RELATIVE-MARGIN REWARD FOR CANDIDATE-SET RANKING IN TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image models generate multiple candidates per prompt, and reward models are used to select among or rank them. Although benchmarks with human preferences over multiple images exist, reward model evaluation still centers largely on pairwise accuracy. Whether this metric adequately reflects a model's ability to recover human rankings over entire candidate sets remains underexplored. We introduce Rank-Bench, a benchmark that recasts reward-model evaluation as tie-aware ranking over candidate sets, comprising 902 rankings over 6,236 images across 11 diagnostic dimensions. Rank-Bench reveals that comparable pairwise accuracy does not imply comparable set-level ranking agreement. We propose MarginRM, a relative-margin reward model trained to predict criterion-level signed ordinal margins and explanations under prompt-specific rubrics. At inference, it scores each candidate relative to a shared anchor under a common rubric, yielding a ranking over the full candidate set. Under all-pairs evaluation on Rank-Bench, MarginRM achieves 61.07% stable accuracy and a Kendall's τ_b of 0.519, compared with 54.47% and 0.487 for the strongest baseline. Single-anchor inference yields a τ_b of 0.485 using only 27.93% of the all-pairs comparison budget. As the reward for GRPO post-training of FLUX.1-dev, it improves GenEval from 63.72 to 72.61 and T2I-CompBench from 49.22 to 54.38, while producing rankings through shared-anchor comparisons without exhaustive pairwise evaluation.

1 INTRODUCTION

Text-to-image generation is increasingly a multi-sample decision problem. Across best-of-N selection, data curation, and reinforcement learning, multiple samples generated from the same prompt must be evaluated against one another (Fan et al., 2023; Black et al., 2024; Wallace et al., 2024). To support such decisions, reward models learned from human preferences are widely adopted to assign preference signals to generated images (Wu et al., 2023b). Their practical value hinges not on judging an isolated pair, but on inducing a reliable ordering over candidates for the same prompt.

Existing reward-model evaluations encompass pairwise preference agreement (Jiang et al., 2024; Hu et al., 2026) and multi-candidate selection (Li et al., 2024). These protocols primarily assess preference discrimination and selection quality; our focus is agreement with an explicit, tie-aware reference ranking over each prompt's full candidate set. To support this evaluation, we introduce Rank-Bench, which organizes candidates by prompt and aggregates human judgments into tie-aware consensus rankings, thereby enabling a holistic evaluation of reward-model fidelity to human-derived rankings. Spanning 11 diagnostic dimensions, Rank-Bench comprises 902 candidate-set rankings over 6,236 images. It reveals that comparable isolated-pair accuracy can correspond to markedly different agreement with human-derived reference rankings. Pair accuracy alone is therefore insufficient for selecting a reward model intended for candidate-set ranking or multi-sample optimization.

This discrepancy between isolated-pair accuracy and candidate-set ranking motivates a deeper question: how do existing reward model paradigms achieve comparability among candidates? Pointwise scalar models score candidates independently, providing a common numerical scale (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023a; Ma et al., 2025). Although the same scoring function is applied to all candidates, its evaluation criteria remain implicit, and the resulting score differences may not faithfully reflect relative quality. Their agreement with human rankings remains an empirical question. Pairwise comparative models make relative judgments explicit by processing two

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.

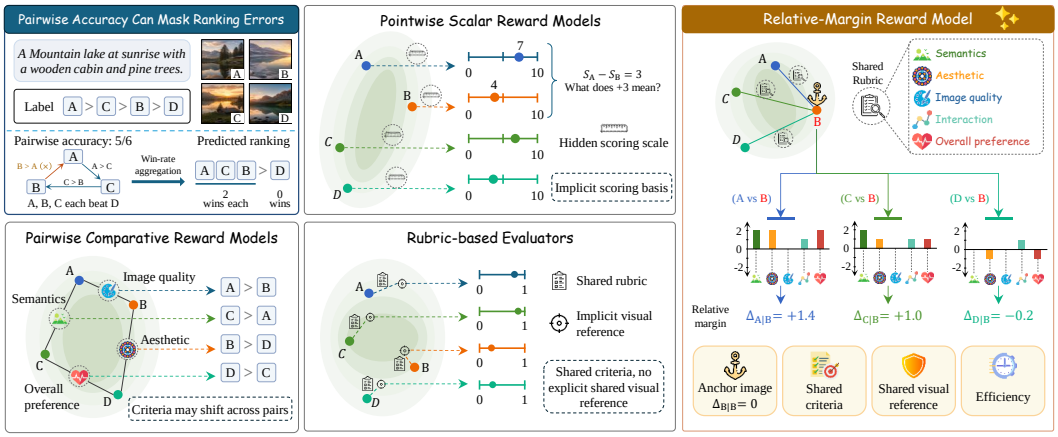


Figure 1: **Candidate-set ranking across reward-model paradigms.** Existing pointwise scalar reward models, pairwise comparative reward models, and rubric-based evaluators fail to jointly fix the evaluation criteria and the comparison reference. MarginRM combines a prompt-specific rubric with a shared visual anchor to produce anchor-relative scores for set-level ranking.

candidates jointly (Wang et al., 2025; 2026c;b), but their comparison criteria may shift with context—one pair distinguished by semantic alignment, another by aesthetics—with nothing enforcing consistency across decisions. Lacking a common criterion-level yardstick, aggregating these pairwise judgments into a set-level ranking can yield inconsistent or non-transitive preferences. Rubric-based evaluators fix the comparison axes through explicit criteria but lack a shared visual reference (Feng et al., 2025; Kao et al., 2026; Qian et al., 2026). Each candidate is assessed against the evaluator's internal prior of an ideal image rather than a common visual anchor. Consequently, two candidates may be scored under the same rubric yet against different implicit baselines. In summary, an explicit rubric specifies common evaluation criteria but does not by itself provide a shared visual reference (see Fig. 1). Conversely, a shared reference does not by itself specify the evaluation criteria. This motivates jointly specifying both for candidate-set ranking.

To unify explicit shared criteria with shared visual anchors, we propose **Relative-Margin Reward Model (MarginRM)**. It places all candidates of a given prompt within a shared comparison frame, defined by a prompt-derived rubric and an anchor sampled from the candidate set. The rubric fixes the axes of evaluation, while the anchor provides a common origin for measuring relative quality. Rather than reducing preference to a binary label, MarginRM predicts criterion-level signed ordinal margins for each candidate relative to the anchor. Each margin encodes the direction and magnitude of the candidate–anchor gap along one rubric dimension. These per-criterion margins can also be aggregated into an anchor-relative scalar reward for ranking. In addition, this shared-reference formulation replaces exhaustive pairwise comparisons with a single anchor comparison for each candidate. MarginRM consistently improves both conventional pairwise evaluation and candidate-set ranking. It reaches 75.2/76.1 accuracy on MMRB2 (Hu et al., 2026)/GenAI (Jiang et al., 2024) and 0.519 Kendall's τ_b on Rank-Bench under all-pairs evaluation, outperforming UnifiedReward-Flex (69.2/73.4 and 0.487, respectively). When used for Flow-GRPO post-training of FLUX.1-dev, MarginRM further improves GenEval from 63.72 to 72.61 and T2I-CompBench from 49.22 to 54.38, showing that its ranking advantages remain effective in downstream generation optimization.

Our contributions are:

- We introduce Rank-Bench, a tie-aware benchmark for evaluating reward models on candidate-set ranking, spanning 11 dimensions, 902 ranking instances, and 6,236 images.
- We propose MarginRM, a novel reward paradigm that explicitly models evaluation dimensions and visual reference points. It scores candidates via criterion-level relative margins defined by prompt-specific rubrics and a shared anchor.
- Experiments demonstrate improved pairwise prediction and all-pairs ranking agreement, a favorable cost–accuracy trade-off under shared-anchor inference, and gains in Flow-GRPO post-training of FLUX.1-dev.

2 Related Work

2.1 Text-to-image reward modeling

Reward models are central to evaluating and improving text-to-image generation. Early work adopts a pointwise scalar paradigm, compressing human preference into a scalar score for each prompt–image pair. HPS, HPSv2, ImageReward, and PickScore learn such scoring functions from preference data for evaluation and data filtering (Wu et al., 2023b; Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023a). HPSv3 and HPSv3++ broaden preference supervision and improve fine-grained discrimination (Ma et al., 2025; Liu et al., 2026b), while DiT-Reward derives stronger reward representations from pretrained diffusion transformers (Yang et al., 2026). At inference, however, these models score each candidate independently rather than comparing candidates jointly. The pairwise comparative paradigm partially alleviates this issue. UnifiedReward makes candidate differences observable to the evaluator (Wang et al., 2025). UnifiedReward-Think adds explicit reasoning prior to preference prediction (Wang et al., 2026c), while UnifiedReward-Flex adapts this reasoning to image content and user context (Wang et al., 2026b). RationalRewards enriches scalar judgment with structured, multi-dimensional critiques (Wang et al., 2026a). These advances make relative preference increasingly explicit, yet each comparison remains local. Rubric-based evaluators decompose preferences into explicit criteria. VisionReward scores visual preference through weighted fine-grained assessments (Xu et al., 2026), while RubricRL makes these criteria prompt-adaptive via dynamic checklists (Feng et al., 2025). AutoRubric-T2I and Auto-Rubric as Reward derive natural-language criteria directly from preference supervision (Kao et al., 2026; Tian et al., 2026). DyCoRM grounds task-relevant criteria before criterion-conditioned judgment (Qian et al., 2026), while RubricRM generates input-specific rubrics for fine-grained comparison (Kan et al., 2026). Yet explicit criteria alone cannot provide a shared visual reference. MarginRM combines explicit prompt-specific criteria with a shared visual anchor across the candidate set.

2.2 Reward-model evaluation and candidate ranking

Reward-model evaluation has largely inherited the pairwise structure of human-preference data, with models typically assessed by their accuracy in recovering held-out human choices. ImageReward and Pick-a-Pic establish this protocol on human-annotated image pairs (Xu et al., 2023; Kirstain et al., 2023). HPSv2 and HPSv3 expand the scale and generator coverage of preference data, yielding increasingly challenging test sets (Wu et al., 2023a; Ma et al., 2025). Subsequent benchmarks broaden what is tested while retaining the pairwise unit: MHP enriches annotations from overall preference to multiple dimensions (Zhang et al., 2024); MJ-Bench evaluates multimodal judges across alignment, visual quality, safety, and bias (Chen et al., 2026); and Multimodal Reward-Bench (MMRB2) extends evaluation to text-to-image generation, image editing, interleaved generation, and multimodal reasoning (Hu et al., 2026). Several recent benchmarks move closer to the candidate-set setting. GenAI-Rank collects human ratings for multiple images and evaluates best-of-N (Li et al., 2024). While this establishes an important ranking precedent, its supervision consists of independently assigned ratings rather than agreement with a human consensus order. DyCoRM likewise evaluates selectors through human rankings of outputs chosen by different reward models, yet its held-out reward-model benchmark remains pairwise (Qian et al., 2026). Consequently, pairwise accuracy captures whether a model makes the correct local choice, but not whether its judgments induce the correct global ordering over all candidates for the same prompt.

3 Methods

This section describes the construction of Rank-Bench and the proposed MarginRM framework. We first explain how pairwise human judgments were collected and consolidated into prompt-level consensus rankings. We then describe how MarginRM constructs prompt-specific evaluation rubrics, converts binary preference data into criterion-level relative-margin supervision, trains the reward models and ranks candidate images through shared-anchor inference.

3.1 Rank-Bench construction

Rank-Bench evaluates whether a reward model can recover the human-consensus ranking of multiple images. The construction process is shown in Fig. 2, with further details in Appendix A.

