RETHINKING QUALITY TUNING FOR RECTIFIED FLOW MODELS FROM A ROTATIONAL PERSPECTIVE

Anonymous authors
Paper under double-blind review

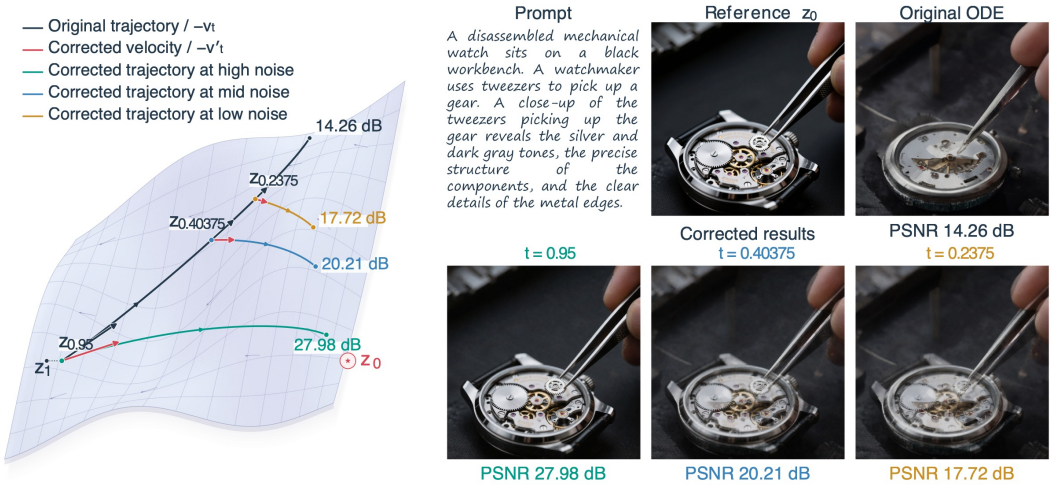


Figure 1: Visual comparison of ODE trajectories under velocity-direction correction with SD3.5-M. Starting from the same noisy latent, we integrate the ODE steps and begin correction at high, middle, or low noise levels. Each corrected step halves the angle to the reference velocity $u_{\text{ref}} = (z_t - z_0)/t$, computed from the clean latent z_0 while preserving the predicted velocity norm. Earlier correction yields higher reconstruction PSNR in this example, although it also affects more steps. This observation motivates rethinking quality tuning for rectified-flow models from a rotational perspective.

ABSTRACT

Quality tuning (QT) is critical for diffusion and rectified-flow models. It adapts pretrained generators to curated high-quality data and supports later post-training. However, the pretraining-QT data gap may require more updates, while overtuning on limited data can cause overfitting and lower generation quality. Standard QT also uses a fixed noise shift, so noise exposure remains unchanged during training. We revisit QT from a rotational perspective, motivated by two observations. First, earlier direction correction improves reconstruction, even though it affects more steps. Second, at similar noise levels, angular errors vary across examples. Based on these observations, we introduce Rotational Quality Tuning (RoQT). At the macro level, a noise curriculum anneals a strong high-noise shift through an intermediate shift toward the reference shift. At the micro level, angular weighting emphasizes difficult examples within each noise region. We further decompose the flow-matching mean-squared error into magnitude and rotational components, and leverage the rotational component to derive an explicit, scale-calibrated direction correction. RoQT improves convergence efficiency and generation quality while mitigating the performance degradation caused by overtuning on limited data. Experiments on SD3.5-Medium and Z-Image demonstrate that RoQT consistently outperforms standard QT on GenEval and Qwen-Image-Bench under the same training data and matched inference settings. We will release the code and data on GitHub.

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

1 INTRODUCTION

Diffusion models have become a leading approach for high-fidelity visual generation, progressing from image synthesis to general text-conditioned image and video generation (Ho et al., 2020; Rombach et al., 2022; Wan Team, 2025). More recent diffusion transformers and flow-matching models further improve resolution, prompt understanding, text rendering, and multimodal generation (Esser et al., 2024; Black Forest Labs et al., 2025; Wu et al., 2025). These advances make the final quality tuning stage of training increasingly important.

Modern visual generative models are commonly developed in three stages. First, large-scale pretraining and subsequent continuous training build broad visual and language knowledge from massive, mixed-quality data. Second, high-quality fine-tuning uses a much smaller, carefully selected dataset to move the output distribution toward visually appealing content. **Emu** formalizes this second stage as *quality tuning* and shows that only a few thousand high-quality images can substantially improve visual preference while retaining broad concept coverage (Dai et al., 2023). Third, preference- or reward-based post-training further improves prompt alignment and human preference (Wallace et al., 2024; Liu et al., 2025; Zheng et al., 2026). These stages are complementary. Post-training builds on the capabilities preserved by high-quality fine-tuning, while quality tuning connects broad pretrained knowledge to the target high-quality distribution. Throughout the rest of this paper, we use supervised fine-tuning (SFT) to refer specifically to this high-quality fine-tuning.

The central difficulty is that the small SFT set may differ substantially from the pretraining distribution. This gap can make adaptation harder, requiring more optimization steps to achieve a comparable fit and increasing the risk of overfitting the limited data. Standard SFT also uses a fixed noise shift. One recent solution organizes curated data into semantic domains, fine-tunes separate domain-specific models, and merges their parameters (Arkhipkin et al., 2025). Although effective, this approach requires multiple full fine-tuning runs and an additional merging stage. We instead ask whether a single model can adapt more effectively by deciding which noise levels and which examples receive stronger supervision during one training process.

We begin by rethinking where and what a flow-matching model should learn during quality tuning. High-resolution rectified-flow training benefits from shifting a logit-normal timestep distribution toward higher noise (Esser et al., 2024). Meanwhile, TACache separates velocity evolution into magnitude and directional components and highlights the importance of directional drift for inference-time caching (Liu et al., 2026). Together, these findings motivate examining noise allocation and velocity direction during quality tuning.

To examine this idea, we perform a training-free direction-correction experiment with the pretrained SD3.5-Medium model (Stability AI, 2024) (Figure 1). Starting from the same noisy latent, we correct the predicted velocity direction using the ground-truth clean latent while keeping the velocity norm unchanged. In this example, beginning correction at higher noise improves final reconstruction, although it also affects more steps. This observation motivates testing whether more high-noise supervision helps quality tuning. Separately, examples at the same noise level can have different angular errors despite comparable velocity magnitudes. This motivates the hypothesis that angular difficulty can guide supervision across examples. We therefore consider both which noise regions to emphasize and which examples within each region need greater attention.

