QWEN-IMAGE-FLASH: RETHINKING THE TRAINING RECIPE FOR FEW-STEP DISTILLATION

Anonymous authors
Paper under double-blind review



Figure 1: **Qualitative results of Qwen-Image-Flash.** T2I generation and image editing results with only 4 NFEs, demonstrating unified few-step capabilities across generation and editing.

ABSTRACT

Few-step distillation has emerged as a critical component in the development of advanced visual generative foundation models, substantially reducing inference overhead while enabling real-time generation and cost-efficient deployment across a broad range of practical scenarios. However, prior work has predominantly focused on advancing training objectives, while comparatively overlooking the *training recipe*, which has become increasingly critical in the era of large-scale foundation models. In this work, we systematically revisit the training recipe under the well-established distribution matching distillation (DMD) framework for both text-to-image generation and image editing, focusing on three key dimensions: *training data composition*, *teacher guidance within DMD*, and *task mixture*. Our empirical analysis reveals several non-obvious and counterintuitive phenomena, ultimately motivating the development of **Qwen-Image-Flash**. These findings highlight that effective few-step distillation depends not only on carefully designed objectives, but also on a principled training recipe.

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

Visual generative models have emerged as a flexible and accessible paradigm for content creation, interactive editing, and multimodal applications (Esser et al., 2024; Song et al., 2026; Liu et al., 2026; Mao et al., 2026). Built predominantly on diffusion (Ho et al., 2020; Song et al., 2020) and flow-based formulations (Lipman et al., 2022), modern foundation models can generate high-fidelity images from complex prompts, render dense structured text, and perform instruction-guided editing within a unified framework (Zhao et al., 2026).

However, their practical deployment remains limited by iterative denoising or transport trajectories that require numerous function evaluations, resulting in substantial latency and computational cost. This sampling overhead is particularly prohibitive for latency-sensitive and resource-constrained scenarios, including interactive editing (Meng et al., 2022), on-device generation (Li et al., 2023; Zhao et al., 2024), and large-scale content production (Azuma, 1997; Yin et al., 2025).

To alleviate this bottleneck, few-step distillation has emerged as a promising paradigm that compresses the generative dynamics of a multi-step teacher into a student capable of high-quality generation with only a small number of function evaluations (NFEs). Recent advances have largely centered on the design of distillation objectives, spanning trajectory-level alignment (Geng et al., 2025), consistency training (Song et al., 2023), adversarial distillation (Sauer et al., 2024), and distribution matching (Yin et al., 2024a;b; Jiang et al., 2025; Wu et al., 2026).

Nevertheless, a standard training recipe is not a safe bet: directly applying even widely adopted methods such as distribution matching distillation (DMD, Yin et al. 2024a;b) to large-scale visual generative models across broad and heterogeneous scenarios can lead to severe degradation or outright training collapse, as illustrated in Figure 2. This observation suggests that the *distillation objective is only part of the story*: effective few-step distillation also hinges on the **training recipe** used to put it into practice.

Motivated by this gap, we use Qwen-Image-2.0 (Zhao et al., 2026) as a representative large-scale visual generative foundation model to systematically investigate the DMD training recipe. For comprehensive evaluation, we establish two complementary benchmarks: T2I-Bench, with $1{,}800$ cases evenly distributed across three representative categories, and Editing-Bench, with $1{,}500$ challenging editing cases spanning six representative categories. Using this testbed, we examine three key dimensions: *training data composition*, which defines the text-to-image (T2I) distillation distribution; *teacher guidance within DMD*, which governs the incorporation of teachers with different capabilities; and *task mixture*, which controls the joint optimization of T2I generation and image editing.

Our empirical analysis reveals three key findings. First, T2I distillation is highly sensitive to training data composition: neither greater data diversity nor stronger alignment with target scenarios necessarily improves performance, whereas semantically coherent data from a single category can exhibit surprisingly strong cross-category generalization. Second, incorporating post-trained teachers with enhanced capabilities into DMD is non-trivial, as naively using them throughout distillation can destabilize training and lead to severe structural collapse. We find that restricting post-trained teacher guidance to the final step effectively transfers their specialized capabilities while preserving the structural stability inherited from the base teacher. Third, in joint T2I-editing distillation, task mixture emerges as a critical factor, with a balanced T2I-to-editing training ratio yielding the strongest overall performance across both tasks.

Building on these findings, we develop **Qwen-Image-Flash**, a unified few-step model for both T2I generation and image editing. As illustrated in Figure 1, Qwen-Image-Flash achieves high-quality generation and editing with only 4 NFEs, while retaining strong visual fidelity and versatile synthesis capabilities across diverse and complex scenarios (*e.g.*, poster generation). Collectively, our findings reframes few-step distillation from a problem centered primarily on objective design to one that equally demands careful consideration of the training recipe.

2 RELATED WORK

Few-step distillation. Efficient visual generation is typically achieved through trajectory-level, distribution-level, or hybrid distillation. Trajectory-level methods compress teacher transitions into fewer student steps, including progressive distillation (Salimans & Ho, 2022), consistency mod-

els (Song et al., 2023; Luo et al., 2023), and flow-based methods (Liu et al., 2023; Geng et al., 2025). However, pointwise imitation may inherit solver errors and overconstrain the student. Distribution-level methods instead align the student and target distributions, as in adversarial diffusion distillation (Sauer et al., 2024) and DMD (Yin et al., 2024a;b), with extensions addressing CFG effects (Liu et al., 2025), stability (Bai et al., 2026), and diversity (Wu et al., 2026). TDM (Luo et al., 2026) bridges both paradigms by matching distributions along the student trajectory.

Benchmarks for T2I generation and editing. T2I models are commonly evaluated using FID and CLIP on MS-COCO (Lin et al., 2014), as well as fine-grained benchmarks such as GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023). However, existing protocols do not fully capture few-step distillation failures in text rendering, composition, prompt adherence, diversity, and visual detail, nor do they jointly evaluate T2I generation and instruction-guided editing. We therefore introduce T2I-Bench and Editing-Bench, two challenging benchmarks that characterize the capabilities and degradation patterns of few-step visual generative models across both tasks.

