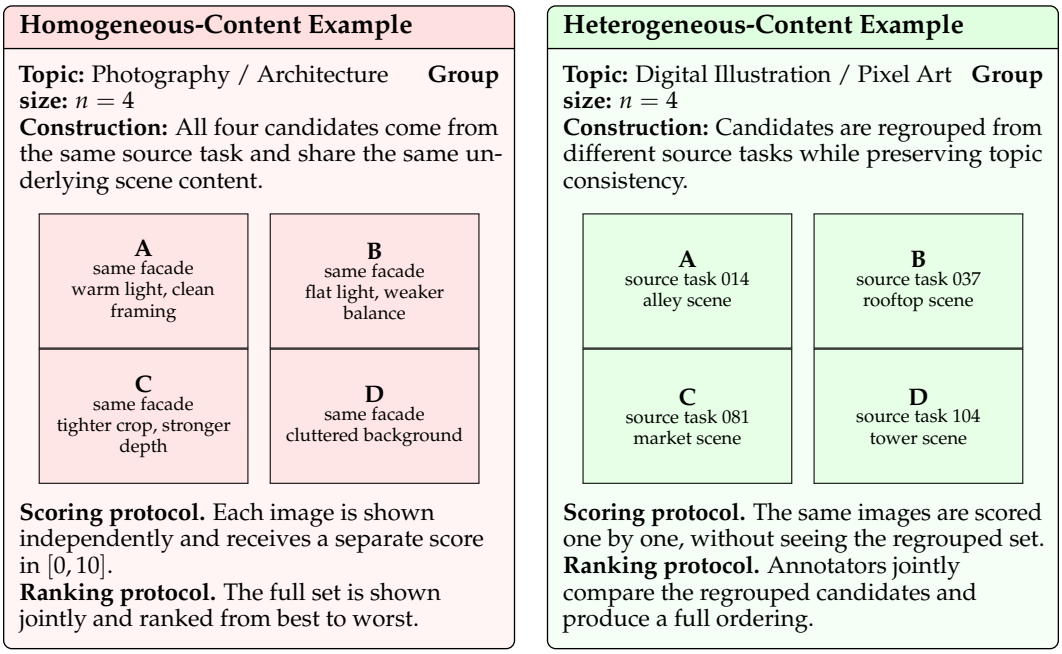


Figure 6: Schematic examples of the two settings used in the human study. **Left**: a homogeneous-content task, where all candidates come from the same source task and share the same underlying content. **Right**: a heterogeneous-content task, where candidates are regrouped from different source tasks. In both settings, the scoring protocol rates images independently, whereas the ranking protocol orders the full set jointly.

where C and D are the numbers of concordant and discordant pairs. We compute τ for each annotator within each task, average across annotators to obtain a task-level value, and report the mean and standard deviation across tasks. We additionally report a top-1 self-consistency (SC) rate: the fraction of annotator–task pairs for which σ_a^{score} and the direct ranking select the same top image.

Reproducibility. We define task-level accuracy acc as the fraction of tasks on which at least five of the eight annotators agree on the relevant label, and report acc_{best}, $\text{acc}_{\text{worst}}$, and $\text{acc}_{\text{best\&worst}}$ for both protocols.

Distinguishability. Restricting to the 119 homogeneous-content tasks, where entropy reflects concentration of aesthetic preference rather than content mismatch, we measure the entropy of the eight annotators' top-1 selections (Shannon, 1948),

$$H = -\sum_i p_i \log_2 p_i,$$

where p_i is the fraction of annotators selecting image i as the top image in a task. Because the maximum entropy depends on task size n, we report the unnormalized quantity so that lower values directly indicate stronger concentration of top-choice judgments.

F.2 Extended Interpretation

Philosophical lens. Our results echo the long-standing debate between *Aesthetic Absolutism*, which posits a context-free, intrinsic standard of beauty, and *Aesthetic Relativism*, which holds that judgments arise from comparison and context. Modern philosophical aesthetics has largely moved toward the relational view: people can reliably tell which image they prefer when alternatives are juxtaposed.

Statistical lens. Scoring introduces an additional measurement layer. Annotators must first form an internal judgment and then translate it onto a numeric scale, after which pairwise preference is inferred from score differences. This pipeline adds calibration variance, scale-use heterogeneity, and extra measurement noise. Ranking measures the relative decision of interest more directly, which helps explain its higher consistency and sharper signal.

Cognitive lens. Comparative judgment is a natural human operation. People can often say which of two images feels better before they can specify by how much. Absolute scoring is cognitively heavier because it requires maintaining an internal reference scale, calibrating intervals, and converting a qualitative impression into a number. That added abstraction likely makes scoring less stable than direct comparison.

F.3 Task Examples

Figure 6 provides schematic examples of the two task settings used in the human study.

F.4 Annotation Protocol

Scoring task. Under the scoring protocol, each annotator independently assigned an absolute score $s_{i,a} \in [0, 10]$ to each of the 400 images in the base image pool. Images were presented individually, without access to the other members of the same task group. The resulting scores were later reused to derive score-derived rankings for both the 119 homogeneous-content tasks and the 107 heterogeneous-content regroupings. Domain-specific single-image scoring rubrics are given below.

Fine Art Single-Image Scoring Rubric

For a single fine art painting, evaluate the degree of artistic completeness and maturity of aesthetic expression based solely on the image itself, and assign an overall score from 0 to 10, accurate to one decimal place.
Reference Scoring Criteria

1. **Composition and Visual Order**
   - Is the primary subject clearly defined and visually grounded?
   - Does the composition feel stable or intentionally dynamic?
   - Is the spatial arrangement natural, without unnecessary congestion or emptiness?
2. **Color and Tonal Relationships**
   - Are color relationships harmonious and layered?
   - Is there effective control of warmth/coolness and light/dark contrast?
   - Does the work avoid muddiness, dullness, or scattered color impressions?
3. **Technical Control**
   - Are the brushstrokes confident, fluid, and rhythmic?
   - Are there signs of hesitation, redundancy, or ineffective marks?
   - Does technique serve the image rather than draw attention to itself?
4. **Degree of Completion**
   - Are there areas that appear unfinished or perfunctory?
   - Are edges and transitions handled naturally?
   - Does the work feel resolved, at a point where the artist could reasonably stop?
5. **Artistic Expression and Overall Character**
   - Is the mood or atmosphere focused and coherent?
   - Does the work convey a unified overall character or presence?
   - Does it demonstrate strong artistic persuasiveness?
6. **Overall Score**

- Based on all criteria above, assign an overall score from 0 to 10, accurate to one decimal place.

Should Not Be Considered

- Personal preference for subject matter.
- Artist background, intent, or any contextual information beyond what is visible in the image itself.

Photography Single-Image Scoring Rubric

For a single photograph, evaluate its overall quality in terms of photographic language and visual expression, based solely on the image itself, and assign an overall score from 0 to 10, accurate to one decimal place.
Reference Scoring Criteria

1. **Composition and Image Structure**
   - Is the subject clearly emphasized?
   - Does the frame feel balanced or intentionally dynamic?
   - Is the background clean and supportive of the subject?
2. **Quality of Light**
   - Is the light natural or expressive?
   - Are highlights and shadows well controlled?
   - Does the image avoid crushed blacks or blown highlights?
3. **Sharpness and Focus**
   - Is focus accurate where it matters most?
   - Is the relationship between sharpness and blur appropriate?
   - Are there any noticeable technical flaws?
4. **Color or Tonal Rendering**
   - Is white balance or color temperature accurate or clearly intentional?
   - Are colors or tonal values consistent and unified?
   - Does the image avoid unwanted color casts or muddy tones?
5. **Moment and Emotional Expression**
   - Does the image capture a meaningful moment?
   - Is the emotional or narrative content clear and compelling?
   - Does the image convey a strong sense of photographic presence?
6. **Overall Score**
   - Based on all criteria above, assign an overall score from 0 to 10, accurate to one decimal place.

Should Not Be Considered

- The personal appeal or attractiveness of the subject.
- Assumptions about shooting difficulty.
- Assumptions about post-processing effort.

Digital Illustration Single-Image Scoring Rubric

For a single digital illustration, evaluate its overall quality in terms of visual design and professional illustration standards, based solely on the image itself, and assign an overall score from 0 to 10, accurate to one decimal place.
Reference Scoring Criteria

1. **Visual Focus and Information Clarity**

- Is the visual focal point clear and well defined?
- Does the visual flow guide the viewer smoothly?
- Is there minimal visual noise or competition for attention?

