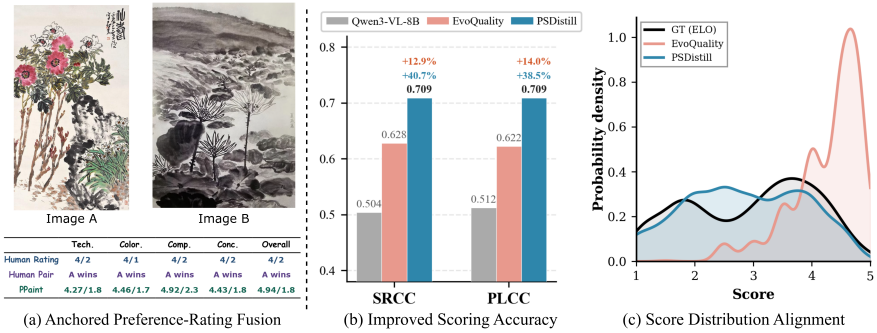
PREFERENCES ORDER, RATINGS ANCHOR: FROM FUSED EXPERT AESTHETIC GROUND TRUTH TO SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pairwise preferences and pointwise ratings are the two dominant annotation protocols in image aesthetic assessment (IAA), yet existing benchmarks adopt only one, leaving their complementarity unmeasured on the same images and dimensions. We propose an approach to expert aesthetic evaluation based on anchored preference-rating fusion. To study this approach, we construct PPAINT, a benchmark of 150 Chinese paintings spanning five aesthetic dimensions. It contains matched ratings and 45,900 pairwise judgments collected from 15 domain experts. The matched design reveals complementary strengths: preferences yield more consistent ordinal rankings, while ratings anchor scores to the expert rating scale. Our approach combines these signals into fused expert ground truth using two distinct preference-to-score constructions, which produce nearly identical scores under shared anchor constraints. Following the same preference-to-score principle, we introduce PSDISTILL, a label-free self-distillation method for VLM aesthetic scoring. It converts VLM pairwise judgments into continuous pseudo-scores via an Elo reference pool, and trains the same VLM with ranking optimization with pair weighting to produce a single-pass aesthetic scorer. Trained on a single painting category, the distilled Qwen3-VL-8B improves mean SRCC from 0.504 to 0.709 across all three categories. Among the evaluated models, it achieves a higher mean SRCC point estimate than the specialized IAA baseline and most closed-source VLMs, using a single forward pass per image. Further experiments on APDDv2 assess the training method's generalization beyond Chinese painting. We will release the training code and provide controlled academic access to the complete PPAINT annotations and score mappings.



(a) Anchored Preference-Rating Fusion (b) Improved Scoring Accuracy (c) Score Distribution Alignment

Figure 1: **PPAINT benchmark and PSDISTILL self-distillation.** (a) Expert preferences and rating anchors yield scores for brush technique, coloration, composition, artistic conception, and overall quality; slash-separated values correspond to images A/B. (b) Mean SRCC/PLCC against these fused expert targets across three categories and five dimensions (higher is better). Percentages show relative gains over Qwen3-VL-8B and EvoQuality. (c) Predicted score distributions versus anchored Elo ground truth, GT (ELO), on the 1–5 scale. PSDISTILL reduces EvoQuality's high-score concentration. Expert labels define evaluation targets; model comparisons supply training pseudo-labels.

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

1 INTRODUCTION

Aesthetic assessment relies on ground-truth labels that faithfully reflect human judgment. Expert judgments are typically elicited through one of two annotation protocols: absolute *ratings* on a numerical scale (e.g., 1–5) or relative *preferences* between pairs of images (A wins/Tie/B wins). The dominant practice constructs ground truth by averaging ratings across annotators (Murray et al., 2012; Kong et al., 2016; Wu et al., 2024), but because annotators calibrate the score range differently, a rating average conflates quality variation with inter-annotator scale usage (Perez-Ortiz et al., 2019; Tripathi et al., 2025). Pairwise preferences mitigate this calibration problem by asking only which of two items is better (Chiang et al., 2024; Li et al., 2026), yet without additional scaling or anchors, preferences do not produce the interpretable scores that downstream models require. We investigate whether preferences provide more consistent orderings and whether rating-based anchors improve cross-method agreement on the resulting expert rating scale.

Chinese painting provides a focused testbed for studying how annotation protocol shapes ground truth because expert evaluation follows explicit, domain-specific aesthetic criteria and remains inherently subjective (Zhang et al., 2024; Yang et al., 2026a). Examining these questions requires ratings and preferences on the same images and dimensions, together with reliability diagnostics. Existing resources do not jointly provide this information (Table 1).

Table 1: Prior resources for aesthetic ground-truth construction. Multi-dim. = multi-dimensional annotation. Reliability = annotator-level data or agreement metrics for independent assessment. Local–global structure = support for local relative comparisons and a global interpretable score scale.

Resource	Multi-dim.	Pointwise	Pairwise	Reliability	Local–global
AADB (Kong et al., 2016)	✓	✓	✗	✓	✗
BAID (Yi et al., 2023)	✗	✓	✗	✗	✗
APDDv2 (Jin et al., 2024)	✓	✓	✗	✗	✗
ArtiMuse (Cao et al., 2025)	✓	✓	✗	✗	✗
FGAesthetics (Yang et al., 2026b)	✗	✗	✓	✗	✗
HanMo-Bench (Yang et al., 2026a)	✓	✓	✗	✗	✗
PPAINT (ours)	✓	✓	✓	✓	✓

We therefore construct PPAINT, a dense expert-curated diagnostic benchmark in which 15 domain experts annotate 150 Chinese paintings from three categories (flower-and-bird, landscape, and figure painting) across five aesthetic dimensions under both rating and preference protocols. The expert pool is shared across categories, with different rating and preference panels within each category.

Under this collection design, pairwise preferences yield higher inter-expert agreement (Kendall's W: 0.795 vs. 0.697), while high-agreement ratings provide anchors for the expert rating scale. **We construct aesthetic scores from relative preferences for evaluation and learning.** For expert evaluation, we develop anchored preference-rating fusion to combine these complementary signals. Two distinct constructions produce nearly identical scores under shared anchor constraints (Figure 1a). The fused scores provide a common target for evaluating aesthetic ranking and score accuracy.

With the expert-derived evaluation target established, we next consider training on the unlabeled corpus. Because expert labels are costly to obtain at scale, we study label-free self-distillation. VLM-generated supervision faces the same preference-to-score issue: raw pointwise scores are poorly calibrated, and majority-vote preferences improve ranking but collapse into a narrow score band (Figure 1c). Following this preference-to-score principle, we introduce PSDISTILL, which converts VLM preferences into continuous pseudo-scores via a small Elo reference pool and trains the same VLM with group relative policy optimization (GRPO) (Shao et al., 2024) using the weighted ranking reward into a single-pass five-dimensional scorer. Trained on one painting category, the distilled Qwen3-VL-8B improves mean SRCC from 0.504 to 0.709 across all three categories (Figure 1b), achieving the highest mean SRCC point estimate among the open-source pointwise scorers in Table 3. It also comes within 0.04 SRCC of Gemini-3.1-Pro, our top closed-source pointwise baseline. We further apply the training method to the oil-painting and sketching subsets of APDDv2 and evaluate both within and across these domains (Jin et al., 2024).

Our contributions can be summarized as follows:

- **Anchored preference-rating fusion.** We propose an anchored fusion that combines preference-derived rankings with rating-based scale calibration, producing evaluation targets with both stable orderings and interpretable scores.
- **Matched dual-protocol benchmark.** PPAINT provides 150 expert-annotated paintings under matched rating and preference protocols, with a high-density 612-image-pair comparison set per category that preserves local comparison information for fine-grained ranking. It also includes approximately 9,000 unlabeled training images for pseudo-label construction, and we will provide raw annotations and fused expert ground truth through controlled academic access, alongside construction and training code.
- **Preference-to-score self-distillation.** PSDISTILL applies the same preference-to-score principle to VLM pseudo-label construction. An Elo-based reference pool converts VLM preferences into continuous pseudo-scores, and GRPO with the weighted ranking reward trains the same VLM into a single-pass five-dimensional scorer, yielding both higher ranking accuracy and lower absolute score errors than pairwise majority-vote distillation.

2 RELATED WORK

Aesthetic assessment and Chinese painting benchmarks. Early IAA work established rating-based supervision through datasets such as AVA (Murray et al., 2012) and AADB (Kong et al., 2016) and models such as NIMA (Talebi & Milanfar, 2018). Recent VLM-based scorers continue to treat aesthetic assessment primarily as pointwise score prediction (Wu et al., 2024; Ke et al., 2023; Zhou et al., 2024). Art-oriented resources extend this setting with richer domains or attributes: BAID (Yi et al., 2023) provides a large-scale artistic IAA dataset with overall aesthetic scores aggregated from online user votes, ArtiMuse (Cao et al., 2025) provides eight-dimensional aesthetic scoring with expert annotations, FGAesthetics (Yang et al., 2026b) provides pairwise comparison annotations within visually similar image series, APDDv2 (Jin et al., 2024) spans 24 painting and drawing categories with ten aesthetic attribute scores per image, and HanMo-Bench (Yang et al., 2026a) targets professional Chinese-painting evaluation for VLMs. These resources broaden domain coverage and annotation detail, but none provides matched pointwise and pairwise annotations on the same images and aesthetic dimensions with the protocol diagnostics studied here.

Pairwise evaluation and quality scale construction. Pairwise evaluation reduces the need for a shared numerical scale. Chatbot Arena (Chiang et al., 2024) and GenArena (Li et al., 2026) demonstrate effective pairwise-preference aggregation for LLM and visual-generation evaluation, respectively, and Lee and Kim (Lee & Kim, 2019) show that pairwise comparisons can unify score regression and classification for image aesthetics. In machine translation, Song et al. (2025) find that side-by-side annotation yields higher inter-annotator agreement than pointwise error marking. These studies show that pairwise rankings can reduce scale-dependent noise, while also indicating that protocol choice introduces its own biases (Tripathi et al., 2025). Bansal et al. (2024) further report that ratings and rankings disagree on a substantial fraction of items even when collected from the same annotators. Statistical methods for converting pairwise outcomes into global quality scales are well established (Thurstone, 1927; Bradley & Terry, 1952; Davidson, 1970; Elo, 1978; Perez-Ortiz et al., 2019). To our knowledge, no prior work provides a matched within-domain expert comparison that quantifies the protocol tradeoff and fuses both signals into a calibrated scale.

Preference-based training for visual scoring. Compare2Score (Zhu et al., 2024) uses multimodal pairwise comparison to derive continuous image quality scores, but focuses on single-dimensional IQA rather than multi-dimensional expert-scale construction. VisualQuality-R1 (Wu et al., 2025) trains a single-dimension IQA scorer via reinforcement learning to rank with human mean opinion scores (MOS) as ground truth, and EvoQuality (Wen et al., 2025) extends this framework to a label-free setting using model-generated majority-vote preferences. Our approach differs in two respects: the offline stage converts VLM pairwise preferences into multi-dimensional pseudo-scores in the range $[1, 5]$, and the distillation uses pair weighting based on pseudo-score gaps to reduce the influence of near-tie comparisons during ranking optimization.

3 PPAINT: MATCHED PROTOCOLS AND FUSED EXPERT GROUND TRUTH

3.1 DATA COLLECTION AND ANNOTATION PROTOCOL

We collect ratings and preferences on the same paintings and aesthetic dimensions to study how they can jointly support score construction. The dataset comprises 150 carefully curated Chinese paintings, with 50 images in each of three categories: flower-and-bird, landscape, and figure painting. Each image is evaluated along five aesthetic dimensions: brush technique, coloration, composition, artistic conception, and overall quality. A pool of 15 domain experts is assigned across categories, with five experts per category for each protocol. Images and dimensions are held fixed across protocols, while the rating and preference panels differ within each category.

Under the rating protocol, experts use a five-point rating scale. They assign an integer score from 1 to 5 to each image and dimension according to the scoring rubric, producing 3,750 raw scores. Under the preference protocol, all five experts in each category panel annotate the same 612 image pairs, selecting A WINS, TIE, or B WINS for each dimension and yielding 45,900 judgments. Annotation details and the pairwise budget analysis are provided in Appendix A.

