DiT-Reward: Generative Representations for Text-to-Image Reward Modeling

Anonymous authors
Paper under double-blind review

Abstract

Can representations learned for image generation also support the evaluation of generated images? We study text-to-image reward prediction as a downstream task of generative representation learning and introduce **DiT-Reward**, which repurposes pretrained diffusion transformers by aggregating image representations conditioned on text across layers. Probing a frozen SD3.5-Large backbone reveals that preference information is already accessible through a lightweight learned head, with stronger predictive features in the middle and later layers and complementary information across stages. Using the same public preference data mixture as HPSv3, both our SD3.5-Large and Qwen-Image variants outperform HPSv3 on all four evaluated benchmarks, with the Qwen-Image variant reaching 79.1% accuracy on HPDv3. Backbone comparisons further indicate a positive scaling trend with generative backbone capacity. When used as the reward for Flow-GRPO, the SD3.5-Large variant produces a policy that wins 75.1% of 2,048 human comparisons against the policy optimized with HPSv3, with 21.6% losses and 3.3% ties, after aggregating five ratings per image pair. With a shared latent space between the policy and reward model, direct latent scoring also achieves a $1.65\times$ reward inference speedup over HPSv3 with comparable peak memory. These results establish generative representations as an effective foundation for reward modeling and policy optimization.

1 Introduction

"What I cannot create, I do not understand." This maxim, attributed to Feynman, suggests a close connection between generation and understanding. Modern text-to-image diffusion transformers raise the question from another direction: if they can generate realistic, complex, and semantically rich images, do the representations learned for generation also contain the information needed to judge image quality and human preference? We investigate whether these generative representations can provide an effective foundation for reward modeling, both to evaluate generated images and to guide the optimization of generative models.

Reward models are an important interface for evaluating and optimizing text-to-image models, supporting candidate ranking, reinforcement learning, and preference optimization (Xu et al., 2023b; Kirstain et al., 2023; Fan et al., 2023; Liu et al., 2025). Mainstream approaches rely on pretrained vision–language representations from CLIP and BLIP, or on more powerful VLMs (Radford et al., 2021; Li et al., 2022; Wu et al., 2023; Ma et al., 2025). Recent studies have established the feasibility of using diffusion backbones for reward modeling (Zhang et al., 2025; 2026; Liu et al., 2026; Mi et al., 2025), but have yet to demonstrate consistent superiority over strong VLM reward models across preference benchmarks or a clear positive scaling trend with backbone capacity (Liu et al., 2026). Meanwhile, how preference information in pretrained generative representations is distributed across network depth remains insufficiently characterized. These gaps motivate us to study the performance and representational foundations of reward modeling from the perspective of generative representation learning.

We introduce **DiT-Reward**, which treats reward prediction as a downstream task of generative representation learning. The method aggregates image representations conditioned on text from multiple layers of a pretrained diffusion transformer and maps the pooled features to a scalar reward through a lightweight prediction head. Probing a frozen SD3.5-Large backbone reveals

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.

that generative pretraining already provides information useful for preference prediction: predictive performance varies substantially with depth, is strongest in the middle and later layers, and benefits from complementary features across stages. After adapting the backbone to preference prediction, both the default DiT-Reward based on SD3.5-Large (Stability AI, 2024) and DiT-Reward-Heavy based on Qwen-Image (Wu et al., 2025a) outperform HPSv3 on all four evaluated benchmarks using the same public preference data mixture. The default model achieves these gains with a backbone of comparable scale to that of HPSv3. Relative to the default model, replacing the backbone with the smaller SD3.5-Medium reduces performance across all four benchmarks, whereas the Qwen-Image variant improves three benchmarks while matching the remaining one. These results indicate a positive scaling trend with generative backbone capacity.

Beyond predicting preferences over static images, we examine whether a generative reward model can effectively guide policy optimization. Echoing the use of a common pretrained model family for policies and reward models in language model alignment (Touvron et al., 2023), we initialize both the policy and DiT-Reward from SD3.5-Large and train them separately for generation and reward prediction. Under matched Flow-GRPO settings (Liu et al., 2025), GPT-5 and Gemini-3-Flash consistently assign higher overall scores to the policy optimized with DiT-Reward than to its counterpart optimized with HPSv3. Human evaluation of 2,048 image pairs further shows that the DiT-Reward policy wins 75.1% of comparisons after aggregating five ratings per pair, with the clearest advantages in visual quality and detail richness. These results show that the benefits of reward modeling extend to improvements in generation perceived by human raters. Moreover, because the policy and reward model use the same latent space, rewards can be computed directly from generated latents without image decoding or additional pixel encoding, yielding a $1.65\times$ reward inference speedup over HPSv3 with comparable peak memory.

Our contributions are threefold. First, we introduce DiT-Reward and demonstrate that its SD3.5-Large and Qwen-Image variants outperform HPSv3 on four preference benchmarks under the same public preference supervision. Second, through probing a frozen backbone, we characterize the distribution of preference information across depth and the complementarity of features across stages, while comparisons of generative backbones reveal benefits from greater capacity. Third, external model evaluations and human judgments establish the practical utility of DiT-Reward for text-to-image reinforcement learning, alongside the efficiency advantage of direct reward inference in a shared latent space.

2 Related Work

Human preference modeling and text-to-image evaluation. Human preference modeling provides reward signals for evaluating and aligning text-to-image models. Common approaches rely on pretrained visual or vision–language representations, including those from CLIP and BLIP (Xu et al., 2023b; Kirstain et al., 2023; Wu et al., 2023), as well as more powerful VLM reward models (Ma et al., 2025; Xu et al., 2026; He et al., 2024; Wu et al., 2025b). Recent studies further investigate generative backbones for reward modeling (Zhang et al., 2025; 2026; Liu et al., 2026; Mi et al., 2025), but several questions remain unresolved. Existing diffusion reward models for images have generally not surpassed strong VLM reward models across all reported preference benchmarks. Evaluations across different generative backbones have also shown no clear improvement with increasing parameter count (Liu et al., 2026). Comparisons of pretrained and random initialization in DRM demonstrate the value of generative pretraining, but do not directly identify how preference information is distributed across the layers of a frozen backbone (Zhang et al., 2026). Using the same mixture of preference data as HPSv3, we outperform it on all four benchmarks, characterize the distribution of preference information across depth by probing frozen representations, and observe improved performance with larger generative backbones.

Feedback learning for diffusion and flow models. Another line of work optimizes diffusion or flow generators with preference and reward signals, including reward-gradient guidance, policy-gradient formulations, KL-regularized online RL, and diffusion variants of DPO (Prabhudesai et al., 2023; Black et al., 2024; Fan et al., 2023; Wallace et al., 2024). Recent methods extend this direction to flow matching and rectified-flow models, such as Flow-GRPO and DiffusionNFT (Liu et al., 2025; Zheng et al., 2026). These works mainly ask how to optimize a generator given a reward. Our focus is complementary: we study how the reward model itself should be constructed.

