# PAIR2POINT: RELIABLE AS PAIRWISE, EFFICIENT AS POINTWISE REWARDS FOR HIGH-QUALITY TEXT-TO-IMAGE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent multimodal reward models (RMs), especially those built on powerful vision-language models, have shown great promise in aligning vision models with human preferences. However, they face two main challenges. First, their training images often come from earlier generations of text-to-image models and no longer match the quality of current generators. Second, it is difficult to combine the reliable judgments of pairwise comparison with the efficiency of pointwise scoring. To address these challenges, we propose Pair2Point. We build three resources with recent generators for training and evaluation, including P2P-CoT-16K, P2P-Image-70K, and P2P-Bench. On top of these resources, a two-stage framework first trains a pairwise model with structured comparison labels from Gemini 3.1 Pro, so that it evaluates two images along relevant dimensions and explains its judgment; it then distills this model into a pointwise model through multi-opponent aggregation. Experiments show that Pair2Point achieves the highest accuracy among the evaluated dedicated reward models across multiple reward benchmarks. When used for downstream reinforcement learning, it consistently improves both text–image alignment and visual quality. The resources and model weights will be released publicly after review.

## 1 INTRODUCTION

Text-to-image generation has advanced rapidly, and reward models (RMs) play an important role in aligning generators with human preferences. They provide training signals for preference optimization, and guide reinforcement learning (Black et al., 2023; Rafailov et al., 2023; Li et al., 2026). Early RMs such as ImageReward, PickScore, and the HPS family use an encoder to map a prompt and an image to a scalar reward (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023; Ma et al., 2025). They are commonly trained on preference pairs with a Bradley–Terry loss, which encourages the preferred image to receive a higher score. At inference, each image can be scored independently and efficiently. However, such scalar rewards provide little explanation and can be vulnerable to reward hacking (Hong et al., 2026).

A first challenge is the gap between RM training data and current text-to-image models. Many widely used preference datasets and benchmarks mainly contain images from earlier generators. As generation quality improves, these images no longer represent the fine-grained differences among recent models (Liu et al., 2026). Figure 2 (top right) contrasts earlier HPDv3 samples with newly collected outputs from recent generators, and Figure 1 shows that existing RMs can assign inconsistent scores to same-prompt images from contemporary models. Updated preference data and evaluation benchmarks are therefore essential for training and measuring modern RMs.

The rise of vision-language models (VLMs) brings stronger visual understanding and makes detailed reward reasoning possible (Wang et al., 2025b; 2026c;b). It also creates a second challenge: how should a generative VLM produce a scalar reward. Adding an MLP reward head keeps pointwise inference fast, but the reward head and the language output use separate interfaces. Generating a score token uses the native language-modeling interface, but a decoded digit is coarse and discards the uncertainty in the token distribution. Asking the VLM to compare two images gives it direct



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

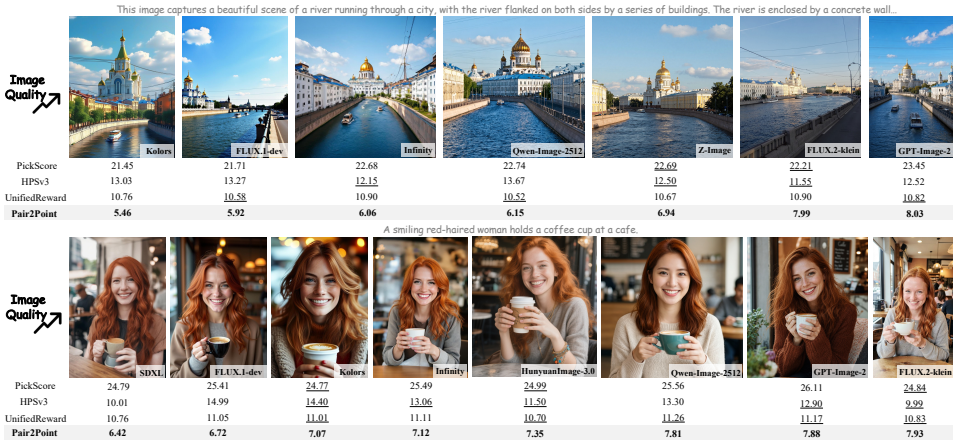


Figure 1: Reward scores on same-prompt images from text-to-image models. Existing reward models often assign scores that contradict the visible quality differences.

contrastive evidence and often leads to a more reliable judgment, but its result is conditioned on the opponent and cannot directly serve as an absolute reward for one image.

Recent work explores different ways to close this gap. JRM uses reasoning supervision during training and a discriminative reward head at inference (Yang et al., 2026), while Z-Reward distills reasoning-conditioned score distributions into an efficient student (Jin et al., 2026). RewardDance represents the reward as the probability that the VLM answers "yes" when asked whether one image is better than a reference (Wu et al., 2025). This design aligns the reward with next-token prediction and avoids relying on one decoded score. However, the resulting reward still depends on the selected reference. These designs reveal the need to combine the reliable comparison and rich reasoning of pairwise RMs with the speed and reference-free use of pointwise RMs.

To address these challenges, we first build three complementary resources on recent generators. We collect about 21K prompts from diverse web sources, and gather their available images from a broad set of text-to-image models and sampling settings. We sample about 24K same-prompt image pairs for Gemini 3.1 Pro annotation; after three human reviewers screen the labels, P2P-CoT-16K provides about 16K structured preference pairs for training the pairwise model. For pointwise distillation, we sample about 10K prompts from the same 21K collection and render them with recent text-to-image models, yielding about 70K unlabeled images. The resulting P2P-Image-70K provides standardized same-prompt candidate pools for multi-opponent aggregation. Finally, P2P-Bench contains 1.2K held-out pairs with unanimous human preferences for evaluating RMs.

On top of these resources, Pair2Point uses a two-stage training strategy. In the first stage, we train a pairwise reasoning model with structured comparison labels produced by Gemini 3.1 Pro. This model predicts the preferred image and also gives visual evidence, per-image scores, and importance weights over the relevant dimensions. In the second stage, we apply the pairwise model to multiple same-prompt opponents and aggregate its outputs into a target for each image. A pointwise model learns these targets without generating long explanations. At inference, it uses the full score-token distribution and scores one image in a single forward pass, while keeping an interpretable dimension-level output.