2. **Form and Proportion**
   - Are the structures of figures or objects accurate?
   - Are poses and silhouettes aesthetically refined?
   - Does the image avoid stiffness, distortion, or unnatural forms?
3. **Color Design**
   - Is the color palette unified and thematically coherent?
   - Are value and lighting hierarchies clear?
   - Is contrast and visual rhythm used effectively?
4. **Detail Control and Finish**
   - Are details handled appropriately, neither excessive nor careless?
   - Are materials and lighting consistent throughout the image?
   - Are edges, masks, and gradients clean and well controlled?
5. **Stylistic Consistency and Purpose Fit**
   - Does the illustration align with its intended stylistic direction?
   - Is quality consistent within a unified style?
   - Is the image suitable for its potential use case?
6. **Overall Score**
   - Based on all criteria above, assign an overall score from 0 to 10, accurate to one decimal place.

Should Not Be Considered

- Personal preference for illustration style.
- Software or tools used.
- Hypothetical or imagined client requirements.

When converting absolute scores into induced rankings, no exact within-task score ties were observed in the collected annotations, so no tie-breaking rule was required.

Ranking task. Under the ranking protocol, each annotator provided a complete best-to-worst ordering of all n images presented simultaneously within a group. Annotators were instructed to rank the images by overall aesthetic quality using the following domain-specific comparative rubrics.

Fine Art Ranking Rubric

For a set of fine art paintings that share the same style, subject matter, and medium, and differ only subtly in execution, rank the images from best to worst according to artistic completeness and maturity of aesthetic expression.
Reference Comparison Criteria

1. **Composition and Visual Order**
   - Is the primary subject more clearly defined?
   - Does the composition feel more stable or intentionally dynamic?
   - Is the spatial arrangement natural, without unnecessary congestion or emptiness?
2. **Color and Tonal Relationships**
   - Are the color relationships more harmonious or layered?
   - Is there stronger control over warmth/coolness and light/dark contrast?

- Does the work avoid muddiness, dullness, or scattered color impressions?

3. **Technical Control**
   - Are the brushstrokes more confident, fluid, and rhythmic?
   - Are there signs of hesitation, redundancy, or ineffective marks?
   - Does the technique serve the image rather than draw attention to itself?
4. **Degree of Completion**
   - Are there any areas that appear unfinished or perfunctory?
   - Are edges and transitions handled naturally?
   - Does the work feel resolved, at a point where the artist could reasonably stop?
5. **Artistic Expression and Overall Character**
   - Is the mood or atmosphere more focused and coherent?
   - Does the work convey a stronger overall character or presence?
   - Does it have greater artistic persuasiveness?
6. **Final Selection**
   - Considering the set as a whole, determine the relative ordering of the images from strongest to weakest.

Should Not Be Considered

- Personal preference for subject matter.
- Any contextual or background information beyond what is visible in the image itself.

Photography Ranking Rubric

For a set of photographs taken of similar subjects, in similar scenes, and under closely matched conditions, rank the images from best to worst in terms of photographic language and visual expression.
Reference Comparison Criteria

1. **Composition and Image Structure**
   - Is the subject more clearly emphasized?
   - Is the frame more balanced or intentionally dynamic?
   - Is the background cleaner and more supportive of the subject?
2. **Quality of Light**
   - Is the light more natural or more expressive?
   - Are highlights and shadows better controlled?
   - Does the image avoid crushed blacks or blown highlights?
3. **Sharpness and Focus**
   - Is focus more accurate where it matters?
   - Is the relationship between sharpness and blur more appropriate?
   - Are there any technical errors that detract from the image?
4. **Color or Tonal Rendering**
   - Is white balance or color temperature more accurate or more intentional?
   - Are colors or tonal values more consistent and unified?
   - Does the image avoid unwanted color casts or muddy tones?
5. **Moment and Emotional Expression**
   - Does the image capture a stronger or more meaningful moment?
   - Is the emotional or narrative content more compelling?
   - Does it carry a stronger sense of photographic presence?

6. **Final Selection**
   - Considering the set as a whole, determine the relative ordering of the photographs from strongest to weakest.

Should Not Be Considered

- The personal appeal or attractiveness of the subject.
- Assumptions about shooting difficulty.
- Assumptions about post-processing effort.

Digital Illustration Ranking Rubric

For a set of digital illustrations with similar purpose, style, and level of completion, rank the images from best to worst according to visual design and professional illustration quality.

Reference Comparison Criteria

1. **Visual Focus and Information Clarity**
   - Is the visual focal point clearer?
   - Does the visual flow guide the viewer more effectively?
   - Is there less visual noise or competition for attention?
2. **Form and Proportion**
   - Are the structures of figures or objects more accurate?
   - Are poses and silhouettes more aesthetically refined?
   - Does the image avoid stiffness, distortion, or unnatural forms?
3. **Color Design**
   - Is the color palette more unified and thematically coherent?
   - Are value and lighting hierarchies clearer?
   - Is there effective use of contrast and visual rhythm?
4. **Detail Control and Finish**
   - Are details handled appropriately, neither excessive nor careless?
   - Are materials and lighting consistent throughout?
   - Are edges, masks, and gradients clean and well controlled?
5. **Stylistic Consistency and Purpose Fit**
   - Does the illustration better align with its intended stylistic direction?
   - Is quality more consistent within a unified style?
   - Is the image more suitable for its potential use case?
6. **Final Selection**
   - Considering the set as a whole, determine the relative ordering of the illustrations from strongest to weakest.

Should Not Be Considered

- Personal preference for illustration style.
- Software or tools used.
- Hypothetical or imagined client requirements.

F.5 Notation

Symbol	Meaning
$A = 8$	number of annotators
n	number of images in a task, $n \in \{2, 3, 4, 5\}$
$s_{i,a}$	absolute score assigned to image i by annotator a, $s_{i,a} \in [0, 10]$
σ_a^{score}	score-derived ranking for annotator a, obtained by sorting images in descending order of $s_{i,a}$
ρ_a^{same}	direct ranking provided by annotator a for a homogeneous-content task
ρ_a^{diff}	direct ranking provided by annotator a for a heterogeneous-content task
y_a^{best}, y_a^{worst}	annotator a's best and worst labels in a task
SC	top-1 self-consistency rate between score-derived and direct rankings
acc	fraction of tasks on which annotators reach majority agreement on the relevant label

Task-level accuracy. For the majority-agreement comparison, task-level accuracy is defined as follows:

- acc_{best}: fraction of tasks on which at least five of the eight annotators select the same best label.
- $\text{acc}_{\text{worst}}$: fraction of tasks on which at least five of the eight annotators select the same worst label.
- $\text{acc}_{\text{best\&worst}}$: fraction of tasks on which both majority conditions are satisfied simultaneously.

F.6 Statistical Tests

We report two-sided paired tests over tasks to assess whether the ranking and scoring protocols differ systematically. Unless otherwise noted, p-values are uncorrected for multiple comparisons.

McNemar's test for majority-agreement comparisons. For each task and each metric acc_{best}, $\text{acc}_{\text{worst}}$, and $\text{acc}_{\text{best\&worst}}$, we construct paired binary outcomes indicating whether the protocol achieves a five-of-eight majority hit. We then apply an exact two-sided McNemar test (McNemar, 1947; Fagerland et al., 2013) to the discordant pairs.

- Homogeneous-content ($N = 119$ tasks): $p = 1.0 \times 10^{-11}$ for acc_{best}, $p = 2.4 \times 10^{-9}$ for $\text{acc}_{\text{worst}}$, and $p = 1.2 \times 10^{-14}$ for $\text{acc}_{\text{best\&worst}}$.
- Heterogeneous-content ($N = 107$ tasks): $p = 1.2 \times 10^{-7}$ for acc_{best}, $p = 2.4 \times 10^{-5}$ for $\text{acc}_{\text{worst}}$, and $p = 1.0 \times 10^{-9}$ for $\text{acc}_{\text{best\&worst}}$.

Wilcoxon signed-rank test for entropy comparison. We compare the paired per-task entropies H_{rank} and H_{score} over the 119 homogeneous-content tasks using a two-sided Wilcoxon signed-rank test (Wilcoxon, 1945), which yields $p = 1.7 \times 10^{-15}$.

Multiple comparisons. After Bonferroni correction (Bonferroni, 1936) over the three metrics in each setting, all six tests remain significant at $p < 0.001$.

F.7 Task Lists

Homogeneous-Content Tasks (119 Tasks)

Each homogeneous-content task groups images that share the same underlying content but differ in aesthetic treatment. The 119 tasks span four topics across three domains.

Table 42: Homogeneous-content tasks (119 tasks).

