Qwen-Image-2.0-RL Technical Report

Yixian Xu*, Kaiyuan Gao*, Yuxiang Chen*, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, Jiahao Li, Jie Zhang, Liang Peng, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, Yan Shu, Yanran Zhang, Yi Wang, Yu Wu, Yujia Wu, Zekai Zhang, Zhendong Wang, Xiao Xu, Kun Yan, Chenfei Wu†

https://qwen.ai

Abstract

We present **Qwen-Image-2.0-RL**, a post-training pipeline that applies reinforcement learning from human feedback (RLHF) and on-policy distillation (OPD) to improve both the visual quality and instruction-following capability of the Qwen-Image-2.0 diffusion model. To provide reliable reward signals, we construct task-specific composite reward models by fine-tuning vision-language models with a pointwise scoring paradigm and chain-of-thought reasoning. For text-to-image generation, the reward models cover alignment, aesthetics, and portrait fidelity dimensions. For image editing tasks, the reward system addresses instruction-following accuracy and face identity preservation. Building on this reward system, we develop a scalable GRPO-based RL training framework, incorporating a hybrid classifier-free guidance (CFG) strategy to preserve pre-trained knowledge, prompt curation via intra-group reward range filtering, and per-category reward weight calibration. To merge the task-specialized RL policies for T2I and editing, we propose on-policy distillation as the final training stage, which consolidates multiple teachers into a single student model through trajectory-level velocity matching. Extensive evaluation shows that Qwen-Image-2.0-RL achieves 57.84 overall score on Qwen-Image-Bench (+2.61 over the base model), Elo ratings of 1193 in text-to-image arena (+78) and 1349 in image edit arena (+93), demonstrating consistent gains in aesthetic quality, prompt adherence, and editing accuracy.

1 Introduction

Diffusion and flow-based generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) have achieved remarkable success in high-fidelity image generation. The field has progressed to latent diffusion models (Rombach et al., 2022; Podell et al., 2024) with scalable Transformer-based architectures (Peebles &Xie, 2023; Chen et al., 2024; Esser et al., 2024; Ma et al., 2024). More recent systems (BlackForest, 2024; Labs, 2025; Wu et al., 2025a; Cai et al., 2025; Cao et al., 2025) have further adopted vision-language foundation models as conditional encoders, whose stronger semantic grounding and multimodal world knowledge enable more precise instruction following and text-image alignment. Meanwhile, commercial systems (Gao et al., 2025; Seedream et al., 2025; OpenAI, 2025; Google, 2025) have pushed the frontier of generation quality, and unified architectures (Wu et al., 2025a; Labs et al., 2025) have extended these capabilities to image editing, enabling a single model to serve both generation and editing tasks within a shared framework.

Despite impressive pre-training results, a persistent gap remains between the outputs of supervised-trained diffusion models and human aesthetic expectations. Supervised training optimizes the denoising score matching objective that does not directly capture the human preference such as compositional harmony, texture, prompt faithfulness, and stylistic coherence. Reinforcement learning from human feedback (RLHF), which has achieved remarkable success in aligning large language models (Shao et al., 2024), offers a principled approach to close this gap. Based on reward signals that encode human preferences, the RLHF paradigm directly optimizes the model with respect to the reward signals.

However, extending the RLHF paradigm to diffusion models introduces distinct challenges. First, reliable reward signals must capture diverse quality dimensions across fundamentally different tasks such as text-to-image (T2I) generation and image editing, spanning global aesthetics and prompt adherence for T2I, and fine-grained identity preservation for editing. This necessitates a composite, task-aware reward design. Second, existing RL frameworks for diffusion models have primarily been validated under LoRA fine-tuning settings (Liu et al., 2026a; Zheng et al., 2025; Wang et al., 2025a). Real-world scenarios involving multiple reward signals, diverse task types, and full-parameter training at scale

*Equal contribution.
†Corresponding author.

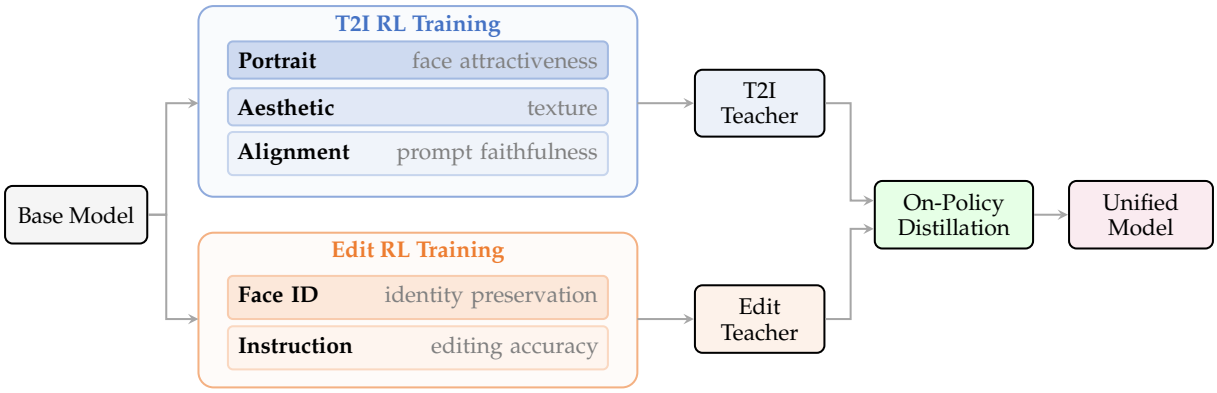


Figure 1: Overview of the Qwen-Image-2.0-RL training pipeline. Starting from a shared base model, we train two task-specialized RL policies with dedicated reward compositions: T2I generation uses a layered reward design progressing from prompt faithfulness to texture quality to portrait-specific optimization, while editing focuses on instruction accuracy and identity preservation. The resulting teachers are merged into a unified model via on-policy distillation.

remain underexplored. Third, practical deployment demands consolidating task-specialized RL policies into a single model without sacrificing per-task quality.

In this work, we address these challenges through a unified post-training pipeline (Fig. 1) applied to the Qwen-Image-2.0 (Zhao et al., 2026) foundation model. Our contributions are summarized as follows:

1. **VLM-based composite reward models (Sec. 3).** We construct task-specific reward suites by fine-tuning Qwen series VLM with chain-of-thought reasoning enabled, adopting a pointwise scoring paradigm that we find empirically superior to pairwise training. For T2I tasks, the rewards follow a layered design: an alignment reward ensures prompt faithfulness, an aesthetic reward improves texture and composition, and a portrait reward targets facial attractiveness. For image editing tasks, we combine an instruction-following reward with a dedicated face identity consistency scorer to capture global structural preservation and subtle facial identity shifts.
2. **Scalable RL training framework (Sec. 4).** We adopt a GRPO-based RL framework with multi-reward advantage computation (Shao et al., 2024; Liu et al., 2026b). In addition, we introduce a hybrid CFG strategy that applies guidance during rollout sampling but excludes it from the policy optimization objective, balancing training stability with preservation of pre-trained knowledge. We further propose a strategy to filter training prompts and per-category reward weight adjustment to balance optimization across various visual domains.
3. **On-policy distillation (Sec. 4.3).** Once the RL policies are trained for specific tasks including T2I and image editing, we propose on-policy distillation (OPD) to unify task-specialized RL teachers into a single student model via trajectory-level velocity matching. OPD avoids cross-task optimization conflicts and eliminates reward model dependency.