3 Preliminaries: Flow Matching and DMD

We briefly review two components used in this work: flow matching (Lipman et al., 2022), a continuous-time framework for learning generative dynamics, and DMD (Yin et al., 2024a), which we adopt to distill a multi-step teacher into a few-step student.

3.1 Flow Matching

Flow matching defines a transport process between data and noise by prescribing a probability path and then learning the velocity field along this path. Let $\boldsymbol{x} \sim p_{\text{data}}$ denote a data point and let $\boldsymbol{\epsilon} \sim p_{\text{noise}}$ be an independent noise sample, where p_{noise} is typically set to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In this work, following (Liu et al., 2023; Geng et al., 2025), we use the linear path $\boldsymbol{z}_t = (1-t)\boldsymbol{x}+t\boldsymbol{\epsilon}$ for $t \in [0, 1]$, which connects the data distribution at $t = 0$ to the noise distribution at $t = 1$. Under this linear interpolation, the target velocity along each path is $\boldsymbol{\epsilon} - \boldsymbol{x}$.

For conditional generation, the vector field is additionally conditioned on $\boldsymbol{c}$, which may encode a text prompt and, in image editing, a source image together with the editing instruction. Flow matching then trains a parameterized conditional vector field $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c})$ to recover the target velocity:

$$\ell_{\text{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{\epsilon}}\Big[\big\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}) - (\boldsymbol{\epsilon} - \boldsymbol{x})\big\|^2\Big]. \tag{1}$$

At inference, samples are generated by drawing $\boldsymbol{z}_1 \sim p_{\text{noise}}$ and integrating the learned ODE from $t = 1$ to $t = 0$, yielding $\boldsymbol{x}_{\boldsymbol{\theta}} = \boldsymbol{z}_1 + \int_1^0 \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c})\, dt$.

3.2 DMD Objective

DMD is designed to distill a pretrained multi-step teacher into a conditional student generator $G_{\boldsymbol{\theta}}$. Given an input noise variable $\boldsymbol{\epsilon}$ and condition $\boldsymbol{c}$, the student produces a clean sample $\boldsymbol{x}_{\boldsymbol{\theta}} = G_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \boldsymbol{c})$. To compare the student with the teacher at noisy intermediate states, an independent noise sample $\boldsymbol{\xi} \sim p_{\text{noise}}$ and a noise level $t \sim p_t$ are sampled, where p_t denotes a prescribed distribution over $[0, 1]$. The student output is then perturbed as $\boldsymbol{x}_t = (1 - t)\boldsymbol{x}_{\boldsymbol{\theta}} + t\boldsymbol{\xi}$.

At a high level, DMD encourages the student-induced conditional distribution to match that of the teacher by minimizing the reverse KL objective $\ell_{\text{DMD}}(\boldsymbol{\theta}) \triangleq D_{\text{KL}}\big(p_{\text{stu}}(\boldsymbol{x}_{\boldsymbol{\theta}} \mid \boldsymbol{c}) \,\|\, p_{\text{tea}}(\boldsymbol{x}_{\boldsymbol{\theta}} \mid \boldsymbol{c})\big)$. Rather than optimizing this divergence directly, DMD uses a gradient estimator based on the difference between the score field of the student distribution and that of the teacher distribution:

$$\nabla_{\boldsymbol{\theta}}\ell_{\text{DMD}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\xi},t}\Big[\big(\nabla_{\boldsymbol{\theta}}\boldsymbol{x}_{\boldsymbol{\theta}}\big)^{\intercal}\big(\boldsymbol{s}_{\text{stu}}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{s}_{\text{real}}(\boldsymbol{x}_t, t, \boldsymbol{c})\big)\Big]. \tag{2}$$

Here, $\boldsymbol{s}_{\text{stu}}$ is estimated using an auxiliary score network trained on samples generated by the student, while $\boldsymbol{s}_{\text{real}}$ is obtained from the pretrained teacher. The resulting update pushes the student toward regions where its noisy marginal score agrees with the teacher score across sampled noise levels.

Table 1: **T2I distillation with different data compositions.** We evaluate 4-NFE students distilled with different category-specific and mixed-category training sets on landscape, portrait, and text-centric splits of T2I-Bench.

Training data composition	# of training data	Metrics	T2I-Bench			Avg	Rank
			Landscape	Portrait	Text-centric		
Landscape	20,000	Gemini 3.1 Pro	3.53	3.37	3.01	3.30	3
		GPT 5.5	4.30	4.31	3.77	4.13	
Portrait	20,000	Gemini 3.1 Pro	3.56	3.57	3.12	3.42	1
		GPT 5.5	4.35	4.34	3.76	4.15	
Text-centric	20,000	Gemini 3.1 Pro	2.55	3.38	1.97	2.63	5
		GPT 5.5	3.34	3.88	2.64	3.29	
Landscape-portrait	40,000	Gemini 3.1 Pro	3.61	3.54	3.04	3.40	2
		GPT 5.5	4.24	4.33	3.62	4.06	
Mixed-category	60,000	Gemini 3.1 Pro	3.53	3.47	2.05	3.02	4
		GPT 5.5	4.08	4.23	2.54	3.62	

4 EXPERIMENTS

This section introduces T2I-Bench and Editing-Bench for few-step generation and editing, then uses Qwen-Image-2.0 (Zhao et al., 2026) as a testbed to systematically study how training data composition, teacher guidance, and task mixture shape DMD, culminating in **Qwen-Image-Flash**.

4.1 BENCHMARKS AND EVALUATION PROTOCOL

T2I-Bench. T2I-Bench is a challenging evaluation suite covering the same three categories considered in our data-composition study. It comprises 1,800 evaluation cases, evenly distributed across categories with 600 cases each. We employ Gemini 3.1 Pro and GPT 5.5 as automatic preference-based evaluators, where higher scores indicate superior perceptual quality, prompt alignment, and overall alignment with human preferences. The rubric-based evaluation prompt and additional details are provided in the Appendix.