Task ID	n
fine-art_landscape-color_61	4
fine-art_landscape-color_64	2
fine-art_landscape-color_65	4
fine-art_landscape-color_66	4
fine-art_landscape-color_68	5
fine-art_landscape-color_69	5
fine-art_landscape-color_70	5
fine-art_landscape-color_71	5
fine-art_landscape-color_72	5
fine-art_landscape-color_73	5
fine-art_landscape-color_74	5
fine-art_landscape-color_75	5
fine-art_landscape-color_76	5
fine-art_landscape-color_79	5
fine-art_landscape-color_80	5
fine-art_landscape-color_81	5
fine-art_landscape-color_82	5
fine-art_landscape-color_83	5
fine-art_landscape-color_84	2
fine-art_landscape-color_86	5
fine-art_landscape-color_87	5
fine-art_landscape-color_88	5
fine-art_landscape-color_89	2
fine-art_landscape-color_90	5
illustration_pixel-art_1	4
illustration_pixel-art_10	4
illustration_pixel-art_12	4
illustration_pixel-art_13	4
illustration_pixel-art_17	4
illustration_pixel-art_18	4
illustration_pixel-art_19	4
illustration_pixel-art_20	4
illustration_pixel-art_21	4
illustration_pixel-art_22	4
illustration_pixel-art_23	4
illustration_pixel-art_24	4
illustration_pixel-art_25	4
illustration_pixel-art_27	4
illustration_pixel-art_28	4
illustration_pixel-art_29	4
illustration_pixel-art_3	4
illustration_pixel-art_30	4
illustration_pixel-art_31	4
illustration_pixel-art_32	4
illustration_pixel-art_33	4
illustration_pixel-art_34	4
illustration_pixel-art_35	4
illustration_pixel-art_36	4
illustration_pixel-art_37	4
illustration_pixel-art_38	4
illustration_pixel-art_39	4
illustration_pixel-art_4	4
illustration_pixel-art_40	4
illustration_pixel-art_7	4
illustration_pixel-art_8	4
illustration_pixel-art_9	4
photograph_architecture_103	3
photograph_architecture_105	2

Continued on next page

Table 42 (continued)

Task ID	n
photograph_architecture_106	2
photograph_architecture_108	2
photograph_architecture_109	2
photograph_architecture_110	2
photograph_architecture_112	2
photograph_architecture_113	2
photograph_architecture_114	2
photograph_architecture_116	2
photograph_architecture_119	2
photograph_architecture_12	3
photograph_architecture_121	2
photograph_architecture_122	2
photograph_architecture_123	2
photograph_architecture_124	2
photograph_architecture_126	2
photograph_architecture_129	2
photograph_architecture_13	3
photograph_architecture_133	2
photograph_architecture_134	2
photograph_architecture_140	2
photograph_architecture_143	2
photograph_architecture_146	2
photograph_architecture_148	2
photograph_architecture_149	2
photograph_architecture_15	3
photograph_architecture_153	2
photograph_architecture_154	2
photograph_architecture_20	3
photograph_architecture_22	3
photograph_architecture_25	3
photograph_architecture_26	3
photograph_architecture_32	3
photograph_architecture_35	3
photograph_architecture_36	3
photograph_architecture_39	3
photograph_architecture_4	3
photograph_architecture_41	3
photograph_architecture_44	3
photograph_architecture_45	3
photograph_architecture_50	3
photograph_architecture_6	3
photograph_architecture_7	3
photograph_landscape_1	3
photograph_landscape_10	3
photograph_landscape_15	3
photograph_landscape_18	3
photograph_landscape_19	3
photograph_landscape_2	3
photograph_landscape_22	3
photograph_landscape_26	3
photograph_landscape_27	3
photograph_landscape_29	3
photograph_landscape_3	3
photograph_landscape_48	3
photograph_landscape_59	3
photograph_landscape_62	3
photograph_landscape_68	3
photograph_landscape_75	3
photograph_landscape_81	3
photograph_landscape_83	3

Continued on next page

Table 42 (continued)

Task ID	n
photograph_landscape_86	3

Heterogeneous-Content Tasks (107 Tasks)

Each heterogeneous-content task is formed by regrouping images drawn from distinct homogeneous-content tasks, typically within the same topic. Of the original 119 regrouped tasks, 12 are excluded because the score-based Kendall's W (Kendall & Smith, 1939) is at least 0.825, indicating that the regrouped task already exhibits near-saturated consensus under the scoring protocol. The remaining 107 tasks are retained for analysis. The Source Task IDs column lists the originating homogeneous-content task for each image in the group, in label order A, B, C, ...

Table 43: Heterogeneous-content task construction (107 tasks, after saturation filter).

ID	n	Source Task IDs
1	3	illustration_pixel-art_19, illustration_pixel-art_23, illustration_pixel-art_17
2	3	photograph_landscape_27, photograph_landscape_2, photograph_landscape_3
3	3	illustration_pixel-art_35, illustration_pixel-art_4, illustration_pixel-art_12
4	4	fine-art_landscape-color_81, fine-art_landscape-color_82, fine-art_landscape-color_70, fine-art_landscape-color_69
5	4	illustration_pixel-art_31, illustration_pixel-art_40, illustration_pixel-art_8, illustration_pixel-art_3
6	2	photograph_landscape_86, photograph_landscape_81
7	2	illustration_pixel-art_22, illustration_pixel-art_8
8	3	illustration_pixel-art_20, illustration_pixel-art_36, illustration_pixel-art_19
9	5	illustration_pixel-art_38, illustration_pixel-art_7, illustration_pixel-art_10, illustration_pixel-art_32, illustration_pixel-art_12
10	3	illustration_pixel-art_13, illustration_pixel-art_12, illustration_pixel-art_24
11	3	photograph_architecture_32, photograph_architecture_110, photograph_architecture_106
12	4	fine-art_landscape-color_79, fine-art_landscape-color_71, fine-art_landscape-color_61, fine-art_landscape-color_88
13	5	photograph_architecture_26, photograph_architecture_6, photograph_architecture_41, photograph_architecture_134, photograph_architecture_149
14	2	photograph_landscape_62, photograph_landscape_1
15	5	illustration_pixel-art_34, illustration_pixel-art_17, illustration_pixel-art_30, illustration_pixel-art_36, illustration_pixel-art_18
16	3	fine-art_landscape-color_90, fine-art_landscape-color_74, fine-art_landscape-color_79
19	3	fine-art_landscape-color_76, fine-art_landscape-color_69, fine-art_landscape-color_90
20	4	illustration_pixel-art_32, illustration_pixel-art_31, illustration_pixel-art_18, illustration_pixel-art_33
21	3	illustration_pixel-art_30, illustration_pixel-art_1, illustration_pixel-art_7
22	5	fine-art_landscape-color_87, fine-art_landscape-color_80, fine-art_landscape-color_68, fine-art_landscape-color_76, fine-art_landscape-color_70
23	4	fine-art_landscape-color_73, fine-art_landscape-color_83, fine-art_landscape-color_86, fine-art_landscape-color_81
24	2	illustration_pixel-art_20, illustration_pixel-art_29

Continued on next page

Table 43 (continued)

ID	n	Source Task IDs
26	4	illustration_pixel-art_17, illustration_pixel-art_33, illustration_pixel-art_38, illustration_pixel-art_34
27	3	photograph_architecture_39, photograph_architecture_103, photograph_architecture_45
28	3	illustration_pixel-art_35, illustration_pixel-art_29, illustration_pixel-art_18
29	3	photograph_landscape_18, photograph_landscape_29, photograph_landscape_1
30	3	illustration_pixel-art_40, illustration_pixel-art_1, illustration_pixel-art_34
32	5	fine-art_landscape-color_73, fine-art_landscape-color_79, fine-art_landscape-color_90, fine-art_landscape-color_70, fine-art_landscape-color_84
33	2	photograph_architecture_124, photograph_architecture_26
34	3	photograph_architecture_106, photograph_architecture_113, photograph_architecture_4
35	4	fine-art_landscape-color_68, fine-art_landscape-color_79, fine-art_landscape-color_83, fine-art_landscape-color_75
36	3	fine-art_landscape-color_68, fine-art_landscape-color_81, fine-art_landscape-color_66
37	4	fine-art_landscape-color_72, fine-art_landscape-color_73, fine-art_landscape-color_76, fine-art_landscape-color_89
38	4	photograph_landscape_2, photograph_landscape_22, photograph_landscape_26, photograph_landscape_10
39	3	photograph_architecture_35, photograph_architecture_112, photograph_architecture_7
40	4	fine-art_landscape-color_75, fine-art_landscape-color_66, fine-art_landscape-color_81, fine-art_landscape-color_86
42	5	fine-art_landscape-color_87, fine-art_landscape-color_89, fine-art_landscape-color_79, fine-art_landscape-color_65, fine-art_landscape-color_86
43	3	photograph_landscape_83, photograph_landscape_18, photograph_landscape_3
44	4	photograph_landscape_62, photograph_landscape_59, photograph_landscape_75, photograph_landscape_3
45	4	photograph_architecture_45, photograph_architecture_13, photograph_architecture_140, photograph_architecture_26
46	4	illustration_pixel-art_28, illustration_pixel-art_20, illustration_pixel-art_4, illustration_pixel-art_39
47	4	illustration_pixel-art_25, illustration_pixel-art_10, illustration_pixel-art_8, illustration_pixel-art_28
48	3	illustration_pixel-art_9, illustration_pixel-art_21, illustration_pixel-art_39
49	4	illustration_pixel-art_30, illustration_pixel-art_18, illustration_pixel-art_4, illustration_pixel-art_8
50	2	photograph_architecture_114, photograph_architecture_36
52	5	photograph_architecture_44, photograph_architecture_153, photograph_architecture_116, photograph_architecture_129, photograph_architecture_39
53	3	photograph_landscape_48, photograph_landscape_83, photograph_landscape_59
54	4	photograph_architecture_123, photograph_architecture_103, photograph_architecture_12, photograph_architecture_13
56	2	photograph_landscape_27, photograph_landscape_75
57	2	photograph_architecture_6, photograph_architecture_39
58	4	illustration_pixel-art_1, illustration_pixel-art_33, illustration_pixel-art_19, illustration_pixel-art_40
59	2	illustration_pixel-art_22, illustration_pixel-art_27
60	3	fine-art_landscape-color_75, fine-art_landscape-color_87, fine-art_landscape-color_81