The resulting model, **Qwen-Image-2.0-RL**, achieves strong performance across multiple evaluation settings (Sec. 5): a 57.84 overall score on Qwen-Image-Bench (+2.61 over the base model), and Elo ratings of 1193 in text-to-image arena (+78) and 1349 in image edit arena (+93), demonstrating consistent improvements in aesthetic quality, prompt adherence, and editing accuracy for both T2I and image editing tasks.

2 Backgrounds

2.1 Diffusion Models

Diffusion models have become the dominant paradigm for high-fidelity image generation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Liu et al., 2022; Lipman et al., 2022). To model the high-dimensional data distribution p_{data}, diffusion models define a forward noising process that progressively corrupts data samples and learn a reverse process for generation. Following the Flow Matching framework (Lipman et al., 2022; Liu et al., 2022), the forward path is defined by

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}, \quad t \in [0,1], \tag{1}$$

where $\boldsymbol{x}_0 \sim p_{\text{data}}$ is a clean data sample and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is standard Gaussian noise. The time derivative of this interpolation defines the conditional velocity field $\boldsymbol{v} := \boldsymbol{\epsilon} - \boldsymbol{x}_0$. A neural network $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c)$ is trained to approximate this velocity field by minimizing the flow matching objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim p_t, \boldsymbol{x}_0\sim p_{\text{data}}, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})} \left\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0)\right\|^2, \tag{2}$$

where p_t is the distribution of the training timestep. Once trained, samples are generated by solving the probability flow ordinary differential equation (ODE) backward from $t = 1$ to $t = 0$:

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c), \quad \boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{3}$$

2.2 Reinforcement Learning for Diffusion Models

Recent work has explored extending reinforcement learning to flow matching models for aligning generation quality with human preferences. Flow-GRPO (Liu et al., 2026a) extends Group Relative Policy Optimization (GRPO, Shao et al. 2024) to flow matching models by formulating the multi-step denoising trajectory as a Markov decision process (MDP). Given a prompt c, the policy π_θ generates a group of G images $\{\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(G)}\}$, and a reward model $R(\boldsymbol{x}_0, c)$ evaluates each sample. The advantage of the i-th sample is computed by group-level normalization:

$$A(\boldsymbol{x}_0^{(i)}, c) = \frac{R(\boldsymbol{x}_0^{(i)}, c) - \mu_c}{\sigma_c}, \tag{4}$$

where μ_c and σ_c are the mean and standard deviation of rewards within the group. A central challenge is that flow matching relies on a deterministic ODE for generation, which hinders the direct application of GRPO. Flow-GRPO addresses this by using an equivalent stochastic sampler:

$$\mathrm{d}\boldsymbol{x}_t = \left[\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) + \frac{\sigma_t^2}{2t}\left(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)\right)\right]\mathrm{d}t + \sigma_t \mathrm{d}\boldsymbol{w}_t, \tag{5}$$

where $\sigma_t \geq 0$ controls the noise scale and $\boldsymbol{w}_t$ denotes the standard Wiener process. Under Euler-Maruyama discretization, the transition density becomes Gaussian, enabling tractable computation of the importance sampling ratio $r_t^{(i)}(\theta) = \pi_\theta(\boldsymbol{x}_{t-1}^{(i)}|\boldsymbol{x}_t^{(i)}, c) / \pi_{\theta_{\text{old}}}(\boldsymbol{x}_{t-1}^{(i)}|\boldsymbol{x}_t^{(i)}, c)$. The policy is then optimized via a clipped surrogate objective:

$$\mathcal{L}_{\text{Flow-GRPO}}(\theta) = -\mathbb{E}_{c\sim\mathcal{D}, \boldsymbol{x}_{0:T}\sim\pi_{\theta_{\text{old}}}(\cdot|c)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\left[\min\left(r_t^{(i)}(\theta)\, A(\boldsymbol{x}_0^{(i)}, c),\ \hat{r}_t^{(i)}(\theta) A(\boldsymbol{x}_0^{(i)}, c)\right)\right], \tag{6}$$

where $\hat{r}_t^{(i)}(\theta) := \text{clip}\left(r_t^{(i)}(\theta), 1-\epsilon, 1+\epsilon\right)$ is the clipped version of importance sampling ratio.

DiffusionNFT (Zheng et al., 2025) proposes an alternative formulation that uses the forward diffusion process for policy optimization. For a noisy state $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}$, three velocity predictions are computed: the current policy $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c)$, the old policy $\boldsymbol{v}_{\theta_{\text{old}}}(\boldsymbol{x}_t, t, c)$, and the reference policy $\boldsymbol{v}_{\theta_{\text{ref}}}(\boldsymbol{x}_t, t, c)$. The method constructs positive and negative velocity predictions defined by

$$\boldsymbol{v}_\theta^+ = \beta \cdot \boldsymbol{v}_\theta + (1-\beta)\cdot \boldsymbol{v}_{\theta_{\text{old}}}, \quad \boldsymbol{v}_\theta^- = (1+\beta)\cdot \boldsymbol{v}_{\theta_{\text{old}}} - \beta \cdot \boldsymbol{v}_\theta, \tag{7}$$

where β is the interpolation strength. Then the training objective of DiffusionNFT is given by

$$\mathcal{L}_{\text{NFT}} = \mathbb{E}_{c\sim\mathcal{D}, \boldsymbol{x}_0\sim\pi_{\theta_{\text{old}}}(\cdot|c), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[r(\boldsymbol{x}_0, c)\|\boldsymbol{v}_\theta^+(\boldsymbol{x}_t, t, c) - \boldsymbol{v}\|^2 + (1 - r(\boldsymbol{x}_0, c))\|\boldsymbol{v}_\theta^-(\boldsymbol{x}_t, t, c) - \boldsymbol{v}\|^2\right], \tag{8}$$

where $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}$, $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$, $r = \frac{\text{clip}(A, -A_{\max}, A_{\max})}{2A_{\max}} + 0.5 \in [0, 1]$ is a rescaled version of the group-relative advantage. To prevent the policy from deviating too far from the pre-trained reference, a KL penalty is added:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{c\sim\mathcal{D}, \boldsymbol{x}_0\sim\pi_{\theta_{\text{old}}}(\cdot|c), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c) - \boldsymbol{v}_{\theta_{\text{ref}}}(\boldsymbol{x}_t, t, c)\right\|^2. \tag{9}$$

3 Reward Modeling

The reward signal capturing human preference is the primary component of the RL training. To align the pretrained image generative model with the human preferences, we construct task-specific composite reward models for different evaluation dimensions. Our reward system combines VLM-based scorers for semantic and aesthetic assessment with model-based scorers for fine-grained identity preservation, tailored to both T2I and image editing tasks.