Generative backbones as transferable representation learners. A growing body of work shows that pretrained generative models are also transferable visual representation learners. Diffusion features have been used for semantic correspondence, segmentation, depth estimation, and related dense prediction tasks, with quality depending on layer, timestep, noise level, and conditioning (Tang et al., 2023; Luo et al., 2023; Xu et al., 2023a; Zhao et al., 2023; Stracke et al., 2025). Meanwhile, text-to-image backbones have scaled from LDMs and DiTs to large multimodal diffusion and flow transformers (Rombach et al., 2022; Peebles & Xie, 2023; Podell et al., 2023; Esser et al., 2024; Black Forest Labs, 2024; Wu et al., 2025a; Cao et al., 2025). DiT-Reward connects these threads by using the generative backbone as the representation source for reward modeling.

3 METHOD

3.1 PROBLEM FORMULATION

We consider the problem of learning a reward model for text-to-image generation. Given a text prompt p and a pair of generated images (x^+, x^-), where x^+ is preferred over x^- according to human judgments, our goal is to learn a reward function $r_\theta(x, p)$ that assigns higher scores to preferred images. Formally, we seek to learn r_θ such that $r_\theta(x^+, p) > r_\theta(x^-, p)$ for preference pairs in the training data. The learned reward model can then be applied to image ranking or as a reward signal for downstream reinforcement learning optimization of text-to-image models.

3.2 GENERATOR-NATIVE REWARD MODELING WITH DIT

Most existing text-to-image reward models are built on discriminative vision-language encoders, such as CLIP-style or VLM backbones. DiT-Reward instead uses the pretrained text-to-image generator itself as the reward backbone. The motivation is that a DiT-based generator must internally model visual quality, prompt-image alignment, composition, and generation feasibility in order to synthesize images. Rather than introducing a separate vision-language encoder, we reuse these generator-native representations for reward prediction.

Given a prompt-image pair (p, x), the frozen text encoders encode p, while the VAE maps x to a clean latent z_0. Because the DiT is trained on noisy flow states rather than exactly clean latents, we construct a near-clean input:

$$z_\tau = (1 - \tau)z_0 + \tau\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad 0 < \tau \ll 1, \tag{1}$$

A small nonzero τ reduces this input-distribution mismatch while preserving the image content; larger noise would instead make the reward reflect a substantially perturbed image.

The near-clean image latent and text embeddings are jointly processed by the pretrained MMDiT backbone. Let $\mathcal{S}$ denote the selected DiT layers, and let $H_\ell(x, p) \in \mathbb{R}^{N \times d}$ be the image-token hidden states at layer ℓ, where N and d denote the number of image tokens and hidden dimension, respectively. To combine reward-relevant information across depths, we mean-pool the image tokens at each selected layer, concatenate the resulting features, and map them to a scalar reward:

$$\bar{h}_\ell(x, p) = \frac{1}{N} \sum_{i=1}^{N} H_\ell^{(i)}(x, p), \qquad r_\theta(x, p) = \mathrm{MLP}\big(\mathrm{Concat}_{\ell \in \mathcal{S}}\, \bar{h}_\ell(x, p)\big)\,. \tag{2}$$

3.3 PREFERENCE LEARNING OBJECTIVE

We train DiT-Reward with the standard Bradley-Terry preference objective. Given a prompt p and a preferred/rejected image pair (x^+, x^-), the two images are scored by the same reward model.

We then optimize

$$\mathcal{L}(\theta) = -\log \sigma\left(r_\theta(x^+, p) - r_\theta(x^-, p)\right), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. This loss increases the reward margin between human-preferred and rejected images. In our main setting, the text encoders and VAE remain frozen, while the DiT backbone and reward head can be fine-tuned for preference prediction; we also study a frozen-backbone variant in the ablation section to probe how much reward-relevant information already exists in the pretrained generator.

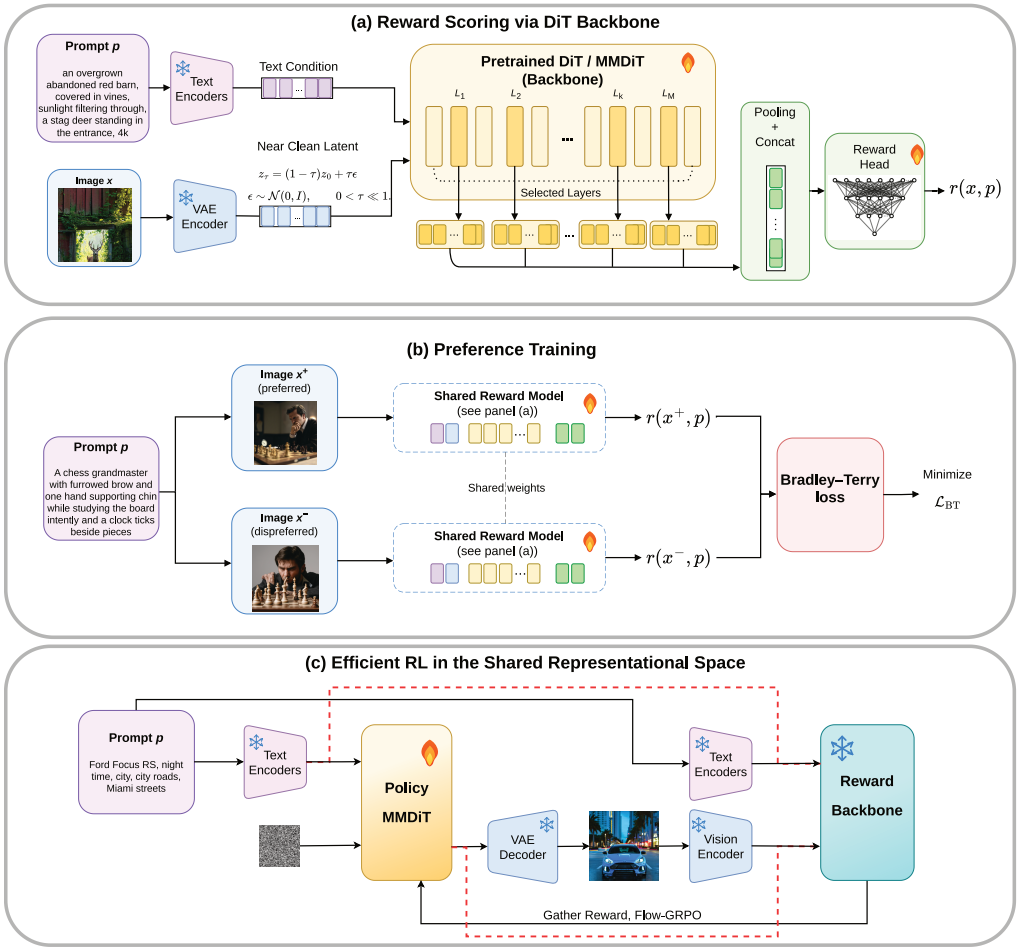


Figure 1: Model architecture, preference training, and reinforcement learning interface of DiT-Reward. **(a)** Given a prompt and image pair, the VAE maps the image into latent space, and a small amount of noise produces a nearly clean latent. A pretrained DiT or MMDiT jointly processes the text condition and image latent. Image token representations from selected layers are pooled, concatenated, and mapped to a scalar reward by a lightweight reward head. **(b)** For preferred and rejected images under the same prompt, DiT-Reward with shared parameters predicts two rewards and learns their ordering through the Bradley–Terry loss. **(c)** During reinforcement learning, a conventional pixel interface requires VAE decoding followed by visual encoding. When the policy and DiT-Reward share a latent space, the red path sends the generated latent directly to the reward model, avoiding additional decoding and encoding while reducing inference cost and peak memory.