Based on these hypotheses, we propose **Rotational Quality Tuning (RoQT)**. Its core is a macro-micro view of supervision allocation. **Macro-Level Noise Curriculum Learning** starts with a strong high-noise shift, moves through an intermediate shift, and gradually returns to the reference shift. This schedule first emphasizes high-noise states associated with global structure, then broadens coverage toward middle- and low-noise states for local detail. **Micro-Level Angular Quality Weighting** compares examples only within the same noise region and uses $1 - \cos\varphi_t$ as online difficulty, giving relatively hard examples larger flow-matching weights. Finally, because the flow-matching mean-squared error can be decomposed into magnitude and rotational components, RoQT makes the rotational component explicit as a scale-calibrated direction correction. RoQT preserves the original path and velocity target with one trainable-model forward pass.

We evaluate RoQT on two different diffusion-transformer backbones, SD3.5-Medium and Z-Image, using the same quality-tuning data and matched inference settings within each backbone. The evaluation covers compositional accuracy, visual quality, aesthetics, and text-image alignment. RoQT

improves the overall scores on both benchmarks over standard SFT for both backbones. Our contributions are threefold.

- We identify noise allocation and velocity direction as two connected factors in quality tuning, supported by high-noise direction-correction and same-noise sample diagnostics.
- We introduce a macro coarse-to-fine noise curriculum and micro angular difficulty weighting to allocate supervision across noise levels and examples.
- We derive a scale-calibrated direction correction from the rotational component of flow-matching MSE and integrate all three components into a one-pass objective.

2 PRELIMINARIES

2.1 CONDITIONAL RECTIFIED FLOW

Flow matching and rectified flow define a continuous transport between data and noise and have shown strong performance in high-resolution image synthesis (Liu et al., 2023; Lipman et al., 2023; Esser et al., 2024). Let $(x_0, c) \sim p_{\text{data}}$ denote a clean latent vector and its condition, and let $\epsilon \sim \mathcal{N}(0, I)$ be standard Gaussian noise. Throughout the paper, $t \in [0, 1]$ denotes the actual noise level, after any timestep shift. Rectified flow uses the straight probability path

$$x_t = (1 - t)x_0 + t\epsilon, \qquad u = \epsilon - x_0. \tag{1}$$

The network predicts $v_t = v_\theta(x_t, t, c)$. Its conditional flow-matching objective is

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{\substack{(x_0,c)\sim p_{\text{data}},\, t\sim p(t) \\ \epsilon\sim\mathcal{N}(0,I)}} \left[\|v_t - u\|_2^2 \right]. \tag{2}$$

The condition c can represent text or other control signals. This tractable conditional objective has the same gradient as the corresponding marginal flow-matching objective up to a constant independent of θ (Lipman et al., 2023).

2.2 RESOLUTION-DEPENDENT TIMESTEP SHIFTING

We sample the unshifted timestep $\tilde{t}$ from a logit-normal distribution

$$\xi \sim \mathcal{N}(0, 1), \qquad \tilde{t} = \text{sigmoid}(-\xi). \tag{3}$$

Higher-resolution images contain more pixels and therefore require more noise to remove a comparable amount of signal (Esser et al., 2024). Let $n = H\,W$ be the number of pixels at a reference resolution and $m = H'W'$ the number at a higher target resolution. Under SD3's constant-image approximation, the signal uncertainty scales as $\sigma(\tilde{t}, n) = \tilde{t}/((1 - \tilde{t})\sqrt{n})$. Matching this uncertainty across the two resolutions, $\sigma(\tilde{t}, n) = \sigma(t, m)$, gives the final noise level

$$t = \frac{\alpha\tilde{t}}{1 + (\alpha - 1)\tilde{t}}, \qquad \alpha = \sqrt{\frac{m}{n}}. \tag{4}$$

For $m > n$, we have $\alpha > 1$ and $t > \tilde{t}$, shifting the timestep distribution toward higher noise. The constant-image assumption is only an approximation, so SD3 selects the shift empirically. For 1024×1024 synthesis, it reports $\alpha = 3.0$ for both training and sampling (Esser et al., 2024). We use α to denote the model's reference shift for standard quality fine-tuning.

For the rectified-flow path, the log signal-to-noise ratio is $\lambda(t) = \log((1 - t)^2/t^2)$. Substituting Equation 4 gives $\lambda(t) = \lambda(\tilde{t}) - 2\log\alpha = \lambda(\tilde{t}) - \log(m/n)$. Thus, resolution-dependent shifting lowers log-SNR by a constant and increases the probability of high-noise training states. SD3 uses this transform with a fixed coefficient. In the next section, we retain the same path and apply the transform once, but vary its coefficient over training to allocate more early supervision to high noise.

Velocity-angle variation along the sampling trajectory. TACache analyzes consecutive rectified-flow velocities by decomposing their discrete acceleration into a component parallel to the current velocity and an orthogonal residual (Liu et al., 2026). The parallel component describes magnitude evolution, whereas the orthogonal residual describes directional turning. Its trajectory analysis and

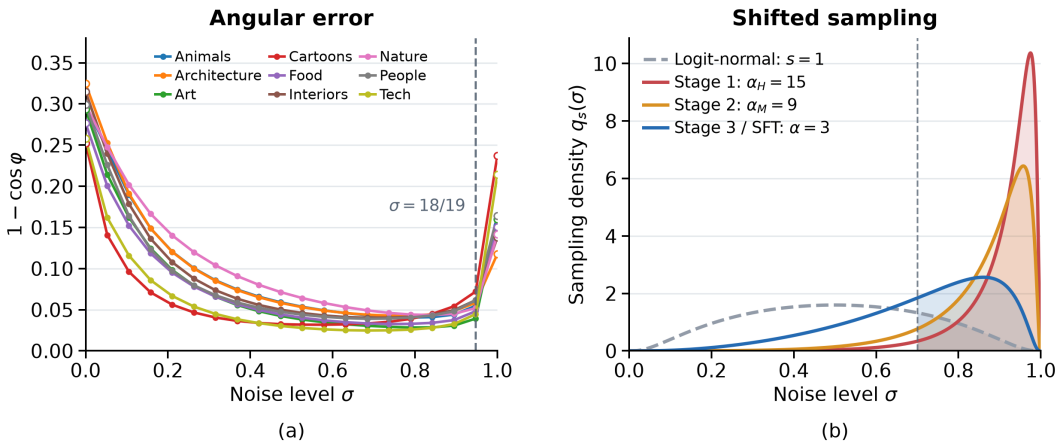


Figure 2: Angular difficulty and noise curriculum. (a) SD3.5-M exhibits category-dependent angular errors across noise levels. (b) The curriculum relaxes the sampling shift from 15 to 9 to the SFT reference of 3. Shaded areas represent sampling probabilities for $\sigma > 0.7$, corresponding to $96.86\%, 91.15\%$, and 59.92%, respectively, compared with 19.84% for the unshifted distribution.