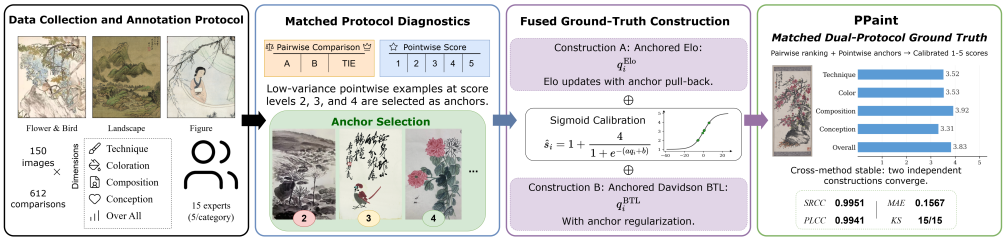


Figure 2: Expert preference-rating fusion for PPAINT. Low-disagreement ratings at levels 2, 3, and 4 anchor quality estimates derived from pairwise preferences. Anchored Elo and anchored Davidson Bradley–Terry (BTL) share anchors and sigmoid calibration to produce scores on the 1–5 expert rating scale. SRCC/PLCC and MAE measure cross-method score agreement, averaged over 15 category–dimension groups. KS counts groups where equal score distributions are not rejected by a two-sample Kolmogorov–Smirnov test ($\alpha = 0.05$). Anchored Elo defines evaluation targets.

3.2 MATCHED PROTOCOL DIAGNOSTICS

We examine inter-expert agreement and ranking stability using four diagnostics (Table 2). Results are averaged over 15 category–dimension groups, with five experts per protocol and category. Definitions of all four protocol diagnostics are provided in Appendix B.1.

Table 2: Protocol diagnostics under matched annotation (averages over 15 category–dimension groups). Higher values indicate stronger inter-expert agreement or ranking stability. Relative gap is $(P - R)/R \times 100\%$, where P and R denote Preferences and Ratings.

Diagnostic	Preferences	Ratings	Relative gap
Kendall's W (expert consensus)	0.795	0.697	+14.1%
Split-half Spearman (ranking stability)	0.836	0.757	+10.4%
Triplet separation (weak-label ordering)	0.92	0.86	+7.0%
To-consensus PRA (annotator–group agreement)	0.787	0.694	+13.4%

Table 2 shows higher average agreement and ranking stability for preferences than for ratings. The Kendall's W difference is positive in 14 of the 15 groups. Additional analyses in a common judgment space and under annotator subsampling support this pattern under the collection design (Appendices B.4 and B.5). However, this average advantage does not imply agreement on every image pair. Among pairs with directional disagreement in the rating panel, about 44% receive a clear preference majority and about 31% remain split (Appendix B.6).

Pairwise preferences specify relative orderings but do not determine scores on the expert rating scale. We use low-disagreement ratings as anchors to supply this scale information (§3.3).

3.3 ANCHORED PREFERENCE-RATING FUSION

We construct the fused expert ground truth by combining preference-derived rankings with low-disagreement rating anchors (Figure 2). **Preferences supply relative quality evidence, while rating anchors align the resulting scores with the expert rating scale.** We implement this construction with anchored Elo and anchored Davidson Bradley–Terry, using the same anchors to examine consistency between these two distinct preference-based ranking estimators.

All constructions share the same anchor set and final calibration. Let $\mathcal{A}$ denote the anchor set and $\bar{s}_i$ the mean expert rating for anchor image i. The calibrated aesthetic score is then obtained by a shared sigmoid mapping,

$$\hat{s}_i = 1 + 4 \cdot \sigma(aq_i + b), \tag{1}$$

where $q_i \in \mathbb{R}$ is the latent quality of image i estimated from expert preferences, and a and b are fitted to anchors pooled across all category–dimension groups to align scores with the expert rating scale.

Construction A: Anchored Elo. We estimate latent qualities through iterative Elo updates (Elo, 1978). After each full pass over the observed pairs, anchor images are regularized toward their latent-space targets:

$$q_i \leftarrow q_i + \alpha(\bar{q}_i - q_i), \quad i \in \mathcal{A}, \tag{2}$$

where $\bar{q}_i$ is the level-mapped Elo target corresponding to $\bar{s}_i$ and α controls anchor strength. The calibrated aesthetic score is then obtained by Eq. (1).

Construction B: Anchored Davidson Bradley–Terry. To assess sensitivity to the ranking estimator, we fit a Davidson extension of Bradley–Terry modeling TIE outcomes and uses the same anchor set and sigmoid calibration. Appendix C provides implementation details for both constructions and the anchor selection procedure, together with cross-method comparisons and sensitivity analyses for Elo hyperparameters and anchor selection.

Anchored Elo and anchored Davidson Bradley–Terry produce nearly identical rank orders ($\mathrm{SRCC} = 0.9951$), calibrated scores ($\mathrm{PLCC} = 0.9941$), and induced pairwise decisions (98.5% agreement) under the same anchor constraints. Without anchors, the mean absolute difference between the Elo and Davidson Bradley–Terry scores is 0.282. With anchors, it is 0.157 (Appendix C.3). We adopt the anchored Elo output as the canonical fused expert ground truth for subsequent experiments.

4 PSDISTILL: PREFERENCE-TO-SCORE SELF-DISTILLATION

Because expert annotation is costly at training scale, we use expert ratings and preferences only for evaluation. Training pseudo-labels come from the base VLM's own comparisons of unlabeled images. Following the preference-to-score principle of §3, PSDISTILL has two stages (Figure 3): an offline Elo-based reference pool converts VLM preferences into continuous pseudo-scores (§4.1), and online GRPO uses the weighted ranking reward to train the same VLM into a single-pass scorer (§4.2). This design reduces the score-scale collapse observed in Figure 1c.

4.1 PREFERENCE-TO-SCORE BRIDGE VIA ELO REFERENCE POOL

We adopt a reference-pool strategy inspired by Compare2Score (Zhu et al., 2024), but estimate the pool scores from the VLM's own pairwise judgments rather than from human-rated reference images.

We first sample $N{=}50$ unlabeled training images and perform exhaustive pairwise comparison within this pool, yielding $\binom{50}{2}{=}1{,}225$ pairs (Figure 3, Stage 1). Standard Elo updates (Elo, 1978) produce a five-dimensional rating for every pool image (Appendix C.1). The pool ratings are then frozen and serve as reference scores for placing new images on the same Elo-derived scale. Unlike the rating-derived anchors in §3.3, these references contain no information from the expert rating scale. Each remaining image is compared with only $K{=}10$ randomly sampled pool references, and we estimate its Elo rating from these sparse comparisons using Bayesian posterior inference.

Finally, we map raw Elo ratings to continuous pseudo-scores in the range $[1, 5]$ using a sigmoid:

$$\tilde{s}_i^{(d)} = 1 + \frac{4}{1 + \exp\left(-\left(\lambda \frac{q_i^{(d)} - q_{\min}^{(d)}}{q_{\max}^{(d)} - q_{\min}^{(d)}} - \lambda/2\right)\right)}, \quad \lambda = 6. \tag{3}$$

where $q_i^{(d)}$ is the raw Elo rating for image i on dimension d, and $q_{\min}^{(d)}$, $q_{\max}^{(d)}$ are the extrema across the training set. The factor 4 sets the width of the output range $[1, 5]$. We set $\lambda = 6$ so that the normalized Elo range is mapped to the sigmoid input interval $[-3, 3]$, which keeps the extrema away from the exact score boundaries. The mapping preserves the Elo ordering and supplies continuous score gaps for the training weights. It does not by itself align these scores with the expert rating scale. It requires $O(N^2+nK)$ calls (approximately 31K versus 4.5M for exhaustive comparison), with the complete procedure in Algorithm 1 (Appendix D.1).

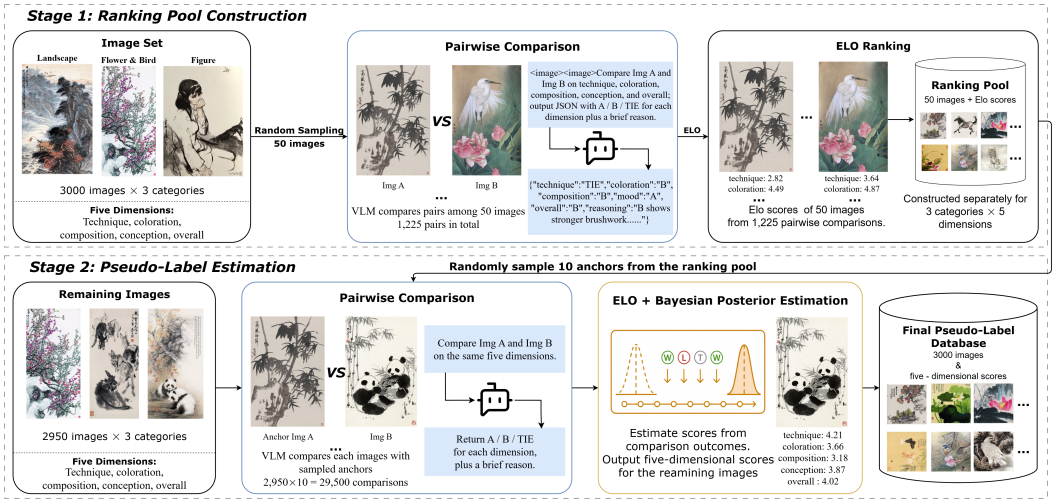


Figure 3: Offline pseudo-score construction in PSDISTILL. Stage 1 aggregates the base VLM's exhaustive pairwise comparisons among $N = 50$ sampled training images into five-dimensional Elo reference ratings. Stage 2 compares each remaining image with $K = 10$ pool references, then applies Bayesian inference and sigmoid mapping to obtain pseudo-scores in $[1, 5]$. The pool references, labeled anchors in the diagram, use no expert annotations. The procedure is applied separately to each training category; the main experiment uses flower-and-bird paintings.

4.2 SELF-DISTILLATION WITH A WEIGHTED FIDELITY REWARD

The continuous pseudo-scores provide a usable scale, but their point estimates remain noisy because each image is observed through only K reference comparisons. Rather than regress to these point estimates directly, we induce relative orderings and optimize a ranking-fidelity reward that preserves the more reliable ordinal signal while reducing sensitivity to pseudo-score noise.

Motivated by the ranking-fidelity objective of VisualQuality-R1 (Wu et al., 2025), we train the model to match target orderings via GRPO. The full objective is given in Appendix D.2.

Each candidate response produces a D=5 dimensional score vector. For a minibatch pair (i, z), we estimate the probability that sample k for image i ranks above image z under a Thurstone model:

$$p_{\text{pred}}^{(d)} = \Phi\left(\frac{\hat{s}_{i,k}^{(d)} - \bar{s}_z^{(d)}}{\sqrt{\sigma_i^{2(d)} + \sigma_z^{2(d)} + \epsilon}}\right), \tag{4}$$

where Φ is the standard normal CDF. The target probability is defined from the ordering of the bridge pseudo-scores:

$$p_{\text{gt}}^{(d)} = \begin{cases} 1 & \tilde{s}_i^{(d)} > \tilde{s}_z^{(d)}, \\ 0 & \tilde{s}_i^{(d)} < \tilde{s}_z^{(d)}, \\ 0.5 & \tilde{s}_i^{(d)} = \tilde{s}_z^{(d)}. \end{cases} \tag{5}$$

We measure agreement using the Bhattacharyya coefficient (Tsai et al., 2007):

$$R_{\text{fidelity}}^{(d)} = \sqrt{p_{\text{pred}}^{(d)} \cdot p_{\text{gt}}^{(d)}} + \sqrt{(1 - p_{\text{pred}}^{(d)})(1 - p_{\text{gt}}^{(d)})}. \tag{6}$$

We downweight near-tie comparisons using the absolute pseudo-score gap, with the weight reaching one at a gap of $\tau_w > 0$:

$$w_{iz}^{(d)} = \min\left(\frac{|\tilde{s}_i^{(d)} - \tilde{s}_z^{(d)}|}{\tau_w},\ 1\right), \tag{7}$$

We aggregate the weighted fidelity across comparison partners and dimensions:

$$R_{\text{rank}}^{(i,k)} = \frac{1}{D}\sum_{d=1}^{D} \frac{\sum_{z\neq i} w_{iz}^{(d)} \cdot R_{\text{fidelity},iz}^{(d)}}{\sum_{z\neq i} w_{iz}^{(d)}}. \tag{8}$$

The score gap is a heuristic proxy for comparison reliability, rather than a Bayesian posterior confidence estimate. The resulting reward drives GRPO. The reference pool and pairwise VLM queries are used only during offline pseudo-label construction. At inference, the trained model scores each image independently in a single forward pass without consulting the reference pool, so total scoring cost grows linearly with the number of images.

5 Experiments

5.1 Experimental Setup

Datasets and evaluation. We train on the unlabeled flower-and-bird split of PPAINT ($\sim$3,000 images) without human annotations, using landscape and figure painting to test cross-category generalization (Appendix D.5). The pretrained model generates pseudo-labels once before training (§4.1). We evaluate zero-shot on all 150 expert-annotated test images across three categories and five aesthetic dimensions. MD5 hashing confirms no training/test image duplicates. We evaluate predictions against the fused expert ground truth (§3.3), using SRCC to measure agreement in image ordering and PLCC to measure linear association between predicted and target scores. Absolute score errors are assessed separately using MAE and RMSE in Appendix D.8. All evaluations use greedy decoding and a single forward pass per image.