Prompt Taxonomy. The benchmark spans four evaluation groups covering eleven diagnostic dimensions. Prompts were drawn from public benchmarks and supplemented with LLM-drafted

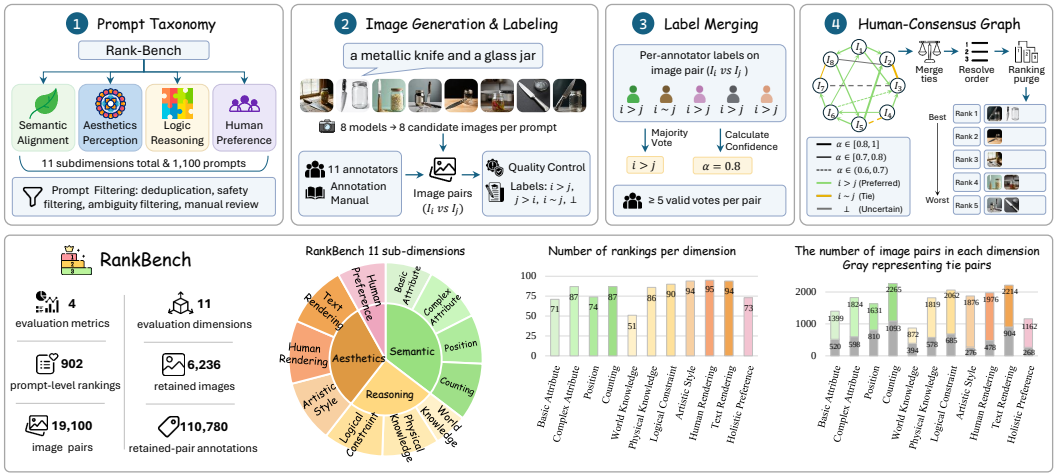


Figure 2: **Overview of Rank-Bench construction.** We construct Rank-Bench through prompt taxonomy, multi-model candidate generation and human pairwise annotation, confidence-aware label merging, and preference-graph-based consensus ranking. The bottom panel summarizes the benchmark scale and distribution across 11 evaluation dimensions.

ones, then manually screened to exclude unsafe content and ambiguous or under-specified visual intent (Ghosh et al., 2023; Huang et al., 2023; Niu et al., 2025).

Image generation and labeling. For each prompt, we generated eight images using eight text-to-image models spanning diverse architectures and training data (Podell et al., 2024; Esser et al., 2024; Cai et al., 2025; Wu et al., 2025; DeepMind, 2025), yielding meaningful within-set variation in fidelity, compositional correctness, and perceptual quality. All unordered pairs within each candidate set were annotated by at least five of eleven trained annotators, each selecting among four outcomes: $I_i \succ I_j$, $I_j \succ I_i$, $I_i \sim I_j$ (tie), or $\perp$ (uncertain). To mitigate model-identity bias, images were presented in randomized order with all metadata stripped.

Label merging. Let $N_{i\succ j}$, $N_{j\succ i}$, and $N_{i\sim j}$ denote the votes for the three valid outcomes, and let N_{ij} be the total number of valid votes. We define the confidence of the highest-vote outcome as:

$$\alpha_{ij} = \frac{\max\{N_{i\succ j}, N_{j\succ i}, N_{i\sim j}\}}{N_{ij}}. \tag{1}$$

A pair was retained only if it had at least five valid votes $N_{ij} \geq 5$ and a unique modal outcome with $\alpha_{ij} \geq 0.8$. Pairs failing either condition received additional annotations, up to eleven distinct annotators, and were discarded if the conditions remained unmet.

Consensus ranking construction. For each prompt, the retained pairwise relations form a directed preference graph $\mathcal{G}$: a strict preference $I_i \succ I_j$ adds an edge $I_i \to I_j$, while a tie $I_i \sim I_j$ adds edges in both directions. An image was removed only if it had no retained relation after filtering. Since ties can induce directed cycles, we computed the strongly connected components (SCCs) of $\mathcal{G}$ and contracted each into a consensus block (Agrawal et al., 2026; Atkinson et al., 2025). Images within a block share a common rank tier, and inter-block edges are preserved, yielding a condensation DAG $\mathcal{G}^*$ that encodes a partial order over consensus blocks. To resolve block pairs left unordered by $\mathcal{G}^*$, we fitted a block-level Davidson model (Davidson, 1970). The final reference ranking is a Davidson-guided topological extension that preserves all inter-block constraints. SCC contraction may absorb strict preferences within a block. All SCCs underwent manual review before the reference rankings were finalized (see Appendix A.3).

3.2 Relative Margin Reward Modeling

MarginRM establishes a novel reward evaluation framework based on explicit evaluation criteria and shared visual anchors. At inference, the rubric establishes a unified standard for evaluation and the shared anchor provides a common visual reference. The model learns to predict signed ordinal margins and corresponding textual explanations between the candidate and anchor images.

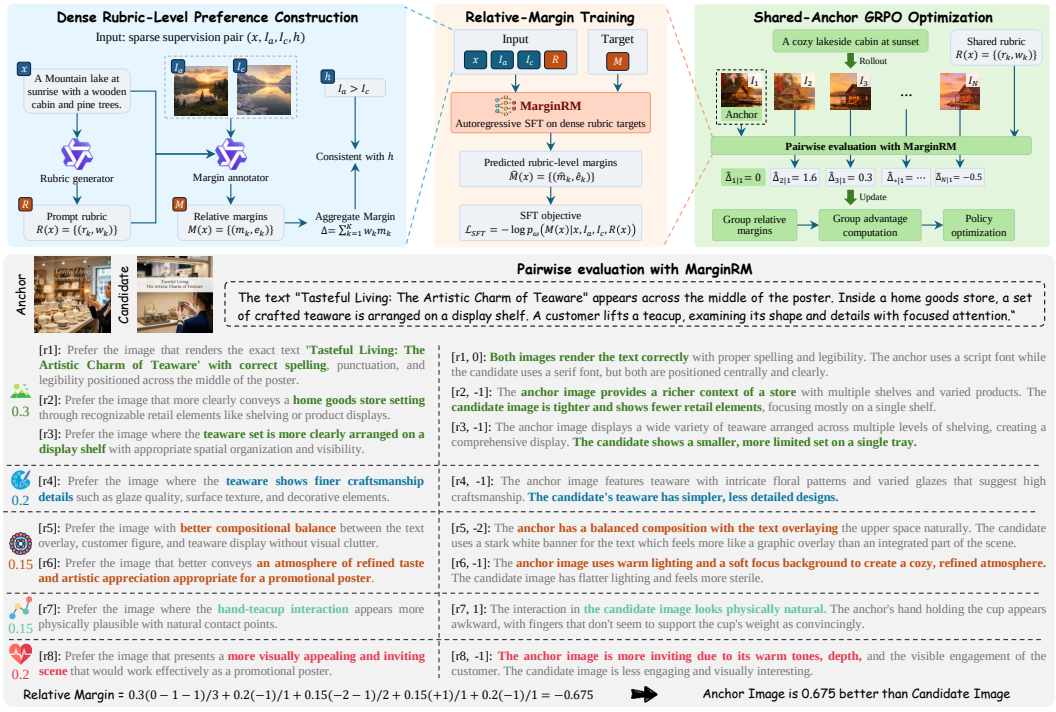


Figure 3: **MarginRM Overview.** From sparse human preferences, we derive prompt-specific rubrics and dense rubric-level margins that aggregate consistently with human judgments. MarginRM learns to predict these margins conditioned on the prompt, image pair, and shared rubric. We integrate MarginRM into GRPO, scoring candidates against a common anchor and converting their relative margins into group advantages for policy optimization. In the bottom panel, MarginRM judgments aggregate into an interpretable signed margin for pairwise preference prediction.

Relative-margin training-data construction. We build our training set from HPDv3 (Ma et al., 2025). A preference sample (x, I_a, I_c, h) comprises a prompt x, an anchor image I_a, a candidate image I_c, and a binary label h indicating whether I_a is preferred over I_c. As shown in Fig. 3, the rubric generator produces a rubric $\mathcal{R}(x)$ from the prompt x alone, independent of any candidate image or preference label. By construction, the rubric cannot be tailored to the outcomes it later judges. The rubric is organized hierarchically (see Appendix B.1 for details). From five predefined dimensions, the generator adaptively activates a prompt-dependent subset of D_x dimensions, where $D_x \in \{2, \ldots, 5\}$. Each activated dimension receives a normalized weight $\lambda_d \geq 0$ ($\sum_{d=1}^{D_x} \lambda_d = 1$) and is decomposed into fine-grained items. Flattening all items into a length-K list yields $\mathcal{R}(x)$:

$$\mathcal{R}(x) = \big((r_1, w_1), \ldots, (r_K, w_K)\big), \qquad w_k \geq 0, \qquad \sum_{k=1}^{K} w_k = 1. \tag{2}$$

Here, w_k denotes the effective weight of criterion r_k. Given $\mathcal{R}(x)$, the margin annotator evaluates each candidate I_c against the reference I_a across all criteria. For each criterion r_k, the evaluation model generates a corresponding criterion-level explanation e_k and a signed ordinal margin $m_k \in \{-2, -1, 0, +1, +2\}$. The margin encodes the comparison direction: positive values indicate that the candidate is preferred, negative values indicate that the anchor is preferred, and zero indicates no discernible difference. Its magnitude distinguishes between minor and substantial differences. We then aggregate these criterion-level margins into a single preference score:

$$\Delta_{c|a} = \sum_{k=1}^{K} w_k m_k. \tag{3}$$

$\Delta_{c|a} > 0$ (< 0) favours the candidate (anchor). We retain only generated supervision whose direction matches the source human preference, augmenting it with prompt-specific criteria, criterion-level explanations, and relative margins. This encoding defines an aggregation rule rather than an empirically calibrated interval scale; Appendix B.3 discusses its interpretation and limitations.

Relative-margin training. We initialize MarginRM from Qwen3.5-9B. Each training sample u_x comprises the prompt x, its rubric $\mathcal{R}(x)$, and the anchor and candidate images I_a, I_c. The supervision target $y_x^\star$ is a criterion-wise sequence of explanations e_k and signed ordinal margins m_k. The model, parameterized by ω, is optimized via teacher-forced cross-entropy over the serialized target tokens. Letting $L_x = |y_x^\star|$ denote the target length, the per-token loss is:

$$\mathcal{L}_{\mathrm{SFT}}(\omega) = -\mathbb{E}_{(u_x, y_x^\star)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\frac{1}{L_x}\sum_{t=1}^{L_x}\log p_\omega\left(y_{x,t}^\star \mid u_x, y_{x,<t}^\star\right)\right], \tag{4}$$

where $\mathcal{D}_{\mathrm{SFT}}$ denotes the fine-tuning dataset. Input and padding tokens did not contribute to the loss.

Shared-anchor GRPO. During the GRPO stage, for each prompt x the policy generates a set of N images $\mathcal{I} = \{I_1, \ldots, I_N\}$ (Liu et al., 2026a). We designate the first image as the shared anchor I_a, fix its relative score to zero ($\Delta_{1|1} = 0$), and evaluate each remaining image against it under the rubric $\mathcal{R}(x)$. For each non-anchor candidate I_n ($n > 1$), MarginRM predicts a textual explanation e_k and a margin m_k for every criterion r_k. These margins are aggregated via Eq. 3 into a relative score $\Delta_{n|1}$ quantifying the quality of I_n against the anchor, placing all images in a unified relative-to-anchor coordinate system. We then convert these relative rewards into group-normalized advantages: for each prompt, the advantage of rollout I_n is its standardized relative reward within the group:

$$\hat{A}_n = \frac{\Delta_{n|1} - \mu_x}{\sigma_x + \epsilon}, \tag{5}$$

where μ_x and σ_x are the mean and standard deviation of the anchor-relative rewards across all N rollouts, respectively, and ϵ is a small constant. Although the anchor has zero raw relative score, its normalized advantage can be nonzero.

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{x,\mathcal{I}\sim\pi_{\theta_{\mathrm{old}}}}\left[\frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\Big(\min\left(\rho_{n,t}\hat{A}_n, \mathrm{clip}(\rho_{n,t}, 1-\varepsilon, 1+\varepsilon)\hat{A}_n\right)\Big)\right], \tag{6}$$

where $\rho_{n,t}$ is the importance ratio along the generation trajectory of I_n, ε is the clipping coefficient, and $\pi_{\theta_{\mathrm{old}}}$ is the behaviour policy used to collect the rollouts (see Appendix C for the full derivation).

4 Experiments

We evaluate MarginRM on pairwise preference prediction and candidate-set ranking, and examine its effectiveness as a reward model for text-to-image post-training. We further investigate the contributions of its key components, sensitivity to anchor choice, inference efficiency, and robustness.

4.1 Experimental Setup

We evaluate reward-model performance on MMRB2, GenAI-Bench, and our proposed Rank-Bench. MMRB2 and GenAI-Bench measure pairwise preference prediction, while Rank-Bench evaluates both pairwise and candidate-set ranking. We compare three reward paradigms: pointwise models rank candidates by scalar scores, whereas general and pairwise models compare all within-prompt pairs and aggregate them into set-level rankings by win rate. MarginRM is initialized from Qwen3.5-9B and fine-tuned on 532K samples (Appendix B.4). For downstream evaluation, we use MarginRM to guide GRPO post-training of FLUX.1-dev and evaluate the generators on GenEval and T2I-CompBench, with CLIP and PickScore as complementary metrics (Appendix C.3).