ablations show that magnitude evolution is comparatively regular, while explicitly controlling directional drift provides the main fidelity gain under step skipping. This indicates that flow-matching trajectories cannot be characterized by velocity magnitude alone. TACache measures direction changes between consecutive sampling velocities at inference time; RoQT instead measures prediction-target alignment at each noise level during training.

3 Method

3.1 Overview

Following the preliminaries, t denotes the shifted noise level, α the reference SFT shift, and $\varphi_t = \angle(v_t, u)$ the angle between the predicted and target velocities. Standard quality tuning uses a fixed shift, although reconstruction sensitivity changes with noise and angular errors vary across examples. RoQT combines "Macro-Level Noise Curriculum Learning" with "Micro-Level Angular Quality Weighting" to allocate supervision across noise levels and relatively difficult examples within each noise region. A scale-calibrated "Direction Correction from the Flow-Matching Error" complements the weighted objective without changing the flow path or velocity target.

3.2 Macro-Level Noise Curriculum Learning

Motivated by the direction-correction results in Figure 1, we allocate more early quality-tuning updates to high-noise states. For a point on the conditional training path, the one-step estimate $\widehat{x}_0 = x_t - tv_t$ satisfies

$$\|\widehat{x}_0 - x_0\|_2^2 = t^2 \, \|v_t - u\|_2^2 . \tag{5}$$

Thus, comparable velocity errors produce larger one-step reconstruction errors at higher noise. Together with the fixed-norm direction-correction results in Figure 1, this observation motivates learning more accurate velocity directions at high noise early in training to improve image quality.

We next measure $1 - \cos\varphi_t$ on high-quality training examples from nine semantic categories, using the same clean–noise pair across noise levels. Figure 2(a) shows a similar profile across categories, with lower angular errors in the middle and higher errors near both endpoints. The high-noise increase indicates that these states remain difficult for the pretrained model and this observation motivates additional high-noise supervision at the start of quality tuning.

Our **macro-level noise curriculum** varies one shared shift over training. Let $p \in [0, 1]$ denote progress over optimizer steps. We sample $\tilde{t}$ as in Equation 3 and replace the fixed shift in Equation 4

with $\alpha(p)$ as follows.

$$t = \frac{\alpha(p)\tilde{t}}{1 + (\alpha(p) - 1)\tilde{t}}, \qquad \alpha(p) : \ \alpha_{\mathrm{H}} \longrightarrow \alpha_{\mathrm{M}} \longrightarrow \alpha, \quad \alpha_{\mathrm{H}} > \alpha_{\mathrm{M}} > \alpha. \tag{6}$$

We start with α_{H} well above the reference α, smoothly decrease it through α_{M}, and finally return to α. Figure 2(b) shows how the larger shift increases sampling in the high-noise region $t \geq 0.7$; reducing it progressively restores coverage of middle and lower noise levels. Our SD3.5 setup uses 15, 9, and 3 for these three shifts, with the recorded schedule and loss components analyzed in Figure 5(a) in the ablation study. The curriculum changes the relative supervision budget across noise levels, not the rectified-flow path or target. Every stage samples a continuous distribution, and all examples share the same shift. With angular weighting and direction correction fixed, the curriculum improves both benchmark Overall scores over a fixed reference shift on SD3.5-M (Table 3(b)). The following micro-level rule separately controls relative example weights.

3.3 Micro-Level Angular Quality Weighting

The macro curriculum determines which noise levels receive more supervision, but it does not distinguish examples at the same noise level. Our second hypothesis is that examples may differ more in prediction–target direction than in velocity magnitude. At a fixed noise level, examples with comparable velocity norms can have clearly different angular errors. These differences are more pronounced in high- and low-noise regions, while angular errors are generally smaller at intermediate noise levels. This motivates a **micro-level angular weighting** rule that gives greater weight to relatively difficult examples within each noise region. Together, the macro curriculum controls noise exposure, while the micro rule controls the relative strength of supervision across examples.

We measure angular difficulty from the prediction–target angle φ_t as

$$\cos\varphi_t = \frac{\langle v_t, u\rangle}{\|v_t\|_2 \|u\|_2}, \qquad a = 1 - \cos\varphi_t. \tag{7}$$

For nonzero velocities, $a \in [0, 2]$, with larger values indicating greater directional mismatch. Since typical angular errors vary with noise level, we compare examples within the same noise region, defined by

$$b(t) = \begin{cases} \mathrm{H}, & t \geq 0.7, \\ \mathrm{M}, & 0.3 \leq t < 0.7, \\ \mathrm{L}, & t < 0.3. \end{cases} \tag{8}$$

Let $\bar{a}_b = \mathbb{E}[a \mid b(t) = b]$ denote the mean angular difficulty in region b, estimated online. We define the loss weight from the relative deviation from this regional reference as

$$w = 1 + \eta \ \mathrm{sg}\left(\frac{a - \bar{a}_{b(t)}}{\bar{a}_{b(t)} + \varepsilon}\right), \qquad 0 \leq \eta < 1, \tag{9}$$

where sg denotes stop-gradient, η controls the weighting strength, and $\varepsilon > 0$ ensures numerical stability. An example with above-average angular difficulty receives $w > 1$, while an easier example receives $w < 1$. With the regional mean as the reference, the expected weight within each region is one. This normalization emphasizes relative difficulty without assigning larger weights to an entire noise region simply because its average angular error is higher. The micro rule changes only the loss weight; it neither resamples examples nor assigns them different shifts. On SD3.5-M, angular weighting improves both Overall scores over the macro-only variant (Table 3(a)).