Continued on next page

Table 43 (continued)

ID	n	Source Task IDs
61	4	illustration_pixel-art_22, illustration_pixel-art_21, illustration_pixel-art_13, illustration_pixel-art_7
62	5	photograph_architecture_105, photograph_architecture_146, photograph_architecture_15, photograph_architecture_36, photograph_architecture_140
63	2	photograph_architecture_20, photograph_architecture_6
64	5	fine-art_landscape-color_73, fine-art_landscape-color_74, fine-art_landscape-color_70, fine-art_landscape-color_68, fine-art_landscape-color_80
66	3	fine-art_landscape-color_64, fine-art_landscape-color_74, fine-art_landscape-color_88
67	4	photograph_architecture_35, photograph_architecture_22, photograph_architecture_32, photograph_architecture_109
68	4	photograph_architecture_133, photograph_architecture_108, photograph_architecture_143, photograph_architecture_103
69	4	illustration_pixel-art_39, illustration_pixel-art_27, illustration_pixel-art_9, illustration_pixel-art_37
70	2	photograph_architecture_129, photograph_architecture_134
71	3	fine-art_landscape-color_90, fine-art_landscape-color_65, fine-art_landscape-color_76
72	3	photograph_landscape_15, photograph_landscape_48, photograph_landscape_1
73	4	illustration_pixel-art_3, illustration_pixel-art_37, illustration_pixel-art_31, illustration_pixel-art_1
75	4	illustration_pixel-art_31, illustration_pixel-art_4, illustration_pixel-art_22, illustration_pixel-art_28
76	5	fine-art_landscape-color_88, fine-art_landscape-color_65, fine-art_landscape-color_74, fine-art_landscape-color_69, fine-art_landscape-color_75
78	4	illustration_pixel-art_3, illustration_pixel-art_17, illustration_pixel-art_32, illustration_pixel-art_40
79	5	photograph_architecture_44, photograph_architecture_22, photograph_architecture_154, photograph_architecture_153, photograph_architecture_32
80	4	fine-art_landscape-color_82, fine-art_landscape-color_68, fine-art_landscape-color_84, fine-art_landscape-color_69
81	3	photograph_landscape_22, photograph_landscape_86, photograph_landscape_15
82	4	fine-art_landscape-color_66, fine-art_landscape-color_75, fine-art_landscape-color_76, fine-art_landscape-color_83
83	3	illustration_pixel-art_37, illustration_pixel-art_39, illustration_pixel-art_33
84	4	photograph_architecture_123, photograph_architecture_41, photograph_architecture_113, photograph_architecture_116
85	3	fine-art_landscape-color_61, fine-art_landscape-color_87, fine-art_landscape-color_70
86	4	illustration_pixel-art_25, illustration_pixel-art_12, illustration_pixel-art_35, illustration_pixel-art_32
87	5	photograph_architecture_126, photograph_architecture_112, photograph_architecture_146, photograph_architecture_25, photograph_architecture_50
88	4	illustration_pixel-art_30, illustration_pixel-art_36, illustration_pixel-art_37, illustration_pixel-art_24
89	3	illustration_pixel-art_25, illustration_pixel-art_20, illustration_pixel-art_10
90	3	photograph_architecture_154, photograph_architecture_110, photograph_architecture_25
91	2	photograph_architecture_20, photograph_architecture_149

Continued on next page

Table 43 (continued)

ID	n	Source Task IDs
92	5	illustration_pixel-art_23, illustration_pixel-art_13, illustration_pixel-art_38, illustration_pixel-art_29, illustration_pixel-art_19
93	5	photograph_architecture_15, photograph_architecture_7, photograph_architecture_41, photograph_architecture_121, photograph_architecture_148
94	2	illustration_pixel-art_23, illustration_pixel-art_10
96	2	illustration_pixel-art_29, illustration_pixel-art_23
97	4	photograph_architecture_121, photograph_architecture_119, photograph_architecture_108, photograph_architecture_22
98	3	photograph_landscape_19, photograph_landscape_81, photograph_landscape_68
99	2	photograph_architecture_7, photograph_architecture_122
100	4	photograph_architecture_45, photograph_architecture_126, photograph_architecture_133, photograph_architecture_122
101	4	photograph_landscape_59, photograph_landscape_26, photograph_landscape_10, photograph_landscape_29
102	3	photograph_architecture_119, photograph_architecture_36, photograph_architecture_148
103	3	illustration_pixel-art_7, illustration_pixel-art_36, illustration_pixel-art_3
104	3	illustration_pixel-art_34, illustration_pixel-art_24, illustration_pixel-art_13
105	4	fine-art_landscape-color_80, fine-art_landscape-color_86, fine-art_landscape-color_90, fine-art_landscape-color_88
106	2	illustration_pixel-art_27, illustration_pixel-art_21
107	3	photograph_architecture_12, photograph_architecture_50, photograph_architecture_15
109	3	photograph_landscape_68, photograph_landscape_62, photograph_landscape_2
110	3	illustration_pixel-art_9, illustration_pixel-art_38, illustration_pixel-art_27
111	4	fine-art_landscape-color_83, fine-art_landscape-color_82, fine-art_landscape-color_72, fine-art_landscape-color_71
112	2	illustration_pixel-art_9, illustration_pixel-art_35
113	3	fine-art_landscape-color_80, fine-art_landscape-color_82, fine-art_landscape-color_87
114	2	photograph_architecture_4, photograph_architecture_12
115	5	fine-art_landscape-color_72, fine-art_landscape-color_88, fine-art_landscape-color_69, fine-art_landscape-color_71, fine-art_landscape-color_86
116	5	photograph_landscape_22, photograph_landscape_27, photograph_landscape_19, photograph_landscape_26, photograph_landscape_29
117	5	photograph_architecture_143, photograph_architecture_25, photograph_architecture_20, photograph_architecture_105, photograph_architecture_124
118	2	illustration_pixel-art_24, illustration_pixel-art_28
119	3	photograph_architecture_109, photograph_architecture_114, photograph_architecture_44

G Fine-Tuning Details

G.1 Training–Evaluation Separation and Scope

The 2,000 expert-annotated training examples are collected separately from the released benchmark. We enforce disjointness at four levels: no evaluation image, prompt, artist or source, or task instance appears in the training set. This separation prevents direct content or provenance overlap between supervision and evaluation. The two sets intentionally share

Table 45: Topic-level fine-tuning Top-1 ap@1 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Open-Weight Models										
GLM-4.6V	21.2	63.0	33.3	26.8	66.7	52.5	54.2	58.3	52.4	42.7
KIMI-K2.5	30.3	85.2	50.0	27.4	100.0	46.5	62.5	50.0	52.4	43.5
QWEN3-VL-235B-A22B	15.2	85.2	50.0	26.8	66.7	48.5	66.7	47.2	57.1	42.4
QWEN3.5-397B-A17B	24.2	96.3	66.7	33.9	100.0	37.4	58.3	43.1	45.2	42.7
Ours										
QWEN3.5-35B-A3B (base)	21.2	85.2	75.0	31.0	83.3	41.4	75.0	37.5	47.6	41.8
KALLISTI-35B-A3B	24.2	81.5	58.3	36.3	66.7	52.5	75.0	47.2	50.0	47.0

the same broad domains and topic taxonomy because the experiment evaluates learning within the same aesthetic skill space; this distributional alignment does not introduce instance-level leakage under the separation criteria above.