4.2 Main Results

Table 1 compares MarginRM against general models, pointwise reward models, and pairwise reward models on MMRB2, GenAI-Bench, and Rank-Bench, where MarginRM achieves the strongest overall performance. On Rank-Bench, pointwise models derive rankings by sorting scalar scores, whereas general-purpose and pairwise models aggregate intra-prompt pairwise preferences into rankings via win rates; MarginRM follows the same all-pairs strategy for fairness. MarginRM consistently outperforms the strongest baseline, UnifiedReward-Flex, achieving 75.2% vs. 69.2% accuracy on MMRB2 and 76.1% vs. 73.4% on GenAI-Bench. On Rank-Bench, it reaches 64.34% vs. 62.05% tie-inclusive accuracy and, more notably, 61.07% vs. 54.47% stable accuracy, indicating greater robustness to input-order reversal. Under the same all-pairs protocol, MarginRM further achieves a higher Kendall's τ_b (0.519 vs. 0.487). Because both models use the same win-rate aggregation, the higher ranking agreement reflects improvements in the pairwise judgments produced

Table 1: **Performance on MMRB2, GenAI-Bench and Rank-Bench.** Stable accuracy is the proportion of image pairs for which the model predicts the correct preference both before and after swapping the input image order. (w/ Tie) includes all annotated relations, while (w/o Tie) excludes ties. Kendall's τ_b measures agreement between model and human consensus rankings.

Model	MMRB2	GenAI-Bench	Rank-Bench			
	Accuracy	Accuracy	Stable Accuracy	Accuracy (w/ Tie)	Accuracy (w/o Tie)	Kendall's τ_b
General Models						
CLIP (Radford et al., 2021)	51.0	50.8	-	35.24	52.21	0.085
LLaVA-8B (Liu et al., 2023)	54.7	51.9	3.12	33.38	51.02	0.047
Qwen3-VL-8B (Bai et al., 2025)	56.3	53.1	31.41	46.71	69.96	0.407
Qwen3.5-9B (Qwen Team, 2026)	60.1	59.7	31.81	47.96	70.28	0.430
Pointwise Scalar Models						
PickScore (Kirstain et al., 2023)	58.6	70.0	-	44.15	67.49	0.293
ImageReward (Xu et al., 2023)	54.0	55.8	-	41.60	63.58	0.228
HPS v2 (Wu et al., 2023a)	54.7	68.8	-	38.85	59.13	0.160
HPS v3 (Ma et al., 2025)	60.2	70.9	-	48.72	69.95	0.401
Pairwise Comparative Models						
UnifiedReward (Wang et al., 2025)	59.8	71.5	29.42	44.74	68.37	0.350
UnifiedReward-Think (Wang et al., 2026c)	66.0	72.3	45.28	51.18	71.46	0.483
UnifiedReward-Flex (Wang et al., 2026b)	69.2	73.4	54.47	62.05	75.18	0.487
Relative-Margin Reward Model						
MarginRM	**75.2**	**76.1**	**61.07**	**64.34**	**77.70**	**0.519**

Table 2: **Ablation study of MarginRM.** Each block varies one component while keeping the remaining training and evaluation settings fixed. MMRB2 and GenAI-Bench report accuracy (%). Rank-Bench reports stable accuracy, pairwise accuracy with and without ties (%), and Kendall's τ_b for agreement with the human consensus ranking.

Analysis	MMRB2	GenAI-Bench	Rank-Bench			
	Accuracy	Accuracy	Stable Accuracy	Accuracy (w/ Tie)	Accuracy (w/o Tie)	Kendall's τ_b
Supervision Design						
Binary SFT	64.6	71.1	58.78	62.39	75.59	0.473
Margin only SFT	70.1	71.7	58.96	63.46	76.73	0.508
Rubric Generator Robustness						
Qwen3-VL-8B	72.6	74.9	58.56	63.40	76.36	0.496
Qwen3.5-9B	73.9	74.8	60.65	64.25	77.58	0.518
Aggregation Weight						
Uniform weights	73.3	74.5	59.34	63.56	76.17	0.512
Backbone						
Qwen3-VL-8B	71.1	74.8	58.68	63.44	76.51	0.501
Full margin supervision + Qwen3.5-397B-A17B rubric + specific weights + Qwen3.5-9B backbone						
MarginRM	**75.2**	**76.1**	**61.07**	**64.34**	**77.70**	**0.519**

by MarginRM. By evaluating candidates within a shared reference frame, MarginRM improves both local preference discrimination and the prompt-level rankings. This also suggests that learning criterion-level margins and explanations helps capture human preferences more accurately.

4.3 Ablation Study

As shown in Table 2, we ablate four key design choices of MarginRM: the supervision formulation, the rubric-generation model, the weight-aggregation formulation, and the backbone. By default, we use explanation-plus-margin supervision, Qwen3.5-397B-A17B for rubric generation, prompt-specific aggregation weights, and Qwen3.5-9B as the backbone.

Supervision Design. Starting from binary SFT, replacing the binary target with a margin-only objective yields substantial gains (MMRB2: $64.6\% \rightarrow 70.1\%$; Rank-Bench Kendall's τ_b: $0.473 \rightarrow 0.508$). These results favor the richer criterion-level supervision over the binary target in this setting. Adding criterion-level explanations further improves performance on all three benchmarks (MMRB2 75.2%, GenAI-Bench 76.1%, Rank-Bench τ_b 0.519). Notably, stable accuracy changes little with margin-only supervision (58.78% to 58.96%) but increases to 61.07% when explanations are included. The two components are thus complementary: margin supervision encodes preference intensity, while explanations ground comparisons in visual evidence and prompt-specific criteria.

Rubric Generator Robustness. We examine whether MarginRM critically depends on a highly capable model for constructing training rubrics by regenerating them with Qwen3-VL-8B and Qwen3.5-9B, keeping the reward-model training procedure fixed. Stronger rubric generators generally yield better downstream performance, but the gap between the largest generator and Qwen3.5-9B is small. Crucially, the small gap between Qwen3.5-9B and Qwen3.5-397B-A17B shows that MarginRM's effectiveness is not tied to a particular large rubric generator: substantially smaller models retain most of the downstream performance.

Aggregation Weight. We further study whether rubric-level judgments should be aggregated uniformly or by prompt-specific importance. Replacing prompt-specific weights with uniform aggregation consistently degrades performance across all benchmarks (MMRB2 and GenAI-Bench drop by 1.9 and 1.6 points). Although smaller than the effect of the supervision formulation, this consistent drop indicates that prompt-specific weighting captures useful prompt-dependent structure, allowing the aggregate margin to emphasize the criteria most diagnostic for the requested content.

Backbone. We further ablate the multimodal backbone by replacing the default Qwen3.5-9B with Qwen3-VL-8B. While the stronger backbone consistently improves performance, it does not account for MarginRM's gains: on Rank-Bench, upgrading the backbone raises Kendall's τ_b by only 0.018, versus 0.046 from replacing binary-preference SFT with full MarginRM supervision. Model capacity and our supervision formulation are thus complementary, with much of the improvement stemming from how the preference signal is represented rather than from backbone scaling alone.

4.4 Inference efficiency and the effect of multiple anchors

Table 3 reports mean per-pair latencies of 2.2s for MarginRM and 4.2s for UnifiedReward-Flex over 200 runs per model. With one random anchor, MarginRM achieves comparable performance at 27.93% of the all-pairs budget; with two, it surpasses UnifiedReward-Flex on all four metrics at 55.85%. Increasing anchors from one to three improves all metrics, including stable accuracy (55.19% to 61.56%) and Kendall's τ_b (0.485 to 0.511). Candidates substantially better than one anchor may receive similarly large margins, obscuring finer differences that another reference could reveal. A single inaccurate margin can also distort multiple pairwise orderings. Averaging across anchors reduces reliance on individual comparisons and, when errors vary across references, can suppress anchor-specific distortions while preserving consistently supported quality differences.

Table 3: Efficiency and ranking performance on Rank-Bench. Time is measured per pair, and budgets are relative to all-pairs evaluation. $\mathrm{Acc_t}$ and $\mathrm{Acc_{nt}}$ denote accuracy with and without ties.

Method	Time (s)	Setup		Rank-Bench ↑			
		Strategy	Budget (%)	Stable Acc.	Acc_t	Acc_{nt}	τ_b
UR-Flex	4.2	All pairs	100.00	54.47	62.05	75.18	0.487
MarginRM	2.2	1 anchor	27.93	55.19	62.15	75.11	0.485
		2 anchors	55.85	58.84	63.36	76.02	0.495
		3 anchors	83.78	**61.56**	**64.37**	**77.83**	0.511
		All pairs	100.00	61.07	64.34	77.70	**0.519**

4.5 MarginRM for post-training

We further employ MarginRM for RL post-training. Within each rollout group, we designate the first generated image as a shared anchor, score the remaining samples relative to it under the same prompt-specific rubric, and derive group-relative advantages from the predicted margins. Table 4 shows that HPSv2 and HPSv3 increase PickScore over FLUX.1-dev but reduce GenEval, T2I-CompBench, and CLIP. Direct PickScore optimization achieves the highest PickScore, yet lowers GenEval and T2I-CompBench. MarginRM consistently improves GenEval (63.72 to 72.61), T2I-CompBench (49.22 to 54.38), and CLIP (35.27 to 38.44), surpassing the strongest baseline, UnifiedReward-Flex. These results support the transfer of MarginRM's gains to downstream policy optimization, with improvements across complementary measures of human preference and text–image alignment.

Table 4: Post-training performance.

Method	**GenEval**	**T2I-CompBench**	**CLIP**	**PickScore**
FLUX.1-dev	63.72	49.22	35.27	22.61
w/ HPSv2	56.93	42.14	31.45	22.82
w/ HPSv3	60.84	45.39	32.12	22.74
w/ PickScore	56.96	44.12	33.48	**23.93**
w/ UnifiedReward	65.94	50.23	35.53	22.97
w/ UnifiedReward-Think	67.30	50.86	35.91	22.62
w/ UnifiedReward-Flex	68.91	51.22	36.49	23.38
w/ MarginRM	**72.61**	**54.38**	**38.44**	23.36

4.6 Anchor Sensitivity Analysis

We examine MarginRM's sensitivity to anchor quality by using each candidate in turn as the anchor and measuring Kendall's τ_b against human consensus rankings. Fig. 4 plots ranking agreement

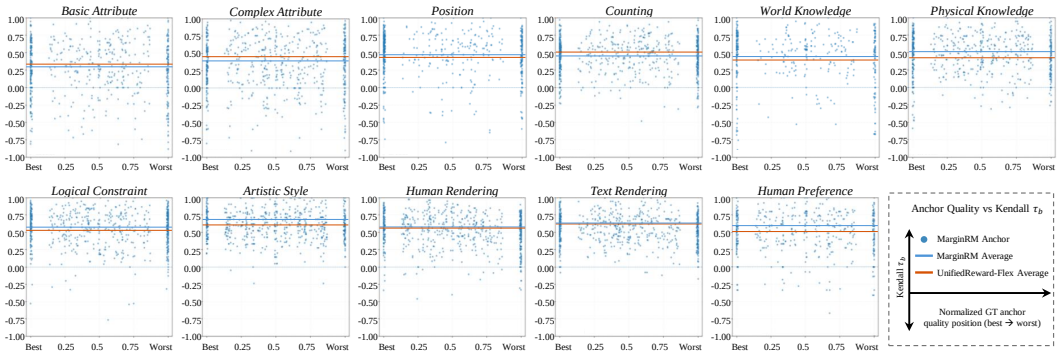


Figure 4: Anchor sensitivity of MarginRM across Rank-Bench dimensions. Each point reports the Kendall's τ_b when a given candidate serves as the shared anchor, plotted against its normalized position in the human consensus ranking (best to worst). The blue line marks MarginRM's mean τ_b over all anchors within each dimension; the orange line shows UnifiedReward-Flex.

against the anchor's normalized ground-truth position across Rank-Bench dimensions. Averaged over all anchors, MarginRM matches or exceeds UnifiedReward-Flex in mean Kendall's τ_b on most dimensions, demonstrating competitive average performance without selecting the best-performing anchor. We observe no clear monotonic relationship between anchor quality and ranking agreement: high agreement occurs across the observed quality range, although performance varies across anchor choices. These findings suggest that an anchor's image quality does not necessarily reflect its usefulness as a shared comparison reference. An anchor's absolute rank may therefore be an incomplete indicator of its usefulness as a comparison reference. Its visual attributes may make certain criterion-level differences more discernible than others, influencing how informative the resulting comparisons are for recovering the candidate-set ranking.

4.7 Reversal Stability

Fig. 5 evaluates the robustness of pairwise judgments to input-order reversal. Stable accuracy requires predictions to match the human label in both presentation orders, after mapping the reversed prediction back to the original orientation. MarginRM achieves the highest stable accuracy of 61.07% and the smallest forward–stable gap of 3.3 percentage points, compared with 5.9–30.3 points for the baselines. This gap measures the proportion of evaluated pairs that are correctly judged in the forward order but incorrectly judged after reversal. With a forward accuracy of 64.34%, approximately 94.9% of MarginRM's forward-correct predictions remain correct when the images are swapped. This robustness is consistent with MarginRM's separation of evaluation criteria from comparison direction. Signed margins make preference direction explicit, while criterion-level explanations ground the comparison in visual evidence.