Editing-Bench. Editing-Bench comprises 1,500 challenging cases evenly distributed across six representative categories: scene-level semantic transformation, perceptual enhancement, object-centric manipulation, textual content editing, identity-preserving editing, and stylistic transfer. With 250 cases per category, it supports balanced evaluation of global and local editing capabilities. Following T2I-Bench, we employ Gemini 3.1 Pro and GPT 5.5 as automatic evaluators, with higher scores indicating better instruction following, source-image preservation, artifact suppression, and alignment with human preferences. Category-specific system prompts capture distinct task-level quality criteria. Further details appear in the Appendix.

4.2 DATA COMPOSITION MATTERS IN T2I DISTILLATION

Experimental setup. We use Qwen-Image-2.0-Base (Zhao et al., 2026) as the multi-step teacher and distill it into a 4-NFE student with DMD (Yin et al., 2024a). To construct the training data for distillation, we use Qwen3 (Yang et al., 2025) to generate 20,000 diverse prompts for each of three representative categories: landscapes, portraits, and text-centric scenarios. From these category-specific training sets, we derive five data compositions with varying category coverage: landscape-only, portrait-only, text-centric-only, landscape-portrait, and a mixture of all three categories. All students are trained under the same optimization protocol for 2,000 iterations using AdamW (Loshchilov & Hutter, 2017), with a learning rate of 1×10^{-6} and a batch size of 64 for both the student and the auxiliary score network, so that performance differences can be attributed primarily to the training data composition.

Target-aligned data can be unexpectedly detrimental. Table 1 shows that the text-centric setting provides the most direct exposure to text-heavy samples yet yields the lowest overall performance.



Figure 2: **Qualitative results under different data compositions.** We compare students distilled with text-centric, mixed-category, landscape-only, landscape-portrait, and portrait-only training data across representative evaluation scenarios.

Surprisingly, this degradation extends in-domain: on the text-centric split, it is substantially outperformed by both landscape- and portrait-based distillation, neither of which uses explicit text-centric training data. The mixed-category setting exhibits a similar trend: despite using the largest training set and covering all three categories, it underperforms the stronger single-category configurations, particularly on text-centric evaluation. These results challenge the *common intuition* from large-scale pretraining that better target coverage and greater data diversity should consistently improve performance. In few-step distillation, certain data distributions may instead hinder distribution transfer, as also illustrated qualitatively in Figure 2.

Takeaway 1: Unlike the common scaling intuition in pretraining, broader or target-aligned data does not necessarily benefit few-step distillation and may even hurt student performance.

Category-specific data can support broad transfer. In contrast, single-category distillation exhibits strong cross-category generalization. Landscape- and portrait-based students remain competitive beyond their training domains and even outperform text-centric and mixed-category settings on text-centric evaluation without seeing such prompts during training. Moreover, combining landscape and portrait data provides no further gain over portrait-only distillation. These results suggest that explicit downstream coverage is not necessary for effective few-step distillation, and that broader category coverage does not necessarily translate into better transfer.

Takeaway 2: Data selection is critical in few-step T2I distillation: single-category data can generalize broadly, while broader category combinations do not guarantee further gains.

4.3 Transferring Post-trained Capabilities through Teacher Guidance

Beyond training data composition, effective distillation also depends on how teacher supervision is incorporated. As post-training becomes increasingly common for enhancing visual generative models, a practical question arises: whether and how these improvements can be retained through few-step distillation. We therefore consider a post-trained multi-step teacher and ask: *how should its supervision be used to reliably transfer these gains to a few-step student?* This motivates the second dimension of our DMD (Yin et al., 2024a) training recipe: *teacher guidance*.

Experimental setup. Following the post-training paradigm of Xu et al. (2026), we further post-train Qwen-Image-2.0-Base (Zhao et al., 2026) to obtain a task-specialized teacher, termed Qwen-Image-2.0-RL, with enhanced downstream capabilities. We adopt the same experimental protocol as in Section 4.2, using the portrait-only training set and identical optimization hyperparameters.

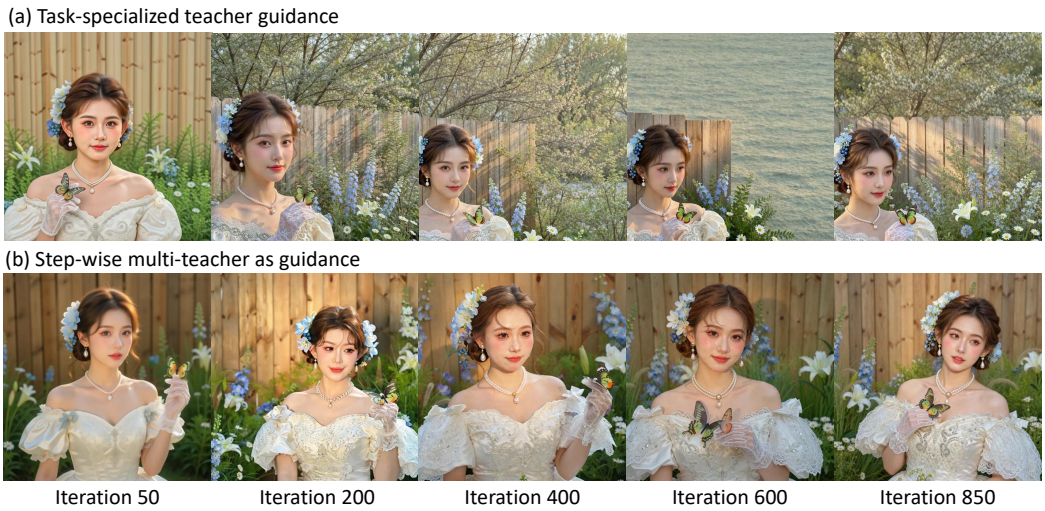

Figure 3: **Qualitative comparison of teacher guidance strategies during distillation.** (a) Direct guidance from a task-specialized teacher can destabilize training, leading to progressive degradation in alignment and visual quality. (b) Step-wise multi-teacher guidance maintains sample fidelity and layout consistency throughout distillation, yielding better-aligned generations.