3.4 Reinforcement Learning in the Latent Space

Like conventional image reward models, DiT-Reward scores completed prompt-image pairs and can be used for reranking or reward-based policy optimization.

Its generator-native design also enables a more efficient pathway when the policy model and reward model share the same latent space. Conventional VLM-based reward models require the policy latent to be decoded into pixels and then processed again by an external vision encoder. In contrast, DiT-Reward can score candidates from the shared latent representation after generation, avoiding unnecessary decoding and re-encoding steps. This latent-space reward evaluation reduces redundant computation and memory usage, which is especially useful for reinforcement learning with large diffusion or flow policies.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

We instantiate our reward models using SD3.5-Large and Qwen-Image, denoted DiT-Reward and DiT-Reward-Heavy, respectively. DiT-Reward also refers to the overall method; unless otherwise specified, experiments use the default SD3.5-Large configuration. Both variants extract image representations from four layers spaced across the middle and later portions of the network, pool and concatenate these features, and predict a scalar reward with a lightweight MLP. Detailed backbone configurations and layer selections are provided in Appendix A.

To focus the comparison on reward-model architecture rather than supervision, we adopt the public HPSv3 training recipe. The training mixture includes HPDv3 pairwise data, a filtered golden subset from HPDv2, sampled Pick-A-Pic data, sampled ImageReward data, and Midjourney user-choice data, without introducing additional data sources.

During training, images are encoded into the latent space of the corresponding backbone. Following prior work on reward modeling from noisy generative representations (Mi et al., 2025), we apply a small Gaussian perturbation with $\tau = 0.005$. We keep the noise level low to preserve image details relevant to preference prediction, as larger perturbations reduce the signal-to-noise ratio of the input latents. The VAE and text encoders remain frozen. The DiT backbone and reward head are trainable in the main benchmark setting. Appendix A provides training details, and Appendix A reports training and benchmark trajectories.

Table 1: Pairwise preference prediction accuracy (%). Each block compares a baseline family with our two diffusion-based variants, separated by a thin rule. Bold and underlined values denote the best and second-best results within each comparison block. DiT-Reward uses SD3.5-Large; DiT-Reward-Heavy uses Qwen-Image. * marks our evaluation of the official HPSv3 checkpoint using the official PickScore protocol.

Model	ImageReward	PickScore	HPDv2	HPDv3
Vision-encoder / VLM baselines				
CLIP ViT-H/14	57.1	60.8	65.1	48.6
Aesthetic Score Predictor	57.4	56.8	76.8	59.9
ImageReward	65.1	61.1	74.0	58.6
PickScore	61.6	**70.5**	79.8	65.6
HPS	61.2	66.7	77.6	63.8
HPSv2	65.7	63.8	83.3	65.3
MPS	**67.5**	63.1	83.5	64.3
HPSv3	66.8	65.1*	85.4	76.9
DiT-Reward	<u>67.0</u>	66.7	<u>85.6</u>	<u>77.6</u>
DiT-Reward-Heavy	<u>67.0</u>	<u>69.8</u>	**85.9**	**79.1**
Diffusion-based baselines				
LRM-SD1.5	59.2	67.2	72.4	54.1
LRM-SDXL	<u>60.4</u>	<u>69.3</u>	71.2	53.8
DiNa-LRM	60.3	59.9	82.1	75.0
DiT-Reward	**67.0**	66.7	<u>85.6</u>	<u>77.6</u>
DiT-Reward-Heavy	**67.0**	**69.8**	**85.9**	**79.1**

4.2 MAIN RESULTS ON PREFERENCE BENCHMARKS

We evaluate four public pairwise-preference benchmarks. ImageReward (Xu et al., 2023b) contains 466 prompts and 6,399 strict pairs after excluding ground-truth ties; HPDv2 (Wu et al., 2023) contains 400 prompts and 15,300 pairs. PickScore (Kirstain et al., 2023) tunes its tie threshold on 500 validation prompts and evaluates 500 test prompts. HPDv3 (Ma et al., 2025) covers recent generators with 14,400 test pairs from 400 prompts. The first three are established benchmarks, while HPDv3 targets modern generators.

Table 1 reports preference prediction accuracy on four benchmarks. Under the same training data mixture, both DiT-Reward variants outperform HPSv3 on all four benchmarks. Notably, the default

DiT-Reward achieves these gains with a backbone of comparable scale to that of HPSv3. DiT-Reward achieves 77.6% on HPDv3, while DiT-Reward-Heavy further improves accuracy to 79.1%, exceeding HPSv3 by 2.2 percentage points. The Heavy variant also improves PickScore accuracy from 66.7% to 69.8%, demonstrating the effectiveness of the approach beyond the default SD3.5-Large backbone.

4.3 Reinforcement Learning with DiT-Reward

We evaluate DiT-Reward as a reward signal for text-to-image reinforcement learning using Flow-GRPO (Liu et al., 2025). We use Stable Diffusion 3.5 Large as the policy and follow the official Flow-GRPO PickScore prompt configuration for both RL training and evaluation. Training images are generated at 512×512 resolution with 10 sampling steps and guidance scale 4.5, while evaluation uses 40 sampling steps. We use a group size of 16 and, through distributed gradient accumulation across 16 GPUs, an effective optimization batch size of 384 samples per update. The policy is optimized with LoRA, with $\beta = 0.01$ controlling the strength of KL regularization. The preceding policy, sampling, and optimization settings are matched between the DiT-Reward and HPSv3 runs, while each reward model retains its native scoring pathway; in particular, DiT-Reward scores samples in latent space.

For external evaluation, we use the 2,048 PickScore prompts from the official Flow-GRPO protocol. GPT-5 and Gemini-3-Flash evaluate each generated image independently at 60-step intervals along four dimensions: prompt alignment, visual quality, realism, and detail richness. The complete prompts and evaluation protocol are provided in Appendix C. For qualitative analysis, all examples from each optimized policy in Figure 3 use its checkpoint at step 900.

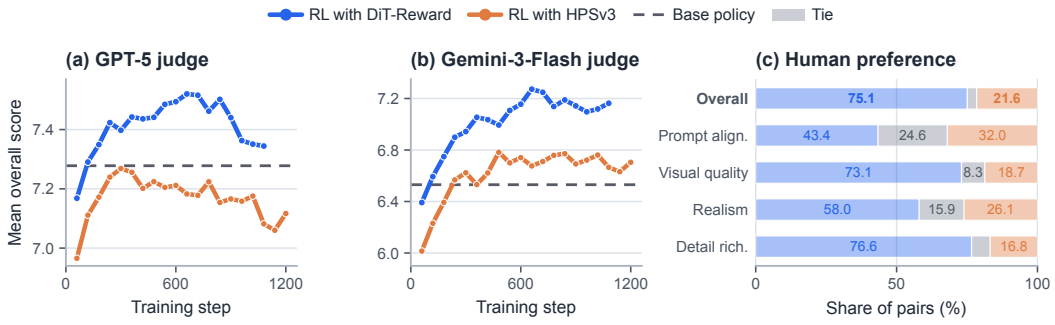


Figure 2: External and human evaluation during Flow-GRPO training. Panels (a) and (b) report mean overall scores assigned by GPT-5 and Gemini-3-Flash, respectively, to policies trained with DiT-Reward (blue) and HPSv3 (orange). Gray dashed lines denote the corresponding base-policy scores of 7.29 and 6.53. Panel (c) reports human preferences at step 660. Each image pair is assigned to the model receiving more preference votes, or to a tie if the counts are equal; individual tie votes do not favor either model. Labels give percentages of image pairs; the unlabeled tie shares for Overall and Detail richness are 3.3% and 6.6%, respectively.