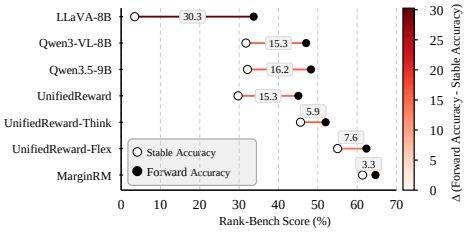


Figure 5: Reversal stability on Rank-Bench. The annotated gap measures the decrease from forward to stable accuracy.

5 Conclusion

We study text-to-image reward modeling through candidate-set ranking and introduce Rank-Bench, a tie-aware benchmark that evaluates agreement with human consensus rankings. Our results show that pairwise accuracy alone does not fully characterize set-level ranking quality. We propose MarginRM, which establishes a shared comparison frame through a prompt-specific rubric and a common visual anchor. It learns to predict criterion-level signed margins with accompanying textual explanations and supports shared-anchor inference to produce directly comparable relative rewards. Experiments demonstrate improvements in pairwise preference prediction, ranking agreement, and stability, as well as gains in generation quality when MarginRM guides Flow-GRPO post-training of FLUX.1-dev. Together, these findings highlight the value of jointly fixing evaluation criteria and visual references for reward modeling in multi-sample selection and optimization.

AI USE STATEMENT

We used generative AI for synthetic data construction and structured training supervision. GPT-5 drafted supplementary prompts for Rank-Bench, all of which underwent human review before inclusion (Appendix A.1). The eight text-to-image systems listed in Appendix A.2 generated the benchmark images. Qwen3.5-397B-A17B generated and validated prompt-specific rubrics and produced criterion-level ordinal margins and explanations for HPDv3 training pairs (Appendices B.1–B.2). Generated supervision was checked for structural validity and retained only when its aggregate preference direction agreed with the source human label. Rank-Bench's pairwise preference annotations were provided by human annotators; its reference rankings were constructed through the aggregation and manual-adjudication procedures described in Appendix A.3, rather than supplied by an LLM. We also consulted ChatGPT for manuscript drafting and revision, as well as for critical feedback on the methodology, mathematical presentation, and interpretation of experimental results. We take responsibility for the final content of this work, including its text, claims, results, and artifacts produced with generative AI assistance.

REFERENCES

Sheshansh Agrawal, Thien Hang Nguyen, and Douwe Kiela. Blitzrank: Principled zero-shot ranking agents with tournament graphs. *arXiv preprint arXiv:2602.05448*, 2026.

Nathan Atkinson, Scott Ganz, Dorit Hochbaum, and James Orlin. The strong maximum circulation algorithm: A new method for aggregating preference rankings. *INFORMS Journal on Optimization*, 7(2):142–155, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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

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.

Google DeepMind. Gemini 2.5 flash and gemini 2.5 flash image model card, 2025.

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

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in neural information processing systems*, 36:79858–79885, 2023.

Xuelu Feng, Yunsheng Li, Ziyu Wan, Zixuan Gao, Junsong Yuan, Dongdong Chen, and Chunming Qiao. Rubricrl: Simple generalizable rewards for text-to-image generation. *arXiv preprint arXiv:2511.20651*, 2025.

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

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36904–36915, 2026.

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

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

Zijian Kan, Wei Wang, Long Luo, Bing Zhao, Xuan Ren, Weixu Qiao, Wenbo Li, Hu Wei, and Lin Qu. Rubricrm: Generative reward modeling via dynamic rubrics for image generation and editing. *arXiv preprint arXiv:2608.26956*, 2026.

Kuei-Chun Kao, Daixuan Huo, Yuanhao Ban, and Cho-Jui Hsieh. Autorubric-t2i: Robust rule-based reward model for text-to-image alignment. *arXiv preprint arXiv:2605.17602*, 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.

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

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

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*, 2026b.

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

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Jiaying Qian, Ziheng Jia, Qian Zhang, Zicheng Zhang, Jiayi Guo, Junqi Zhang, Guangtao Zhai, and Xiongkuo Min. Dycorm: Dynamic criterion-aware reward modeling for text-to-image generation. *arXiv preprint arXiv:2605.25876*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

Juanxi Tian, Fengyuan Liu, Jiaming Han, Yilei Jiang, Yongliang Wu, Yesheng Liu, Haodong Li, Furong Xu, and Wanhua Li. Auto-rubric as reward: From implicit preferences to explicit multimodal generative criteria. *arXiv preprint arXiv:2605.08354*, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 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, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 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*, 2023a.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2096–2105. IEEE, 2023b.

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.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Yuanming Yang, Guoqing Ma, Bo Wang, Yuan Zhang, Wei Tang, Chenyi Li, Haoyang Huang, and Nan Duan. Dit-reward: Generative representations for text-to-image reward modeling. *arXiv preprint arXiv:2606.23626*, 2026.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027. IEEE, 2024.

A RANK-BENCH

Rank-Bench is designed to diagnose whether a reward model can recover human preferences over a prompt-conditioned candidate set, rather than merely predict isolated pairwise labels. The construction pipeline comprises four stages: (i) prompt taxonomy design, (ii) image generation and labeling, (iii) label merging, and (iv) human consensus aggregation. Each stage is detailed below.

A.1 PROMPT TAXONOMY

The benchmark is organized into four evaluation groups—semantic fidelity, reasoning, aesthetics, and holistic human preference—spanning eleven diagnostic dimensions in total. The taxonomy was constructed to satisfy three design criteria:

1. **Coverage of consequential failure modes.** It captures the dominant error types of contemporary T2I systems, ranging from local attribute errors to failures that require world or physical knowledge.
2. **Separation of decision criteria.** It disentangles relatively objective prompt-grounded constraints from perceptual and holistic judgments, so that the same reward model can be analysed under qualitatively distinct decision criteria.
3. **Annotation tractability.** It remains compact enough to support dimension-specific annotation manuals and reliable annotator training, yet fine-grained enough to expose where ranking performance diverges across models.

The initial taxonomy was informed by the prompt and evaluation structures of GenEval, T2I-CompBench, and WISE. Under-represented dimensions and boundary cases were then expanded by prompting GPT-5 to draft additional candidate prompts. GPT-5 was used solely for prompt drafting; it did not provide image-preference labels nor participate in consensus ranking. All machine-generated prompts were manually reviewed prior to inclusion.

The operational definitions of the eleven diagnostic dimensions are given below. These definitions guided both prompt curation and the dimension-specific annotation manuals. Annotators were instructed to prioritize the target dimension when comparing a pair, while treating severe off-target defects that rendered an image unusable as valid evidence against that image.

SEMANTIC FIDELITY

1. *Basic attribute*. The prompt involves a single object instance characterized by simple attributes such as color or shape.
2. *Complex attribute*. Multiple object entities are involved, each with its own attributes (e.g., color, shape, texture) that must be jointly satisfied.
3. *Position*. Explicit spatial relationships and object placement, including left/right, top/bottom, front/back, inside/outside, between, and relative position within the frame.
4. *Counting*. The exact number of objects or repeating elements requested by the prompt. Missing, duplicated, merged, or ambiguous instances are treated as counting failures.

REASONING

1. *World knowledge*. Whether the model can translate implicit factual, cultural, or commonsense knowledge into coherent visual content. Prompts may allude to knowledge—e.g., historical or cultural associations—that is not expressed as an explicit list of visual markers.
2. *Physical knowledge*. The visible consequences of physical conditions and interactions, including states of matter, temperature, gravity, and support. The criterion is consistency between the specified conditions and the depicted physical state, rather than photorealism.
3. *Logical constraint*. Whether the image satisfies structured constraints on motion, deformation, or causal change implied by the prompt.

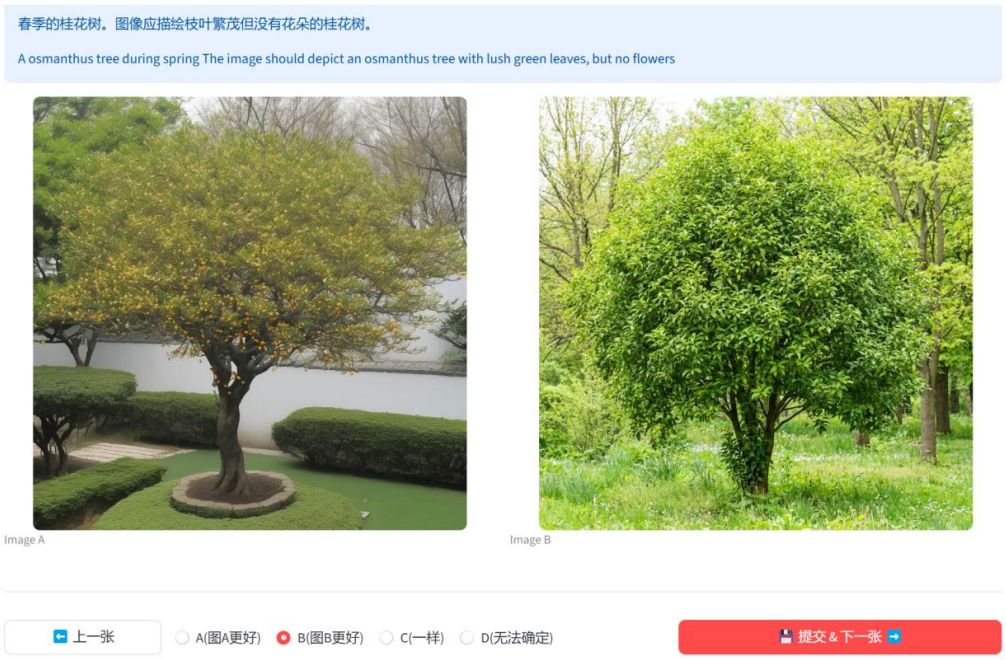


Figure 6: **The Rank-Bench annotation interface.** Annotators are shown the text prompt together with two anonymized candidate images with randomized left–right ordering, and select one of four labels ($\succ$, $\prec$, $\sim$, $\perp$) after a minimum dwell time of 5 s.

AESTHETICS

1. *Human rendering*. The realism and quality of depicted characters, focusing on face, hands and fingers, limbs, posture, body proportions, body contact, occlusion, and the consistency of visible body parts.
2. *Artistic style*. Whether the requested art medium or style is realized, and the quality of its execution.
3. *Text rendering*. Whether the required text is present, legible, correctly spelled, complete, and placed as specified by the prompt.

HOLISTIC HUMAN PREFERENCE

1. *Human preference*. Overall human preference in cases where no single diagnostic constraint alone determines the outcome. Annotators jointly consider prompt alignment, realism or stylistic consistency, composition, artifact salience, and overall attractiveness.

Guided by the taxonomy above, we assembled a prompt pool from three complementary sources. First, prompts targeting basic attributes, position, and counting were selected from GenEval and augmented with T2I-CompBench prompts covering the same dimensions. Second, knowledge-intensive prompts particularly suited to world- and physical-knowledge evaluation were drawn from WISE. Third, for dimension–constraint combinations that remained under-represented after mapping the first two sources onto our taxonomy, GPT-5 was prompted to generate additional candidates.

All candidate prompts underwent three independent rounds of human review. A prompt was discarded if it (i) contained unsafe content, (ii) was a near-duplicate of a retained entry, (iii) required non-visual context unavailable to the annotator, (iv) exhibited substantial ambiguity in the intended scenario, or (v) specified target conditions that could not be verified from a static image. After filtering, the number of retained prompts per sub-dimension was balanced to 100, yielding the final prompt pool used for image generation.

A.2 IMAGE GENERATION AND LABELING

For each retained prompt, one image was generated by each of eight representative T2I systems: SDXL, SD3, FLUX.1-dev, Z-Image, Nano Banana, Qwen-Image, GPT-Image, and Banana-Pro. All images were rendered at a resolution of 512×512 pixels and stored with model identity stripped from filenames and metadata, so that provenance remained hidden throughout the annotation process.

All pairwise judgments were provided by external annotators; in total, 11 annotators participated. Prior to formal annotation, each annotator completed a standardized training programme. A dedicated annotation manual was prepared for every one of the 11 diagnostic dimensions, specifying (i) the objective of the sub-dimension, (ii) the admissible evidence for comparison, (iii) common boundary cases, and (iv) worked examples for each available label. Training included example walk-throughs and item-level feedback on incorrect responses.

Annotators were required to pass a qualification test before being assigned formal batches. The test consisted of 50 pairwise comparisons spanning multiple dimensions, and a score of $\geq 80\%$ was required to qualify. Annotators who failed were re-trained on the misclassified items and permitted to retake the test up to two additional times.