Experiments on GenAI-Bench (Jiang et al., 2024), MMRB2 (Hu et al., 2025), and P2P-Bench show that Pair2Point achieves the highest accuracy among the evaluated dedicated reward models. When used as the reward for downstream Flow-GRPO fine-tuning, it improves both text–image alignment and visual quality.

Our contributions are as follows.

1. We build three resources: P2P-CoT-16K with about 16K structured preference pairs, P2P-Image-70K with about 70K unlabeled images for pointwise distillation, and P2P-Bench with 1.2K annotated pairs.

2. We introduce Pair2Point, a two-stage framework that first learns pairwise verdicts together with dimension-level evidence, scores, and weights, and then transfers this knowledge into an efficient pointwise RM through multi-opponent aggregation.
3. Extensive experiments show that Pair2Point outperforms existing dedicated reward models in preference prediction and downstream reinforcement learning.

## 2 Related Work

### 2.1 Preference Data for Evolving Generators

Public preference datasets used by ImageReward, PickScore, and the HPS family were mainly collected from earlier generator families (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023; Ma et al., 2025). As text-to-image systems improve, static datasets provide less supervision for the fine-grained differences among current models. HPSv3++ (Liu et al., 2026) addresses this problem by calibrating rewards across generator capability levels and RL iterations.

### 2.2 Reward Models for Visual Generation

Encoder-based RMs map a prompt and an image to a scalar and commonly learn from pairwise preferences with a Bradley–Terry objective (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023; Ma et al., 2025). Their pointwise inference is efficient, but the scalar output contains little information about the visual evidence behind a judgment. VLM-based RMs instead use the language-modeling interface to produce judgments and explanations. UnifiedReward (Wang et al., 2025b) jointly learns reward modeling across multimodal understanding and generation tasks. UnifiedReward-Think (Wang et al., 2026c) introduces explicit long-form reasoning, while UnifiedReward-Flex (Wang et al., 2026b) builds evaluation criteria according to the input. RationalRewards (Wang et al., 2026a) uses preference labels to construct reliable rationales and trains a model to reason before scoring. JRM (Yang et al., 2026) uses language supervision during training and a discriminative head for efficient inference.

### 2.3 From Pairwise Judgments to Pointwise Rewards

Bradley–Terry models convert pairwise preferences into pointwise rewards by learning a latent utility whose differences explain comparison outcomes. Generative RMs offer other ways to connect the two settings. RewardDance (Wu et al., 2025) reads the probability that an image beats a reference from the VLM's next-token distribution. This produces a continuous reward through the language-modeling interface, but the value remains conditioned on the selected reference. Averaging over several references can reduce this dependence at the cost of additional calls. Z-Reward (Jin et al., 2026) transfers a reasoning-conditioned score distribution from a teacher to an efficient student, showing the value of separating offline reasoning from online scoring.

### 2.4 Reinforcement Learning for T2I Alignment

Preference optimization for diffusion and flow models includes policy-gradient methods such as DDPO (Black et al., 2023), reward back-propagation through ReFL (Xu et al., 2023), direct preference optimization (Rafailov et al., 2023), and Flow-GRPO (Shao et al., 2024). These methods depend on the quality and robustness of their reward signal. Weak rewards can lead to reward hacking and visual artifacts (Hong et al., 2026; Liu et al., 2026).

## 3 Method

Pair2Point uses a two-stage training strategy. The first stage trains a pairwise reasoning model on structured comparisons from P2P-CoT-16K. The second stage applies this model to the same-prompt candidate pools in P2P-Image-70K and aggregates its multi-opponent judgments into targets for a pointwise model. The two models share the language-modeling interface, while the pointwise model removes long explanations and scores one image per forward pass. Throughout, $p$ denotes a prompt, $\mathcal{C}(p)$ denotes its candidate pool, and $\pi_\theta$ and $\pi_\phi$ denote the pairwise and pointwise language models.

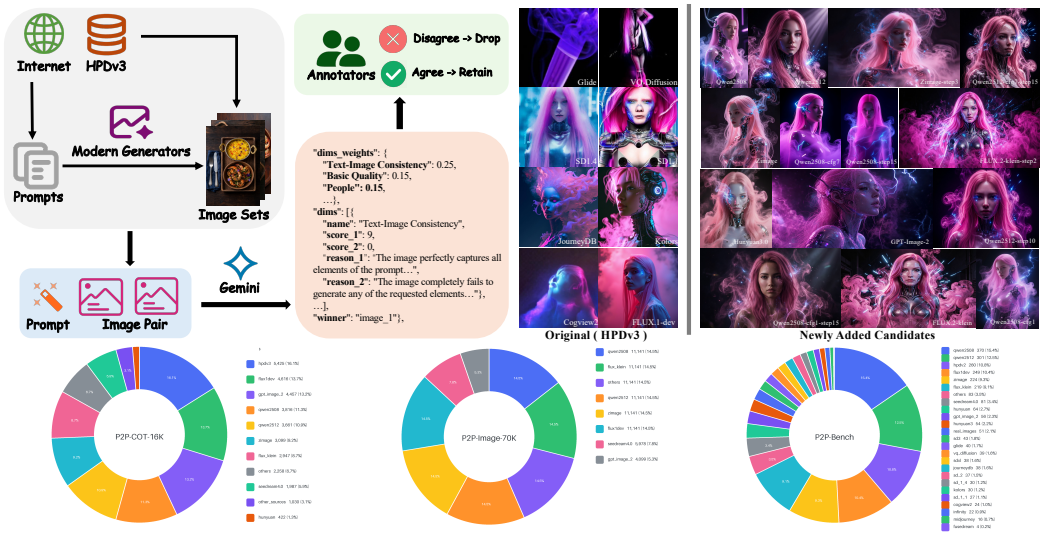


Figure 2: Overview of our data. Top left: data construction pipeline, in which Gemini 3.1 Pro annotates sampled same-prompt image pairs and human review supports quality control for P2P-CoT-16K. Top right: earlier samples from the HPDv3 dataset and newly collected samples from recent text-to-image models, both of which enter the same-prompt image pool used for preference-pair sampling. Bottom: source-model mix of P2P-CoT-16K, P2P-Image-70K, and P2P-Bench.

### 3.1 Preference Data Construction

Figure 2 (top left) illustrates the data construction pipeline, and its bottom row summarizes the source-model mix of the three resources.