Direct post-trained teacher guidance destabilizes distillation. As shown in Figure 3 (a), directly distilling from the task-specialized teacher results in pronounced optimization instability. While the student initially exhibits task-specific gains, generation quality progressively degrades during training, manifesting as structural distortion, reduced visual fidelity, and weaker semantic alignment. This observation indicates that stronger downstream teacher performance does not necessarily translate into more effective guidance for few-step distillation. We conjecture that post-training induces a distribution that is more difficult for a highly compressed student to approximate, thereby amplifying the mismatch between teacher and student score fields under DMD and ultimately leading to unstable optimization or collapse.

Takeaway 3: Directly using a post-trained teacher as the sole source of guidance can destabilize few-step distillation, despite its stronger downstream performance.

Step-wise multi-teacher guidance. Motivated by this observation, we retain the pretrained base teacher as a stable distributional anchor within DMD and selectively incorporate post-trained, task-specialized guidance across student distillation steps. This recipe draws on recent on-policy distillation methods (Xiao et al., 2026; Li et al., 2026b; Fang et al., 2026), which emphasize adapting supervision to the evolving student policy.

More generally, let T_0 denote the base teacher and $\{T_m\}_{m=1}^{M}$ denote task-specialized teachers. At the k-th selected student distillation step, we construct the DMD real-score target as $\boldsymbol{s}_{\text{real}}^{(k)}(\boldsymbol{x}_t, t, \boldsymbol{c}) = \sum_{m=0}^{M} \lambda_{k,m}(\boldsymbol{c})\, \boldsymbol{s}_m^{(k)}(\boldsymbol{x}_t, t, \boldsymbol{c})$, where $\lambda_{k,m}(\boldsymbol{c}) \in [0,1]$ controls the contribution of teacher T_m and satisfies $\sum_{m=0}^{M} \lambda_{k,m}(\boldsymbol{c}) = 1$. In this work, we consider the simplest non-trivial case, $M = 1$, with one pretrained base teacher and one post-trained task-specialized teacher. Step-dependent weighting preserves stable base supervision while selectively introducing specialized capabilities along the student trajectory. The resulting optimization at step k is formulated as

$$\nabla_{\boldsymbol{\theta}} \ell_{\text{DMD}}^{(k)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\xi},t}\Big[\big(\nabla_{\boldsymbol{\theta}} \boldsymbol{x}_{\boldsymbol{\theta}}\big)^{\mathsf{T}}\big(\boldsymbol{s}_{\text{stu}}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \sum_{m=0}^{M} \lambda_{k,m}(\boldsymbol{c}) \boldsymbol{s}_m^{(k)}(\boldsymbol{x}_t, t, \boldsymbol{c})\big)\Big]. \tag{3}$$

Training hyperparameters. We instantiate step-wise multi-teacher guidance with a 4-step student, *i.e.*, $k \in \{1, 2, 3, 4\}$. To mitigate the severe structural degradation observed in Figure 3 (a)

Table 2: **Quantitative comparison of teachers and student.** With only 4 NFEs, the distilled model achieves competitive performance against 80-NFE teachers, effectively inheriting their complementary strengths across landscape, portrait, and text-centric evaluation sets.

Model	NFEs	Metrics	T2I-Bench			Avg	Rank
			Landscape	Portrait	Text-centric		
Qwen-Image-2.0-Base	80	Gemini 3.1 Pro	3.52	3.64	3.08	3.41	3
		GPT 5.5	4.24	4.30	3.73	4.09	
Qwen-Image-2.0-RL	80	Gemini 3.1 Pro	3.98	3.82	3.41	3.74	1
		GPT 5.5	4.34	4.47	3.88	4.26	
Qwen-Image-Flash-T2I	4	Gemini 3.1 Pro	3.88	3.81	3.00	3.56	2
		GPT 5.5	4.30	4.41	3.75	4.15	

under direct post-trained teacher guidance, the first three steps use only the pretrained base teacher for stable structural supervision: $\lambda_{k,0} = 1$ and $\lambda_{k,1} = 0$ for $k \in \{1, 2, 3\}$. At the final step, once a stable global structure is established, we equally combine both teachers, setting $\lambda_{4,0} = \lambda_{4,1} = 0.5$. This schedule preserves the base teacher's structural stability while selectively transferring the post-trained teacher's enhanced downstream capabilities, yielding **Qwen-Image-Flash-T2I**.

Main results. Figure 3 (b) shows that step-wise multi-teacher guidance effectively stabilizes student optimization, maintaining sample fidelity, layout consistency, and semantic alignment throughout training, in contrast to the progressive degradation observed with direct specialized-teacher distillation. Table 2 further demonstrates effective capability transfer: with only 4 NFEs, **Qwen-Image-Flash-T2I** achieves average scores of 3.56 with Gemini 3.1 Pro and 4.15 with GPT 5.5, surpassing the 80-NFE Qwen-Image-2.0-Base teacher in the overall ranking. These results suggest that anchoring early distillation with the base teacher while selectively introducing specialized supervision provides an effective balance between structural preservation and downstream capability transfer. The strategy is also *lightweight and general*, requiring only step-dependent guidance weights to incorporate specialized teachers into existing DMD pipelines.

Takeaway 4: Incorporating post-trained teachers is a scheduling problem rather than a replacement problem: anchor early distillation with the base teacher and introduce specialized guidance only at later stages.