The resulting comparison supports a deliberately scoped conclusion: under the VAB protocol, fine-tuning improves the 35B base model on all four reported metrics. Because this experiment does not evaluate transfer to an external matched-comparison benchmark, we treat it as preliminary within-benchmark evidence rather than a claim of external generalization or a general model-scaling result.

G.2 Hyperparameters

Table 44 lists the hyperparameters used to fine-tune KALLISTI-35B-A3B.

Table 44: Hyperparameters used for fine-tuning KALLISTI-35B-A3B.

Parameter	Value
Base Model	Qwen3.5-35B-A3B
Fine-tuning Method	LoRA
LoRA Rank	64
LoRA Alpha	128
LoRA Target	All layers
Vision Tower	Frozen
MLP Projector	Trainable
Learning Rate	1×10^{-4}
LR Scheduler	Cosine
Warmup Steps	10
Epochs	3
Batch Size	4 per device
Precision	BF16
Max Sequence Length	65,536

G.3 Detailed Results

As in the main-text fine-tuning table, we report topic-level detailed results under all four metrics: Top-1 ap@1, Top-1 pass^3, TB-1 ap@1, and TB-1 pass^3. The full breakdown shows that the gains from fine-tuning are uneven across topics: KALLISTI-35B-A3B improves most clearly on several fine-art topics and does not uniformly dominate its base model in photography.

Top-1 ap@1.

Tables 45, 46, and 47 report topic-level Top-1 ap@1 results for fine art, illustration, and photography, respectively.

Top-1 pass^3.

Tables 48, 49, and 50 report topic-level Top-1 pass^3 results for fine art, illustration, and photography, respectively.

TB-1 ap@1.

Table 46: Topic-level fine-tuning `Top-1 ap@1` results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Open-Weight Models							
GLM-4.6V	51.3	53.3	31.8	37.7	60.3	27.3	43.3
KIMI-K2.5	71.8	56.7	36.4	36.2	39.7	30.3	43.0
QWEN3-VL-235B-A22B	69.2	33.3	30.3	30.4	39.7	36.4	38.3
QWEN3.5-397B-A17B	71.8	53.3	33.3	27.5	50.8	48.5	44.3
Ours							
QWEN3.5-35B-A3B (base)	53.8	46.7	30.3	44.9	41.3	30.3	40.7
KALLISTI-35B-A3B	66.7	50.0	28.8	36.2	44.4	36.4	41.7

Table 47: Topic-level fine-tuning `Top-1 ap@1` results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Open-Weight Models										
GLM-4.6V	35.4	53.3	39.3	58.3	16.7	0.0	100.0	48.7	63.0	43.2
KIMI-K2.5	49.0	80.0	63.7	75.0	50.0	50.0	100.0	61.5	70.4	60.9
QWEN3-VL-235B-A22B	37.5	66.7	54.1	58.3	50.0	50.0	66.7	57.3	66.7	52.5
QWEN3.5-397B-A17B	44.8	66.7	64.4	83.3	50.0	50.0	100.0	72.6	66.7	62.8
Ours										
QWEN3.5-35B-A3B (base)	51.0	46.7	57.0	83.3	50.0	33.3	100.0	65.0	70.4	59.0
KALLISTI-35B-A3B	50.0	73.3	67.4	75.0	50.0	33.3	100.0	66.7	70.4	63.3

Tables 51, 52, and 53 report topic-level `TB-1 ap@1` results for fine art, illustration, and photography, respectively.

`TB-1 pass^3`.

Tables 54, 55, and 56 report topic-level `TB-1 pass^3` results for fine art, illustration, and photography, respectively.

Table 48: Topic-level fine-tuning Top-1 pass^3 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Open-Weight Models										
GLM-4.6V	9.1	55.6	0.0	5.4	50.0	21.2	25.0	33.3	14.3	18.0
KIMI-K2.5	9.1	77.8	25.0	5.4	100.0	18.2	50.0	16.7	28.6	19.9
QWEN3-VL-235B-A22B	9.1	77.8	50.0	8.9	50.0	24.2	50.0	20.8	21.4	22.4
QWEN3.5-397B-A17B	18.2	88.9	50.0	16.1	100.0	6.1	37.5	16.7	14.3	21.1
Ours										
QWEN3.5-35B-A3B (base)	0.0	66.7	75.0	14.3	50.0	15.2	50.0	16.7	28.6	21.7
KALLISTI-35B-A3B	0.0	77.8	50.0	19.6	50.0	27.3	62.5	33.3	14.3	28.0

Table 49: Topic-level fine-tuning Top-1 pass^3 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Open-Weight Models							
GLM-4.6V	23.1	20.0	9.1	17.4	38.1	0.0	19.0
KIMI-K2.5	53.8	30.0	13.6	21.7	9.5	9.1	21.0
QWEN3-VL-235B-A22B	46.2	0.0	4.5	4.3	9.5	18.2	12.0
QWEN3.5-397B-A17B	53.8	10.0	13.6	4.3	33.3	45.5	24.0
Ours							
QWEN3.5-35B-A3B (base)	23.1	20.0	9.1	34.8	14.3	18.2	20.0
KALLISTI-35B-A3B	53.8	10.0	9.1	21.7	19.0	27.3	22.0

Table 50: Topic-level fine-tuning Top-1 pass^3 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Open-Weight Models										
GLM-4.6V	6.2	20.0	15.6	50.0	0.0	0.0	100.0	15.4	33.3	15.8
KIMI-K2.5	18.8	60.0	44.4	75.0	50.0	50.0	100.0	35.9	55.6	38.8
QWEN3-VL-235B-A22B	6.2	60.0	26.7	50.0	50.0	0.0	0.0	28.2	33.3	24.5
QWEN3.5-397B-A17B	31.2	40.0	42.2	75.0	50.0	0.0	100.0	51.3	55.6	43.9
Ours										
QWEN3.5-35B-A3B (base)	21.9	20.0	35.6	75.0	50.0	0.0	100.0	35.9	44.4	33.8
KALLISTI-35B-A3B	21.9	40.0	46.7	50.0	50.0	0.0	100.0	33.3	44.4	36.7

Table 51: Topic-level fine-tuning TB-1 ap@1 results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Open-Weight Models										
GLM-4.6V	21.2	63.0	33.3	17.3	66.7	52.5	54.2	58.3	52.4	39.3
KIMI-K2.5	30.3	85.2	50.0	13.1	100.0	46.5	62.5	50.0	52.4	38.5
QWEN3-VL-235B-A22B	15.2	85.2	50.0	10.7	66.7	48.5	66.7	47.2	57.1	36.9
QWEN3.5-397B-A17B	24.2	96.3	66.7	18.5	100.0	37.4	58.3	43.1	45.2	37.3
Ours										
QWEN3.5-35B-A3B (base)	21.2	85.2	75.0	16.7	83.3	41.4	75.0	37.5	47.6	36.9
KALLISTI-35B-A3B	24.2	81.5	58.3	17.9	66.7	52.5	75.0	47.2	50.0	40.6

Table 52: Topic-level fine-tuning TB-1 ap@1 results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Open-Weight Models							
GLM-4.6V	51.3	50.0	13.6	20.3	28.6	15.2	27.0
KIMI-K2.5	71.8	53.3	18.2	20.3	17.5	18.2	29.0
QWEN3-VL-235B-A22B	69.2	33.3	22.7	14.5	22.2	3.0	25.7
QWEN3.5-397B-A17B	71.8	46.7	18.2	15.9	20.6	0.0	26.0
Ours							
QWEN3.5-35B-A3B (base)	53.8	43.3	4.5	18.8	22.2	9.1	22.3
KALLISTI-35B-A3B	66.7	46.7	19.7	18.8	15.9	0.0	25.3

Table 53: Topic-level fine-tuning TB-1 ap@1 results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Open-Weight Models										
GLM-4.6V	28.1	40.0	28.1	33.3	33.3	0.0	66.7	39.3	48.1	33.1
KIMI-K2.5	29.2	80.0	39.3	33.3	16.7	0.0	66.7	45.3	59.3	40.5
QWEN3-VL-235B-A22B	28.1	53.3	36.3	91.7	33.3	66.7	66.7	42.7	51.9	40.0
QWEN3.5-397B-A17B	31.2	66.7	49.6	83.3	50.0	66.7	0.0	57.3	66.7	50.1
Ours										
QWEN3.5-35B-A3B (base)	41.7	46.7	42.2	50.0	0.0	33.3	0.0	52.1	74.1	46.3
KALLISTI-35B-A3B	38.5	53.3	43.0	41.7	50.0	16.7	33.3	55.6	59.3	46.5

Table 54: Topic-level TB-1 pass^3 fine-tuning results on fine-art tasks. The final column reports the corresponding domain-level average (%).