Each annotation task displayed the text prompt together with two images from which all generator metadata had been removed; the left–right order was randomized independently for every presentation (see Fig. 6). Each unordered image pair was initially assigned to at least five distinct annotators. Let $N_{ij} = N_{i\succ j} + N_{j\succ i} + N_{i\sim j}$ denote the number of valid votes, excluding uncertain responses. Additional judgments were collected from previously unassigned annotators until the pair had at least five valid votes and a unique modal label with $\alpha_{ij} \geq 0.8$, or until all eleven annotators had assessed the pair. A pair was retained only if both conditions were met; otherwise, it was discarded at this limit. The consensus ratio was evaluated only after at least five valid votes were available.

Batch-level quality control Annotation tasks were organized into batches of 300 pairs, each batch homogeneous with respect to diagnostic dimension, and every batch was randomly assigned to five annotators. Within each batch, 20 reference pairs—constructed by the benchmark authors from clearly separable and undisputed samples—were inserted as gold-standard controls. Batches with a reference accuracy below 90% were rejected in full and reassigned for relabelling. Beyond batch-level acceptance, we further monitored the response-time distribution, modal-label share, number of annotators per pair, tie rate, and invalid/uncertain vote rate, stratified by diagnostic dimension. These diagnostics help distinguish genuine preference ambiguity from low-quality labelling.

Consensus threshold selection The retention threshold $\alpha_{ij} \geq 0.8$ used in the main text was set after a pilot on 900 pairs, in which we compared the trade-off between retention rate and agreement with an independently re-labelled hold-out set at $\alpha \in \{0.6, 0.7, 0.8, 0.9\}$. The value 0.8 was the smallest threshold at which hold-out agreement exceeded 95% while retaining more than 90% of the pairs.

A.3 LABEL MERGING

Because ties and local preference relations can jointly induce directed cycles, $\mathcal{G}$ may contain strongly connected regions. In our data, the strict-preference subgraph is acyclic. Cycles arise only after tie relations are incorporated. We therefore computed its strongly connected components (SCCs) and contracted each SCC into a single consensus block. SCC contraction is an aggregation convention, not evidence that every image pair within a block received a consensus tie. Non-transitive ties can make it impossible for a weak order to preserve all retained labels: for example, $A \sim B$, $B \sim C$, and $A \succ C$ cannot all hold in a weak order. Assigning a common tier to each SCC can therefore convert a retained strict preference into an aggregated tie and can tie pairs with no retained relation (Atkinson et al., 2025). Images within the same block were assigned to a common rank tier, and strict edges internal to an SCC were absorbed into that block rather than used to create additional within-block rank distinctions. All preference edges between different blocks were preserved. The resulting condensation graph $\mathcal{G}^*$ is a directed acyclic graph (DAG) whose nodes are consensus blocks and whose edges encode a partial order: an edge from block A to block B requires A to precede B in the final ranking. To resolve pairs of blocks left unordered by $\mathcal{G}^*$, we fitted a Davidson model at the

block level. For any two consensus blocks A and B, the response probabilities are

$$P(A \succ B) = \frac{\theta_A}{d_{AB}}, \qquad P(B \succ A) = \frac{\theta_B}{d_{AB}}, \qquad P(A \sim B) = \frac{\nu\sqrt{\theta_A \theta_B}}{d_{AB}}, \tag{7}$$

where

$$d_{AB} = \theta_A + \theta_B + \nu\sqrt{\theta_A \theta_B}. \tag{8}$$

Here $\theta_A > 0$ and $\theta_B > 0$ denote the latent preference strength of blocks A and B, respectively, and $\nu > 0$ governs the overall propensity for ties. Each strength parameter is therefore attached to an SCC block rather than to an individual image. Only image pairs retained by the consensus rule $\alpha_{ij} \geq 0.8$ were included in the Davidson fit. For each retained pair whose two images belonged to distinct blocks A and B, we aggregated the valid individual votes into $n(A \succ B)$, $n(B \succ A)$ and $n(A \sim B)$. Pairwise comparisons discarded by the consensus filter were excluded entirely, and invalid or uncertain votes were not included in any count. The parameters θ_A and ν were obtained by maximising the log-likelihood

$$\ell(\boldsymbol{\theta}, \nu) \;=\; \sum_{\{A,B\}} \Big\{ n(A \succ B) \log \frac{\theta_A}{d_{AB}} \;+\; n(B \succ A) \log \frac{\theta_B}{d_{AB}} \;+\; n(A \sim B) \log \frac{\nu\sqrt{\theta_A \theta_B}}{d_{AB}} \Big\}, \tag{9}$$

where the sum is taken over all unordered pairs of distinct consensus blocks, subject to $\theta_A > 0$ for every block A, $\nu > 0$, and the identifiability constraint $\sum_A \log \theta_A = 0$.

The final ranking was produced by a strength-weighted topological sort on $\mathcal{G}^*$. At each step, among all blocks with zero in-degree, the one with the largest estimated θ was appended to the ordering, after which the block and its outgoing edges were removed. This procedure resolves only those block pairs left unspecified by the DAG. It can neither reverse a retained ordering constraint nor separate images within the same consensus block. Prompts for which all retained images collapsed into a single consensus block were excluded, as they carry no ranking signal.

Implementation details of the Davidson fit For each prompt, the block-level Davidson log-likelihood defined in Eq. 9 was maximised by L-BFGS on the log-parameterisation $\theta_A = \exp(\eta_A)$, which enforces positivity and improves numerical stability; the tie parameter was analogously re-parameterised as $\nu = \exp(\eta_\nu)$. The identifiability constraint $\sum_A \log \theta_A = 0$ was imposed by mean-centring $\{\eta_A\}$ after each iteration. Convergence was declared when the relative change in the log-likelihood fell below 10^{-6} or after 500 iterations, whichever came first. Optimisation was restarted from three random initialisations per prompt, and the solution with the highest log-likelihood was retained.

Two boundary cases required dedicated handling. (i) If, between two blocks A and B, all valid annotations lay in a single category (e.g., $n(A \succ B) > 0$ but $n(B \succ A) = n(A \sim B) = 0$), a small additive prior of 10^{-3} was applied to the corresponding counts to prevent degenerate boundary estimates of the strength parameters. (ii) If two blocks were already ordered by $\mathcal{G}^*$, that ordering was enforced during the strength-weighted topological sort regardless of the fitted θ values, so the Davidson estimates could never reverse a retained pairwise inter-block constraint.

In addition to the exclusion of prompts whose retained images collapsed into a single consensus block, we further discarded prompts whose condensation graph $\mathcal{G}^*$ contained fewer than three distinct blocks, so that every released ranking supports non-trivial ordinal comparisons.

Audit of reference construction. We distinguish the original vote-based consensus labels, subsequent manual adjudications, and relations determined by topological extension. All counts below concern the prompts and images included in the released benchmark. Manual review revised 62 retained strict-preference labels whose image pairs lay within SCCs to ties. The released tie-label set contains 6,604 image pairs and coincides with the set of all within-tier pairs in the final reference rankings. These include the 62 manually revised labels, which are distinguished from ties supported by the original consensus votes.

Before topological extension, two distinct blocks were considered unordered only if neither was reachable from the other in the condensation DAG. The Davidson-guided extension determined the relative order of 2,170 such block pairs, corresponding to 3,461 image pairs. Relations already implied by DAG reachability were excluded from these counts. The extension preserves all inter-block

ordering constraints encoded in the DAG. Its completed relations form a construction-dependent extension of the observed partial order and do not constitute additional vote-based human consensus labels.

A.4 From raw output to canonical prediction

All evaluations on Rank-Bench used the same baseline prompts, candidate images, and retained pairwise relations. All methods loaded images from the same source files and processed them using the official image processor for each model; the evaluation models were not exposed to generator identities, filenames, ranking tiers, or human labels. For the general VLMs, we reused the scalar utility evaluation prompt from UnifiedReward.

For every retained relation in Rank-Bench, we require each evaluated system to produce a pairwise verdict indicating which image is preferred or whether they are tied. Because the 12 systems emit their outputs in different formats, we adopt a unified mapping rule for each of the three model families. For pointwise scalar reward models, which assign an independent scalar score to each image conditioned on the prompt, the pairwise verdict is determined by comparing the two scores. To avoid treating floating-point noise as a meaningful preference difference, we used an absolute numerical tolerance 10^{-3}. A pair was mapped to a tie when the absolute score difference did not exceed this tolerance; otherwise, the image with the higher score was predicted as preferred. A candidate-set ranking was obtained by sorting all images in descending order of their scalar scores, with images receiving identical scores assigned to the same rank tier. For general-purpose VLMs and pairwise comparative reward models, which take the prompt and both images jointly as input, the model is instructed to return one of three categorical verdicts which is used directly as the pairwise prediction. Constructing a candidate-set ranking required evaluating all unordered pairs in a set of images. We then assigned each image an average pairwise win score. Images were ranked in descending order of win score, with equal scores assigned to the same tier. For MarginRM, the per-criterion relative margins are first aggregated into a single anchor-relative score using the rubric weights. Following the anchor–candidate convention adopted during training, a positive aggregate score indicates that the candidate image is preferred over the anchor, a negative score indicates the opposite, and a zero aggregate score is treated as a tie. To ensure fairness and maintain consistency with pairwise methods, the final ranking is determined based on the win rates derived from the results of all pairs.

A.5 Metric definitions

Let $\mathcal{P}$ denote the retained human-labelled pair relations in Rank-Bench. Each relation $p \in \mathcal{P}$ has a reference label $y_p \in \{-1, 0, +1\}$ and a model prediction $\widehat{y}_p$.

Accuracy with ties Using all retained relations and treating strict preferences and ties as equally valid reference outcomes, we report the fraction of pairs whose prediction matches the reference label:

$$\mathrm{Acc}_{\mathrm{w/\ tie}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}(\widehat{y}_p = y_p)\,. \tag{10}$$

Under this metric, a predicted tie counts as correct if and only if the reference is also a tie.

Accuracy without ties Let $\mathcal{P}_{\mathrm{strict}} = \{p \in \mathcal{P} : y_p \neq 0\}$ denote the subset of pairs with a strict human preference. Accuracy on this subset is

$$\mathrm{Acc}_{\mathrm{w/o\,tie}} = \frac{1}{|\mathcal{P}_{\mathrm{strict}}|} \sum_{p \in \mathcal{P}_{\mathrm{strict}}} \mathbf{1}(\widehat{y}_p = y_p)\,. \tag{11}$$

Reference ties are excluded from the denominator, but predicted ties are not: a predicted tie on a strictly preferred reference pair is counted as incorrect.

Stable accuracy Stable accuracy requires the model to predict the reference relation correctly in *both* presentation orders. Let $\widehat{y}_p^{\rightarrow}$ be the prediction when the images are presented in the canonical

order (I_i, I_j), and $\widehat{y}_p^{\leftarrow}$ the prediction obtained by swapping the image order and mapping the raw output back to the canonical orientation. Then

$$\mathrm{StableAcc} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}\big(\widehat{y}_p^{\rightarrow} = y_p \text{ and } \widehat{y}_p^{\leftarrow} = y_p\big)\,. \tag{12}$$

This metric penalises order-dependent predictions and is therefore always upper-bounded by $\mathrm{Acc}_{\mathrm{w/\ tie}}$.

Stable rate. Stable rate measures whether a model's pairwise prediction is invariant to reversal of the image-presentation order, irrespective of whether the prediction is correct. Let $\widehat{y}_p^{\rightarrow}$ denote the prediction for relation p when the images are presented in the canonical order (I_i, I_j). Let $\widehat{y}_p^{\leftarrow}$ denote the prediction obtained from the reversed presentation (I_j, I_i) after mapping the raw output back to the same canonical orientation. Under the signed-label convention, this mapping reverses the sign of a strict-preference prediction, whereas a predicted tie remains zero. We define

$$\mathrm{StableRate} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}\left(\widehat{y}_p^{\rightarrow} = \widehat{y}_p^{\leftarrow}\right). \tag{13}$$

A pair is therefore counted as stable when the model produces the same canonicalized relation in both presentation orders, including either the same strict preference or a tie in both orders. Stable rate measures reversal consistency rather than correctness: a model may obtain a high stable rate by repeating the same incorrect prediction in both orders. We report this metric only for models that jointly process the two images in an ordered input. For independently scored scalar utility models, reversing the image presentation does not constitute a distinct joint inference and the metric is therefore not reported.