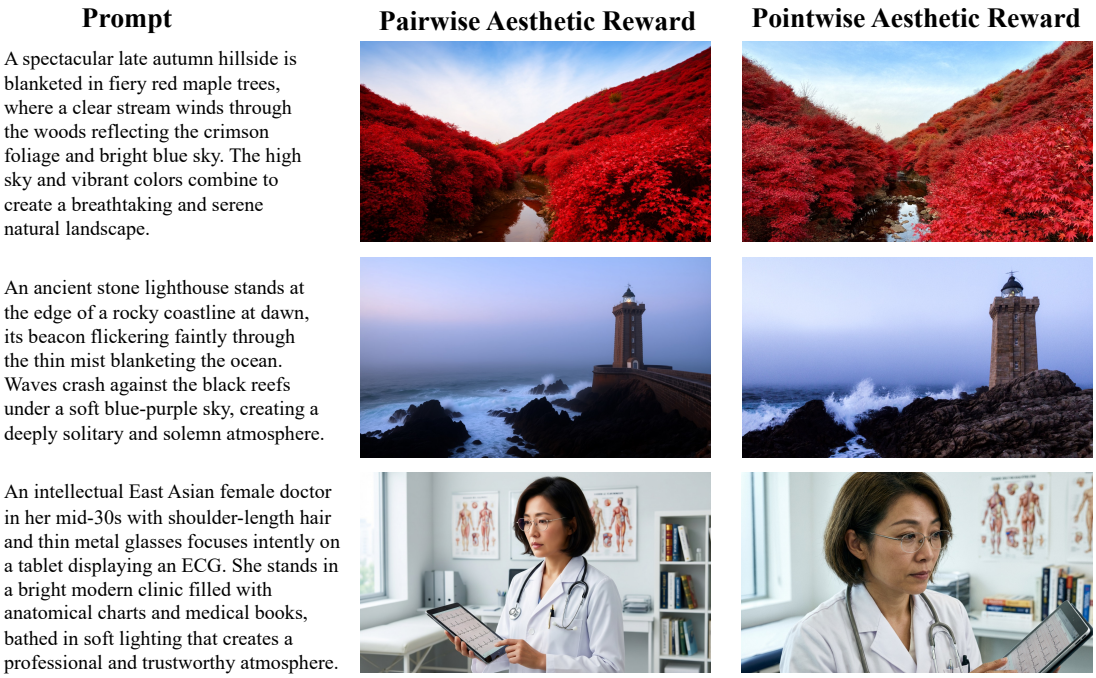


Figure 2: Qualitative comparison of RL training outcomes using pointwise vs. pairwise reward model training paradigms. Both reward models target the same evaluation dimensions (aesthetic quality and visual texture) and are trained on data from the same model pool, isolating the effect of the training paradigm. The pointwise-trained reward model produces images with consistently better visual quality, finer texture detail, and fewer artifacts.

3.1 Reward Model Training Paradigms

We explore two training paradigms for reward model fine-tuning. The first is **pairwise reward training**, where the model is optimized on pairs of images $(\boldsymbol{x}_w, \boldsymbol{x}_l)$ generated from the same prompt c, with $\boldsymbol{x}_w$ preferred by human annotators over $\boldsymbol{x}_l$. The Vision-Language Model (VLM) produces scalar scores $R_\phi(x, c)$ for each image, and the training objective minimizes the Bradley-Terry ranking loss:

$$\mathcal{L}_{\text{pair}} = -\sum_{(\boldsymbol{x}_w, \boldsymbol{x}_l, c)} \log \sigma\big(R_\phi(\boldsymbol{x}_w, c) - R_\phi(\boldsymbol{x}_l, c)\big)\,, \tag{10}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. The second is **pointwise reward training**, where each image $\boldsymbol{x}$ is paired with an absolute human-annotated score $y \in \mathbb{R}$, and the model is trained to directly regress to this score:

$$\mathcal{L}_{\text{point}} = \sum_{(\boldsymbol{x}, c, y)} \big(R_\phi(\boldsymbol{x}, c) - y\big)^2\,. \tag{11}$$

In practice, the reward model is trained to output tokens in discrete score set $\mathcal{S} = \{1, 2, 3, 4, 5\}$, and the reward score is produced by the expectation under the VLM's probability distribution p_ϕ:

$$R_\phi(\boldsymbol{x}, c) = \sum_{s \in \mathcal{S}} s \cdot p_\phi(s | \boldsymbol{x}, c). \tag{12}$$

To compare the two training paradigms, we construct two annotated datasets that deliberately share the same evaluation focus—aesthetic quality and visual texture—while differing only in annotation format. Both datasets draw images from a common pool of state-of-the-art AIGC models, ensuring that the comparison isolates the effect of the training paradigm rather than data distribution.

Pointwise annotation. We collect images datasets with prompts randomly sampled from high-quality portrait reference images and rewritten to diversify linguistic patterns beyond the training distribution. Each image is independently scored by human annotators on a 5-point Likert scale along two structured dimensions: (1) Quality—assessing clarity, lighting, color balance, stylistic coherence, and material texture; and (2) Fidelity—evaluating structural correctness, physical consistency, and absence of AI artifacts such as unnatural smoothing or texture repetition.

Pairwise annotation. We collect image pairs dataset, where two images generated from the same prompt are presented side by side for preference judgment. Annotation follows a strict priority hierarchy: image-text consistency $>$ structural distortion $>$ texture quality $>$ aesthetic appeal. Under this scheme, when both images faithfully depict the prompt content and are free of structural distortions, which is the common case for images from high-quality models, the comparison reduces to a direct judgment of texture quality and aesthetic appeal. Each pair is further labeled with auxiliary attributes including sample validity, text distortion presence, and human figure distortion presence to support fine-grained analysis.

Comparison. We train two reward models by finetuning the same VLM architectures and use each as the RL training reward. As shown in Fig. 2, the pointwise-trained reward model produces images with consistently better visual quality and fewer artifacts. We attribute this to the richer supervisory signal of absolute scores: pointwise annotations encode how good an image is on a calibrated scale, whereas pairwise annotations only capture which is better. Based on this finding, we adopt the pointwise paradigm as the default training objective for all VLM-based reward models in the final system.

3.2 Reward Models for Text-to-Image Generation

Having established the pointwise paradigm as our default training objective (Sec. 3.1), we now describe the specific reward models for T2I generation. All VLM-based T2I reward models are implemented by fine-tuning Qwen series VLM. Our T2I reward model design follows a layered logic. The most fundamental requirement for image generation is faithfulness to the user's prompt: a visually stunning image that ignores the specified content is a failed generation. We therefore begin with an image-text alignment reward that evaluates semantic correspondence without considering aesthetics. Once prompt adherence is established, we layer on an aesthetic reward to enrich texture fidelity. Finally, because human-subject images demand more than generic aesthetic quality, we introduce a dedicated portrait reward that specifically optimizes facial attractiveness and fine-grained skin and hair realism.

Image-text alignment reward. This is the most fundamental reward, measuring the semantic correspondence between the generated image and the input prompt. It explicitly penalizes outputs that omit, misinterpret, or contradict user-specified requirements, without considering aesthetic merit. The VLM evaluates prompt adherence along a priority hierarchy: (1) object presence and count accuracy, (2) attribute correctness (color, size, shape, material), (3) spatial relationship fidelity, and (4) action and pose accuracy. Images that fail the highest-priority criteria are capped at low scores regardless of other qualities.