Flow-GRPO steadily increases the DiT-Reward evaluation score from -0.94 to 5.06, confirming that the policy can optimize the learned reward. As shown in Figure 2, across the 18 matched evaluation checkpoints from steps 60 to 1080, replacing HPSv3 with DiT-Reward increases the average overall score from 7.18 to 7.41 under GPT-5 and from 6.61 to 7.02 under Gemini-3-Flash. DiT-Reward achieves a higher overall score at every matched checkpoint under both judges. The gains are concentrated in visual quality, realism, and detail richness, while HPSv3 retains a modest advantage in prompt alignment, particularly during later training. For the DiT-Reward run, the internal reward continues to rise after the external scores peak. Qualitative inspection indicates that visible reward hacking emerges at a similar stage in both reward-model runs. Appendix C reports complete trajectories for all four dimensions.

Human evaluation confirms that these gains translate into improvements perceived by human raters. We compare 2,048 image pairs generated by the two policies at step 660, with five raters per pair and randomized A/B assignments. After aggregating votes for each pair, DiT-Reward wins 75.1% of comparisons (95% CI: 73.2–76.9%), versus 21.6% for HPSv3, with 3.3% ties. The advantage

is strongest in visual quality and detail richness, while the improvement in prompt alignment is smaller. Appendix D provides the evaluation protocol, detailed results, and robustness analyses.

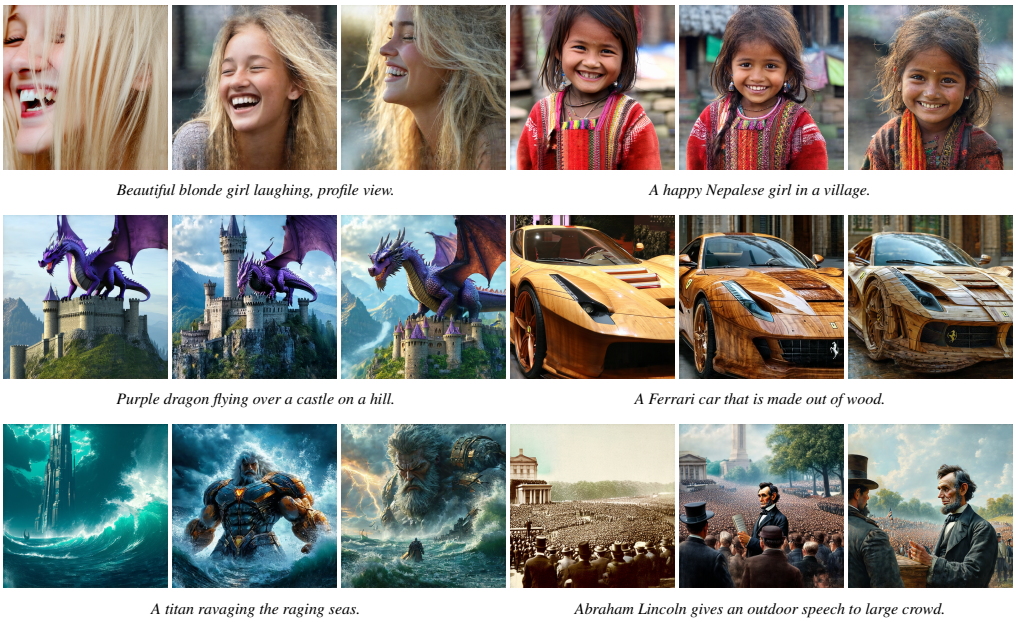


Figure 3: Qualitative comparison at training step 900. Within each prompt group, images show the base policy, RL with HPSv3, and RL with DiT-Reward from left to right. All examples from a given optimized policy are sampled from its step-900 checkpoint.

Figure 3 compares matched-prompt generations from the base policy and policies optimized with HPSv3 and DiT-Reward. Before pronounced reward overoptimization, the policy optimized with DiT-Reward produces richer local details and more realistic portraits, animals, and objects. It also exhibits less drift toward excessive brightness, saturation, and contrast than the policy optimized with HPSv3, avoiding some unnatural colors and distorted shapes.

Both reward models eventually exhibit reward hacking: faces and objects become frontal and rigid, hallucinated details accumulate, and outputs converge toward reward-favored styles at the expense of prompt fidelity. Their failure modes nevertheless differ. HPSv3 tends to amplify saturation and contrast, sometimes producing implausible colors and geometry, whereas DiT-Reward better preserves global structure but may yield muted colors, spurious faces, and weaker prompt alignment. These differences indicate that reward models trained on the same preference data can induce distinct optimization biases because of their pretraining objectives and backbone architectures. A plausible explanation is that the diffusion backbone provides stronger structural priors, while the VLM-based reward retains an advantage in text-instruction following. One possible direction for future work is to incorporate text-token representations from the joint transformer and explicitly model text–image interactions to improve the prompt-alignment sensitivity of DiT-Reward.

Appendix E presents an exploratory cross-family comparison of FLUX.1 [dev] policies optimized with HPSv3 or DiT-Reward on a separate 1,000-prompt challenge set.

4.4 Efficient Reward Inference

A key advantage of DiT-Reward is that it can directly evaluate generated latents when the reward model and policy share the same latent space. This pathway avoids VAE decoding and pixel-space re-encoding by an external vision backbone. We compare its inference efficiency with HPSv3 and the pixel-interface DiT-Reward pathway in Table 2.

Table 2: Reward inference efficiency on 512×512 images measured on one NVIDIA RTX PRO 6000 Blackwell GPU with batch size 1. Each mode uses one warmup iteration excluded from timing. The reward interface indicates whether the model receives decoded pixels or policy latents. Lower values are better.

Reward Model	Reward Interface	Per-Image Latency	Peak Memory
DiT-Reward	Pixel	154 ms	27 GB
HPSv3	Pixel	89 ms	**16 GB**
DiT-Reward	Latent	**54 ms**	**16 GB**

Although the complete DiT-Reward inference stack contains an 8B DiT backbone and approximately 5B additional parameters in its frozen text encoders, the latent-interface pathway achieves the lowest inference latency: 54 ms per image compared with 89 ms for HPSv3, corresponding to a $1.65\times$ speedup. Both the latent-interface pathway and HPSv3 require approximately 16 GB of peak memory. DiT-Reward with the pixel interface instead requires 154 ms per image and approximately 27 GB of memory. Together with the improved reward-model and downstream RL performance reported above, these results indicate that aligning the policy and reward model in a shared latent representation space can improve performance without sacrificing inference efficiency. In particular, the aligned latent interface enables a large generative reward model to achieve faster inference with a memory footprint comparable to HPSv3, highlighting representation-space alignment as both a modeling and systems advantage for text-to-image reinforcement learning.

5 Ablation and Interpretability

We organize our analysis around two objectives. First, we examine whether representations from a pretrained generative DiT can support reward modeling with a frozen backbone and identify which layers are most informative for this task. Second, we investigate whether reward modeling performance exhibits scaling behavior as the parameter count of the generative backbone increases.