3.4 Direction Correction from the Flow-Matching Error

Micro weighting identifies where the current model has greater rotational difficulty, but the standard flow-matching loss still mixes magnitude and direction errors. The direction term follows directly from its mean-squared error. Let $r = \|v_t\|_2$ and $q = \|u\|_2$. For nonzero velocities,

$$\|v_t - u\|_2^2 = (r - q)^2 + 2rq\,(1 - \cos\varphi_t)\,. \tag{10}$$

The first term measures magnitude mismatch, whereas the second is exactly the rotational part of the flow-matching error. RoQT preserves the complete mean-squared error and adds a scaled copy of this rotational component, yielding

$$\mathcal{L}_{\mathrm{RoQT}} = \mathbb{E}\big[w\|v_t - u\|_2^2\big] + \beta\,\mathbb{E}\big[w\,t^2\ \mathrm{sg}(2rq)\,(1-\cos\varphi_t)\big]\,. \tag{11}$$

The factor $2rq$ matches the scale of the rotational MSE component. We detach it so that the added term produces only a directional gradient in velocity space. The factor t^2 follows the reconstruction sensitivity in Equation 5, which gives larger weight to the same velocity error at higher noise. Unlike the macro curriculum, which controls sampling frequency, t^2 controls correction strength at each noise level. The default direction-loss coefficient is $\beta = 0.25$. Adding direction correction yields the highest Overall scores among the component variants in Table 3(a). The complete training procedure is summarized in Appendix D.

4 Experiments

4.1 Experimental Settings

We curate approximately 50K images from Pixabay and LAION's DataComp-HQ dataset (Callaghan, 2023; Schuhmann et al., 2022), filtering by aesthetic quality, sharpness, NIQE, and CLIP-IQA. Category counts and prompt re-annotation details will be provided in Appendix A. We evaluate RoQT on SD3.5-Medium (Stability AI, 2024) and Z-Image (Z-Image Team, 2025). We freeze each pretrained backbone and train rank-32 LoRA adapters with AdamW at a learning rate of 5×10^{-5}. SFT and RoQT share the same curated training data. We evaluate on GenEval (Ghosh et al., 2023) and Qwen-Image-Bench (Li et al., 2026c) at 1024×1024, using matched inference settings within each backbone. Scores use the 0–1 and 0–100 scales, respectively.

4.2 Qualitative Comparison

Figure 3 compares six shared prompts on both backbones. RoQT shows clearer facial and material details, more distinct light and shadow, and more balanced composition, particularly in portraits and indoor scenes. The portrait and study panels make the changes in fine textures and lighting especially visible. These visual gains are consistent with the macro curriculum's emphasis on global structure and the micro angular terms' focus on difficult details. Full prompts appear in Appendix C.

4.3 Quantitative Comparison

Table 1 compares Base, SFT, and RoQT on GenEval. Relative to SFT, RoQT raises GenEval Overall from 0.6814 to 0.6943 on SD3.5-M and from 0.7344 to 0.7562 on Z-Image. It also improves all six GenEval categories on both backbones, including counting and spatial relations. On SD3.5-M, Position also rises from 0.2275 to 0.2525. For Z-Image, Two Object rises from 0.8687 to 0.9091, and Counting rises from 0.7813 to 0.8188. On Qwen-Image-Bench (Table 2), RoQT raises Overall from 43.53 to 44.04 on SD3.5-M and from 42.48 to 44.16 on Z-Image. Both backbones improve over SFT in Aesthetics, Alignment, and Creative Generation. Z-Image gains 2.54 points in Aesthetics and 2.31 points in Quality over SFT. Together, the two benchmarks demonstrate consistent gains in compositional accuracy and VLM-judged generation quality across both backbones, using the same training data and matched inference settings.

4.4 Ablation Study

The ablations test which noise levels and examples should receive stronger supervision. All experiments use SD3.5-M. Table 3(a) adds the macro noise curriculum (Section 3.2), micro angular quality weighting (Section 3.3), and direction correction (Section 3.4) in sequence. Micro improves both Overall scores over Macro alone, and the full method performs best. Table 3(b) varies the noise curriculum with Micro and Dir. fixed. The full $\alpha_{\mathrm{H}} \to \alpha_{\mathrm{M}} \to \alpha$ schedule outperforms fixed α. Its high initial shift prioritizes global structure, the intermediate shift provides a gradual transition, and the reference shift restores broader noise coverage. Keeping α_{H} fixed may limit lower-noise refinement, while removing α_{M} may weaken the transition between stages. Both controls score below the full



(a) SD3.5-M



(b) Z-Image

Figure 3: Qualitative comparison of Base, SFT, and RoQT on six shared prompts for (a) SD3.5-M and (b) Z-Image. Each column uses the same prompt across both backbones. Long prompts are shortened above; full text is provided in Appendix C. Images are resized to uniform square panels for display.

Table 1: GenEval comparison at 1024×1024 with matched inference settings within each backbone. SFT and RoQT use the same training data. Scores use the 0–1 scale; higher is better. Best and second-best scores within each backbone are bold and underlined, respectively.

Backbone	Variant	GenEval						
		Single Object ↑	Two Object ↑	Counting ↑	Colors ↑	Position ↑	Attribute Binding ↑	Overall ↑
SD3.5-M	Base	0.9906	0.8434	0.5750	0.8112	0.1975	**0.5950**	0.6688
	SFT	0.9938	0.8586	0.6063	0.8271	0.2275	0.5750	0.6814
	RoQT	**0.9969**	**0.8712**	**0.6281**	**0.8298**	**0.2525**	0.5875	**0.6943**
Z-Image	Base	**0.9906**	0.8586	0.7500	**0.8351**	0.4075	0.5125	0.7257
	SFT	0.9781	0.8687	0.7813	0.8032	0.4425	0.5325	0.7344
	RoQT	**0.9906**	**0.9091**	**0.8188**	0.8165	**0.4575**	**0.5450**	**0.7562**

three-stage schedule on GenEval and Qwen-Image-Bench, supporting the use of the intermediate stage. Further diagnostics are planned in Appendix B.

Table 2: Qwen-Image-Bench comparison at 1024×1024 with matched inference settings within each backbone. SFT and RoQT use the same training data. Scores use the 0–100 scale; higher is better. Best and second-best scores within each backbone are bold and underlined, respectively.