Aesthetic reward. Building upon prompt-faithful generation, this reward assesses the intrinsic visual quality of generated images, emphasizing compositional balance, realistic illumination, texture fidelity, and overall artistic coherence. The aesthetic reward model is trained on the pointwise-annotated dataset described in Sec. 3.1.

Portrait reward. Generic aesthetic optimization is insufficient for human-subject generation, where facial attractiveness and anatomical correctness are critical. This reward provides a specialized signal for improving facial proportion accuracy, identity-preserving facial details, and fine-grained skin and hair texture realism. The scorer explicitly checks for common failure modes such as incorrect finger counts, distorted facial features, and unnatural body proportions. We train this reward model on a separate portrait-specific dataset with prompts sampled from high-quality portrait reference images. The annotation rubric focuses specifically on facial attractiveness, skin and hair texture realism, capturing the standards of human beauty that generic aesthetic scoring cannot adequately address.

3.3 Reward Models for Text-Guided Image Editing

Building on the success of the pointwise reward paradigm established for T2I, we transfer the same training methodology to image editing tasks. The VLM-based rewards are likewise built by fine-tuning Qwen series VLM, adapting the evaluation rubrics to editing-specific requirements. We additionally introduce a model-based face identity consistency scorer to address fine-grained identity preservation beyond the VLM's capabilities.

Instruction-following reward. This reward evaluates whether user-specified modifications are accurately executed, covering editing operations such as object replacement, attribute modification, and style transfer. The fine-tuned VLM model receives the source image, the editing instruction, and the output image as a triplet, and is prompted to decompose the instruction into core editing requirements and

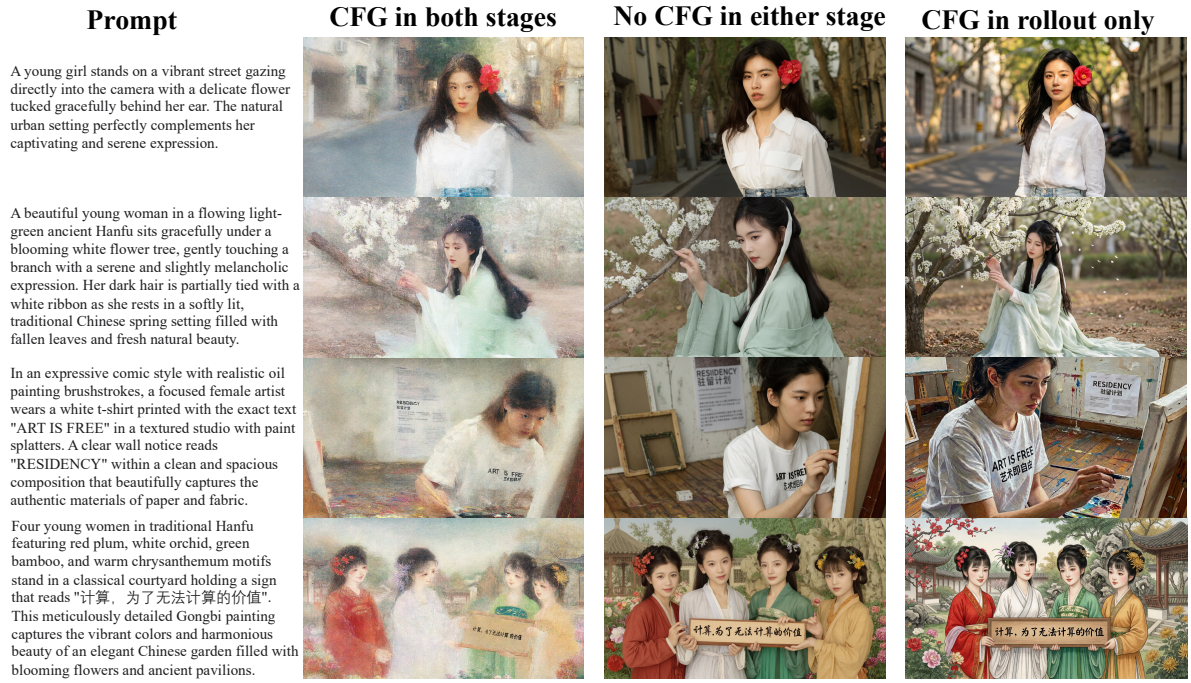


Figure 3: Qualitative comparison of three CFG strategies during RL training. **Left**: CFG applied in both rollout and training leads to training instability and eventual image collapse. **Middle**: removing CFG from both stages causes progressive loss of stylization ability and world knowledge. **Right**: CFG in rollout only (our hybrid strategy) maintains training stability while preserving the pre-trained model's full generative capabilities.

non-core auxiliary requirements. The evaluation follows a structured rubric that assesses: (1) whether the core editing instruction has been fulfilled, (2) whether non-core requirements are addressed, and (3) whether the overall output is visually coherent.

Face identity consistency reward. While the VLM-based visual consistency reward captures global structural preservation, we find that it is insufficient for reliably detecting subtle facial identity shifts. We therefore introduce a dedicated model-based face identity scorer. This model-based reward provides a precise, embedding-level identity preservation signal that complements the VLM's higher-level semantic consistency assessment.

4 Training

With the reward models established in Sec. 3 and the GRPO-based RL framework introduced in Sec. 2.2, we now describe our training pipeline. We first train separate RL policies for T2I generation and image editing using their respective reward compositions (Secs. 3.2 and 3.3), then merge the resulting task-specialized models into a deployable model via on-policy distillation. In Sec. 4.1, we present the shared pipeline infrastructure for T2I and editing task. Sec. 4.2 details the task-specific optimization choices. Finally, Sec. 4.3 introduces on-policy distillation, which unifies the two task-specialized teachers into a single student model through trajectory-level velocity matching.

4.1 Training Pipeline

Hybrid CFG strategy. A key design consideration in diffusion-based RL is whether classifier-free guidance (CFG) (Ho &Salimans, 2022) should be employed during rollout sampling and policy optimization. We systematically evaluate three strategies (see Fig. 3):

- CFG in both rollout and training. Applying CFG during both rollout sampling and the training stage leads to severe training instability. As training progresses, the generated images deteriorate completely, ultimately collapsing into incoherent outputs;
- No CFG in either stage. Although reward scores steadily improve, the model progressively loses stylization ability and world knowledge, failing to reproduce well-known celebrity appearances

and losing the capacity for style-specific generation. We attribute this phenomenon to the base model's reliance on CFG to fully express its pre-trained knowledge during inference;

- CFG in rollout only. CFG is used during the rollout stage to generate high-quality candidates for reward evaluation, while the unconditional branch is excluded from the policy optimization objective.

We adopt the third, hybrid strategy. The CFG-guided rollout fully leverages the pre-trained model's capabilities to produce structurally coherent images that yield reliable reward signals. Meanwhile, the CFG-free training objective avoids the optimization difficulties introduced by jointly optimizing the conditional and unconditional branches, maintaining stable gradient updates and substantially reducing computational overhead.