5.1 Reward-Relevant Representations in Generative DiTs

To test whether a pretrained generative DiT already contains reward-relevant representations, we freeze SD3.5-Large and train only a lightweight reward head with approximately 10M parameters over its intermediate image tokens. The head-only model achieves substantial preference-prediction accuracy across all four benchmarks, including 72.3% on HPDv3 (Table 3), indicating that preference information is already accessible without adapting the generative backbone.

We next examine where this information emerges across depth. The layer-similarity analysis in Appendix B identifies four locally coherent regions, B1–B4.

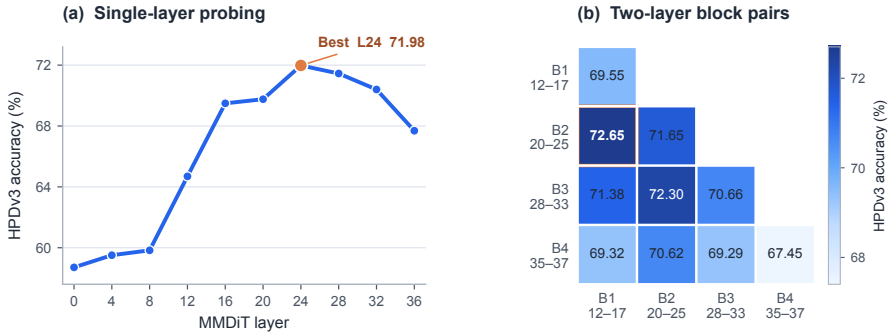


Figure 4: Layer-wise reward probing with a frozen SD3.5-Large backbone; only the lightweight head is trained on HPDv3. **(a)** Individual-layer accuracy evaluated at four-layer intervals. **(b)** Accuracy of two-layer block combinations; the symmetric upper triangle is omitted. B1–B4 are defined by the layer-similarity analysis in Appendix B.

The single-layer probes in Figure 4(a), evaluated at four-layer intervals, show a clear depth-dependent trend: HPDv3 accuracy increases from 58.71% at layer 0 to 71.98% at layer 24, before declining toward the final layers. Figure 4(b) shows that selected cross-stage combinations are complementary: B1+B2 achieves the highest accuracy of 72.65%, followed by B2+B3 at 72.30%. These results suggest that reward-relevant information is strongest at intermediate depth and can benefit from combining representations across neighboring stages.

5.2 Scaling with Generative Backbone Capacity

To study how reward modeling scales with generative backbone capacity, we compare SD3.5-Medium, SD3.5-Large, and Qwen-Image under the same reward-modeling framework. The Medium-to-Large comparison provides a within-family scaling test, whereas Qwen-Image examines whether the trend extends beyond the SD3.5 family. Notably, Qwen-Image and SD3.5 remain highly similar in backbone design: both adopt a dual-stream MMDiT architecture that jointly processes textual and visual representations, making Qwen-Image a meaningful cross-family extension.

Table 3: Effects of backbone adaptation and capacity under identical training data, reward objective, and evaluation protocol. "Frozen" means training only the reward head. SD3.5-Large and Qwen-Image correspond to DiT-Reward and DiT-Reward-Heavy; the Medium variant is denoted DiT-Reward (SD3.5-M).

Backbone	Backbone Training	ImageReward	PickScore	HPDv2	HPDv3
SD3.5-Large (8.1B)	Frozen	57.5	60.9	78.0	72.3
SD3.5-Medium (2.5B)	Full adaptation	57.9	60.9	79.7	71.5
SD3.5-Large (8.1B)	Full adaptation	**67.0**	66.7	85.6	77.6
Qwen-Image (20B)	Full adaptation	**67.0**	**69.8**	**85.9**	**79.1**

As shown in Table 3, scaling from SD3.5-Medium to SD3.5-Large consistently improves performance across all four benchmarks. DiT-Reward-Heavy with Qwen-Image further improves three metrics while matching the remaining one. In contrast, the results reported by DiNa-LRM (Liu et al., 2026) do not show a consistent positive scaling trend and even exhibit negative scaling in some settings. Overall, our architecture demonstrates a clearer and more robust scaling trend as the generative backbone becomes stronger.

6 Conclusion

We presented DiT-Reward, which repurposes pretrained text-to-image diffusion transformers for reward modeling. Probing a frozen backbone shows that generative representations already contain information useful for preference prediction, with stronger predictive features in the middle and later layers and complementary information across stages. Under the same preference data mixture, both the SD3.5-Large and Qwen-Image variants outperform HPSv3 on all four evaluated benchmarks, while backbone comparisons suggest benefits from greater generative capacity. Beyond preference prediction, DiT-Reward supports reinforcement learning that improves human preference over generated images and enables efficient reward inference directly in the shared latent space. Although reward overoptimization and limitations in prompt alignment remain, these findings support generative representations as an effective foundation for both evaluating and improving image generation.

AI Assistance

Generative AI tools, including OpenAI Codex, were used to polish language, support literature retrieval and discovery, and assist in drafting and revising portions of the manuscript. The authors independently verified the cited sources, technical claims, and all AI-assisted revisions. All research decisions, experiments, reported results, and conclusions were produced and validated by the authors, who take full responsibility for the final manuscript.

REFERENCES

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

Black Forest Labs. FLUX.1 [dev] model card. Technical report, Black Forest Labs, 2024.

Siyu Cao, Hangting Chen, Peng Chen, et al. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 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 *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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

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. In *Advances in Neural Information Processing Systems*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Gongye Liu, Bo Yang, Yida Zhi, Zhizhou Zhong, Lei Ke, Didan Deng, Han Gao, Yongxiang Huang, Kaihao Zhang, Hongbo Fu, and Wenhan Luo. Beyond VLM-based rewards: Diffusion-native latent reward modeling. *arXiv preprint arXiv:2602.11146*, 2026. URL https://arxiv.org/abs/2602.11146.

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. In *Advances in Neural Information Processing Systems*, 2025.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Fan Tang. Video generation models are good latent reward models. *arXiv preprint arXiv:2511.21541*, 2025. URL https://arxiv.org/abs/2511.21541.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Stability AI. Introducing stable diffusion 3.5. Technical report, 2024.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Bjorn Ommer. CleanDIFT: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yan Zeng, and Weilin Huang. RewardDance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025b.

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

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023a.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935, 2023b.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11269–11277, 2026.

Jaxon Zhang, Binxin Yang, Hubery Yin, Chen Li, and Jing Lyu. DRM: Diffusion-based reward model with step-wise guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL `https://arxiv.org/abs/2605.25661`.

Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *arXiv preprint arXiv:2502.01051*, 2025. URL https://arxiv.org/abs/2502.01051.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026.

A DiT-Reward Training Details

We instantiate DiT-Reward with Stable Diffusion 3.5 Large (SD3.5-L). In the official diffusers configuration, the MMDiT transformer contains 38 layers and uses 38 attention heads with head dimension 64, corresponding to a hidden dimension of 2432. The joint text-conditioning dimension is 4096, and the pooled projection dimension is 2048.