**P2P-CoT-16K.** From the available generator outputs, we sample about 24K same-prompt image pairs. In addition to cross-generator pairs, we sample pairs from the same generator under different classifier-free guidance (CFG) scales and numbers of denoising steps. These variants share a generator-specific visual style while differing in quality, so the reward model must judge prompt fidelity, structure, and aesthetics rather than latch onto a fixed generator pattern. Without them, the model can overfit generator identity and systematically down-rate outputs that merely resemble a familiar generator distribution. Gemini 3.1 Pro annotates each pair under a fixed instruction with a verdict, dimension-level evidence, scores, and weights. Three human reviewers then screen these labels; across reviewers, agreement with Gemini ranges from 72% to 81% on decisive judgments. We discard pairs that reviewers overturn and retain about 16K structured comparison pairs for training the pairwise model.

**P2P-Image-70K.** For pointwise distillation, we sample about 10K prompts from the same 21K collection and render each prompt with recent generators, including Z-Image, Qwen-Image-2508, Qwen-Image-2512, FLUX.1-dev, FLUX.1-Klein, Seedream 4.0, and GPT-Image-2. This yields about 70K unlabeled images. These images carry no preference labels; after the pairwise model is trained, it compares candidates within pools to construct the pointwise targets.

**P2P-Bench.** We hold out a pool of 5K image pairs from the broader 21K-prompt collection for evaluation. Six annotators judge each pair independently, with the image order randomized for every annotation. The mean inter-annotator agreement on this pool is 76.4%, and 27.4% of pairs receive a unanimous non-skip preference. From these unanimous pairs we keep 1.2K examples. We use only unanimous labels because we believe a reward model must be reliable on comparisons where humans already agree; pairs with human disagreement are inherently ambiguous, and including them would inject label noise into the evaluation. On a 100-pair unanimous subsample, Gemini 3.1 Pro matches the human winner 89% of the time.

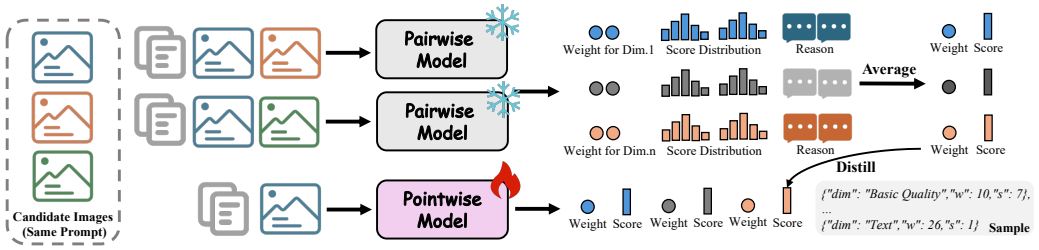


Figure 3: Overview of Pair2Point. The pairwise model produces structured judgments for same-prompt candidates. Multi-opponent aggregation then constructs dimension-level targets for a pointwise model that scores each image in one forward pass.

### 3.2 Pairwise Reasoning Model

Given a prompt and an ordered image pair $(x_i, x_j)$, the pairwise model selects the dimensions relevant to the comparison and predicts the preferred image. For every selected dimension $d$, it provides visual evidence for both images and assigns each image a score $s_{i|j}^{(d)} \in \{0, \ldots, 9\}$. It also predicts a shared pair-level weight $w_{ij}^{(d)} \in [0, 1]$ that represents the importance of dimension $d$ in this comparison, with the selected weights summing to one. The model selects from nine dimensions: basic quality, geometry and physics, color, light and shadow, material and texture, content logic, text–image consistency, people, and rendered text.

We train the language backbone with full-parameter supervised fine-tuning on P2P-CoT-16K, while keeping the vision encoder and projector frozen. The supervised targets use integer scores from Gemini 3.1 Pro. The objective is the language-modeling loss over the structured target sequence:

$$\mathcal{L}_{\text{pair}} = -\sum_{t} \log \pi_\theta(y_t \mid y_{<t}, p, x_i, x_j)\,, \tag{1}$$

where $y$ contains the dimension-level evidence, scores, weights, and final verdict. We randomize image order during training. During evaluation, we query the pairwise model in both presentation orders and report the mean of the forward and reverse accuracies.

### 3.3 Multi-Opponent Distillation to Pointwise Rewards

For a prompt $p$ with candidate pool $\mathcal{C}(p) = \{x_1, \ldots, x_K\}$, we enumerate all unordered image pairs and query the pairwise teacher once per pair with randomized presentation order. When reading the teacher's scores, we do not take the decoded digit. Instead, we normalize the next-token probabilities over the valid score digits $\{0, \ldots, 9\}$ and compute the expected score

$$\mathbb{E}[s] = \sum_{k=0}^{9} k \cdot p(k), \qquad p(k) = \frac{\exp(\ell_k)}{\sum_{r=0}^{9} \exp(\ell_r)}. \tag{2}$$

This teacher-side expectation is used as $s_{i|j}^{(d)}$ in the aggregation below.

Let $\mathcal{P}_i$ be the set of opponents that have a valid comparison with $x_i$, and let $z_{ij}^{(d)} \in \{0, 1\}$ indicate whether dimension $d$ is selected for the comparison between $x_i$ and $x_j$. The opponents that assess $x_i$ on dimension $d$ are

$$\mathcal{J}_i^{(d)} = \{j \in \mathcal{P}_i : z_{ij}^{(d)} = 1\}. \tag{3}$$

Each score is conditioned on an opponent, while each importance weight describes a pair. We aggregate both quantities over the valid opponents:

$$q_i^{(d)} = \frac{1}{|\mathcal{J}_i^{(d)}|} \sum_{j \in \mathcal{J}_i^{(d)}} s_{i|j}^{(d)}, \qquad \bar{w}_i^{(d)} = \frac{1}{|\mathcal{J}_i^{(d)}|} \sum_{j \in \mathcal{J}_i^{(d)}} w_{ij}^{(d)}. \tag{4}$$

Table 1: Preference accuracy (%) across benchmarks. Boldface marks the best result in each column.