Backbone	Variant	Qwen-Image-Bench					
		Quality ↑	Aesthetics ↑	Alignment ↑	Real-world Fidelity ↑	Creative Generation ↑	Overall ↑
SD3.5-M	Base	**48.28**	45.77	40.88	<u>42.79</u>	33.58	43.19
	SFT	<u>47.86</u>	<u>46.25</u>	<u>41.51</u>	42.42	<u>35.13</u>	<u>43.53</u>
	RoQT	47.83	**46.86**	**42.69**	**42.92**	**35.42**	**44.04**
Z-Image	Base	40.14	41.51	43.94	41.65	37.84	41.22
	SFT	<u>40.98</u>	<u>42.96</u>	<u>44.55</u>	**43.00**	<u>40.98</u>	<u>42.48</u>
	RoQT	**43.29**	**45.50**	**45.99**	<u>42.13</u>	**41.87**	**44.16**

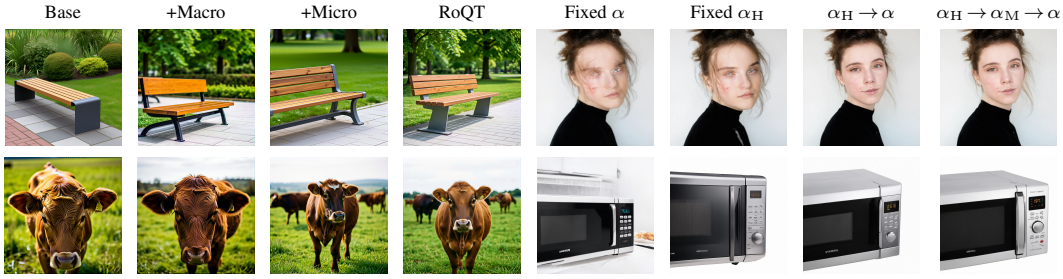
Table 3: SD3.5-M ablations of (a) method components and (b) noise curricula with Micro and Dir. fixed. Best scores are bold.

Variant	Macro	Micro	Dir.	GenEval Overall ↑	Qwen-Image Overall ↑
Base	×	×	×	0.6688	43.19
(i) + Macro	✓	×	×	0.6711	43.61
(ii) + Micro	✓	✓	×	0.6822	43.87
(iii) RoQT	✓	✓	✓	**0.6943**	**44.04**

(a) Component ablation

Shift schedule	GenEval Overall ↑	Qwen-Image Overall ↑
Fixed α	0.6649	43.60
Fixed α_{H}	0.6794	44.03
$\alpha_{\mathrm{H}} \to \alpha$	0.6857	43.91
$\alpha_{\mathrm{H}} \to \alpha_{\mathrm{M}} \to \alpha$	**0.6943**	**44.04**

(b) Noise curriculum



(a) Component ablation **(b) Noise-curriculum ablation**

Figure 4: SD3.5-M visual ablations corresponding to Table 3.

Figure 4 shows representative SD3.5-M outputs for both ablations. As the RoQT components are added, object details, lighting, and composition progressively improve. The staged noise curriculum also reduces residual artifacts in the portrait and recovers finer facial details. These examples complement the quantitative results in Table 3.

Training dynamics. Figure 5(a) shows the loss components in Equation 11 and the recorded macro shift. The FM terms decline around both shift transitions. Figure 5(b) reports Overall scores at matched optimizer steps. RoQT outperforms SFT at every evaluation after initialization on both benchmarks, indicating faster adaptation to the high-quality target distribution. SFT requires more updates to improve its GenEval score, yet its VLM-judged Qwen-Image-Bench score declines after an early peak, even while GenEval continues to rise. This divergence shows how monitoring a single metric can conceal the decline in generation quality caused by over-tuning on limited data. RoQT maintains stronger performance across later evaluations, reducing this overfitting effect.

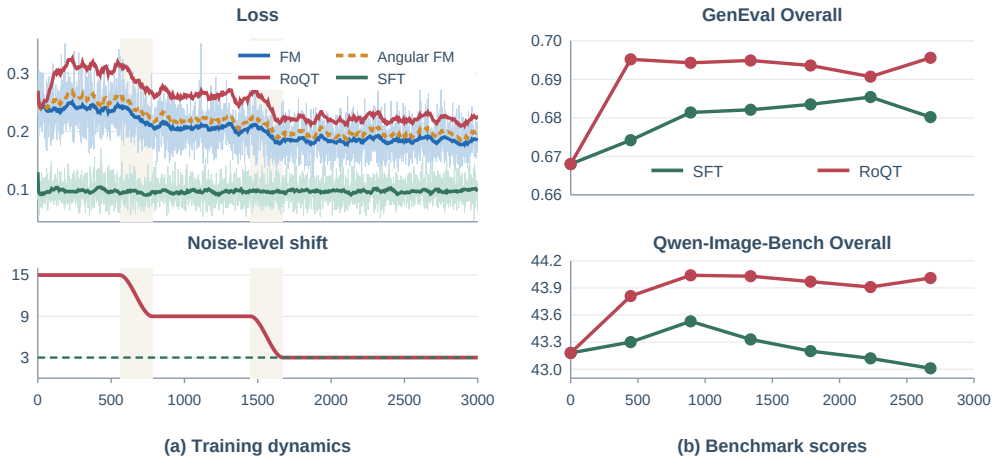


Figure 5: SD3.5-M training losses and noise shift through optimized steps, with GenEval and Qwen-Image-Bench Overall scores.

5 Related Work

Quality tuning and multi-stage training. Emu separates large-scale knowledge learning from a final quality-learning stage and names the latter *quality tuning* (Dai et al., 2023). It shows that a few thousand carefully selected images can strongly improve visual preference, while early stopping is needed to preserve generality. Kandinsky 5.0 further studies distribution imbalance in high-quality SFT data (Arkhipkin et al., 2025). It divides the curated data into nine domains, trains domain-specific models, and merges their parameters into an SFT model soup.

Diffusion models and Flow matching. In recent years, diffusion models (Ho et al., 2020; Song et al., 2020a;b; Nichol & Dhariwal, 2021) have achieved remarkable progress in generative tasks. Flow matching and rectified flow learn continuous transport from noise to data with a direct velocity target (Lipman et al., 2023; Liu et al., 2023). SD3 (Esser et al., 2024) studies timestep sampling for high-resolution training and shifts the logit-normal distribution toward higher noise. Related flow-matching models extend this formulation to image generation, image editing, and video generation (Black Forest Labs et al., 2025; Wan Team, 2025; Chen et al.; Hong et al., 2022; Yang et al., 2024; Li et al., 2026a; Kong et al., 2025; Wang et al., 2026). TACache decomposes velocity changes into magnitude and directional components for inference-time caching and shows the importance of controlling directional drift (Liu et al., 2026). RoQT uses these observations during training. It schedules high-noise exposure at the macro level and measures target-relative angular difficulty at the micro level.