Model	Calligraphy	Chinese Painting	Ink & Wash	Landscape Color	Portrait Color	Portrait Sketch	Quick Sketch	Still Life Color	Still Life Sketch	Avg.
Open-Weight Models										
GLM-4.6V	9.1	55.6	0.0	7.1	50.0	21.2	25.0	33.3	14.3	18.6
KIMI-K2.5	9.1	77.8	25.0	0.0	100.0	18.2	50.0	16.7	28.6	18.0
QWEN3-VL-235B-A22B	9.1	77.8	50.0	0.0	50.0	24.2	50.0	20.8	21.4	19.3
QWEN3.5-397B-A17B	18.2	88.9	50.0	5.4	100.0	6.1	37.5	16.7	14.3	17.4
Ours										
QWEN3.5-35B-A3B (base)	0.0	66.7	75.0	3.6	50.0	15.2	50.0	16.7	28.6	18.0
KALLISTI-35B-A3B	0.0	77.8	50.0	5.4	50.0	27.3	62.5	33.3	14.3	23.0

Table 55: Topic-level TB-1 pass^3 fine-tuning results on illustration tasks. The final column reports the corresponding domain-level average (%).

Model	Anime & Manga	Comic	Concept Art	Digital & AI Art	Pixel Art	Stylized 3D	Avg.
Open-Weight Models							
GLM-4.6V	23.1	20.0	4.5	8.7	0.0	0.0	8.0
KIMI-K2.5	53.8	30.0	4.5	8.7	4.8	0.0	14.0
QWEN3-VL-235B-A22B	46.2	0.0	4.5	0.0	4.8	0.0	8.0
QWEN3.5-397B-A17B	53.8	10.0	0.0	0.0	0.0	0.0	8.0
Ours							
QWEN3.5-35B-A3B (base)	23.1	20.0	0.0	8.7	0.0	0.0	7.0
KALLISTI-35B-A3B	53.8	10.0	0.0	8.7	0.0	0.0	10.0

Table 56: Topic-level TB-1 pass^3 fine-tuning results on photography tasks. The final column reports the corresponding domain-level average (%).

Model	Architecture	Food & Product	Landscape	Macro	Night & Astro	Portrait	Sports	Street & City	Wildlife	Avg.
Open-Weight Models										
GLM-4.6V	0.0	20.0	2.2	0.0	0.0	0.0	0.0	10.3	22.2	5.8
KIMI-K2.5	3.1	60.0	6.7	25.0	0.0	0.0	0.0	15.4	33.3	12.2
QWEN3-VL-235B-A22B	3.1	20.0	11.1	75.0	0.0	0.0	0.0	12.8	22.2	12.2
QWEN3.5-397B-A17B	15.6	20.0	17.8	50.0	50.0	0.0	0.0	30.8	44.4	23.7
Ours										
QWEN3.5-35B-A3B (base)	15.6	20.0	17.8	25.0	0.0	0.0	0.0	20.5	44.4	19.4
KALLISTI-35B-A3B	9.4	20.0	17.8	0.0	50.0	0.0	0.0	20.5	22.2	16.5

H Prompt Templates and Scoring Interfaces

This appendix records the prompt templates and scoring interfaces used in data construction and evaluation. We group them by usage: prompts for the photography pipeline, prompts for the illustration pipeline, prompts for MLLM evaluation, and scoring interfaces for reward-model evaluation.

H.1 Photography Pipeline Prompts

This subsection records the prompts used in the photography agent-edit pipeline described in Section 3. The pipeline consists of three stages: aspect-wise critique generation with ArtiMuse, edit planning with GPT through LiteLLM, and image editing with Gemini-3-Pro-Image (Nano Banana Pro). In Step 1, ArtiMuse is queried eight times independently with the same image, once per aesthetic aspect. The resulting responses are then concatenated into a single review string that is passed to the planning stage.

H.1.1 Step 1: ArtiMuse Aspect-Wise Critique

ArtiMuse Aspect-Wise Critique Calls

The same image is attached in every call. The following eight prompts are sent independently:

1. Please evaluate the aesthetic quality of this image from the aspect of Composition & Design.
2. Please evaluate the aesthetic quality of this image from the aspect of Visual Elements & Structure.
3. Please evaluate the aesthetic quality of this image from the aspect of Technical Execution.
4. Please evaluate the aesthetic quality of this image from the aspect of Originality & Creativity.
5. Please evaluate the aesthetic quality of this image from the aspect of Theme & Communication.
6. Please evaluate the aesthetic quality of this image from the aspect of Emotion & Viewer Response.
7. Please evaluate the aesthetic quality of this image from the aspect of Overall Gestalt.
8. Please evaluate the aesthetic quality of this image from the aspect of Comprehensive Evaluation.

Critique Aggregation Format

The eight responses are concatenated into a single `review` string in the following format:

```
[Composition & Design] <response>
[Visual Elements & Structure] <response>
[Technical Execution] <response>
[Originality & Creativity] <response>
[Theme & Communication] <response>
[Emotion & Viewer Response] <response>
[Overall Gestalt] <response>
[Comprehensive Evaluation] <response>
```

H.1.2 Step 2: Edit Planning

The planning stage is executed with GPT through LiteLLM. Below we reproduce the literal runtime system and user messages. The placeholder {`review`} denotes the full ArtiMuse output obtained by concatenating the eight aspect-wise responses from Step 1. We retain the original runtime field names, including `substyle`.

Planning System Message

```
You are an expert aesthetic analyst and image editing planner for creating
aesthetic preference benchmark datasets.
```

Planning User Message

```
Analyze this photograph and plan a synthetic edit for an aesthetic preference
dataset.
AESTHETIC ASSESSMENT:
{review}
AVAILABLE CLASSIFICATIONS:
Domains: Photo
Substyles:
{
  "Photo": [
    "Portrait",
    "Landscape",
    "StreetCity",
    "Architecture",
    "Wildlife",
    "Macro",
    "FoodProduct",
    "NightAstro",
    "Sports"
  ]
}
TASK:
1. Classify the image: choose the most appropriate domain and substyle based on
image itself
2. Freely decide how to edit this image to create better and worse versions
3. Write a "better" prompt: describe how to edit THIS image to make it
aesthetically better
4. Write a "worse" prompt: describe how to edit THIS image to make it
aesthetically worse
- "Better" means: take the current image and improve its quality
- "Worse" means: take the current image and degrade its quality.
EDIT Types - Do NOT always default to color/tone adjustments. Actively consider
all edit types:
1. **Color & Tonal Adjustments**:
  - White balance, color cast correction or introduction
  - Saturation, vibrance, contrast adjustments
  - Color grading and mood shifts
2. **Perspective & Viewpoint Changes**:
  - Simulate a slightly different camera angle or position
  - Adjust the apparent depth or distance from the subject
  - Change the sense of scale or spatial relationships
3. **Element Removal**:
  - Remove contextual elements that enhance (or clutter) the scene
  - Modify background elements to improve or degrade composition
4. **Compositional Restructuring**:
  - Adjust cropping or framing emphasis
  - Rebalance visual weight distribution
  - Modify leading lines or focal points
5. **Light & Shadow Changes**:
  - Modify shadow depth, highlight recovery, or exposure
  - Change lighting direction or intensity
  - Adjust local brightness to guide visual attention
Choose multiple edit types that best address the specific strengths and weaknesses
identified in the aesthetic assessment.
WORSE PROMPT RESTRICTIONS:
When creating the "worse" prompt, DO NOT use the following editing techniques:
- DO NOT add lens flare or light flares
```

```
- DO NOT add vignetting (darkened corners)
- DO NOT add object outlines, glows, or halos around subjects
These techniques are overused and create unrealistic degradation.
PROMPT WRITING GUIDELINES:
- Describe the scene narratively, don't just list keywords, a narrative,
descriptive paragraph will almost always produce a better, more coherent image
than a list of disconnected words.
- Use photographic and cinematic language terms. Mention camera angles, lens types,
lighting, and fine details to guide the model toward a photorealistic result.
- Explain how changes should integrate into the scene (lighting, style,
composition)
- For adding/removing elements: describe what to add/remove and how it should
blend in
- For style changes: describe both the target style and what to preserve
Return JSON:
{
  "content": {"domain": "Photo", "substyle": "..."},
  "planned_prompts": {
    "better": "...",
    "worse": "..."
  },
  "reasoning": "Explain your substyle classification choice and why you chose
these specific photographic edits for better/worse versions"
}
```

H.1.3 Step 3: Image Editing

The final stage uses Gemini-3-Pro-Image (Nano Banana Pro) to synthesize the edited variants. Round 1 generates the better image from the original image alone. Round 2 generates the worse image from the original image while conditioning on the better image as additional reference context. For Round 2, we report the file-backed with-context template actually used at runtime. This is the operative version loaded by the pipeline and therefore excludes the fallback sentence beginning with "Important: Apply this edit to the ORIGINAL image" that appears only in the default code path.