The text condition is produced by three frozen pretrained text encoders: CLIP-L/14, OpenCLIP bigG/14, and T5-XXL. The two CLIP token embeddings are concatenated along the channel dimension and padded to match the T5 embedding dimension; the resulting CLIP sequence is then concatenated with the T5 sequence along the sequence dimension and fed to the text side of the joint DiT. The pooled CLIP embeddings are also concatenated and used as global conditioning together with the timestep embedding. For image inputs, a frozen 16-channel VAE encoder maps the image into latent tokens, which are patchified and fed to the image side of the joint DiT. Under the rectified-flow formulation, the DiT takes the latent coordinate z, text condition c, and timestep t as inputs and predicts a velocity field in the latent space.

Feature extraction and training configuration. For DiT-Reward, we extract image hidden states from layers 15, 23, 28, and 37, using zero-based indices. The reward head uses a hidden dimension of 1536 and dropout of 0.1. The default full adaptation configuration uses images of size 512×512, a noise level of $\tau = 0.005$, and three training epochs. The backbone and reward head learning rates are 5×10^{-6} and 2×10^{-5}, respectively, with cosine scheduling, a warmup ratio of 0.05, and zero weight decay. The VAE and text encoders are frozen. Training uses bfloat16 precision and gradient checkpointing.

DiT-Reward-Heavy. The Heavy variant uses the Qwen-Image backbone, with approximately 20B parameters, 60 transformer blocks, and a hidden dimension of 3072. It uses the Qwen-Image VAE and Qwen2.5-VL text encoder, both frozen during reward training. Image representations are extracted from layers 29, 39, 49, and 59, with the same pooling and concatenation procedure used for DiT-Reward. The standard full adaptation configurations use a reward head hidden dimension of 1536, dropout of 0.1, images of size 512×512, a maximum text length of 512 tokens, and $\tau = 0.005$. These configurations train for three epochs with cosine scheduling, zero weight decay, bfloat16 precision, and gradient checkpointing.

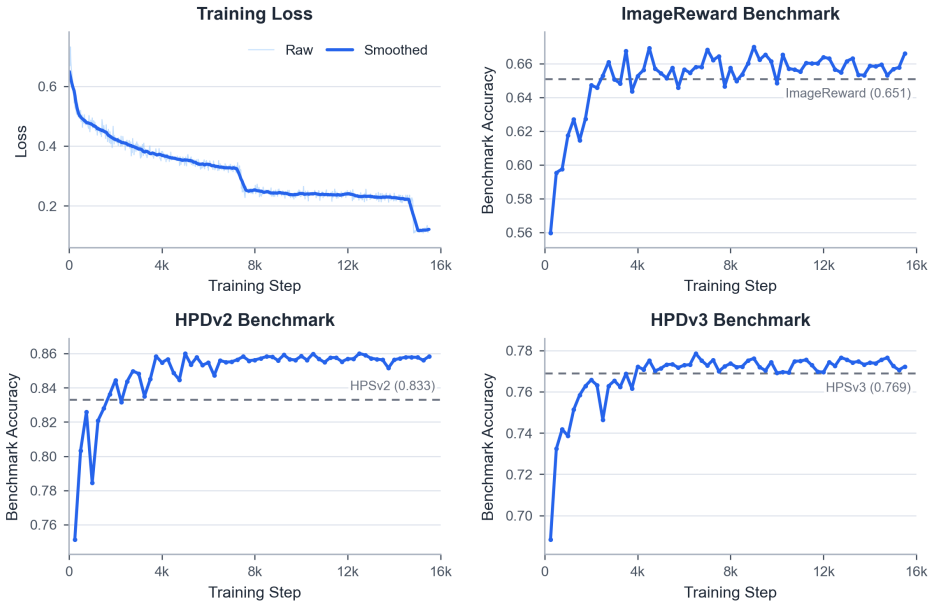


Figure 5: Training dynamics of DiT-Reward. **Top-left**: training loss. **Top-right**: ImageReward benchmark accuracy. **Bottom-left**: HPDv2 benchmark accuracy. **Bottom-right**: HPDv3 benchmark accuracy. Dashed lines indicate the corresponding baseline scores.

B Layer Similarity Across Denoising Steps

Figure 6 provides the layer-similarity analysis used to define the probing blocks in Figure 4. Panel (a) shows prompt-averaged similarity and prompt-wise variability across denoising steps. The annotated Step-39 matrix identifies four locally coherent regions, B1–B4, while the full sequence shows that this block structure is stable throughout denoising. Panel (b) summarizes the same structure: within-block similarity remains near 0.94, whereas across-block similarity is 0.71–0.75. Prompt-wise variability is small overall (mean pairwise standard deviation 0.006–0.016), although it increases at later steps, particularly for interactions involving shallow and deep layers.

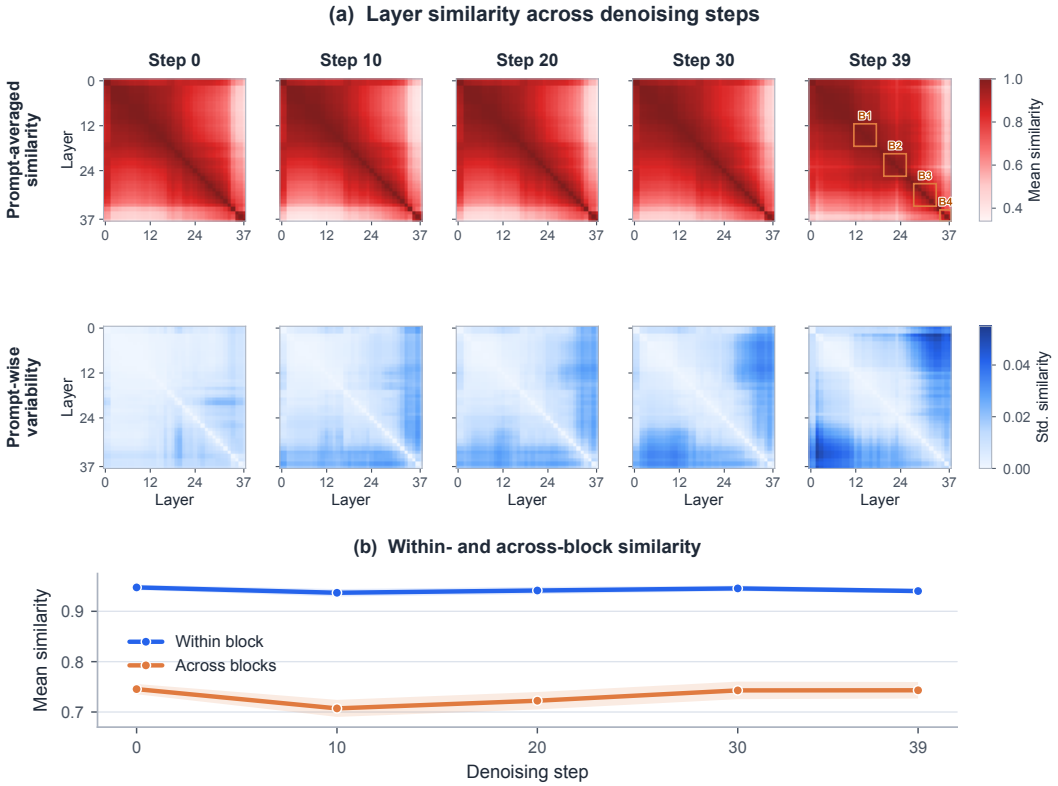
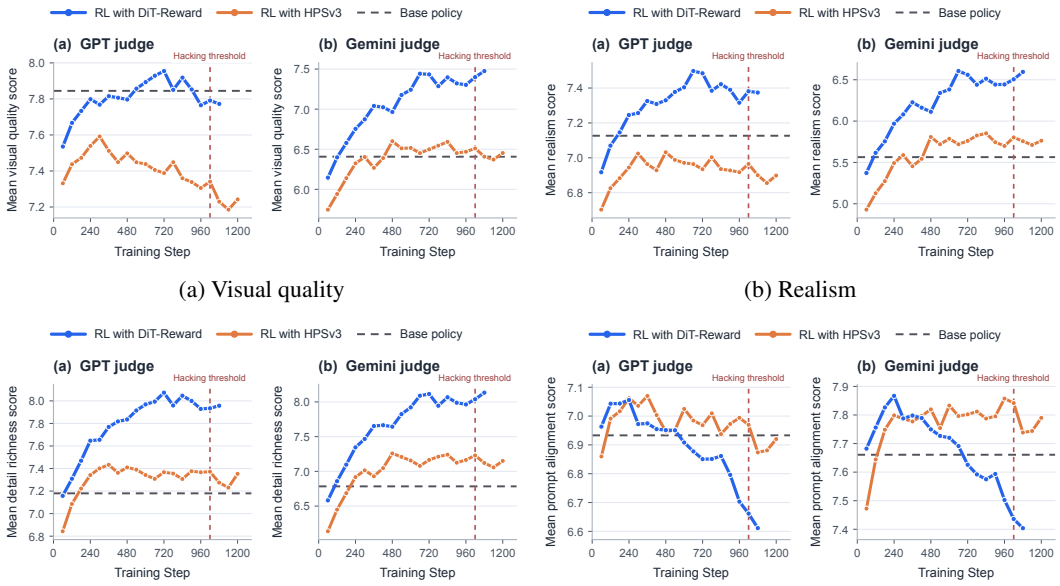