| Category | Model | GenAI-Bench | MMRB2 | P2P-Bench | Avg. |
|---|---|---|---|---|---|
| MLLM as a judge | Qwen3-VL-8B | 61.6 | 58.0 | 77.7 | 65.8 |
| | Qwen3-VL-32B | 63.0 | 60.3 | 80.6 | 68.0 |
| | Gemini 3.1 Pro (commercial) | 71.0 | 74.0 | 90.1 | 78.4 |
| Reward model | HPSv2 | 68.8 | 54.6 | 81.5 | 68.3 |
| | PickScore | 70.0 | 57.4 | 81.1 | 69.5 |
| | HPSv3 | 70.7 | 58.5 | 80.7 | 70.0 |
| | UnifiedReward | 71.5 | 58.5 | 88.3 | 72.8 |
| | UnifiedReward-Think | 72.3 | 63.6 | 87.0 | 74.3 |
| | RationalRewards | 69.8 | 64.2 | 87.4 | 73.8 |
| | **Pair2Point** | **72.9** | **68.2** | **90.2** | **77.1** |

Dimension relevance is the empirical selection frequency

$$\rho_i^{(d)} = \frac{|\mathcal{J}_i^{(d)}|}{|\mathcal{P}_i|}. \tag{5}$$

We retain $\mathcal{D}_i = \{d : \rho_i^{(d)} \geq \tau\}$ with $\tau = 0.5$. Thus, $q_i^{(d)}$ and $\bar{w}_i^{(d)}$ summarize the score and importance of a selected dimension, while $\rho_i^{(d)}$ measures how often that dimension is relevant across opponents. For $d \in \mathcal{D}_i$, we normalize the mean weights onto a 100-point budget and convert them to non-negative integers $\hat{w}_i^{(d)}$ with the largest-remainder method, so that the integer weights sum to exactly 100. The training target for image $x_i$ is then a compact JSON object containing the rounded score $\hat{s}_i^{(d)} = \mathrm{round}(q_i^{(d)})$ and integer weight $\hat{w}_i^{(d)}$ for every retained dimension. It contains no explanation text.

The pointwise model $\pi_\phi$ is initialized from the pairwise checkpoint and adapted with rank-8 LoRA, using token-level cross entropy over this target sequence:

$$\mathcal{L}_{\text{point}} = -\sum_t \log \pi_\phi(\hat{y}_{i,t} \mid \hat{y}_{i,<t}, p, x_i)\,, \tag{6}$$

where $\hat{y}_i$ serializes the retained dimensions and their $(\hat{w}, \hat{s})$ targets. LoRA updates the compact scoring interface while preserving the representations learned in the pairwise stage. At pointwise inference, we apply the same distributional expectation to the student's score-token probabilities. The final scalar reward uses the dimensions and weights predicted by the pointwise model:

$$R(p, x) = \frac{\sum_d w_d\, \mathbb{E}[s_d]}{\sum_d w_d}. \tag{7}$$

Each image is scored once, so online inference has no left/right position, requires no reference image, and costs one call per rollout.

### 3.4 Online Optimization with Pointwise Rewards

We use the pointwise model as the reward for Flow-GRPO (Shao et al., 2024) fine-tuning of a flow matching text-to-image policy. Given a prompt $p$, the policy samples a group of $G$ images $\{x_g\}_{g=1}^G$. The pointwise model independently computes $r_g = R(p, x_g)$ for every rollout. We then form a group-relative advantage

$$A_g = \frac{r_g - \mathrm{mean}_j(r_j)}{\mathrm{std}_j(r_j) + \epsilon}. \tag{8}$$

The policy is updated with the standard clipped Flow-GRPO objective over the sampled flow trajectories. Since the advantage is centered within each prompt group, optimization depends on the relative ordering and separation of the rewards inside the group.

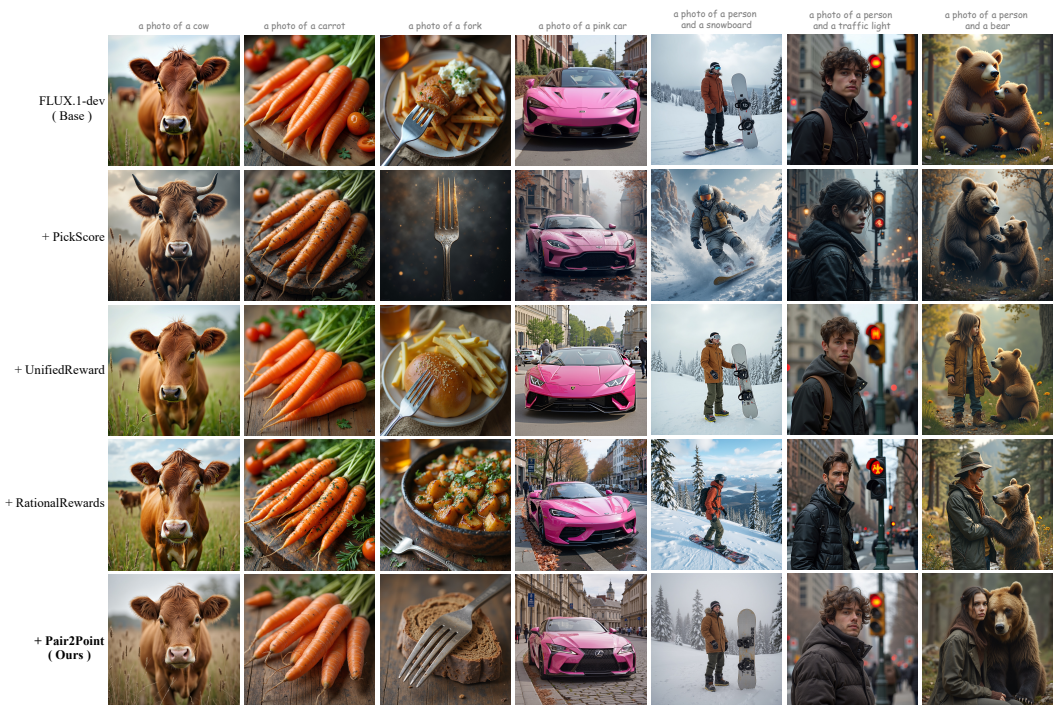


Figure 4: Qualitative comparison before and after Flow-GRPO fine-tuning with Pair2Point and other competitive rewards.

The pointwise interface is important for online training. A pairwise reasoning model would require up to $\binom{G}{2}$ comparisons to cover the group, and every score would remain tied to another rollout. Pair2Point evaluates the same group with $G$ independent calls. This reduces reward computation and gives every rollout one scalar that can be used directly in the Flow-GRPO objective.