Round 1: Better-Image Editing Prompt

You are an expert image editor. Edit the provided image according to these instructions:

`{edit_instruction}`

Round 2: Worse-Image Editing Prompt with Context

You are an expert image editor.

Context: Previously, the following improvement was made to the original image:
"`{context_prompt}`"
(The improved result is shown in the second image for reference)

Now, edit the **ORIGINAL** image (first image) according to these instructions:
"`{edit_instruction}`"

H.2 Illustration Pipeline Prompts

This subsection records the prompt templates used in the prompt-based illustration generation pipeline. We use one template for digital and AI art, concept art, and pixel art, and a separate storyboard-style template for anime and manga and comic generation. The stylized 3D subset is constructed through asset rendering rather than text-to-image prompting.

H.2.1 Digital and AI Art / Concept Art / Pixel Art Generation

For digital and AI art, concept art, and pixel art generation, we use a structured template to control semantic content while allowing variation in visual execution. Each prompt is composed in a fixed order: subject or character, motion or action, environment or background, lighting condition, atmosphere or mood, camera or composition, and art style.

Illustration Structured Prompt Template

Generate 25 concise prompts ready for text-to-image artwork generation on the topic of `[ART CATEGORY]`. Each prompt should describe a clear visual scene using the following order: subject or character, motion, environment/background, lighting condition, atmosphere or mood, camera or composition (view, framing, depth), and art style.

H.2.2 Manga / Comic Storyboard Generation

For anime and manga and comic generation, we adopt a storyboard-style template with stronger structural constraints. Each prompt describes a short multi-panel narrative with consistent characters and environment, while each panel specifies shot type, character action, key props, and scene details.

Illustration Storyboard Prompt Template

Generate 25 concise storyboard prompts for `[ART CATEGORY]` artwork generation. Each prompt should describe a micro-story using panels. Anchors such as character identity and environment must remain consistent. Each panel must specify shot type (wide, medium, close-up), character actions, key props, and environment details. Keep the description visual and concise.

H.3 MLLM Evaluation Prompts

This subsection reports the generic prompts used to evaluate MLLMs on VAB. Depending on the task setting, the model is asked to identify the best image, the worst image, or both the best and the worst images in a candidate set.

H.3.1 Best-Image Selection Prompt

Pick-Best Prompt

You will be shown `{n}` images labeled `{labels}`. Evaluate the overall aesthetic quality of each image, considering composition, color, technique, and artistic expression.

Which image has the best overall aesthetic quality?
`{options}`

You may reason step by step. Then, on the very last line of your response, write your final answer in exactly this format:

`BEST: X`

where `X` is the letter label of the best image.

H.3.2 Worst-Image Selection Prompt

Pick-Worst Prompt

You will be shown `{n}` images labeled `{labels}`. Evaluate the overall aesthetic quality of each image, considering composition, color, technique, and artistic expression.

Which image has the worst overall aesthetic quality?
`{options}`

You may reason step by step. Then, on the very last line of your response, write your final answer in exactly this format:

```
WORST: X
```

where `X` is the letter label of the worst image.

H.3.3 Joint Best-and-Worst Selection Prompt

Pick-Best-and-Worst Prompt

You will be shown `{n}` images labeled `{labels}`. Evaluate the overall aesthetic quality of each image, considering composition, color, technique, and artistic expression.

Which image has the best overall aesthetic quality, and which has the worst?
`{options}`

You may reason step by step. Then, on the very last lines of your response, write your final answers in exactly this format:

```
BEST: X
WORST: Y
```

where `X` and `Y` are the letter labels.

H.4 Reward-Model Scoring Interfaces

This subsection records the model-specific scoring interfaces used to evaluate reward models on VAB. Unlike MLLMs, reward models do not share a single comparative prompt. Instead, each model scores images independently following its released evaluation method, and we convert the resulting per-image scores into within-task rankings. Across all reward models, each image is scored `num_trials` times, the scores are averaged, and the averaged scores are used to determine the predicted best and worst images. We use the following terminology: *explicit prompt* denotes prompt text exposed at the call site, *implicit prompt* denotes prompt text embedded inside the model library, and *no prompt* denotes a pure vision scorer with no language instruction.

H.4.1 ArtiMuse

ArtiMuse uses an *implicit prompt*: the instruction is hardcoded inside the released library rather than passed from our evaluation script.

ArtiMuse Internal Scoring Prompt

Rate the aesthetics score of the image in 0-100.

In the output format, numbers are replaced by 2 corresponding letters, and the mapping relationship is:

score 0 to 25: 0-aa, 1-ab, 2-ac, 3-ad, . . . , 25-az,
score 26 to 50: 26-ca, 27-cb, 28-cc, 29-cd, . . . , 50-cy,
score 51 to 75: 51-da, 52-db, 53-dc, 54-dd, . . . , 75-dy,
score 76 to 100: 76-ea, 77-eb, 78-ec, 79-ed, . . . , 100-ey.

The answer only outputs 2 corresponding letters.

At runtime, the library automatically prepends an image token before the instruction. The model outputs a two-letter code, which is then decoded back to a numerical score in $[0, 100]$.

H.4.2 Q-Align

Q-Align also uses an *implicit prompt*. In our experiments, we use the aesthetic scorer by default; the library additionally exposes a quality scorer.

Q-Align Internal Prompts

Aesthetic scorer (used by default):

```
USER: How would you rate the aesthetics of this image?
<|image|>
ASSISTANT: The aesthetics of the image is
```

Quality scorer (alternative interface):

```
USER: How would you rate the quality of this image?
<|image|>
ASSISTANT: The quality of the image is
```

The scorer reads the last-token logits over the five words `excellent`, `good`, `fair`, `poor`, and `bad`, applies a softmax, and computes a weighted sum with weights $[1, 0.75, 0.5, 0.25, 0]$.

H.4.3 Q-Instruct

Q-Instruct uses an *explicit prompt* passed through the command-line interface.

Q-Instruct Evaluation Prompt

Rate the quality of the image.

The instruction is wrapped in the `mplug_owl2` conversation format with a prepended `DEFAULT_IMAGE_TOKEN`. The default score is the logit difference between `good` and `poor`.

H.4.4 RealQA

RealQA uses an *explicit prompt*. Under the AVA adapter used in our experiments, the default instruction is:

RealQA Evaluation Prompt

Rate the aesthetics and quality of this image briefly.

The model generates free text, from which a numerical score in $[0.0, 10.0]$ is parsed. If parsing fails, the library issues a second constrained prompt:

RealQA Retry Prompt

Output only one numeric score. Range: 0.0 to 10.0. No words, no explanation.

H.4.5 PEAS

PEAS-aes and PEAS-ava use *no prompt*. They are pure vision models with no language instruction. The scorer outputs a 10-bin distribution over scores from 1 to 10, and the final score is the expected value of that distribution.

PEAS Scoring Interface

No textual prompt is used. The model outputs a 10-bin score distribution over the range 1 to 10, and the final scalar score is computed as the expected value under that distribution.

I Topic-Level Annotation Examples

This appendix provides one representative annotation example for each of the 24 topics in VAB, grouped by domain.