Asynchronous reward pipeline. Our reward models are deployed as remote API services, separate from the training process. Since reward scoring involves network I/O to these remote VLM endpoints, synchronous evaluation would bottleneck the training loop. We therefore decouple reward computation from model training through an asynchronous pipeline: after the policy generates a batch of images via GPU inference and gathers them across ranks, a background thread asynchronously submits the images to the remote reward API endpoints. Once the asynchronous reward responses return, all ranks synchronize to gather the raw scores, perform per-prompt-group normalization, and compute advantages for the policy gradient update. This design hides nearly all reward latency behind inference computation, enabling efficient scaling to multiple reward models without proportional increases in training time.

Multi-reward advantage computation. As mentioned in Sec. 3, we use multiple reward models for the training of T2I and image editing task respectively. Inspired by Liu et al. (2026b), the group-relative advantage in Eqn. (4) is calculated by weighted summation with per-prompt-group normalization:

$$A(\boldsymbol{x}_0^{(i)}, c) = \sum_{k=1}^{K} w_k \cdot \frac{R_k(\boldsymbol{x}_0^{(i)}, c) - \mu_k}{\sigma_k}, \tag{13}$$

where R_k is the k-th reward model, w_k is its weight satisfying $\sum_{k=1}^{K} w_k = 1$, and μ_k, σ_k are the mean and standard deviation of R_k computed within each prompt group. This per-prompt-group normalization is critical: it ensures that the composite reward is invariant to absolute scale differences across reward models, preventing any single reward dimension from dominating the advantage signal due to its numerical range.

4.2 Task-Specific Optimization

Timestep sampling. During rollout, images are generated using a 40-step ODE solver. A naive approach would apply the RL training objective at all 40 timesteps. However, we observe that this leads to rapid reward hacking, resulting in degradation within a few iterations. To address this issue, we train on only a subset of the rollout timesteps, with a particular emphasis on high-noise timesteps (i.e., those closer to $t = 1$). High-noise timesteps govern global structure and semantic layout, making them more robust targets for policy optimization. By restricting the training signal to a carefully selected subset, we slow the reward exploitation process and ensure that the model improves comprehensively across quality dimensions.

Prompt curation. Not all prompts contribute equally to policy improvement. We employ the trained reward models to filter the prompt pool before RL training. For each candidate prompt, the base model performs G rollouts and the composite reward is computed for each sample. We then compute the intra-group range (maximum minus minimum reward) within each prompt group. Only prompts whose range exceeds a predefined threshold are retained for training. Prompts with uniformly high or low rewards across all samples provide a weak signal for policy optimization. This filtering step significantly improves training efficiency by concentrating compute on prompts where the policy has room for meaningful improvement.

Per-category reward calibration. We organize the retained prompts into semantic categories (e.g., portrait, landscape, typography, general) and assign category-specific reward weight vectors. For instance, portrait prompts receive higher weight on the portrait fidelity reward, while typography prompts emphasize alignment accuracy. This per-category calibration ensures that the RL objective reflects the distinct quality requirements of each visual domain, preventing the optimization from converging to a single dominant style at the expense of others.

4.3 On-Policy Distillation

While the preceding RL optimization produces models with superior quality on each task, the resulting policies are task-specialized: a T2I-optimized model may exhibit degraded editing performance, and vice versa. To address this limitation, we propose On-Policy Distillation (OPD), which unifies multiple task-specialized RL-trained teachers into a single student model via trajectory-level velocity matching.

Training objective. The student model v_θ (initialized from the pre-trained base model) generates images by solving the reverse ODE from $t = 1$ to $t = 0$ with N discrete steps, using a timestep schedule $\{t_0 = 1, t_1, \ldots, t_N \approx 0\}$. The full trajectory of the student model $\{\boldsymbol{x}_{t_0}, \boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_N}\}$ is saved, where $\boldsymbol{x}_{t_0} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the initial noise sample and $\boldsymbol{x}_{t_N}$ is the denoised output. The student is trained to match the task-appropriate teacher's velocity at each point along its own trajectory:

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{c, \boldsymbol{x}_{[1:N]} \sim \pi_\theta(\cdot|c)} \left[\sum_{n=1}^{N} \left\| \boldsymbol{v}_\theta(\boldsymbol{x}_{t_n}, t_n, c) - \boldsymbol{v}_{\theta^*}(\boldsymbol{x}_{t_n}, t_n, c) \right\|^2 \right], \tag{14}$$

where v_{θ^*} is the task-related teacher model (see Sec. A for a formal derivation). The OPD objective ensures that the student learns to correct its own prediction errors on its own inference trajectories.

Multi-Teacher Distillation. A central advantage of OPD is its ability to distill from multiple task-specialized teachers into a single student. We maintain two teacher models: a T2I teacher optimized for text-to-image generation with aesthetic, alignment, and portrait rewards, and an editing teacher optimized for image editing tasks with instruction-following and face identity preservation rewards. For each training batch, the appropriate teacher is selected based on the task type of the current sample. To manage GPU memory, only the active teacher is loaded onto the GPU at any time; inactive teachers are offloaded to CPU. This dynamic teacher activation mechanism allows training with multiple large teacher models without proportionally increasing GPU memory requirements. As teacher models are originally trained with CFG, we apply CFG during the teacher's velocity prediction in OPD, but keep the student model without CFG. By distilling from specialized teachers rather than jointly training with competing rewards, OPD avoids the optimization conflicts that arise when T2I and editing objectives are optimized simultaneously. In addition, the CFG is also integrated into the student model after OPD.

Comparison with mixed RL training. A natural alternative to the decomposed OPD pipeline is to train a single model with RL on mixed T2I and editing data (Mix-RL), where all task-specific rewards are jointly optimized in one training process. While this approach is simpler, it forces the model to simultaneously satisfy competing optimization objectives from different tasks, leading to suboptimal trade-offs. Fig. 5 presents a three-way qualitative comparison across T2I scenarios among the pre-trained base model (Qwen-Image-2.0-Base), the Mix-RL baseline, and our final Qwen-Image-2.0-RL model produced via OPD. The comparison reveals a clear quality progression: Mix-RL already improves over the base model in texture fidelity, compositional coherence, and overall realism, confirming that RL training with our reward suite effectively enhances generation quality. However, Qwen-Image-2.0-RL consistently outperforms Mix-RL, producing sharper details, more accurate prompt adherence, and better aesthetic quality. A similar trend is observed for image editing: Fig. 6 shows that Qwen-Image-2.0-RL achieves superior face identity preservation and instruction-following accuracy compared to both the base model and Mix-RL, where the latter still suffers from identity drift or incomplete edits under complex instructions. This demonstrates the advantage of our decomposed strategy over jointly optimizing all task rewards in a single RL training process.