When multiple checkpoints are evaluated, we select the checkpoint with the highest mean SRCC on the evaluation set and report its performance on the same set. We also select the reference-pool seed for the main 8B model based on evaluation performance. For APDDv2, selection uses the mean SRCC across the oil-painting and sketching test subsets.

Competing methods. We evaluate seven closed-source and four open-source general-purpose VLMs by direct pointwise scoring (Table 3), including the Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Zhu et al., 2025) families. Closed-source model versions are reported in Appendix D.4. Task-oriented baselines include ArtiMuse (Cao et al., 2025) for multi-dimensional image aesthetic assessment and EvoQuality-style majority-vote preference distillation (Wen et al., 2025) for image quality assessment. We evaluate ArtiMuse using the authors' released checkpoint without fine-tuning, and retrain the EvoQuality-style baseline and PW-Distill on the same Qwen3-VL-8B backbone and unlabeled flower-and-bird corpus as PSDISTILL. As a training-free reference, we evaluate Qwen3-VL-8B with full pairwise Elo aggregation (Elo, 1978), requiring $O(n^2)$ test-time VLM calls.

Implementation details. We adopt Qwen3-VL-8B-Instruct (Bai et al., 2025) as the policy backbone and fine-tune it via GRPO with DeepSpeed ZeRO-3 on 8$\times$NVIDIA A100 80GB GPUs. We sample G=6 responses per image. Optimization uses AdamW with learning rate 1×10^{-6} and gradient clipping at 1.0. The effective batch size is 64 images per step. Training runs for 10 epochs with maximum completion length 512 tokens. Prompt templates appear in Appendix D.3.

5.2 Main Results

PSDISTILL reaches a mean SRCC of 0.709, a 40.7% relative improvement over its base and the highest point estimate among open-source single-pass pointwise scorers with full five-dimensional outputs in this main comparison, compared with the dedicated aesthetic-assessment model ArtiMuse (0.686) and the larger Qwen3-VL-32B (0.663). ArtiMuse performs strongly on flower-and-bird paintings, but its mean PLCC is also lower (0.638 vs. 0.709). PSDISTILL comes within 0.02 SRCC of Gemini-3.1-Pro on flower-and-bird (0.727 vs. 0.744) and has a higher point estimate than all

closed-source pointwise models on figure painting, while also surpassing the training-free pairwise Elo reference in mean SRCC (0.638) at single-pass inference cost. The remaining point-estimate gap to the closed-source leader is concentrated in landscape, absent from training data, yet landscape SRCC still rises from 0.453 to 0.686 (51.4% relative improvement) and figure painting from 0.568 to 0.713 (25.5% relative improvement), indicating cross-category generalization of the distilled signal.

Table 3: Aesthetic scoring on PPAINT against expert targets combining preferences and rating anchors. SRCC/PLCC measure rank/linear correlation (higher is better), averaged over five dimensions per category and then across categories. HN/RW/SS denote flower-and-bird/figure/landscape. Brackets give 95% CIs for mean SRCC from 2,000 category-stratified image bootstrap resamples, retaining all dimensions. $*$: base model. †: pairwise Elo with quadratic inference cost. ‡: shared Thurstone-based GRPO with different pseudo-label sources. **Bold** marks the best closed-source result; underlining marks the best open-source or self-distillation result with full five-dimensional scoring.

	SRCC ↑					PLCC ↑			
Method	**HN**	**RW**	**SS**	**Mean**		**HN**	**RW**	**SS**	**Mean**
Closed-source VLMs (direct pointwise)									
Gemini-3.1-Pro	**0.744**	0.688	**0.807**	**0.746**	[0.649, 0.803]	**0.816**	**0.774**	0.826	**0.805**
Gemini-3-Flash	0.734	0.666	0.801	0.733	[0.629, 0.795]	0.779	0.741	**0.834**	0.785
Claude-Sonnet-4.6	0.669	**0.698**	0.732	0.700	[0.613, 0.758]	0.710	0.711	0.720	0.714
Qwen3.6-Plus	0.718	0.566	0.660	0.648	[0.549, 0.709]	0.770	0.627	0.686	0.694
Doubao-Seed-1.6	0.534	0.631	0.598	0.587	[0.479, 0.665]	0.541	0.633	0.601	0.591
GPT-5.4-Mini	0.492	0.622	0.487	0.534	[0.427, 0.614]	0.501	0.630	0.462	0.531
Qwen3-VL-Plus	0.558	0.647	0.516	0.574	[0.484, 0.639]	0.503	0.635	0.441	0.526
Open-source models (direct pointwise)									
Qwen3-VL-32B	0.644	0.734	0.610	0.663	[0.576, 0.728]	0.660	0.714	0.619	0.664
ArtiMuse	0.797	0.587	0.675	0.686	[0.607, 0.759]	0.758	0.564	0.592	0.638
Qwen3-VL-8B*	0.492	0.568	0.453	0.504	[0.405, 0.583]	0.485	0.580	0.472	0.512
InternVL3.5-38B	0.322	0.555	0.233	0.370	[0.261, 0.455]	0.314	0.527	0.248	0.363
InternVL3.5-8B	0.259	0.176	0.329	0.255	[0.151, 0.347]	0.249	0.171	0.288	0.236
Pairwise aggregation (training-free, $O(n^2)$ inference)									
Qwen3-VL-8B†	0.665	0.641	0.609	0.638	[0.555, 0.696]	0.692	0.617	0.583	0.630
Self-distillation (single-pass inference)									
EvoQuality‡	0.644	0.692	0.549	0.628	[0.537, 0.692]	0.643	0.680	0.542	0.622
PW-Distill‡	0.523	0.689	0.459	0.557	[0.475, 0.617]	0.526	0.683	0.459	0.556
PSDISTILL (ours)	0.727	0.713	0.686	0.709	[0.637, 0.755]	0.743	0.693	0.692	0.709

Within the self-distillation block of Table 3, the three pseudo-label sources exhibit a clear ordering: Calibrated Elo attains 0.709 mean SRCC, majority vote 0.628, and PW-Distill 0.557. This ordering mirrors the protocol complementarity observed in expert annotation (§3.2): majority vote preserves the ordinal direction of preferences but discards score magnitude, whereas PW-Distill retains a nominal scale yet inherits the calibration noise of raw pointwise outputs. Calibrated Elo combines a global preference-derived ordering with an explicit score-scale mapping, producing pseudo-labels of substantially finer granularity than either alternative (Appendix D.1, Table 16).

5.3 ABLATION STUDIES

Generalization beyond Chinese painting. To test whether the training method applies beyond Chinese painting, we train PSDISTILL separately on the oil-painting and sketching training subsets of APDDv2 (Jin et al., 2024) and evaluate each model on both test subsets (Appendix E.1). Mean SRCC improves by 6.0% to 6.6% over EvoQuality and by 22.9% to 31.7% over the base model, with gains in both cross-domain directions (Table 4).

Robustness to fused expert ground-truth construction. Beyond the fused expert ground truth above, we evaluate PSDISTILL against pure human pointwise targets to test whether the gains hinge on the choice of target construction. The relative ordering among self-distillation variants is preserved under both targets, with mean SRCC reaching 0.621 against the rating mean and 0.594 against the rating median (Appendix D.8, Table 22). On the expert rating scale, PSDISTILL also achieves a mean absolute error approximately 40% lower than that of the next-best baseline against the fused

Table 4: Cross-domain evaluation on APDDv2. Best epoch per configuration is selected by mean SRCC. All models start from Qwen3-VL-8B-Instruct. **Bold** marks the highest mean SRCC and mean PLCC across all evaluated training configurations.

		SRCC ↑			PLCC ↑		
Method	**Train**	**Oil**	**Sketch**	**Mean**	**Oil**	**Sketch**	**Mean**
Base (pretrained)	–	0.540	0.477	0.508	0.552	0.495	0.524
EvoQuality	Oil	0.683	0.574	0.628	0.670	0.575	0.623
EvoQuality	Skt	0.620	0.560	0.590	0.627	0.533	0.580
PSDISTILL (ours)	Oil	0.752	0.587	**0.670**	0.746	0.574	**0.660**
PSDISTILL (ours)	Skt	0.690	0.559	0.625	0.678	0.551	0.615

expert ground truth (Table 23). The mean and median targets summarize direct expert ratings without pairwise fusion. The gains under both targets indicate that the improvement extends to direct expert ratings, rather than depending solely on the combination of pairwise preferences and rating anchors used to construct the benchmark's fused expert evaluation target.

Effect of training-source category. We train PSDISTILL and the EvoQuality-style majority-vote baseline (Wen et al., 2025) on each category (Appendix D.5, Table 19). PSDISTILL achieves higher per-category and mean SRCC/PLCC for every training source. Its mean SRCC varies by only 0.019 across training sources, compared with 0.050 for majority vote and 0.053 for PW-Distill. Figure-painting training gives the highest mean SRCC (0.715) and strong landscape transfer (0.754).

Effect of pair weighting. The pair weight in Eq. 7 uses the pseudo-score gap as a proxy for reliability and reduces the contribution of nearly tied pairs. To isolate its effect, we train two variants under identical ranking-pool seed and optimization budget, with and without the weight applied. Per-category results are reported in Appendix D.6 (Table 20). Pair weighting raises mean SRCC from 0.660 to 0.709, a relative improvement of 7.4%, and produces gains on all three categories. The largest gain appears on landscape ($+0.073$ SRCC, 11.9% relative), the category for which a larger fraction of pseudo-label pairs falls near the decision boundary. Both variants use identical offline pseudo-scores, so this comparison isolates the effect of pair weighting. The improvement supports downweighting comparisons with small pseudo-score differences, where noise in the pseudo-scores can more easily change the resulting pairwise ordering.

6 CONCLUSION

Under the annotation design studied here, preferences yield more consistent ordinal rankings while high-agreement ratings anchor scores to the expert rating scale, and the two signals carry complementary information for subjective aesthetic evaluation. PPAINT instantiates this insight as a dense matched dual-protocol benchmark for Chinese painting aesthetics, and the proposed anchored fusion produces a calibrated fused expert ground truth on which two distinct preference-to-score procedures converge to nearly identical scores. PSDISTILL extends the same principle to label-free VLM training. Using only model-generated supervision, the distilled Qwen3-VL-8B raises mean SRCC from 0.504 to 0.709 across all three painting categories, outperforming the same-backbone distillation baselines. These gains extend to unseen categories, with one forward pass per image at inference.

Limitations and future work. Collecting dense pairwise preferences and pointwise ratings across five aesthetic dimensions is costly in expert time, limiting the current PPAINT benchmark to 150 Chinese paintings across three categories. More efficient pair selection could support larger collections while preserving ranking stability and cross-method agreement. Within each category, different expert panels provide ratings first and preferences later, confounding protocol effects with panel composition and prior annotation experience. Future studies should control panel composition and counterbalance protocol order. The annotations reflect Chinese painting specialists' judgments under the adopted rubric. Whether the same preference–rating complementarity holds for other artistic traditions and viewer populations remains to be established. Our APDDv2 experiments evaluate self-distillation on oil paintings and sketches, but do not extend the matched expert annotation study to these domains. Collecting both ratings and preferences across more domains and annotator backgrounds would more directly assess how well the fusion approach generalizes.

ETHICS STATEMENT

This work studies how annotation protocols shape aesthetic ground truth and uses Chinese painting as a focused domain for evaluation. The main positive impact is to make subjective evaluation targets more transparent and auditable by separating local preference information from rating-based scale calibration. This can help researchers diagnose when aesthetic models are learning stable expert structure rather than artifacts of a single annotation protocol. Potential risks include over-interpreting model scores as objective aesthetic judgments, using expert-derived scores to suppress legitimate diversity in taste, or transferring a domain-specific scoring rubric to settings where it is not appropriate. We mitigate these risks by stating the scope of the benchmark, documenting the expert-defined scoring rubrics and construction procedure, and treating fused scores as expert-derived evaluation targets rather than universal measures of artistic value.

PPAINT scores reflect judgments by a particular expert panel under a fixed five-dimensional scoring rubric. They should not be interpreted as assessments of a work's overall artistic value or its creator's ability or standing. To reduce reputational harm from decontextualized use, complete annotations and the mappings between works and scores will be provided through application-based controlled academic access. Public materials will not associate identifiable artworks with their PPAINT scores or publish work-level or artist-level rankings. Access terms will prohibit public disclosure or redistribution of these associations and rankings. A correction or removal request channel will be provided, with changes recorded in the release documentation. Image redistribution will follow the applicable copyright, licensing, and source terms described in Appendix A.1. Where redistribution is not permitted, we will provide source links rather than image files. Corresponding annotations and image-to-score mappings will remain subject to the controlled academic access described above.