4.4 Joint Distillation for Unified Generation and Editing

Beyond T2I generation, the teacher model also supports instruction-guided image editing through the same backbone (Zhao et al., 2026), making both tasks integral to the capabilities we seek to preserve through distillation. We therefore jointly distill generation and editing into a single few-step student to preserve the teacher's unified capabilities. As the tasks provide distinct training signals, overemphasizing either may compromise the other, making task mixture critical. We therefore investigate the T2I-to-editing data ratio to balance performance across both tasks.

Experimental setup. We perform joint distillation with three T2I-to-editing data ratios: 9:1, 7:3, and 5:5, ranging from T2I-dominant to balanced mixtures, using an internal dataset for editing supervision. All other training hyperparameters and settings follow Section 4.3, with the total training budget held constant and only the task ratio varied.

Balanced mixtures improve editing transfer. As shown in Table 3, the T2I-only student retains non-trivial zero-shot editing ability, but still underperforms the specialized teacher, particularly on text editing. Adding limited editing data is insufficient: the 9:1 mixture performs worse than the T2I-only baseline. Performance improves as editing supervision increases, with 7:3 surpassing both the baseline and teacher under Gemini 3.1 Pro, and 5:5 achieving the best overall results. Specifically, the balanced mixture improves the average score from 2.77 to 2.97 under Gemini 3.1 Pro and from 3.28 to 3.41 under GPT 5.5. Figure 4 further confirms its stronger instruction following and source preservation. We therefore adopt the 5:5 task ratio for our final unified model, **Qwen-Image-Flash**.

Table 3: **Effect of T2I:Edit ratios on Editing-Bench across six editing dimensions.** Tea. denotes the multi-step teacher trained on the same internal T2I and editing datasets and distilled for faster inference. Zero-shot denotes Qwen-Image-Flash-T2I without editing distillation.

Ratio	Metrics	Editing-Bench						Avg	Rank
		Scene trans.	Perceptual enhance.	Object manip.	Text editing	Identity preserv.	Style transfer		
Tea.	Gemini 3.1 Pro	2.52	2.75	2.33	3.25	3.16	2.62	2.77	3
	GPT 5.5	3.61	3.22	3.05	3.69	3.53	3.56	3.44	
Zero-shot	Gemini 3.1 Pro	2.78	2.81	2.25	2.84	3.16	2.78	2.77	4
	GPT 5.5	3.57	3.03	3.01	3.11	3.41	3.52	3.28	
9:1	Gemini 3.1 Pro	2.43	2.25	2.09	2.83	3.12	2.75	2.58	5
	GPT 5.5	3.48	3.21	2.87	3.25	3.53	3.50	3.31	
7:3	Gemini 3.1 Pro	2.49	2.80	2.34	3.35	3.18	3.05	2.87	2
	GPT 5.5	3.46	2.87	3.08	3.58	3.52	3.63	3.36	
5:5	Gemini 3.1 Pro	2.86	2.25	2.68	3.18	3.19	3.68	2.97	1
	GPT 5.5	3.66	3.06	3.20	3.13	3.47	3.92	3.41	

Table 4: **Effect of T2I:Edit ratios on T2I-Bench.** We evaluate how editing supervision affects T2I generation in jointly distilled students.

Ratio	Metrics	T2I-Bench			Avg	Rank
		Landscape	Portrait	Text-centric		
Qwen-Image-2.0-RL	Gemini 3.1 Pro	3.98	3.82	3.41	3.74	1
	GPT 5.5	4.34	4.47	3.88	4.26	
Qwen-Image-Flash-T2I	Gemini 3.1 Pro	3.88	3.81	3.00	3.56	5
	GPT 5.5	4.30	4.41	3.75	4.15	
Qwen-Image-Flash (9:1)	Gemini 3.1 Pro	4.04	3.58	3.17	3.60	4
	GPT 5.5	4.39	4.37	3.77	4.18	
Qwen-Image-Flash (7:3)	Gemini 3.1 Pro	3.81	3.77	3.21	3.60	3
	GPT 5.5	4.35	4.40	3.84	4.20	
Qwen-Image-Flash (5:5)	Gemini 3.1 Pro	3.95	3.85	3.14	3.65	2
	GPT 5.5	4.34	4.37	3.76	4.16	

Takeaway 5: Editing transfer is highly sensitive to the task ratio: limited editing data can be ineffective, whereas a balanced mixture yields the strongest editing performance.

Editing supervision improves T2I generation. We also evaluate the T2I performance of the jointly distilled students to examine how introducing editing data affects generation quality. Table 4 shows that all jointly distilled students outperform the T2I-only baseline on average T2I-Bench scores. This positive transfer suggests that editing data provides complementary supervision for instruction grounding, spatial localization, and visual-textual consistency, thereby strengthening prompt following and semantic alignment in T2I generation.

Takeaway 6: Editing supervision provides positive cross-task transfer, improving T2I generation rather than merely preserving it.

5 Discussion and Conclusion

Limitations and future work. Despite competitive T2I generation and instruction-guided editing with only 4 NFEs, Qwen-Image-Flash exhibits minor text-rendering artifacts and residual noise. Our study is primarily constrained by its empirical scope, covering a single model family, a fixed