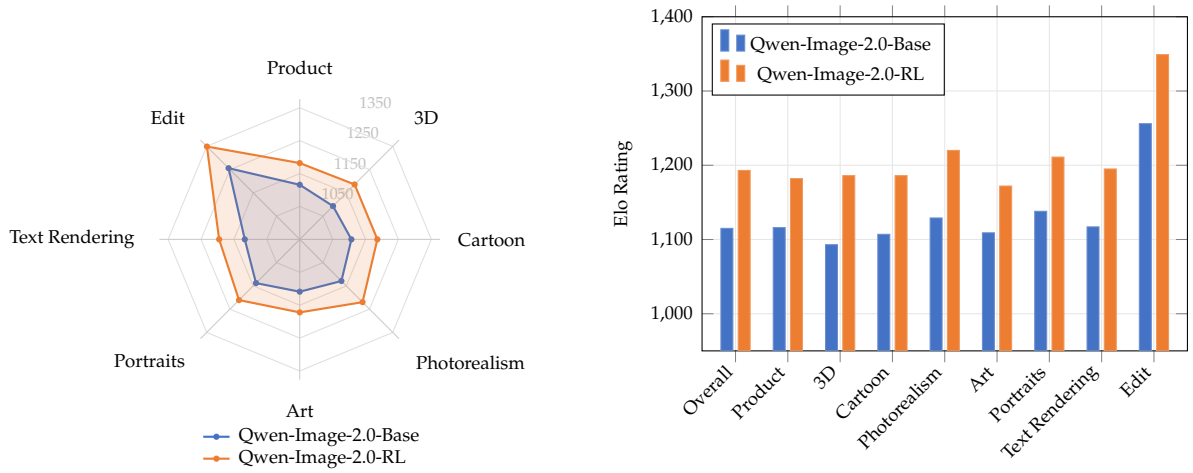
5 Evaluation

We evaluate Qwen-Image-2.0-RL from automated quality metrics on standardized benchmarks to human preference rankings on competitive arenas.

Text-to-image generation results. We assess the effectiveness of our RL training framework on T2I using Qwen-Image-Bench (Li et al., 2026a), a creator-centric benchmark designed to evaluate T2I models across five first-level pillars: Quality, Aesthetics, Alignment, Real-world Fidelity, and Creative Generation. Evaluation is conducted by Q-Judger, a unified judge model trained on over 130K human-labeled image-prompt pairs annotated by 80 professional artists. Tab. 1 presents the performance of Qwen-Image-2.0 before and after RL training alongside strong baselines on Qwen-Image-Bench. RL training yields consistent improvements across all five evaluation pillars, raising the overall score from 55.23 to 57.84. The most substantial gains emerge in Creative Generation (6.72 improvement) and Real-world Fidelity (4.29 improvement).

Table 1: Performance comparison on Qwen-Image-Bench (Li et al., 2026a). Scores are on a $[0, 100]$ scale, aggregated bottom-up from 56 third-level facets through a three-level taxonomy. Baseline models are sorted by overall score in ascending order.

Model	Quality	Aesthetics	Alignment	Real-world Fidelity	Creative Gen.	Overall
GLM Image	49.26	50.64	47.90	44.69	45.23	48.19
Kling Image 2.1	49.11	50.15	49.18	44.74	44.67	48.26
Qwen Image	48.44	52.25	50.72	43.16	47.30	49.23
Imagen 4.0	50.16	52.68	51.64	44.84	47.94	50.29
HunyuanImage 3.0	50.35	53.57	52.00	44.31	49.12	50.81
Imagen 4.0 Ultra	50.90	54.25	54.02	45.59	51.14	51.99
Qwen Image 2512	51.76	54.74	52.72	47.00	50.19	52.06
GPT Image 1	52.34	55.09	56.28	48.14	55.78	54.07
FLUX 2 Pro	52.30	56.94	57.01	47.29	56.18	54.57
FLUX 2 Max	53.64	56.85	57.35	49.35	56.50	55.33
Seedream 4.0	54.01	58.81	56.64	51.05	58.15	56.21
Seedream 4.5	54.41	58.72	57.31	51.69	60.64	56.78
Seedream 5.0	52.55	58.40	58.90	51.92	65.29	57.22
Nano Banana Pro	55.67	60.26	61.25	54.07	66.23	59.45
GPT Image 1.5	55.14	60.88	61.72	53.95	66.35	59.65
Nano Banana 2.0	54.77	61.08	62.40	54.28	67.05	59.82
GPT Image 2	**58.65**	**67.53**	**65.85**	**57.38**	**75.23**	**64.69**
Qwen-Image-2.0-Base	52.29	57.10	57.64	47.54	58.22	55.23
Qwen-Image-2.0-RL	**54.39**	**58.67**	**59.28**	**51.83**	**64.94**	**57.84**



(a) Elo rating profile across categories.

(b) Grouped Elo rating comparison.

Figure 4: Human preference evaluation on arena. RL training consistently improves Elo ratings across all eight sub-categories and the overall score.

Human preference evaluation. Beyond automated benchmarks, we evaluate Qwen-Image-2.0-RL on text-to-image and image edit arena, where users vote between anonymized image pairs from competing models. Fig. 4 presents the Elo ratings of Qwen-Image-2.0 before and after RL training across eight sub-categories. RL training yields substantial Elo rating improvements across all dimensions, with the overall T2I rating rising from 1115 to 1193 (+78). The most pronounced gains appear in 3D Modeling (+93), followed by Photorealism (+91), reflecting improved structural consistency and fine-grained detail rendering. Consistent with the T2I evaluation above, the performance in image edit arena also improves from 1256 to 1349 (+93), further corroborating the effectiveness of our editing-specific RL training.

6 Related Works

Reinforcement learning for diffusion models. Aligning diffusion models with human preferences via RL has gained significant attention. Flow-GRPO (Liu et al., 2026a) extends GRPO to flow matching models by treating the multi-step generation as a Markov decision process. GRPO-Guard (Wang et al., 2025a) identifies implicit over-optimization issues in flow matching RL and proposes regulated clipping to mitigate them. AWM (Xue et al., 2025) and DiffusionNFT (Zheng et al., 2025) introduces online

reinforcement learning with a forward process formulation. Our work builds upon these methods, introduce a hybrid CFG strategy specifically designed for the Qwen-Image-2.0 architecture.

Reward models for diffusion models. Reward modeling is central to aligning diffusion models with human preferences. Early CLIP-based approaches (Radford et al., 2021) are limited by fixed architectures and poor cross-task generalization. Subsequent regression-based methods ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), and HPSv3 (Ma et al., 2025) train scalar predictors via Bradley–Terry ranking loss. More recently, UnifiedReward (Wang et al., 2025b; 2026) and RewardDance (Wu et al., 2025b) reformulate scoring as token prediction probability, enabling scaling along both model size and context dimensions including task-specific instructions and chain-of-thought (CoT) reasoning.

On-policy distillation. On-policy distillation (OPD) consolidates heterogeneous capabilities by having students learn on self-generated trajectories under teacher supervision. GKD (Agarwal et al., 2024) established this framework for LLMs, and frontier models have since adopted multi-teacher OPD to avoid the seesaw effect of multi-reward RL. Two concurrent works Flow-OPD (Fang et al., 2026) and DiffusionOPD (Li et al., 2026b) extend OPD to flow matching models by showing that the KL divergence between Gaussian transition kernels reduces to a velocity-field MSE loss. Both of these works focus on consolidating single-reward T2I teachers. Our approach differs by specializing teachers by task type (T2I and editing), deriving the objective from a W_2 upper bound.