AI USE STATEMENT

Generative AI tools assisted with literature searches, drafting and language editing, LaTeX formatting, and writing and running analysis code to check the numerical consistency of reported results. The authors take responsibility for the final text, references, analyses, claims, and research artifacts.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631.

Hritik Bansal, John Dang, and Aditya Grover. Peering through preferences: Unraveling feedback acquisition for aligning large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dKl6lMwbCy.

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

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025. doi: 10.48550/arXiv.2507.14533.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

Xin Jin, Qianqian Qiao, Yi Lu, Huaye Wang, Heng Huang, Shan Gao, Jianfei Liu, and Rui Li. APDDv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3274.

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. VILA: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10041–10051, 2023. doi: 10.1109/CVPR52729.2023.00968.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision – ECCV 2016*, volume 9905 of *Lecture Notes in Computer Science*, pp. 662–679. Springer, 2016. doi: 10.1007/978-3-319-46448-0_40.

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.

Ruihang Li, Leigang Qu, Jingxu Zhang, Dongnan Gui, Mengde Xu, Xiaosong Zhang, Han Hu, Wenjie Wang, and Jiaqi Wang. GenArena: How can we achieve human-aligned evaluation for visual generation tasks? *arXiv preprint arXiv:2602.06013*, 2026. doi: 10.48550/arXiv.2602.06013.

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, 2012. doi: 10.1109/CVPR.2012.6247954.

Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafal K. Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019. doi: 10.1109/TIP.2019.2936103.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300.

Yixiao Song, Parker Riley, Daniel Deutsch, and Markus Freitag. Enhancing human evaluation in machine translation with comparative judgement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20536–20551, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1002. URL https://aclanthology.org/2025.acl-long.1002/.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. doi: 10.1109/TIP.2018.2831899.

Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. Pairwise or pointwise? evaluating feedback protocols for bias in LLM-based evaluation. In *Proceedings of the Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=uyX5Vnow3U.

Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 383–390. ACM, 2007. doi: 10.1145/1277741.1277808.

Wen Wen, Tianwu Zhi, Kanglong Fan, Yang Li, Xinge Peng, Yabin Zhang, Yiting Liao, Junlin Li, and Li Zhang. Self-evolving vision-language models for image quality assessment via voting and ranking. *arXiv preprint arXiv:2509.25787*, 2025. doi: 10.48550/arXiv.2509.25787.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025. doi: 10.48550/arXiv.2505.14460.

Hongji Yang, Yucheng Zhou, Wencheng Han, Songlian Li, Xiaotong Zhao, and Jianbing Shen. HanMoVLM: Large vision-language models for professional artistic painting evaluation. *arXiv preprint arXiv:2603.10814*, 2026a. doi: 10.48550/arXiv.2603.10814.

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. *arXiv preprint arXiv:2603.03907*, 2026b. doi: 10.48550/arXiv.2603.03907. Accepted to CVPR 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. doi: 10.1109/CVPR52729.2023.02144.

Wei Zhang, Jian-Wei Zhang, Kam-Kwai Wong, Yi-Fang Wang, Ying-Chao-Jie Feng, Lu-Wei Wang, and Wei Chen. Computational approaches for traditional chinese painting: From the "six principles of painting" perspective. *Journal of Computer Science and Technology*, 39(2):269–285, 2024. doi: 10.1007/s11390-024-3408-x.

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*, 2024. doi: 10.48550/arXiv.2404.09619.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*, 2024. doi: 10.48550/arXiv.2405.19298.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. doi: 10.48550/arXiv.2504.10479.

A Dataset and Annotation Protocol Details

A.1 Dataset Composition

PPaint consists of 150 expert-annotated paintings, with 50 paintings drawn from each of three categories (flower-and-bird, landscape, figure painting). Within each category, paintings are stratified into three quality tiers (good, medium, poor) based on a preliminary expert screening, and tier-internal selection covers a range of styles, periods, and techniques. Each painting receives pointwise ratings from five experts on five aesthetic dimensions, together with pairwise comparisons over 612 image pairs per category, also judged by five experts per pair. Aggregate annotation statistics are summarized in Table 5.

In addition to the 150-painting expert-annotated set, PPaint includes a pool of approximately 9,000 unlabeled paintings (about 3,000 per category) used to construct the pseudo-labels for self-distillation in §4. The public dataset page will provide documentation and an application for academic research access. Approved users will receive the complete raw expert ratings and preferences, majority-with-tie labels, fused scores, and image-to-score mappings for all 150 paintings. Images and corresponding

annotations will be provided for 51 coursework paintings and 69 paintings from museum open-access collections, subject to the applicable permissions and source terms. For the remaining 30 paintings, source or download links will be provided with the corresponding annotations. The unlabeled training pool is subject to the applicable source terms and redistribution permissions. Construction, training, and evaluation code will be provided separately.

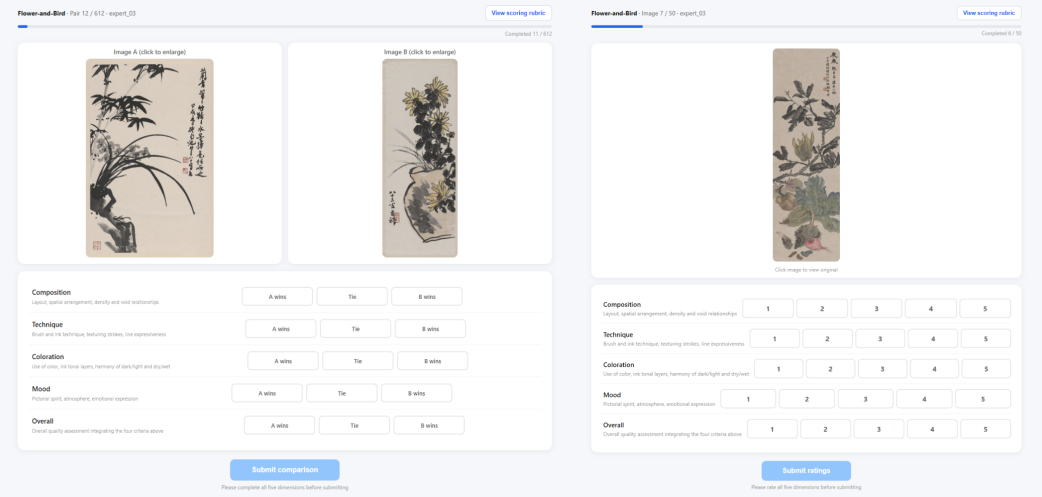
Table 5: PPAINT expert annotation counts and unlabeled image pool.

Quantity	Value
Expert-annotated paintings	150
Painting categories	3
Aesthetic dimensions per painting	5
Experts per category per protocol / total	5 / 15
Pointwise raw scores	3,750
Pairwise image pairs annotated	1,836
Pairwise expert judgments	45,900
Unlabeled image pool for pseudo-labels	∼9,000

A.2 ANNOTATION PROTOCOL

Annotators. A total pool of 15 domain experts is assigned across painting categories. Each category has five annotators per protocol, with different panels for rating and preference annotation. All annotators are research specialists in traditional Chinese painting affiliated with fine-arts academies, with extensive research or practical painting experience. Assignments are randomized across categories.

Interface. Both annotation tools are implemented as Flask web applications, illustrated in Figure 4. The pairwise interface (Figure 4a) displays two paintings from the same category side by side, and the annotator selects one of three options per dimension, A wins, Tie, or B wins. The pointwise interface (Figure 4b) displays one painting and asks the annotator to assign an integer score from 1 to 5 on each of the five aesthetic dimensions, with dimension descriptions shown alongside the buttons. Both tools enforce a minimum viewing time before submission is allowed (5 seconds per image for pointwise and 10 seconds per pair for pairwise), and require the annotator to read the scoring rubric before the first task (10-second minimum stay on the guidelines page).



(a) Pairwise annotation interface. **(b) Pointwise annotation interface.**

Figure 4: Expert annotation interfaces for Chinese painting aesthetics. (a) Pairwise preferences: A wins, Tie, or B wins for two paintings from the same category. (b) Pointwise ratings: integer scores from 1 to 5, with higher scores indicating better quality. Both cover brush technique, coloration, composition, artistic conception, and overall quality and provide category-specific scoring rubrics.

Pointwise score-scale calibration. Pointwise rating requires each annotator to internalize the expert rating scale, while pairwise comparison only needs relative judgments. To equalize the scale information available across the two protocols, each annotator previews a set of reference paintings before the pointwise phase begins. These references anchor the five levels of the expert rating scale for every aesthetic dimension on the assigned category and are drawn from outside the 150-image dataset to avoid contamination. They remain accessible through a side panel throughout the pointwise session, allowing annotators to revisit the anchor levels at any time. The preview provides explicit scale guidance for pointwise rating, reducing the burden of internalizing the expert rating scale.

Workflow and quality control. Annotation proceeds in two phases separated by a two-day rest period. The pointwise phase is completed first for all images, followed by the pairwise phase on the same image set. This ordering was chosen to collect absolute ratings before pairwise comparisons. Because experts are assigned across categories and the order is fixed, the design does not isolate protocol effects from panel composition or prior annotation experience. Three quantitative quality-control mechanisms are applied throughout. First, repeat annotations are inserted to measure intra-annotator consistency (10 repeated images per annotator with a minimum gap of 5 intervening items in the pointwise phase, and 30 repeated pairs per annotator with a minimum gap of 10 in the pairwise phase). Repeat-pair agreement rates range from 78.3% to 93.3% across annotators. Second, transitivity is verified for each annotator's pairwise judgments. The fraction of triples violating transitivity ranges from 0.10% to 4.47%, well below the chance level. Third, annotators whose rankings deviate most strongly from the group consensus are flagged for review. After the review, no annotator is excluded from the final dataset.

Annotation effort. Each judgment is logged with a millisecond-resolution timestamp. The per-annotator median time is 12 to 46 seconds per painting for pointwise rating and 21 to 28 seconds per pair for pairwise comparison. Aggregated across all 15 annotators, the total annotation effort amounts to approximately 115 expert-hours, of which roughly 7 hours are spent on pointwise rating and 108 hours on pairwise comparison.

Annotator consent and compensation. Annotators participated under an agreed expert-annotation arrangement and were informed that their judgments would be used for research and dataset construction. Compensation followed the agreed professional annotation rate, and the released dataset does not contain annotator identities or other personal information.

A.3 PAIRWISE ANNOTATION BUDGET

A complete pairwise protocol over the 50 paintings in each category requires $\binom{50}{2} = 1{,}225$ comparisons. The pairwise budget determines how densely expert preferences cover local quality relations within a category. Exhaustive annotation of all 1,225 pairs is costly, but reducing the budget too far removes comparisons needed to distinguish nearby-quality paintings and makes the recovered ranking depend heavily on indirect inference. PPAINT adopts a budget of 612 pairs per category, approximately half of the complete set, to retain dense local preference information while keeping expert workload manageable. This subsection provides empirical justification for the choice and characterizes the resulting comparison set.

As a budget validation check, we test whether this reduced budget retains sufficient local information for Elo ranking, using pairwise judgments from Qwen3-VL-8B as a proxy. For each category-dimension group, we first aggregate the full 1,225 pairwise judgments into a reference Elo score. We then sample comparison graphs at fractions ranging from 10% to 80% of the complete graph, re-run Elo aggregation on each subsample, and compare the resulting scores with the full-graph reference using SRCC and PLCC. As shown in Figure 5, the curves saturate well before the budget point. At 612 pairs the subsampled Elo scores reach 0.986 SRCC and 0.989 PLCC on average across category-dimension groups and random subsampling seeds, indicating that the chosen budget recovers nearly the full-graph ranking signal at roughly half the annotation cost.

We further characterize the connectivity of the 612-pair comparison graph by its diameter, defined as the longest shortest path between any two images. Across all three categories, the graph has diameter 2 and an average shortest path length of 1.47. Any two images are therefore connected by at most two

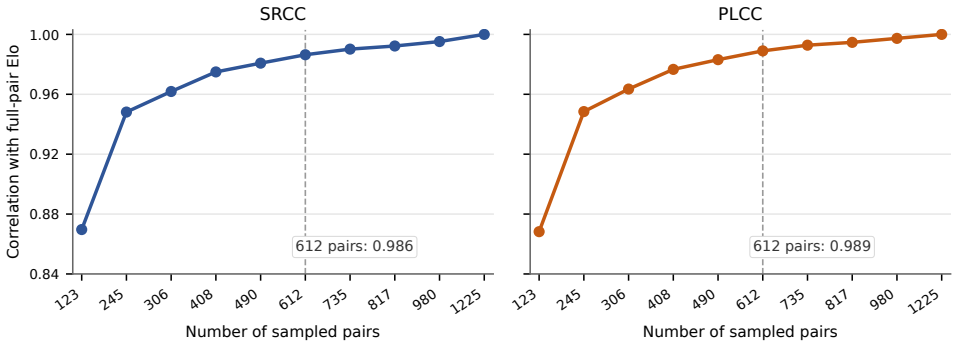


Figure 5: Pairwise annotation budget pilot. Using Qwen3-VL-8B pairwise judgments as a proxy, we subsample different fractions of the complete 1,225-pair graph, run Elo aggregation, and compare the resulting scores with the full-graph Elo reference. Curves report means over category-dimension groups and random subsampling seeds. The dashed line marks the 612-pair PPAINT budget.

intermediate comparisons, which limits the propagation of error along long transitive chains during Elo aggregation.