Preference-based diffusion post-training. Preference optimization further aligns pretrained generators with human judgments. Diffusion-DPO adapts direct preference optimization to diffusion models and learns from pairwise comparisons (Wallace et al., 2024). Flow-GRPO introduces online reinforcement learning for flow-matching models by enabling stochastic exploration along the sampling process (Liu et al., 2025). DiffusionNFT instead applies online reinforcement signals through the forward process and contrasts positive and negative generations inside a supervised-style objective (Zheng et al., 2026). HP-Edit (Li et al., 2026b) applies GRPO to align diffusion-based image editing models with human preferences.

6 Conclusion

We present Rotational Quality Tuning (RoQT), a macro–micro framework for fine-tuning rectified-flow models on high-quality data. RoQT combines a macro noise curriculum with micro angular difficulty weighting and derives its direction correction directly from the rotational component of the flow-matching mean-squared error. RoQT improves overall GenEval and Qwen-Image-Bench scores over SFT on SD3.5-Medium and Z-Image, while adapting faster and avoiding the late-stage quality degradation observed with SFT. In the future, we will extend noise allocation and velocity rotation to online and offline reinforcement learning and improve the performance of post-training.

AI USE STATEMENT

We used large language models (LLMs) for language polishing and code debugging. All LLM-generated suggestions were reviewed and verified by the authors. The authors take full responsibility for the final content of this manuscript.

REFERENCES

Vladimir Arkhipkin et al. Kandinsky 5.0: A family of foundation models for image and video generation. *arXiv preprint arXiv:2511.14993*, 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.

Michael D Callaghan. Pixabay. com image repository. In *Angular for Business: Awaken the Advocate Within and Become the Angular Expert at Work*, pp. 277–280. Springer, 2023.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, et al. Boogu-image-0.1: Boosting open agentic multimodal generation via understanding under a minimal budget. *arXiv preprint 2607.13125*.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Patrick Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *CVPR*, pp. 12809–12819. IEEE, 2025.

Bingchen Li, Zhixin Wang, Fan Li, Jiaqi Xu, Jiaming Guo, Renjing Pei, Xin Li, and Zhibo Chen. Colorflux: A structure-color decoupling framework for old photo colorization. In *CVPR*, 2026a.

Fan Li, Chonghuinan Wang, Lina Lei, Yuping Qiu, Jiaqi Xu, Jiaxiu Jiang, Xinran Qin, Zhikai Chen, Fenglong Song, Zhixin Wang, Renjing Pei, and Wangmeng Zuo. Hp-edit: A human-preference post-training framework for image editing. In *CVPR*, 2026b.

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*, 2026c.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online reinforcement learning. *arXiv preprint arXiv:2505.05470*, 2025.

Xiao Liu, Kai Liu, Naiyang Guan, Hongliang Lu, Zhixin Wang, Zhikai Chen, Renjing Pei, and Yulun Zhang. Accelerating rectified flow models via trajectory-aware caching. *arXiv preprint arXiv:2605.16789*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Stability AI. Stable diffusion 3.5 medium model card. https://huggingface.co/stabilityai/stable-diffusion-3.5-medium, 2024.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Chonghuinan Wang, Zhikai Chen, Chunwei Wang, Yecong Wan, Junwei Yang, Zhixin Wang, Wei Zhang, Jiaqi Xu, Renjing Pei, Xiaohe Wu, et al. Illuminating unified multimodal model for free-form interleaved text-image generation. *ECCV*, 2026.

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*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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*, volume 2026, pp. 134129–134150, 2026.

A Data Curation and Prompt Re-annotation

A.1 Dataset composition

The curated dataset contains 50,000 image–text examples, with 49,800 in the training split and 200 in the validation split. The raw captions are then re-captioned by a closed-source VLM API. Table 4 reports all 11 flat category labels, including *text* and *other*. These full-dataset counts should be distinguished from the nine semantic categories used for the angular diagnostic in Figure 2(a). The stored label *tech* corresponds to technology.

Table 4: Category counts in the 50K dataset. Percentages use all 50,000 examples.

Category	Train	Validation	Total	Share (%)
People	13,269	53	13,322	26.644
Text	8,561	45	8,606	17.212
Nature	6,731	20	6,751	13.502
Architecture	5,474	19	5,493	10.986
Tech	3,652	13	3,665	7.330
Interiors	3,152	16	3,168	6.336
Animals	2,349	7	2,356	4.712
Food	1,944	9	1,953	3.906
Art	1,884	5	1,889	3.778
Cartoons	1,425	7	1,432	2.864
Other	1,359	6	1,365	2.730
Total	49,800	200	50,000	100.000

A.2 Prompt statistics

We measure prompt length from the dataset's `text` field as the number of Unicode characters after removing whitespace, including spaces and line breaks; punctuation is retained. This is a character count, not a word count or a tokenizer-based length. Across all 50,000 prompts, the mean and median lengths are 576.39 and 425 characters, respectively, and lengths range from 6 to 2,911. Figure 6 shows the category composition and the full length distribution in 10-character bins. Table 5 summarizes length by category. P90 denotes the 90th percentile, meaning that 90% of prompts in a category have lengths at or below the reported value.

The supplied audit reports no null or whitespace-only prompts, and confirms that IDs and text match the original records after category labeling. It identifies 25,305 prompts containing Chinese characters (50.61% of the dataset). Character presence is not a strict language classification and does not imply that each such prompt is exclusively Chinese.

Table 5: Prompt lengths by category across both splits. All values are in non-whitespace Unicode characters.