## 4 Experiments

We first evaluate preference accuracy against published reward models, then test the pointwise reward in Flow-GRPO fine-tuning, and finally ablate the components.

### 4.1 Setup

We compare against two groups of baselines. The first consists of multimodal LLMs prompted to judge preference without task-specific training: Qwen3-VL at 8B and 32B, together with Gemini 3.1 Pro as a proprietary reference. For these judges, we reuse the same structured comparison prompt used by Gemini 3.1 Pro when annotating P2P-CoT-16K, and decode at temperature 0. The second consists of dedicated reward models: HPSv2 (Wu et al., 2023), PickScore (Kirstain et al., 2023), HPSv3 (Ma et al., 2025), UnifiedReward (Wang et al., 2025b), UnifiedReward-Think (Wang et al., 2026c), and RationalRewards (Wang et al., 2026a). For each dedicated reward model, we follow its official prompting and scoring protocol under a shared serving configuration.

**Reward benchmarks.** We report agreement with ground-truth preferences on GenAI-Bench (Jiang et al., 2024), the text-to-image split of MMRB2 (Hu et al., 2025), and P2P-Bench. GenAI-Bench and MMRB2 measure transfer to external data, while P2P-Bench measures discrimination among recent generators. Pairwise models are evaluated in both presentation orders, and we report the mean of the forward and reverse accuracies. Pointwise models score the two images independently and select the image with the higher reward. We report pair-level accuracy on each benchmark.

Table 2: Flow-GRPO results with different rewards on FLUX.1-dev and Qwen-Image-2512. Only the reward model varies across rows; boldface marks the best result in each column.

| Reward | FLUX.1-dev | | | | Qwen-Image-2512 | | | |
|---|---|---|---|---|---|---|---|---|
| | Semantic Consistency | | Image Quality | | Semantic Consistency | | Image Quality | |
| | GenEval | UniGenBench | PickScore | HPSv3 | GenEval | UniGenBench | PickScore | HPSv3 |
| None (base policy) | 66.11 | 60.37 | 23.35 | 8.72 | 73.64 | 78.34 | 23.51 | 9.08 |
| PickScore | 55.81 | 55.63 | 23.56 | 9.03 | 73.65 | 78.35 | 23.51 | 9.09 |
| UnifiedReward | 67.88 | 62.04 | 23.33 | 8.75 | 74.35 | 78.44 | 23.66 | 9.16 |
| RationalRewards | 67.54 | 59.19 | 23.58 | 9.13 | 74.12 | 78.89 | 23.74 | 9.31 |
| **Pair2Point** | **68.21** | **63.94** | **23.60** | **9.27** | **74.41** | **79.10** | **23.89** | **9.39** |

**Training configuration.** The pairwise model is initialized from Qwen3-VL-8B-Instruct. We fully fine-tune the language backbone on P2P-CoT-16K for 5 epochs with learning rate $1\times10^{-5}$, cosine decay, and a 10% warmup, while freezing the vision encoder and projector. Images are capped at $1024\times1024$ pixels. The pointwise model is initialized from the pairwise checkpoint selected by validation loss and adapted with rank-8 LoRA for 5 epochs at learning rate $1\times10^{-4}$ under the same scheduler. Stage-2 LoRA updates the compact scoring interface while keeping the reasoning ability acquired in stage 1 from degrading under pointwise distillation. Both stages use per-device batch size 1 with gradient accumulation 8, bf16 training, and a 10% held-out validation split evaluated every 100 steps; we pick the checkpoint with the best validation loss.

### 4.2 Preference Accuracy

Table 1 reports preference accuracy. Pair2Point achieves the highest accuracy among the evaluated dedicated reward models on all three benchmarks. GenAI-Bench and MMRB2 mainly reflect earlier generator distributions and can have limited annotator agreement, so the absolute gaps among dedicated reward models are small and these benchmarks are close to saturation. Transfer to these external sets remains informative, while P2P-Bench provides clearer discrimination among outputs from recent generators.

### 4.3 Downstream Reinforcement Learning

Table 2 evaluates whether reward-model accuracy transfers to policy optimization. We fine-tune FLUX.1-dev with Flow-GRPO under LoRA (rank 64, $\alpha$=128). For every reward setting, we keep the same base policy, English training prompts, group size $G$=8, resolution 1024, 20 denoising steps with guidance scale 3.5, and learning rate $3\times10^{-4}$. We change only the reward model among PickScore, UnifiedReward, RationalRewards, and Pair2Point, and include the unmodified policy as a reference. Unless a run collapses early from reward hacking, we evaluate the checkpoint at 400 update steps with fixed prompts and sampling settings (28 denoising steps, guidance 3.5). We repeat the same protocol on Qwen-Image-2512. GenEval (Ghosh et al., 2023) measures compositional prompt following, while UniGenBench (Wang et al., 2025a) measures broader instruction compliance. We additionally report HPSv3 and PickScore to capture complementary changes in image quality and visual preference, so that gains are not judged by the optimized reward alone. Across both base generators, Flow-GRPO with Pair2Point improves semantic consistency and image quality over the corresponding base policy and yields stronger cross-metric results than the alternative rewards. Gains on both prompt-following and preference-oriented metrics indicate that the pointwise reward supports broad quality improvement beyond a single evaluator.

Figure 4 shows qualitative samples. PickScore tends to add dense detail with darker tones, but the extra detail can appear as floating artifacts. UnifiedReward improves semantic consistency, yet the images often remain overly smooth and synthetic. RationalRewards strengthens texture, but with high contrast and saturation that look greasy. Pair2Point adds moderate texture and more natural tones while avoiding oversaturation and that greasy look.