Kendall's τ_b To assess how well each model recovers the full prompt-level ranking, we report Kendall's τ_b, which accommodates ties in both the human and predicted orderings. Considering all unordered candidate pairs, C_x is the number of pairs that are strictly ordered in the same direction in both orderings, whereas D_x is the number of pairs that are strictly ordered in opposite directions. $T_{H,x}$ is the number of pairs tied in H_x but strictly ordered in $\widehat{H}_x$, and $T_{\widehat{H},x}$ is the number of pairs strictly ordered in H_x but tied in $\widehat{H}_x$. Pairs tied in both orderings contribute to none of these four counts. Kendall's τ_b for prompt x is then

$$\tau_b(x) = \frac{C_x - D_x}{\sqrt{(C_x + D_x + T_{H,x})(C_x + D_x + T_{\widehat{H},x})}}, \tag{14}$$

with $\tau_b(x) \in [-1, +1]$. The benchmark-level value is the mean over prompts,

$$\tau_b = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \tau_b(x), \tag{15}$$

We evaluated every model on the same fixed set of prompts $\mathcal{X}$. When the denominator in Eq. (14) was zero, we assigned that prompt an evaluation score of zero rather than excluding it. Because every reference ranking contains at least three distinct tiers, this occurs when the predicted ranking places all candidates in a single tier. The benchmark score was

$$\overline{\tau}_b = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \widetilde{\tau}_b(x), \qquad \widetilde{\tau}_b(x) = \begin{cases} \tau_b(x), & \text{if the denominator in Eq. (14) is nonzero,} \\ 0, & \text{otherwise.} \end{cases}$$

This zero assignment is an evaluation convention for undefined cases; all prompts remain in the benchmark-level average.

B Relative Margin Reward Modeling

This section documents the full pipeline that turns raw pairwise preference records into the criterion-level relative-margin supervision used to train the Relative-Margin Reward Model (MarginRM).

The pipeline consists of five stages: (i) sourcing and filtering pairwise preference data; (ii) generating a prompt-specific hierarchical rubric $\mathcal{R}(x)$ for every unique text prompt x; (iii) producing, for each retained pair, a signed relative margin and a short explanation for every criterion in $\mathcal{R}(x)$; (iv) direction-consistency filtering that enforces agreement between the aggregate margin and the original human preference; and (v) serialising the retained supervision into training examples and fine-tuning MarginRM. Steps (ii) and (iii) use the same teacher model, Qwen3.5-397B-A17B, with the prompts reproduced below. Each stage strictly preserves the direction of the original human preference; the teacher is used only to provide a structured decomposition.

B.1 PROMPT-SPECIFIC RUBRIC GENERATION

For each unique prompt x, the teacher model generated a prompt-specific hierarchical rubric $\mathcal{R}(x)$ together with per-criterion aggregation weights $\{w_k\}_{k=1}^{K_x}$ satisfying $w_k \geq 0$ and $\sum_{k=1}^{K_x} w_k = 1$. The rubric consists of three levels: activated anchor dimensions with normalized weights, activated sub-dimensions under each anchor, and observable pairwise rubrics under each sub-dimension.

Crucially, the rubric generator received only the text prompt. It saw neither the candidate images nor the source preference direction, which prevents the criteria from being tailored to any observed pairwise outcome and ensures that $\mathcal{R}(x)$ depends on x alone. The system and user prompts used for rubric generation are reproduced in Supplementary Listing 1 and Supplementary Listing 2.

```
SYSTEM_PROMPT =
You are an expert evaluation-plan designer for image generation reward modeling.

Your task is NOT to score any image.
Your task is to construct a reusable, prompt-specific hierarchical evaluation plan for
↪ comparing future candidate images generated from the same text prompt.

The evaluation plan must be built in 3 levels:

Level 1: activate a set of high-level anchor dimensions and assign normalized weights across
↪  them.
Level 2: under each activated anchor dimension, activate the most relevant subdimensions.
Level 3: under each activated subdimension, generate specific rubrics that are observable,
↪ pairwise-comparable, and suitable for later margin labels in {-2, -1, 0, +1, +2}.

Anchor dimensions and candidate subdimensions:

1. semantic_alignment
- subject_presence_and_identity
- attribute_binding_and_state
- action_pose_and_event
- counting_and_quantity
- spatial_layout_and_localization
- relation_and_interaction_accuracy
- scene_context_and_background
- text_symbol_and_sign_rendering

2. aesthetics_and_artistry
- style_faithfulness
- medium_and_technique_match
- composition_and_framing
- color_palette_and_harmony
- mood_atmosphere_and_tone
- visual_impact_and_subject_emphasis

3. image_quality
- anatomy_structure_and_proportion
- geometry_perspective_and_3d_consistency
- lighting_shadow_and_reflection_consistency
- color_consistency_and_material_rendering
- texture_detail_and_local_clarity
- artifact_suppression
- depth_focus_and_occlusion_coherence

4. reasoning_and_interaction
- constraint_and_negation_following
- commonsense_physics_and_function
- interaction_dynamics
- emotional_expression_and_resonance

5. human_preference
- overall_scene_appeal
- prompt_intent_impression
- emotional_atmospheric_impact
- visual_cohesion_and_flow
- polished_final_image_feel

Definition of human_preference:
```

```
- human_preference captures global, prompt-specific residual preferences that human raters
↪ are likely to care about after concrete semantic alignment, technical image quality,
↪ aesthetic style, and reasoning constraints have already been considered.
- human_preference should evaluate the whole-image viewing experience, not local correctness
↪  of individual objects, attributes, poses, anatomy, spatial relations, or physical
↪ plausibility.
- A human_preference rubric should usually remain meaningful even if both images already
↪ satisfy the main hard constraints of the prompt.

Important rules:
1. Focus only on image generation.
2. Prompt-specific semantic requirements must be prioritized over generic image quality,
↪ aesthetics, or human preference.
3. Do not score any actual image.
4. Do not mention candidate A or candidate B.
5. Do not output a final winner or a final score.
6. Avoid vague criteria such as "overall quality", "general preference", "better image", or
↪ duplicated criteria.
7. Use exactly the anchor and subdimension names from the provided inventory.
8. Do not invent new anchor dimensions or subdimensions.
9. If a criterion is already covered by one rubric, do not restate it elsewhere with
↪ different wording.
10. Prefer rubrics phrased as pairwise preference rules beginning with "Prefer the image
↪ that ...".
11. Output valid JSON only.
```

Listing 1: System prompt for prompt-specific rubric generation.

```
USER_TEMPLATE =
Task: Create a reusable hierarchical evaluation plan for pairwise comparison of future
↪ images generated from the following text prompt.

Text prompt: {PROMPT}

Your goal is to build a 3-level evaluation plan:

Level 1:
- Activate 2 to 5 high-level anchor dimensions from the provided inventory.
- Assign normalized weights across activated anchor dimensions.
- Anchor weights must be numeric values and must sum to 1.0 across all activated anchor
↪ dimensions.
- Semantic_alignment and human_preference should usually be activated because image
↪ generation prompts almost always contain content requirements.
- if human_preference is activated, its weight should usually be between 0.15 and 0.30.

Level 2:
- Under each activated anchor dimension, activate 1 to 5 relevant subdimensions.
- Activate a subdimension only if removing it would cause meaningful loss of evaluation
↪ coverage for this prompt.

Level 3:
- Under each activated subdimension, generate 1 to 3 concrete rubrics.
- Do not use rubrics that require metadata, hidden intent, intermediate generation steps,
↪ model internals, or evidence not visually inferable from the prompt and final image.
- Each rubric must be observable from a final image, pairwise-comparable, and suitable for
↪ later labels in {-2, -1, 0, +1, +2}.

First, classify the prompt complexity as one of:
- Simple_object_prompt
- Moderate_scene_prompt
- Complex_compositional_prompt

Use these guidelines:
- Simple_object_prompt: usually a single main object or entity, with only a few essential
↪ attributes or placement constraints, and no meaningful narrative, counting, text-heavy,
↪ or multi-relation requirements.
- Moderate_scene_prompt: multiple important constraints, richer scene context, or more than
↪ one meaningful relation.
- Complex_compositional_prompt: multiple entities, counting, spatial reasoning, text
↪ rendering, narrative interaction, or dense compositional constraints.

Then scale the evaluation plan accordingly:
- Simple_object_prompt: prefer 2 to 3 anchors, 4 to 8 rubrics total.
- Moderate_scene_prompt: prefer 3 to 4 anchors, 6 to 12 rubrics total.
- Complex_compositional_prompt: prefer 4 to 5 anchors, 8 to 14 rubrics total.

Requirements:
- Focus only on image generation.
- If the prompt contains explicit count, position, left/right, foreground/background,
↪ attribute, relation, visible text, geometry, negation, exclusion, or commonsense
↪ constraints, reflect them explicitly.
- Activate reasoning_and_interaction only when negation, exclusions, commonsense constraints
↪ , interaction dynamics, or emotional relations are central to prompt success.
- Prefer criteria that directly reference concrete requirements in the prompt.
- Avoid introducing generic criteria unless they are necessary to distinguish likely
↪ generation failures for this prompt.
- Do not output any winner or score for an actual image.
```

```
- Rubric ids must match the pattern ^[a-z][a-z0-9_]*$ and must be unique within the full
↪ JSON output.

Human-preference requirements:
- human_preference is a global residual preference anchor, not a local correctness anchor.
- Write human_preference rubrics as whole-image preference rules.
- A human_preference rubric should compare the overall viewing experience, impression,
↪ emotional atmosphere, cohesion, final-image feel, or prompt-specific tradeoff balance.
- Do not create human_preference rubrics that focus on isolated objects, body parts, poses,
↪ counts, spatial relations, physical plausibility, exact attributes, or scene elements
↪ unless they are discussed as part of the whole-image impression.

Good human_preference rubric examples:
- Prefer the image that presents the requested cozy bedroom scene in a way that feels more
↪ inviting and naturally livable, beyond merely including the required objects.
- Prefer the image that balances the requested surreal elements in a more intentional and
↪ visually coherent way, without making the scene feel accidental or cluttered.

Notation:
- The total number of rubrics across all subdimensions must be between 4 and 14.
- Prefer fewer rubrics if they can already capture the main failure modes.
- Do not create multiple evaluation criteria that rely on substantially identical visual
↪ evidence unless each criterion captures a meaningfully different comparison signal.
- Each rubric must justify its existence by answering: "If this rubric is removed, would we
↪ lose important ability to distinguish between two candidate images?" If not, remove it or
↪  merge it with another rubric.
- Prefer merging related criteria into a single stronger rubric rather than creating
↪ multiple weak and overlapping rubrics.
- prompt_summary should be a brief 1 to 2 sentence summary of the prompt, not a verbatim
↪ copy of the full prompt.
- hard_constraints are requirements whose violation should clearly count against prompt
↪ success.
- soft_preferences are desirable qualities or stylistic preferences that matter only when
↪ they are explicitly requested or strongly implied by the prompt.
- Each activated anchor dimension must contain at least 1 activated subdimension.
- Each activated subdimension must contain at least 1 rubric.

Critical output constraint:
- Return a single valid JSON object only.
- Do not output markdown fences.
- Do not output any explanation before or after the JSON.

Return valid JSON only with the following schema:

{
  "prompt_analysis": {
    "prompt_summary": "...",
    "hard_constraints": ["..."],
    "soft_preferences": ["..."],
    "likely_failure_modes": ["..."]
  },
  "activated_anchor_dimensions": [
    {
      "anchor_dimension": "semantic_alignment",
      "weight": 0.45,
      "why_activated": "...",
      "activated_subdimensions": [
        {
          "subdimension": "attribute_binding_and_state",
          "why_activated": "...",
          "rubrics": [
            {
              "id": "ab1",
              "preference_rule": "Prefer the image that ..."
            }
          ]
        }
      ]
    },

    ...,

    {
      "anchor_dimension": "human_preference",
      "weight": 0.15,
      "why_activated": "...",
      "activated_subdimensions": [
        {
          "subdimension": "prompt_intent_impression",
          "why_activated": "...",
          "rubrics": [
            {
              "id": "hp1",
              "preference_rule": "Prefer the image that ..."
            }
          ]
        }
      ]
    }
  ],
  "normalization_check": {
    "anchor_weights_sum_to_1": true
```

```
  }
}
```

Listing 2: User template for prompt-specific rubric generation. The placeholder {PROMPT} is replaced with the text prompt x.

Rubric validation and regeneration Every generated rubric was passed through an automatic validator implemented as a second teacher call, together with lightweight schema-level checks. The validator inspected: (a) JSON structural correctness and adherence to the anchor inventory; (b) specificity of each rubric with respect to the concrete constraints in x; (c) visual observability from a final image; (d) non-redundancy across rubrics; and (e) discriminative power, i.e., whether removing the rubric would meaningfully reduce the ability to separate plausible failure modes. A rubric set that contained unobservable criteria, near-duplicate rubrics, omitted salient prompt constraints, or weights that could not be normalized was rejected and regenerated, up to a maximum of two additional attempts.