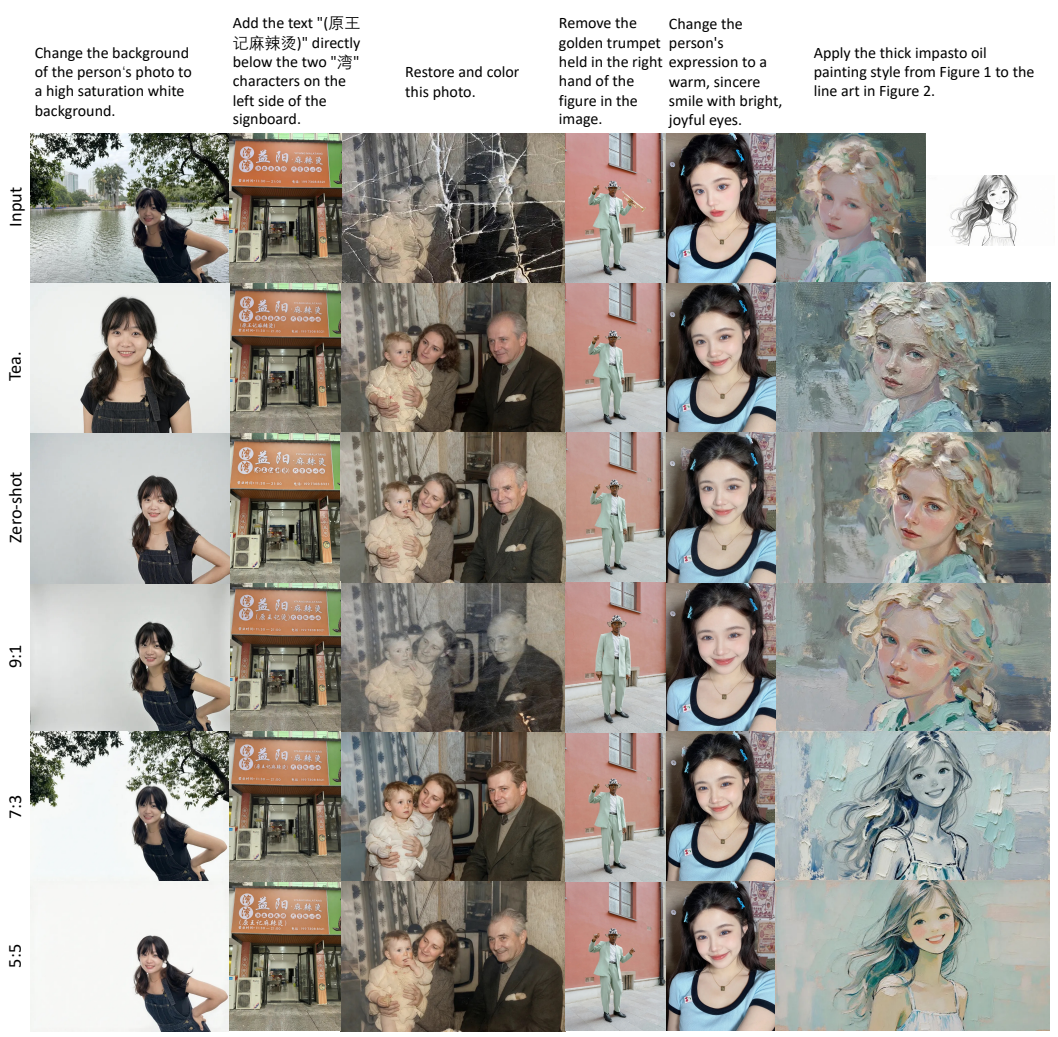

Figure 4: **Qualitative comparison under different T2I:Edit ratios.** Across six editing categories, we compare the editing teacher, T2I-only zero-shot student, and jointly distilled students trained with 9:1, 7:3, and 5:5 mixtures. The balanced 5:5 ratio best improves instruction following while preserving fidelity, identity, and style.

sampling budget, and limited teacher schedules. Inspired by the step-specific roles in DP-DMD (Wu et al., 2026), future work will explore combining trajectory matching for structural preservation with distribution matching for fine-detail refinement to improve training stability and visual fidelity under aggressive sampling compression.

Conclusion. In this work, we systematically revisit the training recipe for few-step distillation of visual generative foundation models, using Qwen-Image-2.0 as an empirical testbed within the DMD framework. Our study identifies *data composition, teacher guidance, and task mixture* as critical determinants of effective capability transfer: broader or target-aligned data can hinder distillation, step-wise multi-teacher guidance reconciles structural stability with post-trained capabilities, and balanced task supervision preserves unified generation and editing. These insights culminate in **Qwen-Image-Flash**, which retains strong T2I generation and instruction-guided editing performance with only 4 NFEs. Collectively, our findings broaden the perspective on few-step distillation by demonstrating how the training recipe shapes the realization of a distillation objective. As visual foundation models grow in capability and scope, principled **training recipe** design will be essential to translating their advances into efficient, stable, and broadly capable generative systems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research code implementation. We have not used generative AI tools to develop conceptual frameworks, formulate hypotheses, design experiments, or interpret results; tasks related to theoretical proofs are not applicable to this work. Additionally, we used generative AI tools to polish the manuscript for clarity, readability, and linguistic consistency. We have carefully reviewed all AI-assisted work, including manually inspecting the generated code for correctness and reviewing text revisions for accuracy and consistency with our intended scientific meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work advances inference acceleration for visual generative models, improving the efficiency and accessibility of image generation and editing. Our methods are intended for legitimate research and beneficial applications, not harmful or deceptive uses. We advocate responsible use that respects privacy, intellectual property, and ethical standards.

REPRODUCIBILITY STATEMENT

The main paper describes the distillation objective, training data compositions, teacher guidance schedule, task mixtures, and optimization settings used in our experiments. Benchmark descriptions and evaluation protocols are provided in the main text, with detailed evaluator prompts and category-specific rubrics in the Appendix. These details support reproducibility, although full replication requires access to the underlying teacher models and internal editing data.

REFERENCES

Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997.

Lichen Bai, Zikai Zhou, Shitong Shao, Wenliang Zhong, Shuo Yang, Shuo Chen, Bojun Chen, and Zeke Xie. Optimizing few-step generation with adaptive matching distillation. *arXiv preprint arXiv:2602.07345*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow Transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

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.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, pp. 78723–78747, 2023.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

Zhangqi Jiang, Zheng Sun, Xianfang Zeng, Yufeng Yang, Xuanyang Zhang, Yongliang Wu, Wei Cheng, Gang Yu, Xu Yang, and Bihan Wen. GEditBench v2: A human-aligned benchmark for general image editing. *arXiv preprint arXiv:2603.28547*, 2026.

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.

Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36:20662–20678, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pp. 740–755, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Machine Learning*, 2022.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, et al. Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025.