Table 3: Ablation study of key components. We report preference accuracy (%) when varying the training setting, backbone scale, distillation targets, and adaptation strategy.

| Variant | GenAI-Bench | MMRB2 | P2P-Bench | Variant | GenAI-Bench | MMRB2 | P2P-Bench |
|---|---|---|---|---|---|---|---|
| *Training setting* | | | | *Backbone scale* | | | |
| Regression head | 67.9 | 63.6 | 90.0 | Qwen3-VL-4B | 70.7 | 64.5 | 88.1 |
| Pairwise teacher | 72.3 | 65.5 | 90.4 | Qwen3-VL-8B | 72.9 | 68.2 | 90.2 |
| Pointwise student | 72.9 | 68.2 | 90.2 | Qwen3-VL-32B | 73.1 | 67.3 | 90.9 |
| *Distill setting* | | | | *Adaptation* | | | |
| Single opponent | 64.5 | 56.3 | 77.9 | Pair LoRA + Point LoRA | 70.3 | 66.3 | 88.8 |
| Without relevance | 69.2 | 63.7 | 88.1 | Pair LoRA + Point full | 70.5 | 66.8 | 89.4 |
| Without weights | 68.4 | 64.6 | 87.5 | Pair full + Point full | 71.9 | 66.5 | 89.0 |
| Full aggregation | 72.9 | 68.2 | 90.2 | Pair full + Point LoRA (ours) | 72.9 | 68.2 | 90.2 |

### 4.4 Ablation Studies

Table 3 ablates four factors under a shared evaluation protocol. In the training-setting block, we compare the pairwise teacher, the distilled pointwise student, and a regression-head alternative that attaches an MLP to predict a scalar instead of generating dimension-level scores. The regression head is initialized from Qwen3-VL-8B-Instruct and trained on the 16K preference pairs of P2P-CoT-16K with a standard Bradley–Terry loss. The student stays within a small margin of the teacher on all three benchmarks and is stronger on the external sets, showing that multi-opponent distillation preserves comparative judgment quality while reducing scoring to a single-image forward pass. The regression head lags behind both generative variants, which supports keeping an interpretable dimension-level interface for the student. The backbone-scale block varies the Qwen3-VL size used by the student under the same distillation recipe. Moving from 4B to 8B improves accuracy on every benchmark, whereas scaling further to 32B brings only marginal changes. We therefore deploy the 8B student as a favorable accuracy–cost trade-off for online reward computation. The distill-setting block isolates how targets are constructed. A single-opponent baseline that assigns each image one randomly sampled pairwise judgment collapses on all three sets, confirming that multi-opponent aggregation is essential for stable pointwise supervision. Dropping relevance filtering or importance weights also hurts accuracy, while the full aggregation—relevance-aware averages of scores and importance weights over all same-prompt opponents—gives the strongest student. Finally, the adaptation block compares full-parameter updates and LoRA at each stage. Our default uses full-parameter supervised fine-tuning for the pairwise model and rank-8 LoRA for the pointwise model: stage 1 learns structured comparative reasoning thoroughly, and stage 2 then adapts the compact scoring interface without overwriting the representations acquired from P2P-CoT-16K.

## 5 Conclusion

We presented Pair2Point to address two challenges in multimodal reward modeling for text-to-image generation: the gap between RM training data and recent generators, and the tension between reliable pairwise comparison and efficient pointwise scoring. To close the data gap, we build three resources on contemporary models—P2P-CoT-16K for structured pairwise supervision, P2P-Image-70K for same-prompt distillation pools, and P2P-Bench for evaluation with unanimous human preferences. On top of these resources, a two-stage framework first trains a pairwise reasoning model that predicts verdicts together with visual evidence, dimension-level scores, and importance weights, then distills it into a pointwise reward through multi-opponent aggregation. The resulting student scores one image in a single forward pass, keeps an interpretable dimension-level interface, and is ready for online optimization. Experiments show that Pair2Point achieves the highest accuracy among the evaluated dedicated reward models, while matching the judgment quality of its pairwise teacher at much lower inference cost. Used as the reward in Flow-GRPO, it improves both semantic consistency and image quality across the evaluated generators.

### AI use statement

We used large language models (LLMs) primarily to assist with and polish the writing of this paper. Gemini 3.1 Pro was also used as an annotator for P2P-CoT-16K and as an evaluation baseline.

## REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023. doi: 10.52202/075280-2270. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Paper-Datasets_and_Benchmarks.pdf.

Yunqi Hong, Kuei-Chun Kao, Hengguang Zhou, and Cho-Jui Hsieh. Understanding reward hacking in text-to-image reinforcement learning. *arXiv preprint arXiv:2601.03468*, 2026.

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

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

Xin Jin, Huanqia Cai, Zhen Li, Zechao Zhan, Dengyang Jiang, Aiming Hao, Yuming Jiang, Chunle Guo, Peng Gao, Ming-Ming Cheng, et al. Beyond scalar rewards by internalizing reasoning into score distributions. *arXiv e-prints*, pp. arXiv–2606, 2026.

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

Xiaomin Li, Qian Liang, Yinan Li, Ying Zhang, Chen Li, Jing Lyu, Huchuan Lu, and Xu Jia. Portraitgen: Exemplar-driven grpo with dual-reward guidance for photorealistic portrait generation. *arXiv preprint arXiv:2606.26930*, 2026.

Yijun Liu, Jie Huang, Zeyue Xue, Yuming Li, Ruizhe He, Haoran Li, Shijia Ge, and Siming Fu. Hpsv3++: Scaling reward models across the full spectrum of diffusion model capabilities. *arXiv preprint arXiv:2606.14657*, 2026.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhe Wang, Cong Wei, Weiming Ren, Jiaming Liu, Fangzhen Lin, and Wenhu Chen. Rationalrewards: Reasoning rewards scale visual generation both training and test time. *arXiv preprint arXiv:2604.11626*, 2026a.

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

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

Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *arXiv preprint arXiv:2602.02380*, 2026b.

Yibin Wang, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, Jiaqi Wang, et al. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *Advances in neural information processing systems*, 38:159130–159157, 2026c.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

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

Yankai Yang, Yancheng Long, Hongyang Wei, Wei Chen, Tianke Zhang, Kaiyu Jiang, Haonan Fan, Changyi Liu, Jiankang Chen, Kaiyu Tang, et al. Joint reward modeling: Internalizing chain-of-thought for efficient visual reward models. *arXiv preprint arXiv:2602.07533*, 2026.

# APPENDIX

## A PER-DIMENSION DOWNSTREAM METRICS

The main paper reports overall GenEval and UniGenBench averages after Flow-GRPO fine-tuning of FLUX.1-dev. Tables 4 and 5 expand those results into the official per-dimension scores. The experimental protocol is unchanged: the same base policy, English prompts, group size, resolution, sampling settings, and evaluation checkpoint are used for every reward, so differences across rows isolate the effect of the reward signal.