7 Conclusion

We presented Qwen-Image-2.0-RL, an image generation system that combines RLHF with OPD to substantially improve visual quality and instruction-following capabilities. Our approach makes three contributions: (1) a VLM-based composite reward system tailored to both T2I and TI2I tasks, with structured evaluation rubrics covering aesthetic quality, prompt adherence, portrait fidelity, instruction-following, and visual consistency; (2) an adapted GRPO training framework featuring a hybrid CFG strategy and asynchronous reward pipeline that enables efficient large-scale RL training of flow matching models; and (3) an OPD mechanism that unifies task-specialized RL-trained teachers into a single deployment model via trajectory-level velocity matching, eliminating reward model dependency while preserving the quality gains of each specialized teacher. Qualitative and quantitative evaluations confirm that OPD not only matches but surpasses a mixed RL baseline that jointly optimizes all task rewards, validating the decomposed training strategy. The resulting system significantly improves performance across human preference benchmarks for both image generation and editing tasks.

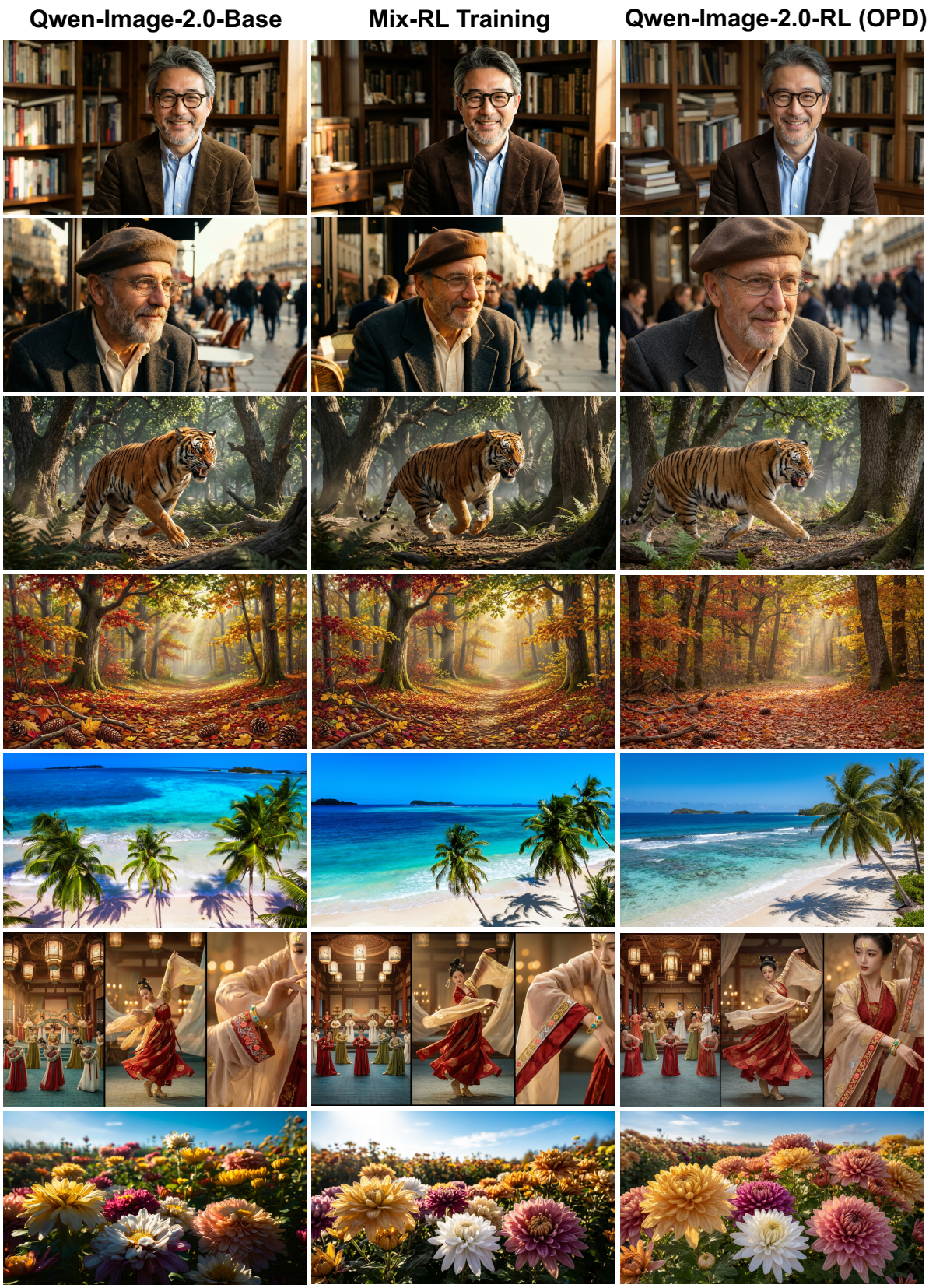


Figure 5: Qualitative comparison across T2I generation scenarios among three model variants: pre-trained Qwen-Image-2.0-Base, Mix-RL (jointly trained on T2I and editing tasks with mixed RL rewards), and Qwen-Image-2.0-RL (task-specialized RL teachers distilled via on-policy distillation). The progression Qwen-Image-2.0-Base → Mix-RL → Qwen-Image-2.0-RL demonstrates that RL training improves visual quality over the pre-trained baseline, and that OPD further surpasses mixed RL by avoiding cross-task optimization conflicts.