A.4 AESTHETIC DIMENSIONS AND SCORING RUBRIC

PPAINT evaluates each painting along five dimensions, namely brush technique (*jifa*), coloration (*shese*), composition (*goutu*), artistic conception (*yijing*), and overall quality. The five-dimension design is motivated by three considerations.

First, decomposing aesthetic evaluation into explicit dimensions reduces the ambiguity of a single holistic score. When annotators are asked for one overall rating, different experts implicitly weight different internal criteria, with one emphasizing brush technique while another emphasizes artistic conception, which produces apparent disagreement that reflects weighting differences rather than genuine quality disagreement. Five explicit dimensions make each criterion concrete and reduce ambiguity about the scoring rubric.

Second, multi-dimensional annotation provides a fair comparison between the two protocols. With five structured dimensions, pointwise annotators receive the same level of guidance as pairwise annotators. This provides a common scoring rubric for comparing the two protocols under the observed collection conditions.

Third, per-dimension analysis reveals where pairwise advantages are strongest. As reported in Appendix B.2, the most subjective dimensions benefit most from pairwise comparison, a finding that a single-dimension design could not surface.

The five dimensions are grounded in traditional Chinese painting theory, and each is defined by a category-specific scoring rubric with five anchored levels. As an illustrative example, the highest level for coloration in flower-and-bird painting is described as exhibiting "highly personal color use, whether the refreshing clarity of Yun Shouping or the vivid boldness of Qi Baishi, with color and ink fused in exceptional aesthetic harmony."

B PROTOCOL DIAGNOSTICS AND INTER-ANNOTATOR AGREEMENT

This section provides extended evidence for the matched-protocol findings reported in §3.2, organized into diagnostic definitions, per-dimension diagnostics, annotator-level calibration, per-judgment agreement, stability under annotator subsampling, and disagreement across protocols.

B.1 Diagnostic Definitions

Kendall's W. For Kendall's W, each category and dimension group is represented as five expert-specific rankings over the same 50 paintings. Rating rankings are induced directly from each expert's 1 to 5 scores, whereas preference rankings are obtained by aggregating that expert's pairwise judgments with Elo. Thus, the statistic compares protocol-induced expert rankings with identical items and panel sizes. The aggregation step is part of how pairwise annotations define a global ranking rather than an additional post hoc correction.

Split-half Spearman correlation. Split-half Spearman repeatedly splits the five experts into two halves and correlates the resulting half-group rankings, quantifying how stable the consensus ranking is under expert subsampling.

Weak-label triplet separation. Triplet separation uses weak quality tiers (high, medium, low) elicited from expert discussion on the overall dimension and measures how often the protocol-induced ranking correctly separates triplets sampled across these three tiers.

To-consensus pairwise ranking accuracy. To-consensus PRA builds a group consensus ranking from all five experts and then averages, across experts, the pairwise ranking accuracy of each expert's protocol-induced ranking against this consensus.

B.2 Aggregate and Per-Dimension Diagnostics

This subsection complements Table 2 (main paper) with two views, an aggregate visualization across categories and groups (Figure 6) and a per-dimension breakdown (Table 6).

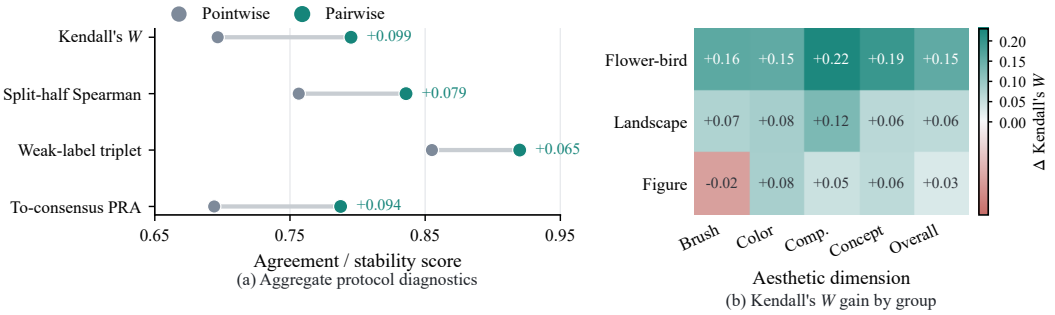


Figure 6: Protocol effects under matched annotation. **(a)** Pairwise judgments exceed pointwise scores on all four aggregate diagnostics reported in Table 2. **(b)** The Kendall's W advantage is broadly distributed, with 14 of 15 category-dimension groups showing a positive pairwise gain. Only figure-painting brush technique shows a negligible reversal.

Table 6 reports the mean pairwise-minus-pointwise gap in Kendall's W for each dimension, averaged across categories and computed against a per-image majority-with-tie reference derived from the five expert ratings. The pairwise advantage is positive for all five dimensions and for all 15 category-dimension groups. The magnitude of the gap varies, with overall quality and artistic conception showing the largest values and coloration the smallest. The resulting ordering, overall quality, artistic conception, composition, brush technique, and coloration, in descending order, is consistent with the intuition that dimensions with greater ambiguity in the scoring rubric benefit more from relative comparison than from absolute scoring.

B.3 Inter-Annotator Agreement and Calibration

We characterize annotator-level agreement under both protocols using three complementary measures: Spearman and Kendall correlation between annotator-induced rankings, the mean rank displacement between each annotator and the group consensus, and the dispersion of these quantities across annotators. Pairwise annotation produces higher inter-annotator similarity, lower mean rank displacement, and tighter dispersion than pointwise annotation, as summarized in Table 7.

Table 6: Mean pairwise-minus-pointwise Kendall's W gap by dimension, computed against a per-image majority-with-tie reference derived from the five expert ratings and averaged over the three painting categories. The gap is positive for every dimension across all 15 category-dimension groups.

Dimension	Mean gap	Median gap
Overall quality	0.275	0.221
Artistic conception	0.258	0.260
Composition	0.206	0.181
Brush technique	0.194	0.149
Coloration	0.181	0.175

The annotator-to-consensus Spearman ρ ranges from 0.777 to 0.923 under pairwise (spread of 0.146) versus 0.714 to 0.870 under pointwise (spread of 0.156). The standard deviation of the per-annotator rank error is also smaller under pairwise (0.99 versus 1.34). Beyond ranking-level agreement, pointwise annotators exhibit substantial scoring-pattern heterogeneity. The mean assigned score ranges from 2.30 to 3.65 across annotators, and the fraction of extreme scores (1 or 5) ranges from 10% to 81%. Pairwise annotators show less variation, with A-win rates ranging from 52.8% to 70.0% and tie rates from 0% to 25.7%. Higher tie rates correlate with lower consensus agreement, although this association does not establish whether ties reflect uncertainty or differing aesthetic judgments.

Table 7: Inter-annotator agreement and calibration under the two protocols. All quantities are averaged over 15 category-dimension groups, with ranges reported across the 5 annotators per panel.

Metric	Pairwise	Pointwise
Annotator-pair Spearman ρ	0.739	0.621
Annotator-pair Kendall τ	0.552	0.532
Mean rank error to consensus (out of 50)	5.38	6.64
Std of rank error across annotators	0.99	1.34
Annotator-to-consensus ρ range	[0.777, 0.923]	[0.714, 0.870]

B.4 Per-Judgment Agreement

The aggregate metrics in Table 7 compare ranking-level agreement after each protocol has been reduced to a per-image ranking. To measure agreement at the level of individual annotation events, we report two complementary analyses. The first projects both protocols onto a common pairwise judgment space by converting each annotator's pointwise scores into induced pairwise outcomes on the same 612 pairs (Table 8). The second compares raw 5-annotator agreement in each protocol's native annotation unit, that is, per-image ratings on five levels versus per-pair preferences on three outcomes (Table 9).

In the induced-pair view, preferences achieve higher agreement than induced pointwise judgments across all four metrics, with Fleiss' κ of 0.373 versus 0.306 (12 of 15 category-dimension groups positive). In the native-unit view, pairwise annotators reach unanimous agreement on 42.4% of pairs whereas pointwise annotators assign identical scores on only 5.7% of images. The label space differs between the two protocols (3 outcomes for preferences and 5 scores for ratings), so the native-unit numbers are reported as descriptive context rather than a strict comparison. The strict, label-space-matched comparison is the induced-pair view in Table 8.

Table 8: Per-judgment agreement on the same 612 pairs (averages over 15 category-dimension groups). Pointwise scores are converted to induced pairwise outcomes (A wins / Tie / B wins) before computing agreement, placing both protocols in a common label space.

Metric	Preferences	Ratings (induced)	Relative gap
Fleiss' κ	0.373	0.306	+22.1%
Krippendorff's α	0.373	0.309	+20.7%
Unanimous (5/5)	42.4%	28.7%	+47.7%
$\geq$4/5 agreement	65.6%	52.7%	+24.5%

Table 9: Raw 5-annotator agreement in each protocol's native annotation unit. Preferences are evaluated per-pair (3 labels: A wins / Tie / B wins) and ratings are evaluated per-image (5 labels: scores 1–5). Native-unit numbers are descriptive only because the label spaces differ. Table 8 compares the two protocols in a common pairwise label space.

Protocol	Unit	Labels	Unanimous (5/5)	$\geq$4/5
Preferences	per-pair	3	42.4%	65.6%
Ratings	per-image	5	5.7%	23.1%

B.5 Stability Under Annotator Subsampling

To evaluate how robust each protocol is to the size of the annotator panel, we conduct two subsampling analyses. The first removes one annotator at a time and measures the Spearman correlation between the leave-one-out ranking and the full-panel ranking. The second varies the subsample size $k \in \{1, 3, 5\}$ and computes the mean Spearman correlation between subsample rankings and the full-panel ranking, averaged over all category-dimension groups and all $\binom{5}{k}$ subsets.

Table 10 reports the results. Pairwise rankings are more robust to single-annotator removal, with mean Spearman $\rho = 0.975$ versus 0.953 for pointwise and a higher Top-10 overlap (0.856 versus 0.707). Pairwise rankings have higher agreement with their full-panel reference at both the single-annotator and three-annotator settings.

Table 10: Stability of induced rankings under annotator subsampling. Leave-one-out values are averaged over 15 category-dimension groups and 5 leave-out trials each. Subsample values are averaged over all $\binom{5}{k}$ subsets for $k \in \{1, 3\}$. The k=5 case is omitted because it trivially reproduces the full-panel reference for pointwise (deterministic averaging) and reaches ρ=0.991 for pairwise (Elo aggregation has minor run-to-run non-determinism).

Stability metric	Pairwise	Pointwise
Leave-one-out Spearman ρ	0.975	0.953
Leave-one-out Top-10 overlap	0.856	0.707
1-annotator subsample, mean ρ	0.860	0.791
3-annotator subsample, mean ρ	0.947	0.910

B.6 Disagreement Across Protocols

For each annotated image pair and dimension, we compare the rating panel's induced ordering with the preference panel's direct votes. A panel is classified as directionally split when each direction receives at least two votes and the majority accounts for at most 60% of non-tie votes. A clear majority requires at least 80% of non-tie votes to favor one direction. Cases with only tie votes are classified separately, and the remaining cases are intermediate. These definitions describe panel agreement on a common set of pairs without assuming that the same experts supplied the two protocols.

Across 9,180 pairs and dimensions, the rating panel is directionally split in 6.3% of cases. Among these cases, the preference panel reaches a clear majority in about 44% and remains split in about 31%, with the rest intermediate. The fraction that remains split ranges from 27% to 34% across the five dimensions. The mean majority fraction among non-tie preference votes also falls from 0.91 on pairs with a clear rating majority to 0.76 on pairs where ratings are split. Thus, the aggregate score coexists with observable expert disagreement.

These patterns are consistent with differences in numerical scale use accounting for part of the rating disagreement, while other differences persist under direct comparison. Since the two protocols use different panels within each category, this analysis does not identify the cause of any individual disagreement. The fused score represents an aggregate under the specified scoring rubric, and the per-expert annotations retain the information needed to study differing judgments.