On GenEval (Table 4), PickScore lowers nearly every compositional subscore relative to the base FLUX.1-dev policy, consistent with the overall drop in the main paper. UnifiedReward and RationalRewards recover most dimensions and improve the average, with UnifiedReward strongest on position and counting, and RationalRewards strongest on color attributes. Pair2Point obtains the highest overall GenEval score, driven mainly by gains on single-object and two-object composition, while remaining competitive on the remaining axes.

On UniGenBench (Table 5), the same pattern appears at a finer grain. PickScore again underperforms the base policy on most dimensions. UnifiedReward and RationalRewards improve several categories, but Pair2Point yields the strongest average and leads on world knowledge, attribute, action, relationship, compound, grammar, logical reasoning, entity layout, and text generation. These per-dimension gains indicate that the improvements reported in the main paper are not concentrated in a single subcategory, but distribute across prompt-following skills that UniGenBench is designed to probe.

Table 4: Per-dimension GenEval scores (%) after Flow-GRPO on FLUX.1-dev. Boldface marks the best result in each column.

| Model | Position | Colors | Color Attr. | Counting | Single Obj. | Two Obj. | Avg. |
|---|---|---|---|---|---|---|---|
| FLUX.1-dev | 22.75 | **80.05** | 45.00 | 71.88 | 98.44 | 78.54 | 66.11 |
| w/ PickScore | 15.00 | 62.50 | 29.25 | 58.75 | 96.88 | 72.47 | 55.81 |
| w/ UnifiedReward | **26.75** | 79.79 | 48.50 | **72.19** | 98.75 | 81.31 | 67.88 |
| w/ RationalRewards | 24.25 | 76.06 | **50.50** | 70.00 | 99.06 | 85.35 | 67.54 |
| **w/ Pair2Point** | 25.75 | 74.47 | 49.75 | 70.00 | **99.38** | **89.90** | **68.21** |

Table 5: Per-dimension UniGenBench scores (%) after Flow-GRPO on FLUX.1-dev. Boldface marks the best result.

| Model | Style | World Know. | Attr. | Action | Rel. | Comp. | Gram. | Logic | Layout | Text | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FLUX.1-dev | **87.00** | 87.82 | 68.27 | 63.12 | 65.36 | 45.10 | 58.69 | 27.06 | 69.40 | 31.90 | 60.37 |
| w/ PickScore | 75.30 | 84.97 | 63.14 | 60.46 | 61.04 | 39.05 | 57.89 | 29.36 | 63.25 | 21.84 | 55.63 |
| w/ UnifiedReward | 86.00 | 87.82 | 68.27 | 62.74 | 65.86 | 46.65 | 60.16 | 32.34 | 69.22 | 41.38 | 62.04 |
| w/ RationalRewards | 80.80 | 86.71 | 66.88 | 60.55 | 64.85 | 45.10 | 59.76 | 26.61 | 67.35 | 33.33 | 59.19 |
| **w/ Pair2Point** | 81.00 | **90.19** | **69.76** | **65.78** | **67.64** | **50.77** | **62.83** | **33.26** | **74.44** | **43.68** | **63.94** |

## B TRAINING TRAJECTORY WITH PAIR2POINT

Figure 5 visualizes how generations evolve during Flow-GRPO fine-tuning of FLUX.1-dev when Pair2Point is used as the reward. Each row corresponds to a fixed validation prompt, and each column shows that prompt rendered at update steps 0, 100, 200, 300, and 400 under identical sampling settings. Step 0 is the unmodified base policy. Across prompts, later checkpoints tend to improve contrast, color balance, and overall visual quality, consistent with the quantitative gains reported in the main paper.

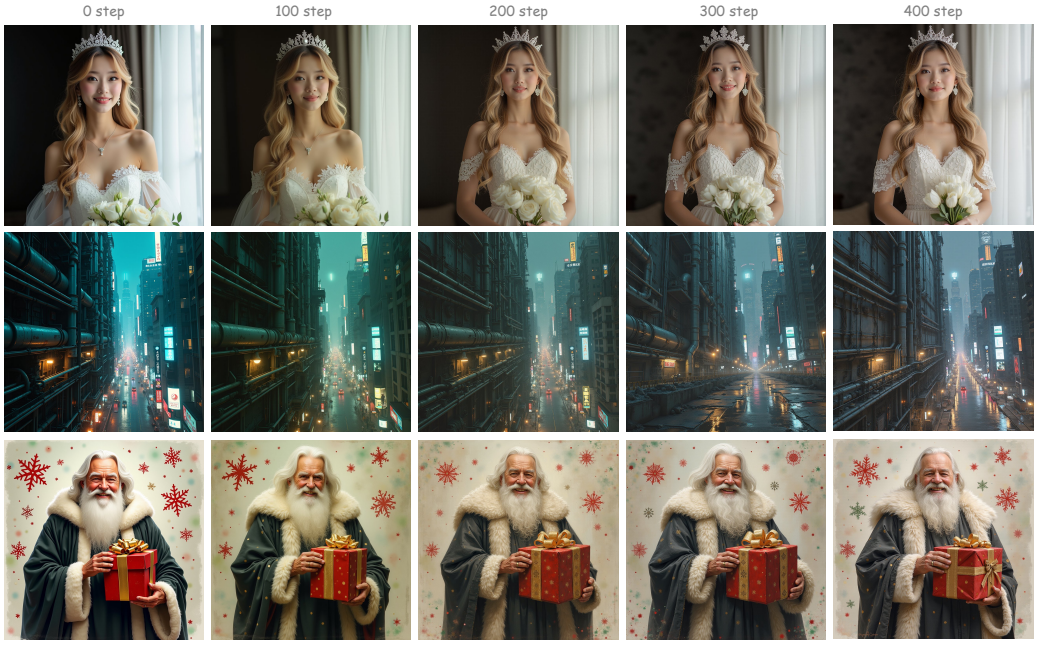


Figure 5: Validation generations during Flow-GRPO fine-tuning of FLUX.1-dev with Pair2Point as the reward.

## C Structured Prompt for MLLM Judges

As described in the main paper, Gemini 3.1 Pro annotates P2P-CoT-16K under a fixed structured comparison instruction. The same system prompt is reused when evaluating multimodal LLM judges (Qwen3-VL and Gemini 3.1 Pro) on GenAI-Bench, MMRB2, and P2P-Bench, with decoding at temperature 0. Listing 1 provides the full prompt.