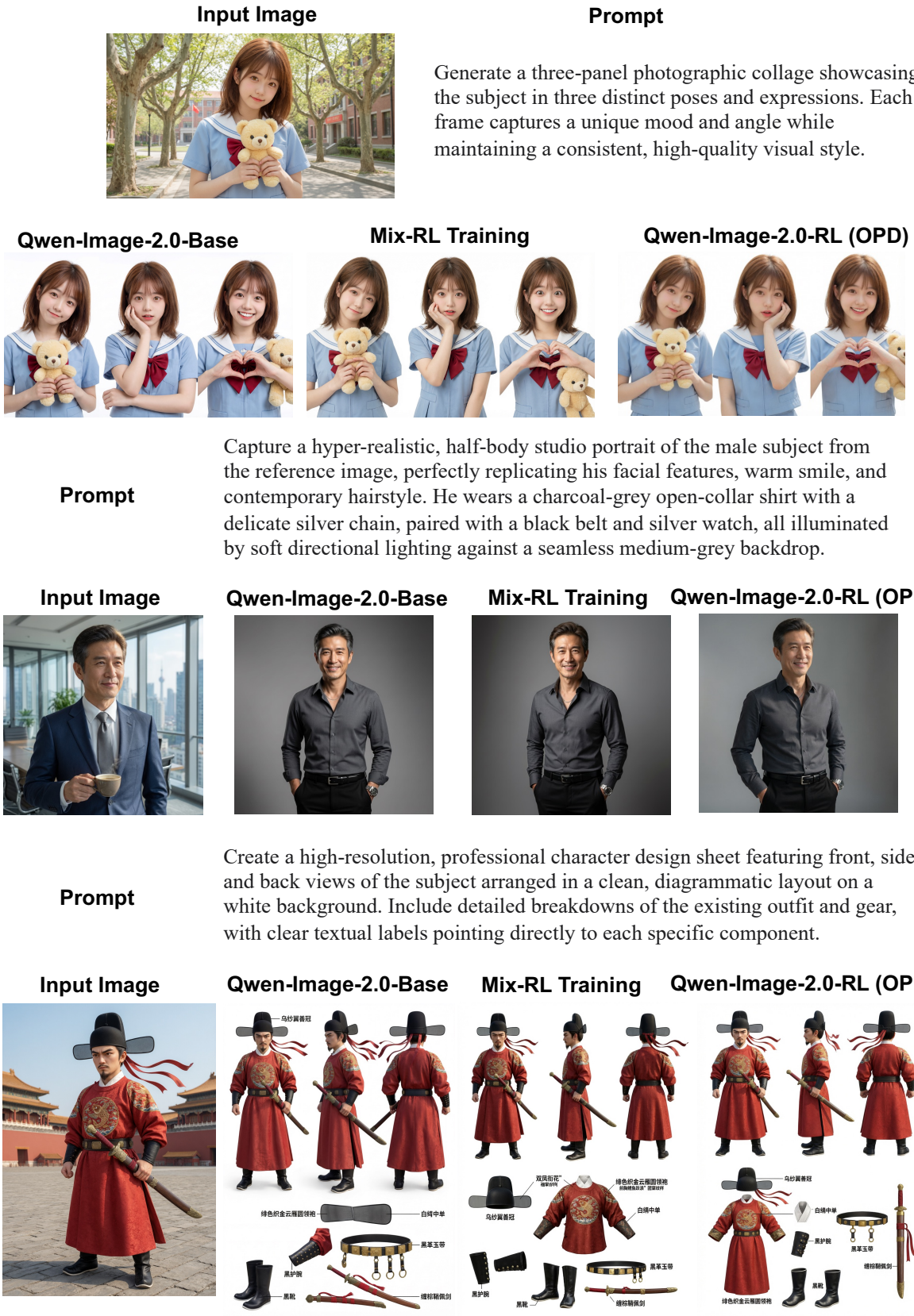


Figure 6: Qualitative comparison across portrait editing scenarios among three model variants. Qwen-Image-2.0-RL achieves the best face identity preservation and instruction-following accuracy, while Mix-RL improves over the base model but still exhibits noticeable identity drift or incomplete edits under complex instructions.

A Derivation of the On-Policy Distillation Objective

In this section, we provide a formal derivation of the OPD loss (Eqn. (14)) from the perspective of distributional distance minimization.

Goal. Let $v_{\theta^\star}$ denote a task-specialized RL-trained teacher and v_θ a student model (initialized from the pre-trained base model). Both share the same initial noise distribution $p_1 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and generate images by solving the reverse ODE (Eqn. (3)) from $t = 1$ to $t = 0$. The goal of OPD is to find θ such that the student's output distribution p_0^θ is as close as possible to the teacher's $p_0^{\theta^\star}$.

Analogy to LLM distillation. In large language model (LLM) distillation, the standard approach minimizes the forward Kullback–Leibler (KL) divergence between the student's and teacher's output distributions:

$$\min_\theta \ \mathrm{KL}(p_\theta(\cdot|c) \,\|\, p_{\theta^*}(\cdot|c)), \tag{15}$$

where p_θ and p_{θ^*} are the student's and teacher's next-token distributions conditioned on context c. The KL divergence admits a tractable decomposition over the autoregressive factorization, making it a natural choice for sequential discrete models. For continuous-space diffusion models, however, the output distributions are defined implicitly through an ODE solve, and the KL divergence between path measures is generally intractable.

Wasserstein-2 distance. We instead consider the 2-Wasserstein distance as the distributional metric:

$$W_2(p_0^\theta, p_0^{\theta^*}) = \left(\inf_{\gamma \in \Gamma(p_0^\theta, p_0^{\theta^*})} \int \|\boldsymbol{x} - \boldsymbol{y}\|^2 \, \mathrm{d}\gamma(\boldsymbol{x}, \boldsymbol{y}) \right)^{1/2}, \tag{16}$$

where $\Gamma(p_0^\theta, p_0^{\theta^*})$ denotes the set of all couplings between the two distributions. However, directly computing W_2 distance requires solving an optimal transport problem, which is intractable in high dimensions. We therefore seek to minimize an efficiently computable upper bound.

Benton et al. (2023) establish a rigorous connection between the velocity field approximation error and the distributional distance of the corresponding flows. Their result relies on the following assumptions:

(A1) *Existence and uniqueness of smooth flows.* For each initial point $\boldsymbol{x}_1 \in \mathbb{R}^d$, both the student flow (under v_θ) and the teacher flow (under v_{θ^*}) admit unique, continuously differentiable solutions.

(A2) *Lipschitz regularity of teacher velocity field.* For each $t \in (0, 1)$, there exists a constant L_t such that $v_{\theta^*}(\cdot, t)$ is L_t-Lipschitz in $\boldsymbol{x}$.

Under these assumptions, Theorem 1 of Benton et al. (2023) yields the following bound on the W_2 distance between the student's and teacher's generated distributions:

$$W_2(p_0^\theta,\ p_0^{\theta^\star}) \ \le \ \left(\int_0^1 \mathbb{E}_{\boldsymbol{x}_t \sim p_t^\theta} \|v_\theta(\boldsymbol{x}_t, t) - v_{\theta^*}(\boldsymbol{x}_t, t)\|^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \cdot \exp\left(\int_0^1 L_t \, \mathrm{d}t \right). \tag{17}$$

Continuous-time optimization objective. The exponential factor $\exp\left(\int_0^1 L_t \, \mathrm{d}t\right)$ depends on the Lipschitz regularity of the teacher's velocity field, which is controlled by the network architecture and does not depend on the training objective. Therefore, minimizing the W_2 upper bound in Eqn. (17) with respect to θ reduces to minimize the following trajectory-level velocity matching objective:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \int_0^1 \mathbb{E}_{\boldsymbol{x}_t \sim p_t^\theta} \|v_\theta(\boldsymbol{x}_t, t) - v_{\theta^*}(\boldsymbol{x}_t, t)\|^2 \, \mathrm{d}t. \tag{18}$$

In practice, the student's ODE is solved with N discrete steps using a timestep schedule $\{t_0 = 1, t_1, \ldots, t_N \approx 0\}$. The student's own trajectory $\{\boldsymbol{x}_{t_0}, \boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_N}\}$ provides samples from $p_{t_n}^\theta$ at each timestep. Approximating the continuous integral in Eqn. (18) by a sum over the discrete trajectory points recovers the practical OPD loss:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{c, \boldsymbol{x}_{[1:N]} \sim \pi_\theta(\cdot|c)} \left[\sum_{n=1}^{N} \|v_\theta(\boldsymbol{x}_{t_n}, t_n) - v_{\theta^*}(\boldsymbol{x}_{t_n}, t_n)\|^2 \right]. \tag{19}$$

This is precisely the training objective in Eqn. (14).

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

BlackForest. Flux. https://github.com/black-forest-labs/flux, 2024.

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

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024.

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.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

Google. Nano Banana Pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026a.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

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

OpenAI. GPT Image 1.5. https://developers.openai.com/api/docs/models/gpt-image-1.5, 2025.

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. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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, 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.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 2025a.

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

Yibin Wang, Yuhang Zang, 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, 2026.

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*, 2025a.

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

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.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

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. *arXiv preprint arXiv:2509.16117*, 2025.