Figure 6: Layer similarity and block definition throughout denoising. **(a)** Prompt-averaged similarity among all 38 MMDiT layers (top) and the corresponding standard deviation across prompts (bottom) at steps 0, 10, 20, 30, and 39. The Step-39 matrix marks the B1–B4 ranges used for block-wise probing. **(b)** Mean similarity within and across these blocks; shading denotes prompt-wise variability.

C Fine-Grained External VLM Evaluation During RL Training

We use GPT-5 and Gemini-3-Flash as external judges. At each evaluation checkpoint, every generated image is evaluated independently against its text prompt. For each policy and checkpoint, we average each numeric field across the 2,048 independently evaluated images. The overall score is returned directly by the judge rather than recomputed from the four dimensions; the rationale field in the JSON response is not used in the quantitative analysis. Figure 7 reports the resulting trajectories for visual quality, realism, detail richness, and prompt alignment.



(a) Visual quality (b) Realism

(c) Detail richness (d) Prompt alignment

Figure 7: Fine-grained external evaluation throughout RL training. Each panel reports scores from the GPT-5 and Gemini-3-Flash judges.

Across the 18 matched checkpoints from steps 60 to 1080, DiT-Reward consistently outperforms the HPSv3 baseline in visual quality, realism, and detail richness. Under GPT-5, the average gains in these three dimensions are 0.39, 0.38, and 0.49 points, respectively; under Gemini-3-Flash, the corresponding gains are 0.69, 0.62, and 0.68 points. All three dimensions favor DiT-Reward at every matched checkpoint under both judges. Prompt alignment follows a different trajectory: the two policies are comparable during early training, but DiT-Reward falls behind HPSv3 at later checkpoints, yielding average differences of -0.09 under GPT-5 and -0.10 under Gemini-3-Flash. Together, these results show that the overall advantage of DiT-Reward is driven by stronger visual quality, realism, and detail richness, alongside a trade-off in prompt alignment.

Both judges receive the same system and user prompts. The system prompt is:

```
You are an expert evaluator for text-to-image generation
models.  Your task is to assess how well a generated image
matches a given text prompt.  You must return a JSON object
and nothing else.
```

The user prompt is shown below, where {prompt} is replaced by the corresponding text prompt for each image:

```
Evaluate the generated image against the following text
prompt:

TEXT PROMPT: "{prompt}"

Score the image on 4 dimensions (1--10, use the full range):
1.  prompt_alignment --- Does the image depict what the
prompt describes?  (subject, count, attributes, scene)
2.  visual_quality --- Is the image sharp, well-composed,
free of artifacts?
3.  realism --- Does it look like a plausible, coherent
scene?
4.  detail_richness --- Are relevant details present and
well-rendered?
```

```
Also give an overall score (1--10, one decimal) that
holistically reflects all four dimensions.

Return ONLY valid JSON (no markdown, no explanation):
{
  "prompt_alignment": <int 1-10>,
  "visual_quality": <int 1-10>,
  "realism": <int 1-10>,
  "detail_richness": <int 1-10>,
  "overall": <float 1-10, one decimal>,
  "rationale": "<one sentence>"
}
```

D HUMAN EVALUATION

We compare images generated at step 660 by the SD3.5-Large policies optimized with DiT-Reward and HPSv3, using the same 2,048 PickScore evaluation prompts as in the external VLM evaluation. Each prompt corresponds to one image pair. Ten raters participate in the evaluation, with five ratings collected for each pair. Each rater selects image A, image B, or a tie for prompt alignment, visual quality, realism and plausibility, detail richness, and overall preference. The raters received training before the evaluation and were blinded to model identities. We shuffled the evaluation records before sending them to the evaluation team, and the team also shuffled their presentation order. Model assignments to A and B are randomized independently for each pair with seed 42. DiT-Reward appears as A in 989 pairs and as B in 1,059 pairs. The records contain five complete evaluations for every pair, yielding 10,240 evaluations and 51,200 individual judgments across the five dimensions. We verified that all pair identifiers and image links match the evaluation inputs and that the 2,048 prompts are distinct.

For each pair and dimension, let n_D and n_H denote the numbers of votes for DiT-Reward and HPSv3. We assign a DiT-Reward win when $n_D > n_H$, an HPSv3 win when $n_H > n_D$, and a tie otherwise. Individual tie votes contribute to neither count. This rule compares the two models' support and does not require agreement from at least three raters. All percentages in Figure 2(c) use the full set of 2,048 pairs as the denominator. Table 4 also reports the conditional win rate after excluding aggregate ties, which has a different denominator.

Table 4: Human preferences aggregated by image pair. D and H denote the policies optimized with DiT-Reward and HPSv3. Win rates and confidence intervals are percentages; intervals are 95% Wilson intervals. The last column excludes aggregate ties.

Dimension	D wins	Ties	H wins	D win rate (95% CI)	Excluding ties (95% CI)
Overall preference	1,538	67	443	75.1 [73.2, 76.9]	77.6 [75.8, 79.4]
Prompt alignment	889	503	656	43.4 [41.3, 45.6]	57.5 [55.1, 60.0]
Visual quality	1,497	169	382	73.1 [71.1, 75.0]	79.7 [77.8, 81.4]
Realism and plausibility	1,188	325	535	58.0 [55.9, 60.1]	68.9 [66.7, 71.1]
Detail richness	1,568	135	345	76.6 [74.7, 78.3]	82.0 [80.2, 83.6]

We use image pairs, rather than individual votes, as the units for statistical inference. A two-sided exact sign test on the 1,981 decisive overall comparisons gives $p = 4.7 \times 10^{-141}$ under the null of equal win probabilities. All five dimensions remain significant at the 0.05 level after Bonferroni correction for five tests; the largest unadjusted p-value is 3.4×10^{-9} for prompt alignment. These tests and the Wilson intervals assume independent image pairs and characterize uncertainty conditional on the participating raters; they do not capture the full uncertainty associated with recruiting a different rater population.