Category	Mean	Median	P90	Minimum	Maximum
People	549.28	411.00	1,301.00	18	2,647
Text	569.20	425.00	1,175.00	19	2,305
Nature	579.12	392.00	1,321.00	20	2,282
Architecture	611.99	429.00	1,364.60	24	2,455
Tech	595.41	450.00	1,347.60	6	2,174
Interiors	607.50	452.50	1,394.30	6	2,537
Animals	528.89	388.50	1,253.00	16	2,143
Food	601.29	483.00	1,290.80	18	2,048
Art	603.76	451.00	1,335.20	18	2,305
Cartoons	652.37	512.50	1,281.90	29	2,911
Other	535.14	408.00	1,272.80	7	2,528

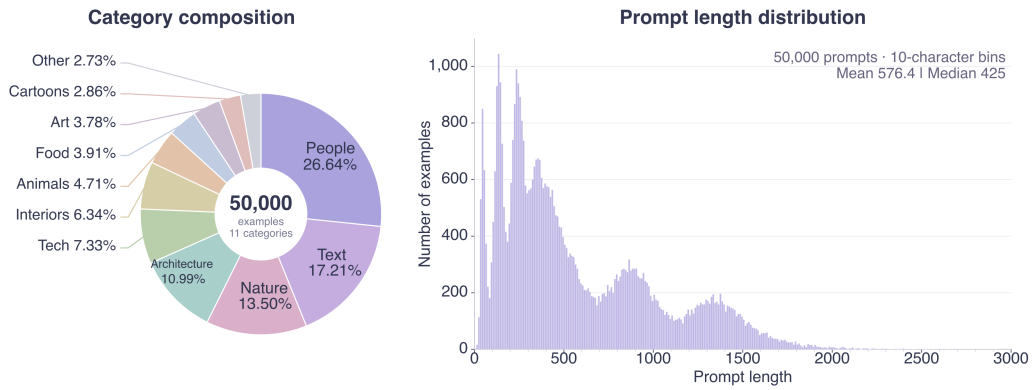


Figure 6: Category composition and prompt-length distribution of all 50,000 examples. Prompt length counts non-whitespace Unicode characters; the histogram uses 10-character bins over the full observed range.

B Noise-Curriculum Diagnostics

Beyond the stage-count ablation, we plan to measure flow-matching error and angular error on a fixed validation set at uniformly spaced noise levels. We will report high-, middle-, and low-noise errors at 25%, 50%, and 100% of training. The hypothesis is that the larger initial shift improves high-noise direction learning early, while the return to the reference shift improves middle- and low-noise coverage later. These diagnostics test the proposed mechanism; they are not completed experimental results.

C Prompts for Qualitative Comparison

The prompts below correspond to the six columns of Figure 3, from left to right. The first two are from Qwen-Image-Bench (Li et al., 2026c); the remaining four are from GenEval (Ghosh et al., 2023).

Column 1. Portrait. realistic half-body close-up portrait of a European model wearing a traditional Quanzhou-style hairpin, combining silk flowers with silver ornaments. The floral elements are layered yet not overly elaborate; the makeup is restrained, and the skin's texture is natural, revealing subtle details; side-backlighting creates contour lighting around the hair, and the silver ornaments reflect light realistically without overexposure; the model has a gentle, confident smile, looking directly at the camera; the background is a light-colored wall with a slight blur. The hairpin should have a reasonable structure, with petal edges that don't melt, and the style should be consistent with a VOGUE cover magazine aesthetic—a high-fashion portrait.

Column 2. Study. realistic restoration of a Ming Dynasty scholar's study: The interior of the Ming-style study is realistic, with reasonable mortise and tenon wooden furniture (official's hat chair, long table); there is Xuan paper, inkstone, writing brush, and paperweight on the table, and natural light from the window casts clear light and shadow on the table; the materials are required to be real (hardwood grain, ink color, paper fiber), the perspective is correct, and the overall color tone is elegant and low saturation; a scholar's profile or back can be seen, with a calm and focused expression.

Columns 3–6. Single objects. The complete prompts are "a photo of a bus", "a photo of a couch", "a photo of a bicycle", and "a photo of a zebra", respectively.

D TRAINING ALGORITHM

Algorithm 1 combines the macro curriculum, micro angular weighting, and direction correction using one trainable-model forward pass.

Algorithm 1 Rotational Quality Tuning (RoQT)

Require: Pretrained rectified-flow model v_θ, high-quality data, and total optimizer steps K

1: **for** $k = 0, \ldots, K-1$ **do**
2: Sample $(x_0, c) \sim p_{\text{data}}$ and $\epsilon \sim \mathcal{N}(0, I)$; set $u \leftarrow \epsilon - x_0$
3: Compute progress p and select the shared macro shift $\alpha(p)$ by Equation 6
4: Sample $\tilde{t}$ by Equation 3 and obtain t with the current shift in Equation 6
5: Form $x_t \leftarrow (1-t)x_0 + t\epsilon$ and predict $v_t \leftarrow v_\theta(x_t, t, c)$
6: Compute angular distance and the micro weight w by Equations 7–9
7: Evaluate $\mathcal{L}_{\text{RoQT}}$ in Equation 11 and update θ
8: **end for**
9: **return** Fine-tuned RoQT model

E MORE VISUAL COMPARISONS

We supplement more viusual comparison across two models SD3.5-M and Z-Image and all these results are from GenEval and Qwen-Image-Bench.

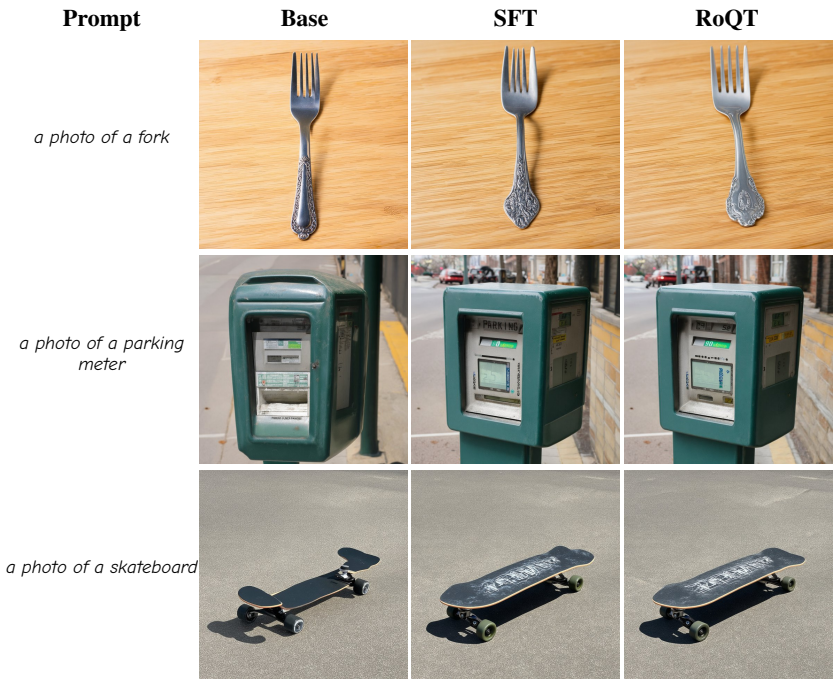


Figure 7: Qualitative comparisons on SD3.5-Medium.