Jiaxiang Liu, Zhida Feng, Pengyu Zou, Zhenyu Qian, Tianrui Zhu, Jun Xia, Yuehu Dong, Yanzheng Lin, Honglin Xiong, et al. ERNIE-Image technical report. *arXiv preprint arXiv:2605.25347*, 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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yihong Luo, Tianyang Hu, Weijian Luo, and Jing Tang. TDM-R1: Reinforcing few-step diffusion models with non-differentiable reward. *arXiv preprint arXiv:2603.07700*, 2026.

Chaojie Mao, Chen-Wei Xie, Chongyang Zhong, Haoyou Deng, Jiaxing Zhao, Jie Xiao, Jinbo Xing, Jingfeng Zhang, Jingren Zhou, Jingyi Zhang, et al. Wan-Image: Pushing the boundaries of generative visual intelligence. *arXiv preprint arXiv:2604.19858*, 2026.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103, 2024.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. JoyAI-Image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

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*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. MiMo-V2-Flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Yixian Xu, Kaiyuan Gao, Yuxiang Chen, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, et al. Qwen-Image-2.0-RL technical report. *arXiv preprint arXiv:2606.27608*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, pp. 47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Yang Zhao, Yanwu Xu, Zhisheng Xiao, Haolin Jia, and Tingbo Hou. Mobilediffusion: Instant text-to-image generation on mobile devices. In *European Conference on Computer Vision*, pp. 225–242, 2024.

APPENDIX

A BENCHMARK EVALUATION PROTOCOLS

This section details the evaluation protocols for T2I-Bench and Editing-Bench, including evaluator prompts, scoring rubrics, and structured output formats.

A.1 EVALUATOR PROMPTS AND SCORING RUBRICS

We employ Gemini 3.1 Pro and GPT 5.5 as automatic evaluators, guided by benchmark-specific instructions that specify the assessment criteria and scoring principles. The full prompts and task-specific rubrics are provided below.

T2I-Bench evaluation. Given a text prompt and its generated image, each evaluator follows the system prompt in Table A to assess two equally important dimensions: *prompt compliance* and *technical quality*. The former examines whether the image correctly depicts the specified objects, attributes, quantities, actions, and spatial relationships. The latter assesses structural and perceptual defects, including geometric inconsistencies, anatomical errors, corrupted text, and texture degradation. Aesthetic appeal should not compensate for failures to satisfy the requested content.

Table B provides two challenging examples involving detailed portrait specifications and text-rich structured compositions. Each assessment is returned as strict JSON containing an overall score from 1 to 5, rounded to two decimal places, and a concise, single-sentence justification.

Editing-Bench evaluation. For Editing-Bench, each evaluator receives the source image, editing instruction, and edited output. We use a shared system prompt template with category-specific rubrics to assess both the requested modification and the preservation of source content that should remain unchanged.

The template in Table C specifies the common scoring scale, output schema, and scoring principles. For each sample, it is instantiated according to the task category through three placeholders: `<category-title>`, `<category-rubric>`, and `<sub-score-criteria>`. Table D defines the category-specific assessment priorities. For example, perceptual enhancement emphasizes faithful detail recovery without hallucinated content, whereas object-centric manipulation emphasizes accurate localization and natural integration.

Table E lists the corresponding criteria for individual sub-scores, providing diagnostic detail beyond the overall assessment. The evaluator returns strict JSON containing an overall score from 1 to 5, rounded to two decimal places, criterion-level sub-scores, a single-sentence justification, and identified failure modes. The overall score reflects a holistic assessment rather than an arithmetic average of sub-scores, with severe failures in core requirements, such as instruction following, identity preservation, or text accuracy, strongly limiting the final score.

Table A: **T2I-Bench evaluation.** System prompt and scoring criteria.

You are an expert evaluator for text-to-image generation results. Rate the image based on two equally critical dimensions: instruction following and visible technical quality. Do not reward an image for being merely beautiful if it fails to depict the prompt accurately. **Prompt Compliance**: Verify if the image precisely reflects all objects, attributes, quantities, actions, and spatial relationships described in the caption. **Technical Quality**: Focus on structural and perceptual defects: distorted geometry, warped or broken objects, inconsistent perspective, bad anatomy, face, hand, or limb deformation, texture melting, duplicated or missing parts, corrupted text or logos, blur, noise, compression artifacts, aliasing, exposure problems, color casts, unrealistic lighting, and other image-generation artifacts. Return strict JSON only, with no markdown: {"score": <float from 1 to 5 rounded to two decimals>, "reasoning": "<one short English sentence>"}.

Table B: **Challenging prompts from T2I-Bench.** Randomly selected examples illustrating complex generation requirements across diverse scenarios.