The overall advantage remains under a stricter aggregation rule that requires at least three of the five raters to prefer a model. Under this rule, DiT-Reward wins 1,509 pairs (73.7%), HPSv3 wins 435 (21.2%), and 104 (5.1%) have no majority for either model. Table 5 compares this result with the

original aggregation and with the distribution of individual votes. Across the 10,240 overall judgments, there are 6,737 votes for DiT-Reward, 3,065 for HPSv3, and 438 ties. Thus, the 75.1% result describes the fraction of image pairs won, not the fraction of individual votes received. Assigning half credit to individual ties gives a mean preference score of 67.9%, with a 95% percentile bootstrap interval of 66.8–69.1%. We obtain this interval from 10,000 resamples of image pairs, retaining all five votes for each sampled pair, with random seed 20260917.

Table 5: Overall preference under alternative aggregation rules and evaluation subsets. Percentages use all units in the corresponding row. The middle column denotes aggregate ties, individual tie votes, or absence of a strict majority, depending on the row. Subset rows use the original aggregation rule.

Aggregation or subset	Units	D wins (%)	Ties / no majority (%)	H wins (%)
Original aggregation	2,048 pairs	75.1	3.3	21.6
Strict majority	2,048 pairs	73.7	5.1	21.2
Individual votes	10,240 votes	65.8	4.3	29.9
DiT-Reward assigned to A	989 pairs	78.3	3.0	18.7
DiT-Reward assigned to B	1,059 pairs	72.1	3.5	24.4

The overall win rate is higher when DiT-Reward is assigned to A than to B. A two-sided Fisher exact test comparing wins and losses across the two assignments gives $p = 0.0017$. DiT-Reward leads in both assignments. Although model-to-position assignments were randomized and the evaluation records were shuffled, the A and B subsets contain different image pairs rather than the same pairs evaluated in both orders. The observed difference is consistent with a position effect but does not by itself isolate one for each image pair; evaluating the same pairs again with swapped positions would provide a more direct check.

Individual raters do not always agree despite the aggregate preference for DiT-Reward. For overall preference, the mean agreement between two raters evaluating the same pair is 59.9%, and Fleiss' κ over the three response categories is 0.158. These results support a preference advantage averaged over the evaluated pairs, rather than a claim of uniformly strong agreement among raters. Together with the smaller improvement in prompt alignment, they indicate that the clearest benefits of DiT-Reward in this evaluation lie in visual quality and detail richness.

E Exploratory Cross-Policy Optimization on Challenging Prompts

To test cross-family policy optimization, we train FLUX.1 [dev] with Flow-GRPO using either HPSv3 or DiT-Reward; all other settings are matched. Training uses 18,000 prompts from a pool of 20,000 covering human and animal structure, motion and sports, spatial counting, optics and lighting, text rendering, complex scenes, and general aesthetics. We shuffle the pool with seed 42 and reserve the remaining 2,000 prompts for evaluation during training.

Qualitative evaluation uses a disjoint set of 1,000 unique challenge prompts in 47 fine-grained categories, including complex anatomy and motion, multi-object spatial relations, counting and negation, text rendering, and physical constraints. There is no exact prompt overlap with the training or training-time evaluation splits. Both runs train rank-64 LoRA adapters on 16 GPUs at 512×512 resolution with 20 sampling steps, guidance scale 3.5, group size 16, learning rate 3×10^{-5}, and BF16 precision. We use CPS sampling at noise level 0.4 and an SDE window of three steps within the first ten sampling steps. Each epoch contains 768 sampled images; the effective optimization batch size is 384.

For Figure 8, we use the step-660 EMA LoRA checkpoint from each run. For every challenge prompt, each policy generates one 1024×1024 image with 28 inference steps and guidance scale 3.5. The two policies use the same prompt and random seed, with seed $51000+i-1$ for prompt index i. The figure shows six selected prompt pairs. Because we do not report an aggregate comparison over the challenge set, we treat these examples as qualitative evidence only.



A Brazilian jiu-jitsu ground submission: one athlete lying flat on their back with their right arm extended straight up and being locked between the opponent's thighs — opponent is perpendicular on the ground, both legs wrapped around the arm, hips elevated for leverage. The trapped arm's elbow is at the apex of the lock, wrist pointing skyward, fingers open. The defender taps the mat with their free left hand, all five fingers spread flat against the floor.

A figure drawing class scene with a male model in a contrapposto stance: the right hip is elevated and the left hip is lower, right leg is the weight-bearing leg and left leg is slightly forward and relaxed. The right arm hangs at the side with fingers loosely curled. The left arm is bent at the elbow with the hand resting lightly against the thigh. The model's head is turned 45° to the left. Students are visible in the background with easels, their hands holding charcoal or brushes.

Solo dancer frozen in a floor-based rolling moment transitioning to standing: the right shoulder blade is the current contact point with the floor, the spine is curled into a C-shape, knees drawn to the chest. The left hand is pressing lightly on the floor to guide the roll direction. The right leg is beginning to extend as the dancer prepares to use momentum to rise. The head is turned toward the direction of travel, chin tucked, using the curve of the skull to continue the roll.



A sign language interpreter mid-conversation showing the letter "A" in American Sign Language — right hand clenched in a fist with thumb resting against the side of the index finger rather than tucked inside. All four fingers are tightly curled, fingernails visible on the outer knuckles. The wrist is relaxed, palm facing the viewer, forearm at a comfortable mid-chest height. The interpreter's face shows a neutral conversational expression, eyes attentive.

A pianist in the middle of a forte chord: both hands pressing keys simultaneously, fingers arched at the second knuckle in a proper piano technique curve, wrists level. The left hand spans an octave plus a third, with thumb on C and pinky on E — the stretch visibly straining the web between fingers. The right hand plays a treble chord, all five fingers down, middle finger slightly depressed deeper than the others on the melody note. Arms and shoulders slightly elevated from the keys.

Pair in synchronized swimming frozen in a bent-knee vertical with rocket boost: the base swimmer is vertical in the water, legs below the surface driving an eggbeater kick. The second swimmer stands on the base's feet, propelled upward — she is 60 centimeters above the waterline, body straight, arms in a streamlined position overhead, toes pointed, chin level. The base swimmer's hands are above water steadying the second swimmer's ankles. Both faces are aligned, both expression controlled.

Figure 8: Qualitative comparison for cross-policy reinforcement learning on challenging prompts. FLUX.1 [dev] is used as the base policy. Within each prompt pair, the left image is optimized with HPSv3 and the right image with DiT-Reward. All images are sampled from the step-660 EMA LoRA checkpoints at 1024×1024 resolution using 28 inference steps and guidance scale 3.5; the two images in each pair share the same prompt and random seed. In these selected examples, the first row shows differences in body structure during complex actions and interactions, while the second row shows differences in fine-grained anatomical details such as finger count and hand shape.