I.1 Fine Art

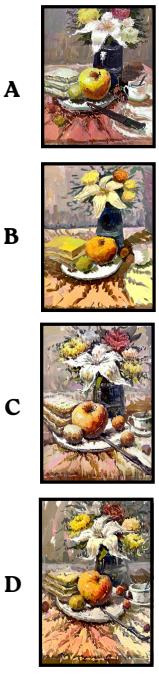
Landscape Color





Expert 1	3rd	4th	1st	2nd
Expert 2	2nd	4th	1st	3rd
Expert 3	1st	4th	2nd	3rd
Expert 4	4th	3rd	1st	2nd
Expert 5	3rd	4th	1st	2nd
Expert 6	1st	4th	3rd	2nd
Expert 7	3rd	2nd	1st	4th
Expert 8	4th	2nd	1st	3rd
Expert 9	2nd	4th	1st	3rd
Expert 10	2nd	3rd	4th	1st

Still Life Color



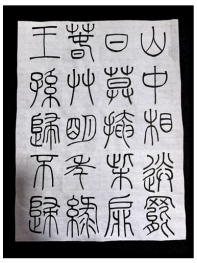


Expert 1	3rd	4th	2nd	1st
Expert 2	3rd	4th	1st	2nd
Expert 3	2nd	4th	3rd	1st
Expert 4	2nd	4th	1st	3rd
Expert 5	3rd	4th	2nd	1st
Expert 6	3rd	4th	2nd	1st
Expert 7	2nd	4th	3rd	1st
Expert 8	3rd	4th	1st	2nd
Expert 9	3rd	4th	2nd	1st
Expert 10	2nd	4th	3rd	1st

Calligraphy

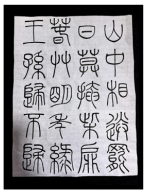


A



B





	1st	2rd
Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	1st	2rd
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	1st	2rd
Expert 10	1st	2rd

Chinese Painting



A



B





	1st	2rd
Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	2rd	1st
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	1st	2rd
Expert 10	1st	2rd

Ink and Wash

A



B







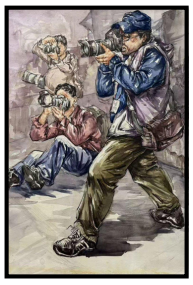
Expert 1	1st	2rd
Expert 2	2rd	1st
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

Portrait Color

A



B





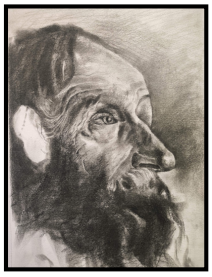


Expert 1	1st	2rd
Expert 2	2rd	1st
Expert 3	1st	2rd
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

Portrait Sketch



A



B





Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	1st	2rd
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	1st	2rd
Expert 10	1st	2rd

Quick Sketch



A



B





Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	1st	2rd
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	2rd	1st
Expert 10	2rd	1st

Still Life Sketch



A



B





Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	1st	2rd
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	1st	2rd

I.2 Photography

Architecture







A



B

Expert 1	2rd	1st
Expert 2	2rd	1st
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

Food and Product

A



B







Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	1st	2rd
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	1st	2rd
Expert 10	1st	2rd

Landscape

A



B



C









Expert 1	1st	2nd	3rd
Expert 2	1st	2nd	3rd
Expert 3	2nd	1st	3rd
Expert 4	1st	3rd	2nd
Expert 5	1st	2nd	3rd
Expert 6	1st	2nd	3rd
Expert 7	1st	2nd	3rd
Expert 8	1st	2nd	3rd
Expert 9	1st	2nd	3rd
Expert 10	1st	2nd	3rd

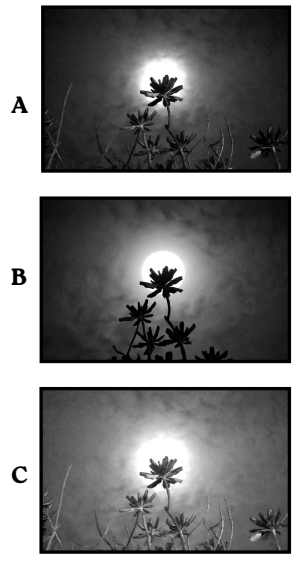
Macro





Expert 1	2nd	1st	3rd
Expert 2	2nd	1st	3rd
Expert 3	1st	2nd	3rd
Expert 4	2nd	1st	3rd
Expert 5	2nd	1st	3rd
Expert 6	1st	2nd	3rd
Expert 7	2nd	1st	3rd
Expert 8	2nd	1st	3rd
Expert 9	2nd	1st	3rd
Expert 10	2nd	1st	3rd

Night and Astrophotography





Expert 1	1st	2nd	3rd
Expert 2	1st	2nd	3rd
Expert 3	1st	2nd	3rd
Expert 4	1st	2nd	3rd
Expert 5	1st	3rd	2nd
Expert 6	1st	2nd	3rd
Expert 7	1st	2nd	3rd
Expert 8	1st	2nd	3rd
Expert 9	1st	2nd	3rd
Expert 10	1st	3rd	2nd

Portrait




A



B



Expert 1	2rd	1st
Expert 2	2rd	1st
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	1st	2rd
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

Sports




A



B



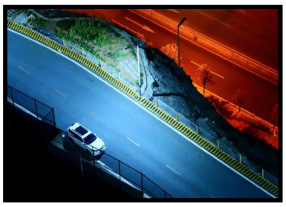
Expert 1	1st	2rd
Expert 2	1st	2rd
Expert 3	1st	2rd
Expert 4	1st	2rd
Expert 5	1st	2rd
Expert 6	1st	2rd
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	2rd	1st
Expert 10	1st	2rd

Street and City

A



B







	2rd	1st
Expert 1	2rd	1st
Expert 2	2rd	1st
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

Wildlife

A



B







	2rd	1st
Expert 1	2rd	1st
Expert 2	2rd	1st
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

I.3 Illustration

Anime and Manga

A



B







Expert 1	2rd	1st
Expert 2	2rd	1st
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	1st	2rd
Expert 8	1st	2rd
Expert 9	2rd	1st
Expert 10	2rd	1st

Comic

A



B

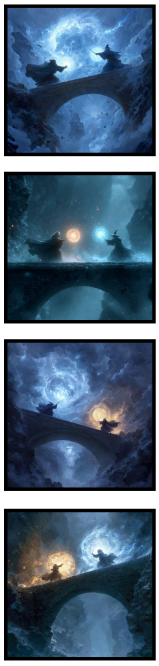


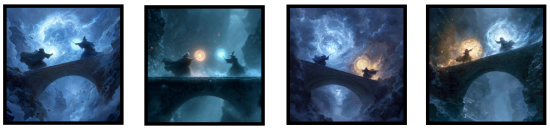




Expert 1	2rd	1st
Expert 2	1st	2rd
Expert 3	2rd	1st
Expert 4	2rd	1st
Expert 5	2rd	1st
Expert 6	2rd	1st
Expert 7	2rd	1st
Expert 8	2rd	1st
Expert 9	2rd	1st
Expert 10	2rd	1st

Concept Art





Expert 1	3rd	4th	2nd	1st
Expert 2	3rd	4th	2nd	1st
Expert 3	4th	3rd	2nd	1st
Expert 4	3rd	2nd	4th	1st
Expert 5	3rd	4th	1st	2nd
Expert 6	2nd	4th	1st	3rd
Expert 7	4th	3rd	2nd	1st
Expert 8	3rd	4th	2nd	1st
Expert 9	3rd	4th	1st	2nd
Expert 10	4th	3rd	2nd	1st

Digital and AI Art





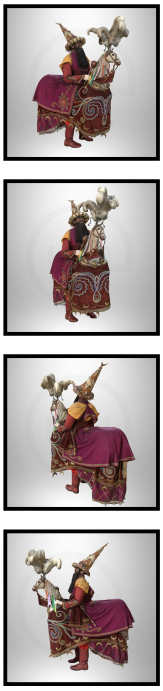
Expert 1	4th	3rd	1st	2nd
Expert 2	4th	3rd	2nd	1st
Expert 3	3rd	4th	2nd	1st
Expert 4	4th	3rd	1st	2nd
Expert 5	3rd	4th	2nd	1st
Expert 6	4th	3rd	1st	2nd
Expert 7	3rd	4th	1st	2nd
Expert 8	2nd	4th	1st	3rd
Expert 9	4th	2nd	1st	3rd
Expert 10	4th	2nd	1st	3rd

Pixel Art



				
Expert 1	1st	3rd	2nd	4th
Expert 2	2nd	3rd	1st	4th
Expert 3	2nd	4th	1st	3rd
Expert 4	3rd	1st	2nd	4th
Expert 5	3rd	2nd	1st	4th
Expert 6	3rd	2nd	1st	4th
Expert 7	4th	1st	2nd	3rd
Expert 8	3rd	2nd	1st	4th
Expert 9	3rd	2nd	1st	4th
Expert 10	2nd	3rd	1st	4th

Stylized 3D



				
Expert 1	1st	2nd	4th	3rd
Expert 2	1st	3rd	2nd	4th
Expert 3	1st	3rd	2nd	4th
Expert 4	1st	3rd	2nd	4th
Expert 5	2nd	1st	3rd	4th
Expert 6	1st	2nd	3rd	4th
Expert 7	2nd	1st	3rd	4th
Expert 8	1st	3rd	2nd	4th
Expert 9	1st	2nd	3rd	4th
Expert 10	1st	2nd	3rd	4th