B.2 Criterion-level relative-margin generation

Conditioned on the text prompt x, its rubric $\mathcal{R}(x)$ produced in Stage (ii), and a source pair consisting of an anchor image I_a and a candidate image I_c, the same teacher model used for rubric generation produced, for every criterion $k \in \{1, \ldots, K_x\}$, a short natural-language explanation e_k and a signed relative margin $m_k \in \{-2, -1, 0, +1, +2\}$ that describes the candidate relative to the anchor on that criterion. The scale is fixed as follows:

- $m_k = +2$: the candidate is substantially better than the anchor on criterion k;
- $m_k = +1$: the candidate is moderately or slightly better;
- $m_k = 0$: no discernible or reliable difference, including cases where both images satisfy or both fail the criterion to a similar degree;
- $m_k = -1$: the candidate is moderately or slightly worse;
- $m_k = -2$: the candidate is substantially worse than the anchor.

This anchor-relative parameterisation is what distinguishes MarginRM supervision from conventional scalar reward labels: each margin is tied to a specific criterion in $\mathcal{R}(x)$ and to a specific reference image I_a, so the resulting signal is both prompt- and anchor-conditioned. The teacher sees both images and the rubric but is never shown the source human preference direction, which prevents it from trivially matching the ground-truth label. The exact criterion-level comparison prompt is reproduced in Listing 3. It fixes the presentation order (first image as anchor, second image as candidate), forbids reference to hidden or non-visual evidence.

```
EVAL_PROMPT_TEMPLATE =
You are an expert evaluator for anchor-based pairwise image preference assessment.

You will compare TWO images generated for the same text prompt:
- The FIRST image is the ANCHOR image.
- The SECOND image is the CANDIDATE image.

Evaluate the CANDIDATE relative to the ANCHOR for EACH rubric independently.

Prompt:
{prompt}

Rubrics:
{rubric_lines}

Scoring scale:
-2 = candidate is significantly worse than anchor
-1 = candidate is slightly worse than anchor
0 = tie, or insufficient evidence
1 = candidate is slightly better than anchor
2 = candidate is significantly better than anchor

Rules:
- Evaluate every rubric independently.
- Use only visible evidence relevant to that rubric.
- Explicitly compare the candidate against the anchor in each explanation.
- Do not speculate.
- If both images are similar or evidence is insufficient, return 0.
- Keep each explanation brief.
```

```
Return VALID JSON only.
Do not output markdown, code fences, or any extra text.
Output a single JSON object.
Use the rubric IDs exactly as provided.
Include every rubric ID exactly once.
Do not add any extra fields.

For each rubric id, output:
{{
  "<rubric_id>": {{
    "rubric": "<original rubric text>",
    "score": <integer: -2 | -1 | 0 | 1 | 2>,
    "explanation": "<brief comparison based only on visible evidence>"
  }}
}}
```

Listing 3: Criterion-level comparison prompt used to produce (e_k, m_k) for each rubric k in $\mathcal{R}(x)$.

Each teacher response was passed through a deterministic parser that enforced the following requirements: (i) exactly one entry per rubric in $\mathcal{R}(x)$; (ii) rubric order preserved as given in the input; (iii) every entry containing both a non-empty explanation e_k and a margin m_k; and (iv) no duplicate rubric identifiers. Responses that violated any of these constraints were regenerated up to two additional times; samples whose output still failed validation on the third attempt were discarded.

Because the teacher model provides only a structured decomposition rather than a new source of ground-truth preference, we require the teacher's aggregate judgement to be consistent with the original human-annotated preference direction. For each converted sample we computed the rubric-weighted aggregate margin

$$\Delta_{c|a} \;=\; \sum_{k=1}^{K} w_k \, m_k, \tag{16}$$

using the per-criterion weights $\{w_k\}_{k=1}^{K}$ produced in Stage (ii). A sample was retained if and only if the sign of $\Delta_{c|a}$ agreed with the source preference:

- if the source preference was $I_c \succ I_a$, retain when $\Delta_{c|a} > 0$;
- if the source preference was $I_a \succ I_c$, retain when $\Delta_{c|a} < 0$.

Samples with $\Delta_{c|a} = 0$, as well as samples whose aggregate margin had a sign inconsistent with the source preference, were discarded. This filter has two effects. First, it removes conversions in which the rubric-level margins, though individually plausible, fail to reproduce the original human preference and are therefore poor supervision for a reward model. Second, it enforces a strict non-inversion property: MarginRM is never trained on a teacher-reconstructed preference that contradicts the human label, so the teacher is used purely to explain and decompose the existing relation, not to relabel it. After this stage, 532,507 samples remained. These were passed to the MarginRM training pipeline described below.

B.3 Interpretation and limitations of ordinal margins

The five margin outputs are ordered preference-strength categories. Their numerical coding and rubric weights define an operational scoring rule; the category spacing has not been calibrated against human judgments of difference magnitude. Equal aggregate scores therefore need not imply equal perceptual quality. Filtering teacher annotations by agreement with the source preference checks the aggregate direction for retained training pairs, but does not validate the spacing between categories.

Fixing the rubric and anchor holds the explicit comparison context constant, but does not guarantee calibration across model calls, anchor-invariant rankings, antisymmetry, or additive consistency. Our training objective supervises individual pairs and does not explicitly enforce these properties. For a fixed anchor, sorting the scalar scores yields a complete preorder, with equal scores tied. This algebraic property also holds for pointwise scalar rewards and does not establish consistency among independently predicted pairwise judgments.

The shared-anchor formulation is evaluated empirically through the anchor analysis in Figure 4 and the efficiency–performance comparison in Table 3. These experiments characterize its practical utility and dependence on the comparison reference.

B.4 MARGINRM FINE-TUNING SETUP

We trained two MarginRM variants (Qwen3VL-8B and Qwen3.5-9B) that share the serialisation format and supervision signal. Both variants were fine-tuned with LoRA under a supervised next-token-prediction objective computed on the target output only (the prompt and rubric portions of the sequence were masked from the loss). Optimisation used AdamW with a cosine learning-rate schedule and a linear warm-up, in `bf16` precision with FlashAttention-2. At inference, MarginRM consumes a text prompt, its rubric $\mathcal{R}(x)$, and an anchor/candidate image pair in the same template order used during training, and emits one `[rubric_id,` m_k`]:` e_k line per criterion.

C FLOW-GRPO POST-TRAINING

We post-trained FLUX.1-dev with online reward feedback using Flow-GRPO. For each training prompt, the behaviour policy sampled a group of image-generation trajectories. A frozen reward evaluator assigned one scalar reward to each terminal image, the rewards were standardized within the rollout group to obtain group-relative advantages, and the flow policy was updated with a clipped policy-ratio objective.

C.1 FLOW-MATCHING FORMULATION

Let x denote a text prompt, let $\mathbf{z}_0 \in \mathbb{R}^d$ be a clean image latent, and let $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ be an independent Gaussian latent. Following the rectified-flow parameterization, we define the interpolation path for $t \in [0, 1]$ as:

$$\mathbf{z}_t = (1 - t)\,\mathbf{z}_0 + t\,\mathbf{z}_1, \tag{17}$$

with the corresponding target velocity:

$$\mathbf{v}^\star = \mathbf{z}_1 - \mathbf{z}_0. \tag{18}$$

Under this convention, $t = 0$ denotes the data endpoint and $t = 1$ the noise endpoint. A velocity model $\mathbf{v}_\theta(\mathbf{z}_t, t, x)$ is conventionally trained by minimizing:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\Big[\|\mathbf{v}_\theta(\mathbf{z}_t, t, x) - \mathbf{v}^\star\|_2^2\Big]. \tag{19}$$

For generation, the sampler integrates the learned vector field from the noise endpoint ($t = 1$) toward the data endpoint ($t = 0$). Let $\mathcal{T}_g = (\mathbf{z}_{g,0}, \ldots, \mathbf{z}_{g,L})$ denote the discrete reverse trajectory of the g-th rollout, where $\mathbf{z}_{g,\ell} := \mathbf{z}_{t_\ell}$ is evaluated on a strictly decreasing time grid

$$1 > t_0 > t_1 > \cdots > t_{L-1} > t_L = 0, \tag{20}$$

with step sizes $h_\ell = t_\ell - t_{\ell+1} > 0$. The drift and diffusion coefficients of the stochastic sampler are evaluated only at interior times $t_\ell \in (0, 1)$.

C.2 FLOW-GRPO AS A STOCHASTIC POLICY-GRADIENT METHOD

ODE-to-SDE conversion and transition policy. A deterministic ODE sampler admits no non-degenerate transition density, and hence cannot support a policy ratio. Following Flow-GRPO, we replace it with a marginal-preserving stochastic sampler. For $t \in (0, 1)$, we adopt the SDE

$$\mathrm{d}\mathbf{z}_t = \mathbf{b}_\theta(\mathbf{z}_t, t, x)\,\mathrm{d}t + \sigma_t\,\mathrm{d}\mathbf{W}_t, \tag{21}$$

with drift

$$\mathbf{b}_\theta(\mathbf{z}_t, t, x) = \mathbf{v}_\theta(\mathbf{z}_t, t, x) + \frac{\sigma_t^2}{2t}\big[\mathbf{z}_t + (1 - t)\,\mathbf{v}_\theta(\mathbf{z}_t, t, x)\big], \tag{22}$$

and noise schedule

$$\sigma_t = \eta\sqrt{\tfrac{t}{1-t}}, \qquad \eta > 0. \tag{23}$$

The noise schedule in Eq. 23 diverges as $t \to 1$ and vanishes at $t = 0$. We therefore evaluate stochastic transitions only at interior times and handle the endpoints separately. Under our convention, generation integrates Eq. 21 backward in time from $t = 1$ to $t = 0$. One Euler–Maruyama step then induces a Gaussian transition

$$\pi_\theta(\mathbf{z}_{g,\ell+1} \mid \mathbf{z}_{g,\ell}, t_\ell, x) = \mathcal{N}\big(\mathbf{z}_{g,\ell+1};\, \mathbf{z}_{g,\ell} - h_\ell\,\mathbf{b}_\theta(\mathbf{z}_{g,\ell}, t_\ell, x),\, \sigma_{t_\ell}^2 h_\ell\,\mathbf{I}_d\big), \tag{24}$$

which we treat as the stochastic policy. For brevity we denote its mean by $\boldsymbol{\mu}_{\theta,g,\ell}$. The resulting MDP has state $\mathbf{s}_{g,\ell} = (x, t_\ell, \mathbf{z}_{g,\ell})$ and action $\mathbf{a}_{g,\ell} = \mathbf{z}_{g,\ell+1}$; the terminal latent $\mathbf{z}_{g,L}$ is decoded to $I_g = \mathcal{D}(\mathbf{z}_{g,L})$, on which the terminal reward is evaluated.

Group-relative advantage. For each prompt x, the behaviour policy $\pi_{\theta_{\text{old}}}$ samples G trajectories with terminal rewards $\{R_g\}_{g=1}^{G}$. Let the group mean be

$$\mu = \frac{1}{G}\sum_{g=1}^{G} R_g, \tag{25}$$

and the (population) standard deviation

$$\sigma = \Big[\tfrac{1}{G}\sum_{g=1}^{G}(R_g-\mu)^2\Big]^{1/2}. \tag{26}$$

The group-relative advantage of rollout g is

$$\widehat{A}_g = \frac{R_g-\mu}{\sigma+\varepsilon_A}, \tag{27}$$

with $\varepsilon_A > 0$ a numerical stabilizer. Because the reward is terminal, $\widehat{A}_g$ is shared by all optimized transitions of trajectory g. When all rollouts in a group receive identical rewards, Eq. 27 vanishes and the group produces no policy-gradient signal. We treat $\widehat{A}_g$ as a stop-gradient quantity throughout.

Policy ratio. For a transition sampled under the frozen behaviour-policy snapshot θ_{old}, the transition-level importance ratio is

$$\rho_{g,\ell}(\theta) = \frac{\pi_\theta(\mathbf{z}_{g,\ell+1} \mid \mathbf{z}_{g,\ell}, t_\ell, x)}{\pi_{\theta_{\text{old}}}(\mathbf{z}_{g,\ell+1} \mid \mathbf{z}_{g,\ell}, t_\ell, x)}. \tag{28}$$

Both densities are evaluated on the same stored state–action pair, which is essential when a rollout batch is reused across multiple update epochs.