C Fused Expert Ground-Truth Construction Details

C.1 Anchored Elo Construction

The canonical fused expert ground-truth construction estimates latent qualities through iterative Elo updates (Elo, 1978). For each observed pair (i, j), the expected score is

$$E_{ij} = \sigma\big((q_i - q_j)/\tau\big), \tag{9}$$

where τ is a temperature parameter. After observing outcome $o_{ij} \in \{1,\, 0.5,\, 0\}$ (win, tie, loss for image i), the ratings are updated as

$$q_i \leftarrow q_i + K\,(o_{ij} - E_{ij}), \tag{10}$$

$$q_j \leftarrow q_j + K\,(1 - o_{ij} - E_{ji}), \tag{11}$$

with step size K. After each full pass through all pairs, anchor images are pulled back toward their latent-space targets:

$$q_i \leftarrow q_i + \alpha\,(\bar{q}_i - q_i), \quad \forall\, i \in \mathcal{A}, \tag{12}$$

where α controls the anchor strength and $\bar{q}_i$ is the level-mapped Elo target corresponding to $\bar{s}_i$. This interleaving of Elo updates and anchor pull-back grounds the latent scale while letting pairwise data determine the ranking. The same sigmoid calibration (Eq. (1)) maps the latent scores to the expert rating scale.

In the implementation, the step size K takes a smaller value on pairs that involve at least one anchor image, which prevents anchor ratings from drifting too far from their targets during local updates. The step size also decays geometrically across passes to stabilize the final ratings. The complete set of hyperparameters is listed in Table 11.

Table 11: Hyperparameters of the anchored Elo construction.

Component	Hyperparameter	Value
Initialization	Initial rating r_0	1500
Pair update	Step size on non-anchor pairs K	32
Pair update	Step size on anchor-involving pairs K_a	6
Pair update	Per-pass step decay	0.995
Anchor pull-back	Anchor strength α	0.15
Anchor target	Elo gap between adjacent anchor levels	400
Optimization	Number of passes	150
Optimization	Pair-shuffling random seed	42

Anchor selection. For each category-dimension group, two anchor images are selected at each of the score levels 2, 3, and 4. Levels 1 and 5 are excluded to avoid boundary effects on the calibration target. Within each level, the candidates with the lowest inter-expert standard deviation are chosen. The total number of anchors per group ranges from 3 to 6 depending on data availability. Groups with fewer low-variance candidates at a given level (for example, figure painting composition) are harder to calibrate but still converge across the two anchored constructions.

C.2 Anchored Davidson Bradley–Terry Construction

As an independent construction, we model pairwise A WINS / TIE / B WINS outcomes with the Davidson extension of Bradley–Terry (Davidson, 1970). For a pair (i, j),

$$P(i \succ j) = \frac{\exp(q_i)}{\exp(q_i) + \exp(q_j) + \nu \exp\big((q_i+q_j)/2\big)}, \tag{13}$$

$$P(i \sim j) = \frac{\nu \exp\big((q_i+q_j)/2\big)}{\exp(q_i) + \exp(q_j) + \nu \exp\big((q_i+q_j)/2\big)}, \tag{14}$$

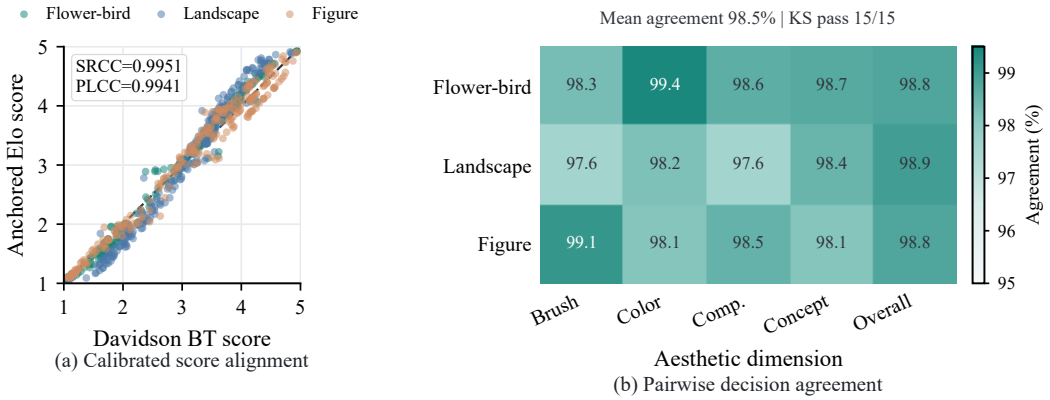
with $P(j \succ i)$ defined symmetrically and ν capturing the propensity of experts to regard two paintings as indistinguishable. The latent qualities are estimated by maximizing the pairwise log-likelihood with an anchor regularizer:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} \sum_{(i,j)} \log P(o_{ij} \mid q_i, q_j, \nu) \;-\; \lambda \sum_{i \in \mathcal{A}} (q_i - \bar{q}_i)^2, \tag{15}$$

where o_{ij} is the observed outcome and λ controls the strength of the anchor constraint. The subsequent sigmoid calibration (Eq. (1)) maps the latent scores to the expert rating scale. The tie parameter ν is estimated jointly with $\mathbf{q}$ during MAP optimization. Anchor targets are obtained by re-centering the corresponding pointwise medians, $\bar{q}_i = \bar{s}_i - 3$, so that the midpoint of the expert rating scale aligns with zero. We set $\lambda = 0.1$ and optimize the objective with L-BFGS-B for at most 2,000 iterations.

C.3 Cross-Method Agreement

This subsection reports cross-method agreement of the fused expert ground truth from three complementary views, an aggregate visualization (Figure 7), a four-method pairwise comparison (Table 12), and a per-group breakdown of the two anchored constructions (Table 13).



(a) Calibrated score alignment

(b) Pairwise decision agreement

Figure 7: Cross-method agreement of the fused expert ground truth. **(a)** Anchored Elo and anchored Davidson Bradley–Terry produce nearly identical calibrated scores (SRCC $= 0.9951$, MAE $= 0.1567$ on the expert rating scale (1–5)). **(b)** Their induced pairwise decisions agree at 98.5% averaged over all 15 painting category-dimension groups.

Table 12 reports the full pairwise agreement among four fused expert ground-truth construction variants, the two anchored methods used in this paper (Anchored Elo and Anchored Davidson Bradley–Terry) together with their unanchored counterparts (Unanchored Elo and Unanchored Davidson Bradley–Terry). The two anchored methods show the strongest agreement, with SRCC = 0.9951, MAE = 0.157, and KS non-rejection in all 15 groups. The comparison between the two unanchored estimators gives MAE = 0.282 and KS non-rejection in 8/15 groups. Comparisons of anchored Elo with unanchored Elo and unanchored Davidson Bradley–Terry retain high rank correlations but yield larger MAE values of 0.343 and 0.517, with KS non-rejection in 12/15 and 3/15 groups, respectively. These comparisons support the role of rating-based anchors in stabilizing agreement on the expert rating scale.

Table 12: Agreement among expert score constructions on PPaint. Anchored variants constrain latent qualities using low-disagreement ratings; unanchored variants omit these constraints. SRCC/PLCC compare score rankings/linear agreement; MAE measures absolute differences on the calibrated 1–5 scale. Decision is agreement on induced A wins/Tie/B wins outcomes. These metrics average over 15 category–dimension groups; higher correlations and Decision, and lower MAE, indicate stronger agreement. KS non-rejection counts groups where a two-sample Kolmogorov–Smirnov test does not reject equality of the two calibrated score distributions ($\alpha = 0.05$).

Method pair	SRCC	PLCC	MAE	Decision	KS non-rejection
Anchored Elo vs Anchored Davidson Bradley–Terry	0.9951	0.9941	0.157	98.5%	15/15
Anchored Davidson Bradley–Terry vs Unanchored Davidson Bradley–Terry	0.9970	0.9899	0.435	98.8%	5/15
Anchored Davidson Bradley–Terry vs Unanchored Elo	0.9967	0.9960	0.262	98.6%	12/15
Anchored Elo vs Unanchored Davidson Bradley–Terry	0.9897	0.9773	0.517	97.5%	3/15
Anchored Elo vs Unanchored Elo	0.9896	0.9899	0.343	97.5%	12/15
Unanchored Davidson Bradley–Terry vs Unanchored Elo	0.9993	0.9918	0.282	99.5%	8/15

Table 13 reports the cross-method agreement between Anchored Elo and Anchored Davidson Bradley–Terry for each of the 15 category-dimension groups. For each of the 15 groups, the two-sample KS test does not reject the null hypothesis of equal distributions at $\alpha = 0.05$. All groups also have SRCC ≥ 0.989 and decision agreement $\geq 97.6\%$, confirming that the close agreement summarized in Figure 7 holds uniformly across categories and dimensions rather than being driven by a small number of well-behaved groups.

Table 13: Per-group cross-method agreement between Anchored Elo and Anchored Davidson Bradley–Terry. The two-sample KS test does not reject equal distributions in any group at $\alpha = 0.05$.

Category	Dimension	SRCC	PLCC	MAE	KS p	Decision
HuaNiao	Composition	0.997	0.995	0.120	0.717	98.6%
HuaNiao	Brush technique	0.996	0.996	0.081	0.869	98.3%
HuaNiao	Overall quality	0.997	0.995	0.120	0.869	98.8%
HuaNiao	Coloration	0.999	0.997	0.162	0.272	99.4%
HuaNiao	Artistic conception	0.997	0.996	0.098	0.967	98.7%
RenWu	Composition	0.994	0.997	0.185	0.272	98.5%
RenWu	Brush technique	0.998	0.998	0.074	0.869	99.1%
RenWu	Overall quality	0.995	0.996	0.119	0.717	98.8%
RenWu	Coloration	0.989	0.995	0.133	0.717	98.1%
RenWu	Artistic conception	0.994	0.994	0.078	0.967	98.1%
ShanShui	Composition	0.994	0.988	0.258	0.396	97.6%
ShanShui	Brush technique	0.989	0.991	0.206	0.112	97.6%
ShanShui	Overall quality	0.997	0.992	0.225	0.717	98.9%
ShanShui	Coloration	0.994	0.992	0.196	0.717	98.2%
ShanShui	Artistic conception	0.996	0.989	0.294	0.396	98.4%

C.4 Elo Gap Calibration Ablation

The anchor Elo gap controls the spread between adjacent anchor levels in the latent target space and therefore acts as a calibration hyperparameter. Table 14 reports cross-method agreement between Anchored Elo and Anchored Davidson Bradley–Terry as the gap is varied. Reducing the gap from 1200 to 400 steadily improves all metrics. The final configuration with global sigmoid calibration and a gap of 400 yields KS non-rejection in all 15 groups and is adopted throughout the paper.

Table 14: Elo anchor spacing and calibration ablation on PPaint. Each row compares anchored Elo with anchored Davidson Bradley–Terry. Gap is the Elo distance between adjacent rating-anchor targets. GSIG fits one sigmoid across all 15 category–dimension groups; otherwise calibration is per group. SRCC/PLCC measure rank/linear agreement and MAE measures absolute score differences on the 1–5 scale. Decision is agreement on A wins/Tie/B wins outcomes. These metrics average over groups (SRCC/PLCC/Decision ↑; MAE ↓). KS non-rejection counts groups where a two-sample Kolmogorov–Smirnov test does not reject equal score distributions ($\alpha = 0.05$).

Config	SRCC	PLCC	MAE	Decision	KS non-rejection
Gap 1200 (baseline)	0.9852	0.9864	0.241	96.9%	12/15
Gap 900	0.9885	0.9911	0.187	97.4%	13/15
Gap 700	0.9917	0.9936	0.144	97.9%	13/15
Gap 500	0.9943	0.9950	0.128	98.4%	13/15
GSIG + Gap 350	0.9950	0.9932	0.185	98.4%	15/15
GSIG + Gap 400	0.9951	0.9941	0.157	98.5%	15/15
GSIG + Gap 450	0.9947	0.9945	0.142	98.4%	15/15

C.5 Sensitivity to Anchor Selection

We further vary the number and selection of expert rating anchors while keeping the remaining construction settings fixed within this analysis. Using constrained Davidson Bradley–Terry followed by sigmoid calibration, the reference selects the two candidates with the lowest inter-expert standard deviation at each of levels 2, 3, and 4. We compare this reference with one or three candidates per level, random selection of two candidates using three seeds, and selection of the two candidates with

the highest standard deviation. Where fewer candidates are available, all available candidates are used.

Table 15: Ranking agreement under alternative expert anchor selection rules. SRCC compares each variant with the default selection rule in the same construction. Mean and minimum values are computed over the 15 category and dimension groups.