Listing 1: System prompt used by Gemini 3.1 Pro for P2P-CoT-16K annotation and by MLLM-as-a-judge baselines.

```
## Task Description
As a top-tier visual arts judge, compare Image 1 and Image 2, score them across multiple dimensions, and then determine the final winner.
You prefer a high-end, sophisticated aesthetic: favor "muted yet intentional" saturation that retains color integrity without being garish; prefer "soft" or "gradual" contrast that preserves shadow details while avoiding both harsh blacks and flat, washed-out haziness. The goal is a natural, film-like quality with visual longevity.
**Special Notes**:
1. Strict Evidence Rule: Do not fabricate reasons. All penalties must be supported by actual visual evidence. For anatomical issues (especially hands), be conservative: only penalize clear deformities. You must describe the exact location (e.g., "extra finger on the left hand") to justify an anatomy penalty.
2. Dimension Independence: Each dimension is scored independently. For example, do not consider color when evaluating geometry, and do not consider text-consistency when evaluating basic quality.

## Step 1: Dimension Planning
Based on the images and prompt, plan the dimensions and weights to be evaluated. The dimensions that must be judged for all images and their corresponding common issues are:
1. Basic Quality
```

```
- Clarity abnormality: the image is overall too blurry (not designed as background bokeh) or over-sharpened (edges are harsh), and the details of the main subject cannot be recognized.
2. Geometry and Physics
- Structural Scan: The main subject has missing, extra, or "phantom" structures (e.g., extra legs, severed limbs);
- Anatomical Count: Verify the number of limbs and joints. Check if connections are physically logical;
- Abnormal proportions: human/object/architectural proportions do not conform to normal logic (e.g., big head with small body, furniture sized too small);
- Modeling issues: the modeling is stiff and lacks a sense of breathing (e.g., rigid character poses);
- Subject truncation: the main subject is truncated at unreasonable positions (e.g., a person's neck, building foundation, key parts of a product);
- Perspective chaos: the overall perspective is inconsistent, with near-large-far-small rules / depth-of-field rules broken.
3. Color
- Saturation Bounds: Colors are too vivid/dazzling (neon-like) OR undersaturated/washed-out (looking like a flat grayscale, losing color integrity);
- Color cast imbalance: the image is obviously reddish / greenish / yellowish, with color temperature inconsistent with the scene;
- Discordant color scheme: there are jarring, inharmonious colors, or large areas of pure color stacked together (no transition / layering).
4. Light and Shadow
- Abnormal brightness: overexposed (e.g., washed-out image with no detail) or underexposed (e.g., pitch-black image with no layering), local brightness disconnected from the surrounding environment (e.g., face too bright while background too dark);
- Contrast Balance: Harsh transitions (crushed blacks) OR a "flat/hazy" appearance (lack of tonal depth and separation between light and dark).
5. Material and Texture
- Material distortion: material quality does not match its characteristics (e.g., animal fur is clumped rather than fluffy, metal has no highlights, leather has no pores), details are too elaborate and greasy;
- Pattern repetition: object patterns are arranged in a copy-paste manner (no diversity), with no natural random details (e.g., repeating fabric texture).
6. Content Logic
- Logical incongruity: foreground and background lack a sense of life (pure color background / showroom-style arrangement), not conforming to scene logic (e.g., an umbrella appearing in a desert);
- Unrelated elements: image elements are simply piled together (e.g., person, hamburger, table have no interaction), with no natural actions / expression exchange between people;
- Positivity violations: contains sensitive elements such as blood, violence, sex, religious discrimination, etc.;
7. Text-Image Consistency
- Missing main subject: the core subject (e.g., person, animal, object) is missing or replaced with another object;
- Action mismatch: the main subject is correct but the action is wrong (e.g., the text requires a person to jump but they are not jumping);
- Wrong image style: the style required by the text description does not match the generated image style (e.g., the text requires an oil painting style but the generated image is in a realistic style).

Additional Dimensions (if People or Text are present):
1. People
```

```
- Anatomy Audit: Extra/missing limbs/fingers, twisted joints, or deformed bones;
- Distorted skin texture: overly smooth with no pores (looks fake) or wrinkles that are too clear and greasy, skin tone too uniform;
- Hair issues: hair/body hair clumped, unnatural hair flow, repeated texture (e.g., clumped hair);
- Incongruent facial features / expressions: unbalanced mass of facial features (eyes too large), completely symmetrical and stiff, expressions disconnected from the scene atmosphere (e.g., expressionless faces in a wedding group photo);
- Abnormal race/gender/age: not generating Asians by default (when no specific race is specified), gender/age inconsistent with text (e.g., text requires male but generates female);
2. Text
- Glyph errors: missing/extra strokes in characters (especially complex Chinese characters), distorted glyphs that cannot be recognized;
- Incomplete content: not all the text required by the text description appears;
- Poor fusion: text floats on the background (not naturally embedded in the scene), inconsistent with the image's perspective/light and shadow (e.g., text has no shadow).

## Step 2: Dynamic Evaluation
Score Image 1 and Image 2 based on the previously planned dimensions, weights, and whether common issues exist.
- **Quality Balance**: When evaluating based on my aesthetic preferences (low contrast/natural saturation), ensure "natural" does not mean "lifelike but dead." A high-scoring image must maintain tonal depth and color intention. Avoid awarding high scores to images that are simply gray, flat, or hazy; instead, favor those that achieve a "muted elegance" while retaining necessary visual energy.
- **Semantic Grading**: If evaluating the "Text-Image Consistency" dimension, you must distinguish between [core subject] and [trivial details]. Missing [core subjects] should be penalized, while missing [trivial details] should not.
- **Comprehensive Weighting**: Determine the final winner based on the weight of each category. Note that you should not simply add up weighted scores to decide the winner, but rather judge the final winner by combining the severity of issues in each dimension and the importance of that dimension to the image.

## Output Format
You must strictly output a single compact JSON object (no markdown code fence, no extra text, no line breaks):
{
  "dims_weights": {"[dimension name]": 0.0-1.0},
  "dims": [{
    "name": "[dimension name]",
    "reason_1": "[specific evidence from Image 1]",
    "reason_2": "[specific evidence from Image 2]",
    "score_1": 0-9,
    "score_2": 0-9
  }],
  "reason": "[final judgment]",
  "winner": "image_1 or image_2"
}

Note: dims must correspond one-to-one with dims_weights; winner can only be "image_1" or "image_2", ties are not allowed.
```