Clipped Flow-GRPO objective. Let $\mathcal{L}_{\text{opt}} \subseteq \{0,\dots,L-1\}$ index the transitions included in the policy update; this covers both full-trajectory optimization and subset-based variants. Treating $\widehat{A}_g$ as a stop-gradient quantity, the per-transition clipped surrogate is

$$\mathcal{S}_{g,\ell}(\theta) = \min\Big\{\rho_{g,\ell}(\theta)\widehat{A}_g,\ \text{clip}\big(\rho_{g,\ell}(\theta),\,1-\epsilon_c,\,1+\epsilon_c\big)\widehat{A}_g\Big\}. \tag{29}$$

The Flow-GRPO objective aggregates this surrogate over rollouts and optimized transitions:

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|\mathcal{L}_{\text{opt}}|}\sum_{\ell\in\mathcal{L}_{\text{opt}}}\Big(\mathcal{S}_{g,\ell}(\theta)\Big)\right]. \tag{30}$$

Here $0 < \epsilon_c < 1$ is the clipping radius. Training minimizes $L_{\text{Flow-GRPO}}(\theta) = -J_{\text{Flow-GRPO}}(\theta)$ without a KL penalty.

SHARED-ANCHOR MARGINRM REWARD

For a prompt x, the behaviour policy produces a rollout group $\mathcal{I}_x = \{I_1,\dots,I_G\}$. We designate the first sampled image as the shared anchor, $I_a = I_1$. This rule selects the anchor by rollout index rather than by its predicted or human-assessed quality. The frozen MarginRM uses the same prompt-specific rubric

$$\mathcal{R}(x) = \{(r_k, w_k)\}_{k=1}^{K}, \qquad w_k \ge 0, \qquad \sum_{k=1}^{K} w_k = 1. \tag{31}$$

For each $g \neq 1$, MarginRM predicts the signed margin vector

$$\widehat{\mathbf{m}}_{g|1} = (\widehat{m}_{g|1,1},\dots,\widehat{m}_{g|1,K}) \in \{-2,-1,0,1,2\}^K, \tag{32}$$

where positive values indicate that candidate I_g is better than the anchor on the corresponding criterion. The raw terminal rewards are

$$R_g^{\text{MarginRM}} = \begin{cases} 0, & g = 1, \\ \displaystyle\sum_{k=1}^{K} w_k \widehat{m}_{g|1,k}, & g \neq 1. \end{cases} \tag{33}$$

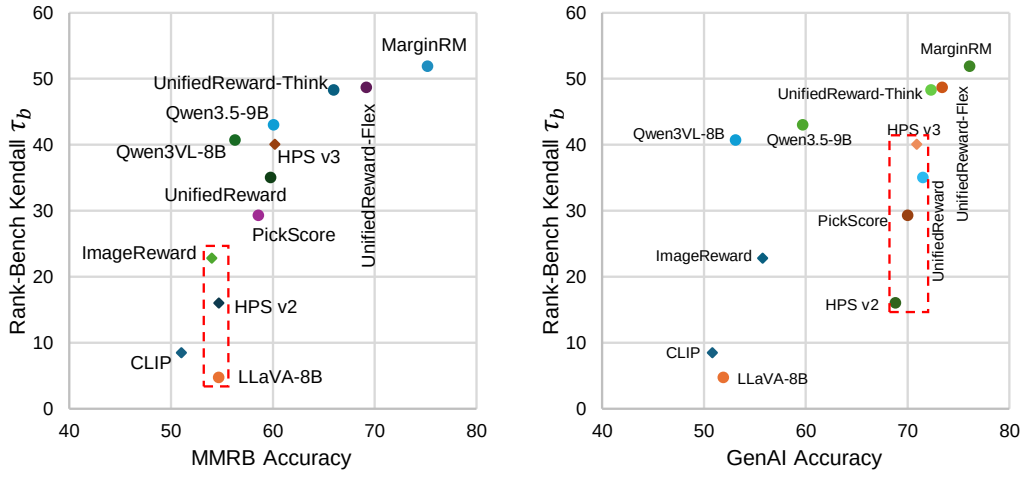


Figure 7: **Traditional pairwise benchmark performance and Rank-Bench ranking performance.** Accuracy on MMRB2 (left) and GenAI-Bench (right) plotted against mean prompt-level Kendall's τ_b on Rank-Bench. τ_b is multiplied by 100 for visualization. Red dashed boxes highlight models with similar pairwise accuracy but substantially different ranking agreement, including cases in which higher pairwise benchmark accuracy corresponds to lower Rank-Bench performance.

Thus, $R_g^{\text{MarginRM}} \in [-2, 2]$. Assigning the anchor a score of zero fixes the origin of the anchor-relative scale. More generally, adding the same constant to all rewards in a group leaves the centered advantages unchanged; it is the relative configuration of the rewards, rather than their common offset, that enters Eq. 27. Computing Eq. 33 requires one MarginRM evaluation for each non-anchor image and therefore $G - 1$ reward-model calls per prompt. The prompt-specific rubric was generated offline and cached. Its construction cost was not included in the reward latency reported in the main text.

C.3 Post-training setup

The post-training prompt set comprised 10,000 prompts from UnifiedReward-Flex. This training set does not contain any prompts that are identical or nearly identical to those in the GenEval and T2I-CompBench evaluation sets. We post-trained FLUX.1-dev at 512×512 resolution with classifier-free guidance scale 3.5, in `bfloat16` mixed precision. Online rollouts use $L_{\text{train}} = 6$ Euler–Maruyama steps with noise level $\eta = 0.9$, while evaluation uses $L_{\text{eval}} = 28$ steps; a fraction of 0.99 of the training steps (rounded down) is included in $\mathcal{L}_{\text{opt}}$. For each prompt we collect $G = 24$ rollouts with independent initial latents, distributed over 8 GPUs at a per-GPU rollout batch size of 3, and draw 48 prompts per RL epoch. Each rollout batch is used for a single inner epoch with gradient accumulation of 24 steps. We use no KL penalty and maintain an EMA of the policy weights for evaluation.

D Beyond isolated-pair accuracy

Isolated-pair benchmark accuracy captures local preference discrimination, but does not fully characterize a model's ability to recover the human ordering of images generated for the same prompt. We evaluated this distinction by comparing model-level accuracy on MMRB2 and GenAI-Bench with mean prompt-level Kendall's τ_b on Rank-Bench. Across general-purpose models, scalar utility models and direct comparative models, higher accuracy on MMRB2 or GenAI-Bench was broadly associated with higher mean Rank-Bench agreement (Fig. 7). However, the substantial variation in Rank-Bench performance among models with similar pairwise accuracy shows that isolated-pair accuracy does not uniquely determine candidate-set ranking quality. For example, LLaVA-8B and HPS v2 had the same rounded MMRB2 accuracy of 54.7%, yet their mean prompt-level τ_b values on Rank-Bench were 0.047 and 0.160, respectively. The discrepancy was more pronounced

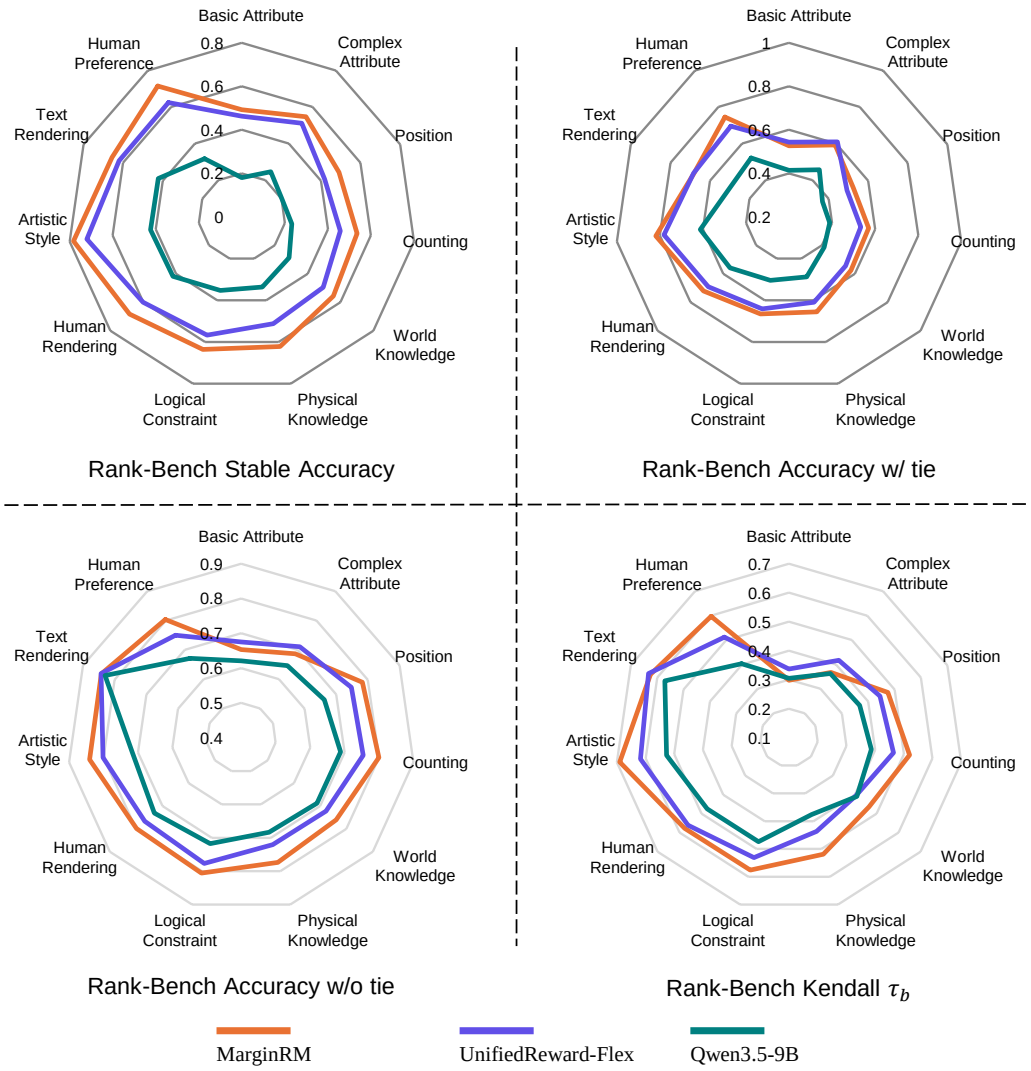


Figure 8: **Dimension-wise performance profiles on Rank-Bench.** Performance across eleven dimensions for MarginRM based on Qwen3.5-9B, compared with the corresponding untrained backbone and UnifiedReward-Flex. Radar plots report stable accuracy (upper left), accuracy with ties (upper right), accuracy without ties (lower left) and Kendall's τ_b (lower right). Radial scales differ across metrics and should therefore be compared only within each radar plot. Larger radii indicate better performance.

on GenAI-Bench: PickScore achieved substantially higher pairwise accuracy than Qwen3-VL-8B (70.0% versus 53.1%), but produced markedly lower agreement with the human candidate-set rankings ($\tau_b = 0.293$ versus 0.407). Thus, a model that is more accurate on independently sampled pairs is not necessarily more consistent with human preferences when all candidates for the same prompt are considered jointly.

E DIMENSION-LEVEL ANALYSIS

MarginRM improved preference evaluation across all eleven Rank-Bench dimensions, indicating that its aggregate advantage was not driven by a small subset of visual judgments. The gains were most consistent in stable accuracy, whereas variation in the remaining ranking metrics was associ-

ated with how effectively the rubric distinguished candidates within each dimension (see Fig. 8). MarginRM achieved higher stable accuracy than UnifiedReward-Flex in all eleven dimensions. Representative gains included physical knowledge, from 51.2% to 62.2%, human preference, from 62.4% to 71.5%, and position, from 41.9% to 49.1%. MarginRM reached its highest stable accuracy on artistic style (78.2%), showing that the shared reference frame remains effective across both semantic and holistic visual criteria. MarginRM exceeded UnifiedReward-Flex in eight of eleven dimensions for accuracy with ties, strict-preference accuracy and Kendall's τ_b, while UnifiedReward-Flex had higher point estimates on basic attribute, complex attribute and text rendering. In dimensions such as basic and complex attributes, multiple candidates may already satisfy the specified semantics to a similar degree. When prompt-relevant differences are small or the discrete margin scale saturates, the remaining human ordering can depend on secondary factors such as visual quality, aesthetics, realism, composition and unrelated generation artifacts. These factors are not always fully represented by the nominal rubric, compressing the ordinal margins that MarginRM predicts within the shared rubric–anchor reference frame. A more flexible evaluator may incorporate such residual holistic cues more freely, which can produce higher point estimates on some dimensions even when its predictions are less stable under input-order reversal. Thus, the dimension-level results support two complementary conclusions. MarginRM applies the intended evaluation criteria more consistently across all eleven dimensions, as reflected by its broad stable-accuracy gains. The remaining variation may reflect incomplete rubric coverage, limitations of the margin predictor, anchor-dependent compression or a combination of these factors.