Anchor selection	Mean SRCC	Minimum SRCC
Lowest standard deviation, 1 per level	0.9910	0.9736
Lowest standard deviation, 3 per level	0.9944	0.9818
Random, 2 per level, seed 1	0.9923	0.9823
Random, 2 per level, seed 2	0.9921	0.9858
Random, 2 per level, seed 3	0.9894	0.9639
Highest standard deviation, 2 per level	0.9875	0.9702

All tested variants retain high ranking agreement with the reference, with mean SRCC between 0.9875 and 0.9944. This supports ranking stability within the tested choices. Agreement between constructions describes sensitivity to those choices, while evaluation against pure human rating targets provides a separate check of model performance. These expert rating anchors are distinct from the model reference images used to construct training pseudo-labels in PSDISTILL.

D Self-Distillation Implementation and Extended Results

D.1 Pseudo-Label Generation

Table 16: VLM inference cost for offline pseudo-label construction. Each forward call returns scores or judgments for all five aesthetic dimensions. n is the number of training images, N the reference-pool size, and K the number of pool comparisons per remaining image. Counts use $n = 3{,}000$, $N = 50$, and $K = 10$. Exhaustive pairwise comparison is theoretical and was not executed.

Pseudo-label source	Total VLM forward calls
Majority vote (32 pairwise trials per image)	$32n = 96{,}000$
PW-Distill (raw VLM scores)	$n = 3{,}000$
PSDISTILL bridge (ours)	$\binom{N}{2} + (n - N)K = 30{,}725$
Exhaustive pairwise reference	$\binom{n}{2} = 4{,}498{,}500$

The offline preference-to-score bridge described in §4.1 produces continuous pseudo-scores for each unlabeled training image on each of the five aesthetic dimensions. Compared with exhaustive pairwise comparison, the bridge yields per-image pseudo-scores at modest VLM inference cost. Table 16 summarizes the offline VLM inference budget required by each pseudo-label source.

The PSDISTILL bridge is about 146 times cheaper than exhaustive pairwise comparison while still using pairwise judgments to construct per-image pseudo-scores. Algorithm 1 gives the complete procedure.

D.2 GRPO Training

We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for the online distillation stage (§4.2). For each image x_i with prompt c, the policy π_θ generates G candidate responses. Let $r_k(x_i)$ denote the reward for the k-th response (computed from $R^{(i,k)}_{\text{rank}}$ in Eq. 8 plus auxiliary format and range terms). The relative advantage is estimated by within-group normalization,

$$a_k(x_i) = \frac{r_k(x_i) - \mu(r(x_i))}{\sigma(r(x_i))}, \tag{16}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation of the G rewards for image x_i. The policy is updated via the clipped surrogate objective

$$\mathcal{L}(\theta) = \frac{1}{BG} \sum_{i=1}^{B} \sum_{k=1}^{G} \min(\rho_k \cdot a_k,\ \text{clip}(\rho_k, 1-\epsilon, 1+\epsilon) \cdot a_k) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}), \tag{17}$$

Algorithm 1 Preference-to-Score Bridge for Pseudo-Label Generation

Require: Unlabeled training images $\mathcal{X} = \{x_1, \ldots, x_n\}$, pretrained VLM $\mathcal{M}$, pool size N, comparisons per image K, aesthetic dimensions D=5

Ensure: Pseudo-scores $\{\tilde{s}_i^{(d)}\}$ for all $x_i \in \mathcal{X}$ and $d \in \{1, \ldots, D\}$

Stage 1: Ranking Pool Construction
Sample N images uniformly from $\mathcal{X}$ to form pool $\mathcal{P}$
for each pair (x_i, x_j) in $\mathcal{P}$, $i < j$ **do**
  Query $\mathcal{M}$ for A wins / Tie / B wins judgments across all D dimensions
end for
for each dimension $d = 1, \ldots, D$ **do**
  Run Elo updates over all $\binom{N}{2}$ pool pairs to obtain pool ratings $\{q_p^{(d)}\}_{p \in \mathcal{P}}$
end for
Freeze pool ratings

Stage 2: Pool-Based Pseudo-Scoring
for each remaining image $x_i \in \mathcal{X} \setminus \mathcal{P}$ **do**
  Sample K pool references without replacement
  for each sampled pool reference x_p **do**
    Query $\mathcal{M}$ for A wins / Tie / B wins judgments across all D dimensions
  end for
  for each dimension $d = 1, \ldots, D$ **do**
    Estimate $q_i^{(d)}$ via Bayesian posterior inference using K pairwise outcomes and pool ratings as prior
  end for
end for

Stage 3: Sigmoid Score Mapping
for each dimension $d = 1, \ldots, D$ **do**
  Compute $q_{\min}^{(d)}, q_{\max}^{(d)}$ across all images
  for each image $x_i \in \mathcal{X}$ **do**
    $\tilde{s}_i^{(d)} \leftarrow 1 + \frac{4}{1+\exp\left(-\left(\frac{6\,(q_i^{(d)} - q_{\min}^{(d)})}{q_{\max}^{(d)} - q_{\min}^{(d)}} - 3\right)\right)}$
  end for
end for
return $\{\tilde{s}_i^{(d)}\}$

where $\rho_k = \pi_\theta(o_k \mid c, x_i)/\pi_{\text{old}}(o_k \mid c, x_i)$ is the importance ratio, ϵ is the clipping threshold, and β controls the strength of KL regularization against the frozen reference policy π_{ref}. Training hyperparameters and the hardware configuration are listed in Table 17.

Table 17: Training configuration for the online GRPO stage.

Component	Setting	Value
Backbone	Base policy	Qwen3-VL-8B-Instruct
Sampling	Group size G	6
Optimization	Optimizer	AdamW
Optimization	Learning rate	1×10^{-6}
Optimization	Effective batch size	64 images per step
Optimization	Gradient clipping	1.0
Optimization	Training epochs	10
Generation	Maximum completion length	512 tokens
Inference	Decoding	greedy, single forward pass
Hardware	Accelerators	8 × NVIDIA A100 80GB
Hardware	Distributed strategy	DeepSpeed ZeRO-3

D.3 PROMPT TEMPLATES

Both the offline pseudo-label construction (§4.1) and the online GRPO training (§4.2) share a single prompt template, with the expert role embedded in the first line of the prompt itself rather than a separate system message. The pointwise variant assigns a continuous score in $[1.00, 5.00]$ on each of the five aesthetic dimensions and returns the answer in a fixed JSON schema. The pairwise variant presents the two paintings as the first and second images and selects A, B, or TIE on each of the same five dimensions. In the prompts and JSON outputs, the field `mood` corresponds to artistic conception. Both variants require the model to first emit its reasoning inside `<think> </think>` tags and then the final answer inside `<answer> </answer>` tags, matching the parser used by the reward function (Appendix D.2).

Pointwise prompt.

```
You are an expert in traditional Chinese painting criticism and
aesthetics.
Evaluate this Chinese painting on five criteria.  Each must be a float
from 1.00 to 5.00, rounded to two decimal places (1.00 = poor, 3.00 =
moderate, 5.00 = excellent).
1) technique  2) coloration  3) composition  4) mood  5) overall
First output the thinking process in ⟨think⟩ ⟨/think⟩ tags and then
output the final answer as valid JSON in ⟨answer⟩ ⟨/answer⟩ tags:
{"technique":  ⟨1.00-5.00⟩, "coloration":  ⟨1.00-5.00⟩, "composition":
⟨1.00-5.00⟩, "mood":  ⟨1.00-5.00⟩, "overall":  ⟨1.00-5.00⟩}
```

Pairwise prompt.

```
You are an expert in traditional Chinese painting criticism and
aesthetics.
Compare two Chinese paintings (A is the first image, B is the second
image) on five criteria.  For each criterion, output "A" if the first
painting is stronger, "B" if the second is stronger, or "TIE" if
genuinely equal.
1) technique  2) coloration  3) composition  4) mood  5) overall
First output the thinking process in ⟨think⟩ ⟨/think⟩ tags and then
output the final answer as valid JSON in ⟨answer⟩ ⟨/answer⟩ tags:
{"technique":  "A"/"B"/"TIE", "coloration":  "A"/"B"/"TIE",
"composition":  "A"/"B"/"TIE", "mood":  "A"/"B"/"TIE", "overall":
"A"/"B"/"TIE"}
```

D.4 CLOSED-SOURCE MODEL VERSIONS

Table 18 lists the closed-source model versions used in the main comparison.

Table 18: Closed-source model versions used in the main comparison.

Model	Version
Gemini-3.1-Pro	`gemini-3.1-pro-preview`
Gemini-3-Flash	`gemini-3-flash-preview`
Claude-Sonnet-4.6	`default`
Doubao-Seed-1.6-Vision	`doubao-seed-1-6-vision-250815`
Qwen3-VL-Plus	`qwen3-vl-plus-2025-12-19`
Qwen3.6-Plus	`qwen3.6-plus-2026-04-02`
GPT-5.4-Mini	`gpt-5.4-mini-2026-03-17`

D.5 CROSS-CATEGORY TRAINING AND SUPERVISION GRANULARITY

Table 19 reports the full cross-category training comparison between PSDISTILL and EvoQuality-style majority-vote distillation (Wen et al., 2025). Each method is trained on a single category and evaluated zero-shot on all three.

Table 19: Effect of training data category. PSDISTILL (ours) versus EvoQuality-style majority-vote distillation and PW-Distill (pointwise pseudo-score distillation). The best checkpoint per configuration is selected by mean SRCC. Top two per column are **bolded** and underlined.

		SRCC ↑				PLCC ↑			
Method	**Train**	**HN**	**RW**	**SS**	**Mean**	**HN**	**RW**	**SS**	**Mean**
PW-Distill	HN	0.523	0.689	0.459	0.557	0.526	0.683	0.459	0.556
	RW	0.601	0.698	0.455	0.585	0.587	0.685	0.500	0.591
	SS	0.582	0.700	0.550	0.610	0.569	0.680	0.537	0.595
EvoQuality	HN	0.644	0.692	0.549	0.628	0.643	0.680	0.542	0.622
	RW	0.606	0.629	0.570	0.602	0.625	0.625	0.548	0.599
	SS	0.663	0.707	0.587	0.652	0.664	0.667	0.598	0.643
PSDISTILL (ours)	HN	**0.727**	0.713	0.686	0.709	**0.743**	0.693	0.692	0.709
	RW	0.676	**0.715**	**0.754**	**0.715**	0.688	**0.706**	**0.738**	**0.711**
	SS	0.720	0.667	0.700	0.696	0.718	0.622	0.677	0.672

Ranking-level versus pair-level supervision. The gap between Calibrated Elo distillation (0.709) and EvoQuality-style majority-vote distillation (0.628) corresponds to a difference in supervision granularity. The majority-vote approach retains only the discrete outcome of each pair, discarding vote margins and imposing no consistency across pairs. The Calibrated Elo pipeline aggregates all pairwise outcomes into a global Elo ranking and maps it to continuous pseudo-scores in the range $[1, 5]$, which yields transitive orderings and finer score resolution. The difference is most visible on landscape painting, where per-pair judgments are noisier (Calibrated Elo achieves 0.686 versus 0.549 for majority vote). Across categories, the Calibrated Elo variant has a cross-category SRCC standard deviation of 0.021, compared with 0.073 for majority vote and 0.119 for PW-Distill, consistent with global aggregation producing more stable supervision than pair-level labels.

Cross-training-source stability. Table 19 further reveals that Calibrated Elo supervision is robust to the choice of training category. Mean SRCC ranges from 0.696 to 0.715 (spread of 2.7% relative) across three training sources. EvoQuality-style majority vote shows a wider range of 0.602 to 0.652 (spread of 8.3% relative), and PW-Distill pseudo-scores exhibit a comparable spread of 0.557 to 0.610 (9.5% relative). The ordering among methods is consistent across all training sources, with Calibrated Elo outperforming majority vote and PW-Distill regardless of the category used for training.

D.6 PAIR WEIGHTING ABLATION

Using the same ranking-pool seed for both variants, pair weighting improves mean SRCC by 7.4% relative over the unweighted Calibrated Elo variant. The gain appears across all three categories and is largest on landscape painting (11.9% relative, from 0.613 to 0.686), supporting the use of pseudo-score gaps to weight pairwise supervision in this setting. Per-category results are reported in Table 20.

Table 20: Effect of pair weighting under Calibrated Elo pseudo-labels. Both variants use the same offline pseudo-scores, ranking-pool seed, and training configuration. The only difference is whether the pair weight $w_{iz}^{(d)}$ (Eq. 7) is applied during online GRPO training.