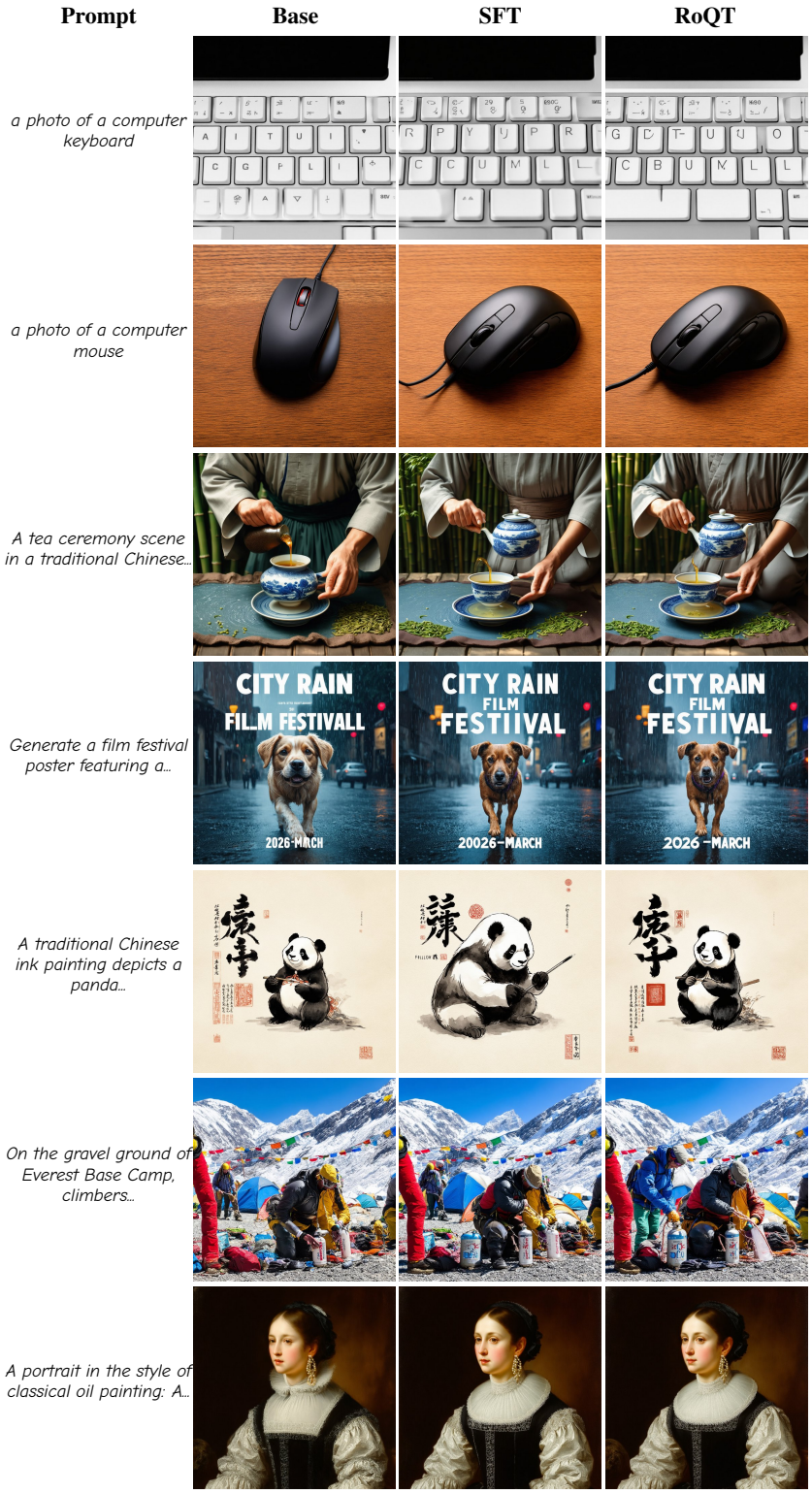


Figure 8: Additional qualitative comparisons on SD3.5-Medium.



Figure 9: Qualitative comparisons on Z-Image.

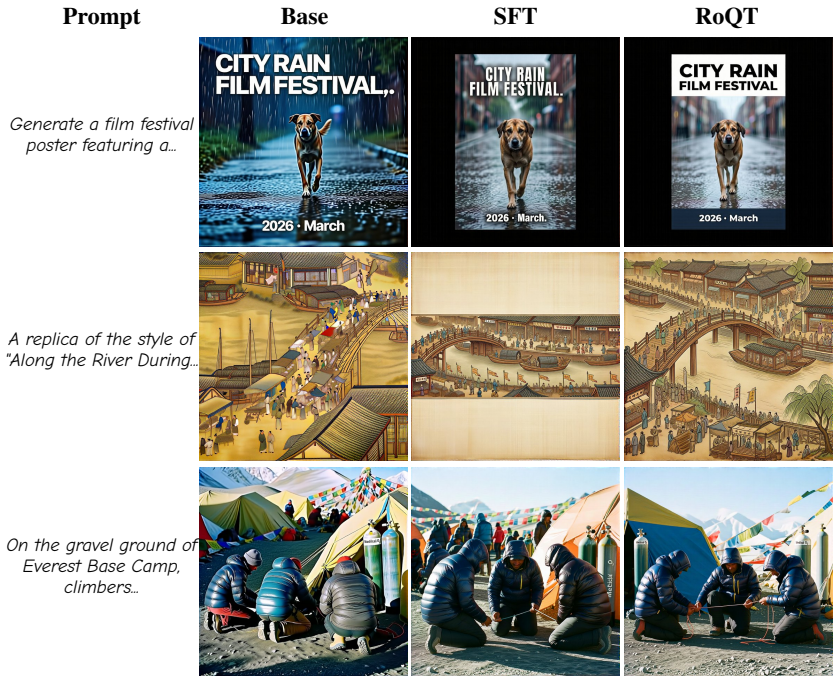


Figure 10: Additional qualitative comparisons on Z-Image.