	SRCC ↑				PLCC ↑			
Pair wt.	**HN**	**RW**	**SS**	**Mean**	**HN**	**RW**	**SS**	**Mean**
✓	**0.727**	**0.713**	**0.686**	**0.709**	**0.743**	**0.693**	**0.692**	**0.709**
✗	0.688	0.677	0.613	0.660	0.692	0.650	0.637	0.660

D.7 SENSITIVITY TO REFERENCE COUNT AND RANKING-POOL SEED

Table 21 studies the number of reference images compared with each non-pool image and the seed used to sample the ranking pool. These references provide model-generated comparisons for pseudo-label construction. The expert ground-truth anchors in §3.3 instead use human ratings to establish the

evaluation scale. All variants here use flower-and-bird training images and the same self-distillation configuration, differing only in the reference count or ranking-pool seed.

Table 21: Sensitivity to reference count and ranking-pool seed in Calibrated Elo pseudo-label construction. The best checkpoint is selected by mean SRCC. Best per column in bold.

	SRCC ↑				PLCC ↑			
Setting	**HN**	**RW**	**SS**	**Mean**	**HN**	**RW**	**SS**	**Mean**
5 references, seed 11	0.710	**0.715**	0.668	0.698	0.602	0.561	0.531	0.565
10 references, seed 11	0.726	0.683	0.680	0.696	0.720	0.678	0.688	0.695
10 references, seed 42	0.727	0.713	0.686	**0.709**	0.743	**0.693**	**0.692**	**0.709**
15 references, seed 11	**0.729**	0.678	**0.702**	0.703	**0.765**	0.634	0.686	0.695

Under the same ranking-pool sampling seed, reducing the number of references from 10 to 5 leaves mean SRCC close to the 10-reference setting, but mean PLCC drops from 0.695 to 0.565. This indicates that fewer comparisons can preserve coarse ordering while reducing linear correlation with the target scores. The corresponding pseudo-labels concentrate many samples near the lower end of the score support, especially for composition, artistic conception, and overall quality, reducing the resolution of the continuous supervision. Increasing the reference count from 10 to 15 yields a small mean-SRCC change (0.696 to 0.703) and nearly identical mean PLCC. We therefore use 10 references as the default setting in the main experiments.

D.8 Robustness to Fused Expert Ground-Truth Construction Method

To test whether the distillation gain is an artifact of shared fused expert ground-truth construction methodology, we re-evaluate all self-distilled models against pure human pointwise targets, computed as the mean (or median) of the five expert pointwise scores per image-dimension pair. This pointwise target is constructed entirely from rating annotations and is therefore independent of the pseudo-label pipeline used in the offline stage (§4.1).

Table 22 reports the results for all three self-distillation methods under the controlled flower-and-bird configuration used in the ablation tables. The method ordering is fully preserved under both the mean and median pointwise targets, with PSDistill > EvoQuality > PW-Distill > Base. PSDistill achieves 0.621 mean SRCC under the mean pointwise target and 0.594 under the median pointwise target, corresponding to 34.7% and 35.6% relative improvements over the base model. Absolute correlations are lower than under the corresponding fused expert ground-truth evaluation (0.709). The preserved method ordering shows that the ranking gains also hold against targets constructed solely from human pointwise ratings.

Table 22: Evaluation against pure human pointwise targets (mean and median of five expert scores). All self-distillation models use the controlled flower-and-bird configuration from the ablation tables. ‡ denotes baselines trained via the same Thurstone-model-based GRPO framework as PSDistill with different pseudo-label sources. The method ordering is preserved under both pointwise targets.

	SRCC ↑				PLCC ↑			
Method	**HN**	**RW**	**SS**	**Mean**	**HN**	**RW**	**SS**	**Mean**
Target = human pointwise mean								
Base (pretrained)	0.515	0.537	0.332	0.461	0.524	0.589	0.395	0.503
PW-Distill‡	0.492	0.584	0.369	0.482	0.478	0.587	0.409	0.491
EvoQuality‡	0.579	0.535	0.488	0.534	0.588	0.586	0.507	0.560
PSDistill (ours)	**0.620**	**0.625**	**0.619**	**0.621**	**0.632**	**0.621**	**0.627**	**0.627**
Target = human pointwise median								
Base (pretrained)	0.506	0.525	0.284	0.438	0.491	0.564	0.344	0.467
PW-Distill‡	0.448	0.565	0.335	0.450	0.425	0.561	0.381	0.455
EvoQuality‡	0.552	0.540	0.465	0.519	0.542	0.564	0.471	0.526
PSDistill (ours)	**0.588**	**0.634**	**0.560**	**0.594**	**0.587**	**0.615**	**0.563**	**0.588**

While Table 22 confirms that the ranking ordering of self-distilled models is preserved across evaluation-target choices, ranking metrics alone cannot detect score-scale collapse. To assess whether

the distilled scores are also calibrated to the target score scale, Table 23 reports the mean absolute error (MAE), root mean squared error (RMSE), and the global predicted score mean for each method, evaluated against both the fused expert ground truth and the human pointwise mean target.

Table 23: Absolute score calibration on PPAINT. Fused denotes preference-derived expert scores with rating anchors; Human denotes the mean of five expert ratings per image and dimension. MAE/RMSE are mean absolute/root mean squared errors on the 1–5 scale (lower is better). Pred. mean averages all 750 image–dimension predictions (150 paintings, five dimensions). Δ Fused is Pred. mean minus the fused target mean, 2.95; positive values indicate overprediction and negative values underprediction.

	MAE ↓		RMSE ↓		Pred. mean	Δ Fused
Method	**Fused**	**Human**	**Fused**	**Human**		
Base (pretrained)	1.54	1.45	1.79	1.75	4.48	$+1.53$
PW-Distill‡	1.36	1.31	1.61	1.60	4.25	$+1.30$
EvoQuality‡	1.38	1.32	1.61	1.59	4.27	$+1.32$
PSDISTILL (ours)	**0.84**	**0.91**	**1.00**	**1.11**	**2.84**	$\mathbf{-0.11}$

The base model and both baselines exhibit severely mis-calibrated predictions despite their nontrivial ranking correlations. Their predicted score means range from 4.25 to 4.48, more than 1.2 points above the fused expert ground-truth mean of 2.95, and their MAE on the expert rating scale ranges from 1.31 to 1.54. PSDISTILL is the only method whose predicted mean (2.84) lies within 0.2 points of either reference mean, and its MAE is roughly 40% lower than the next-best baseline. The pattern provides a quantitative counterpart to the predicted-distribution collapse visualized in Figure 1c.

D.9 PER-DIMENSION DISTILLATION ANALYSIS

Table 24 breaks down the distillation results by aesthetic dimension (averaged over three categories). The improvement pattern mirrors the per-dimension pairwise advantage reported in §B.2. The most subjective dimensions benefit most from Elo-based pseudo-labels. Artistic conception shows the largest relative improvement in SRCC over the base model for PSDISTILL (95.5%), followed by composition (72.8%), coloration (41.5%), and overall quality (20.9%). In contrast, brush technique, where the base model already achieves 0.701 SRCC, improves by 8.5%. This pattern is consistent with the larger pairwise advantage observed on subjective dimensions in human annotation.

Table 24: Per-dimension SRCC and PLCC (averaged over three categories, flower-and-bird training). All self-distillation methods use the same base model. Tech., Color., Comp., Conc., and Over. denote brush technique, coloration, composition, artistic conception, and overall quality. Artistic conception shows the largest gain, mirroring the per-dimension pairwise advantage in human annotation (Table 6).

	SRCC ↑						PLCC ↑					
Method	**Tech.**	**Color.**	**Comp.**	**Conc.**	**Over.**	**Mean**	**Tech.**	**Color.**	**Comp.**	**Conc.**	**Over.**	**Mean**
Base (pretrained)	0.701	0.411	0.431	0.365	0.613	0.504	0.717	0.400	0.452	0.347	0.645	0.512
PW-Distill‡	0.704	0.515	0.627	0.277	0.660	0.557	0.721	0.475	0.603	0.328	0.654	0.556
EvoQuality‡	0.710	0.577	0.575	0.580	0.698	0.628	0.736	0.544	0.595	0.549	0.685	0.622
PSDISTILL (ours)	**0.761**	**0.582**	**0.745**	**0.714**	**0.741**	**0.709**	**0.763**	**0.573**	**0.721**	**0.720**	**0.769**	**0.709**

E CROSS-DOMAIN EVALUATION ON APDDV2

E.1 SETUP

APDDv2 (Jin et al., 2024) contains 24 artistic categories and 10 aesthetic attributes. Our experiments use its oil-painting and sketching subsets to evaluate PSDISTILL and the EvoQuality baseline beyond Chinese painting. For each method, we start from Qwen3-VL-8B-Instruct and train separately on each APDDv2 training subset using model-generated supervision. Each resulting model is evaluated on both APDDv2 test subsets. Oil painting is dominated by color and texture, whereas sketching is dominated by line and composition. Each subset provides 1,000 training images and 371 (oil) or 378 (sketching) test images. We confirm that the test sets contain no image duplicates with the PPAINT corpus by md5 hashing.

APDDv2 provides annotations over multiple aesthetic dimensions. For a controlled train–test evaluation, we use the three dimensions that are consistently available in both the training and test annotations: layout-and-composition, details-and-texture, and overall artistic quality. This setup uses three aesthetic dimensions with scores ranging from 1 to 10, whereas PPAINT uses five dimensions with scores ranging from 1 to 5. Both methods inherit the configuration described in Appendix D.2 with two adaptations. The reward functions are extended to operate over 3 dimensions, and both the score range in the prompt and the output range of the mapping are set to $[1, 10]$. All other hyperparameters remain unchanged.

E.2 CROSS-DOMAIN TRANSFER

The two-domain APDDv2 setup also reveals an asymmetry in cross-domain transfer. Models trained on sketching transfer to oil painting more strongly than the reverse direction. Training PSDISTILL on sketching yields a cross-domain SRCC of 0.690 on oil painting, whereas training on oil painting yields only 0.587 on sketching. A possible explanation is that sketching forces the policy to focus on layout and composition, which transfer well across genres, while oil-painting training admits richer surface cues such as color, texture, and brushwork that are more genre-specific. Per-domain in-domain versus cross-domain results are summarized in Table 25.

Table 25: Per-domain SRCC on APDDv2. Rows are grouped by training source. The diagonal entries report in-domain performance and the off-diagonal entries report cross-domain transfer. **Bold** marks the higher SRCC within each (training source, test domain) cell.

Method	Train	Test = Oil	Test = Sketch
EvoQuality	Oil	0.683 (in-domain)	0.574 (cross)
PSDISTILL	Oil	**0.752** (in-domain)	**0.587** (cross)
EvoQuality	Skt	0.620 (cross)	0.560 (in-domain)
PSDISTILL	Skt	**0.690** (cross)	0.559 (in-domain)

F MODEL SCALE AND FINE-TUNING SETTINGS

We further evaluate PSDISTILL with Qwen3-VL-32B using the same unlabeled flower-and-bird training split and five-dimensional evaluation protocol. The 32B model generates its own pseudo-labels and is adapted with LoRA using rank 32, scaling factor 64, and dropout 0.05. Training uses 8×NVIDIA A100 80GB GPUs, four generations per prompt, and an effective batch size of 64 prompts. The five-dimensional GRPO objective is unchanged.

Table 26: Model scale and fine-tuning settings on PPAINT. The 32B LoRA result uses checkpoint 94 (two epochs), selected by maximum mean SRCC among evaluated checkpoints on 150 test images.

Model	Fine-tuning	Mean SRCC	Mean PLCC
Qwen3-VL-8B	None	0.504	0.512
PSDISTILL 8B	Full-parameter	0.709	0.709
Qwen3-VL-32B	None	0.663	0.664
PSDISTILL 32B	LoRA	0.739	0.742

Table 26 reports a mean SRCC of 0.739 and mean PLCC of 0.742 for 32B LoRA, compared with 0.663 and 0.664 for its pretrained base. This provides additional evidence that PSDISTILL can be applied to a larger backbone with LoRA.

G PAIRED MODEL COMPARISONS

Table 27 reports paired differences in mean SRCC between PSDISTILL and selected comparison models. We use 2,000 bootstrap resamples, sampling images with replacement within each painting category and retaining all five dimensions of each image. Both models are evaluated on the same resampled images in each iteration. The table reports 95% intervals for the paired difference.

Table 27: Paired bootstrap comparisons on PPAINT. A positive difference favors PSDISTILL.

Comparison model	Δ **mean** SRCC	**95% CI**
Qwen3-VL-8B	$+0.204$	$[0.116, 0.292]$
EvoQuality	$+0.080$	$[0.020, 0.142]$
PW-Distill	$+0.152$	$[0.080, 0.220]$
Qwen3-VL-32B	$+0.047$	$[-0.020, 0.110]$
Claude-Sonnet-4.6	$+0.009$	$[-0.064, 0.080]$
ArtiMuse	$+0.023$	$[-0.070, 0.095]$