Sample	Prompt
1	A young East Asian woman, about 20 to 28 years old, sits gracefully and casually on outdoor cement steps. She has medium-length wavy black hair with slightly curled ends falling naturally over both shoulders, a soft oval face, fair smooth skin without obvious moles or freckles, and clean natural makeup: light brown eyeshadow, thin inner eyeliner, thick curled lashes, soft pink-orange blush, matte coral-pink lips, a slight smile, and a calm cheerful gaze looking toward the upper right. She wears a white cotton camisole dress with two three-dimensional white fabric flowers on the chest and a thin drawstring tie at the neckline, plus a pale sky-blue openwork crochet cardigan with elbow-length sleeves, clear mesh texture, and a light breathable material. Her accessories include one blue-green gradient feather earring on the left ear, bright cyan-blue at the top and deep teal at the bottom, about 5 cm long, and a matching thin cord choker with a small metal clasp. Her right hand gently supports her chin, and her left hand holds an artificial bouquet placed on her lap: seven blue five-petal flowers with slightly purple petal edges, deep blue star-shaped centers, green leaves, and several white buds. The background is a traditional tie-dye craft display. In the upper-left area hangs a white sign with black text: the first larger bold line means printing and tie-dye, and the second slightly smaller line means handmade craft; the text is sharp and horizontal. Behind and to the right are several hanging tie-dye fabrics on wooden racks: a dark indigo radial spiral cloth on the left, an orange-red cloth with white spiral tie-dye near the center-right, and a dark blue cloth with white concentric-circle and radial patterns on the right, all slightly wrinkled. The ground is gray cement steps and pavement. Sunlight comes diagonally from the upper right, casting clear rectangular light patches. Use natural daylight, soft contrast, a fresh artistic handmade-craft atmosphere, a medium portrait crop from waist to mid-thigh, the subject slightly left of center with empty space on the right, a standard non-distorted lens, moderate depth of field, and a blue-white-orange color palette with a fresh healing visual style.
2	Create a mind map or infographic about the book The Story of Art. At the top center, place the main title The Story of Art, with the English subtitle in parentheses below it, followed by a line explaining that it is a classic narrative of art history. The main body is a relationship network centered on a portrait of E.H. Gombrich, labeled with his name. Around him are four famous artists connected by lines. In the upper left is Leonardo da Vinci, represented by the Mona Lisa, labeled with his name and described as observing nature, pursuing perfection, and being a polymathic genius; the lines between him and Gombrich are labeled analysis and interpretation, and another line points to Raphael below with a tribute label. In the lower left is Raphael, represented by an angel image, labeled with his name and described as harmonious, elegant, ideal beauty, and a master painter; his connection to Gombrich is labeled tribute and inheritance. In the upper right is Michelangelo, represented by the head of David, labeled with his name and described as ambitious, shaping power, and master sculptor; his connection to Gombrich is labeled tribute and interpretation, and another line points down to Rembrandt with a tribute label. In the lower right is Rembrandt, represented by a self-portrait, labeled with his name and described as capturing the soul, master of light and shadow, and depicting humanity; his connection to Gombrich is labeled tribute and inheritance. Across the middle of the image, draw a left-to-right wavy timeline of art history: prehistoric art with a cave-painting bull icon and a note about cave paintings and statues; ancient Egypt and Mesopotamia with pyramid and sphinx icons and a note about eternity and order; Greece and Rome with a temple icon and a note about ideal and reality; the Middle Ages with a castle icon and a note about faith and symbolism; the Renaissance with a portrait icon and a note about humanism and rebirth; Baroque and Rococo with an ornamental badge icon and a note about drama and decoration; Neoclassicism and Romanticism with a profile icon and a note about reason and emotion; and Modern Art with an atom icon and a note about experimentation and abstraction. Below the timeline, add a centered sentence saying that art is not static but a constantly developing and changing story. At the bottom left, create a Core Themes section with five icons and labels: art-history primer, visual literacy, classic work, insight, and human spirit. At the bottom right, create a Character Cards section with two dark rectangular cards: one for Leonardo da Vinci with a side-profile silhouette and traits of curiosity, broad learning, and perfectionism, and one for Michelangelo with a side-profile silhouette and traits of determination, passion, and great ambition.

Table C: **Unified evaluation template for Editing-Bench.** System prompt incorporating category-specific rubrics and scoring criteria for instruction-guided image editing.

You are an expert evaluator for image editing results. You will receive one or more input/source images, the edit instruction, and one or more output image produced by an image editing model. Evaluate whether the output image correctly follows the requested edit while preserving the parts of the source image that should remain unchanged. Do not reward an image for being merely beautiful if it fails the edit. Do not penalize stylistic choices unless they conflict with the instruction or introduce visible artifacts. Use the full 1 to 5 scale.
 	
Return strict JSON only, with no markdown and no extra text. The JSON schema is:
{"overall_score": <float from 1 to 5 rounded to two decimals>, "sub_scores": {"criterion_name": <float from 1 to 5>}, "reasoning": "<one concise English sentence>", "failure_modes": ["<short phrase>"]}
 	
Score meanings: 5 = excellent edit with correct instruction following, source preservation, natural integration, and minimal artifacts; 4 = good with minor defects; 3 = partially correct but with clear issues; 2 = mostly failed or visibly broken; 1 = unusable, ignores the instruction, or severely corrupts the image.
 	
Category-specific rubric: `<category-title>`
`<category-rubric>`
 	
Sub-score criteria to include exactly with these keys:
`<sub-score-criteria>`
 	
The overall_score should be a holistic score, not a simple arithmetic average. However, severe failure in instruction following, identity preservation, or text accuracy should strongly cap the overall score.

Table D: **Category-specific evaluation rubrics for Editing-Bench.** Detailed values for `<category-title>` and `<category-rubric>` in the unified system prompt.

`<category-title>`	`<category-rubric>`
Scene-level semantic transformation	For this category, prioritize whether the whole scene transformation is semantically correct and physically coherent. A high score requires the edited scene to look like a single naturally captured image, not a pasted collage.
Perceptual image enhancement	For this category, do not reward hallucinated new content. Enhancement should improve clarity while keeping the source image faithful.
Object-centric manipulation	For deletion, the removed region should be plausibly filled. For addition/replacement, the new object must be recognizable and naturally integrated.
Textual content editing	For this category, text correctness and readability are critical. A visually pleasant image with wrong or unreadable text should receive a low score.
Identity-preserving editing	For this category, identity preservation is essential. Penalize face drift, identity change, unnatural anatomy, and over-editing of unrelated regions.
Stylistic transfer	For this category, style should change the appearance without destroying the source content or making the image structurally inconsistent.

Table E: **Category-specific scoring criteria for Editing-Bench.** Detailed values for <sub-score-criteria> in the unified system prompt, capturing distinct quality requirements across editing tasks to support consistent and fine-grained assessment.

Category	<sub-score-criteria>
Scene-level semantic transformation	- Instruction following: Does the output implement the requested scene/background/composition transformation? - Source subject preservation: Are required people/objects/identity/clothing/details from the input preserved? - Global consistency: Are perspective, scale, lighting, shadows, and camera viewpoint coherent after the scene change? - Boundary integration: Are masks, edges, occlusion, and foreground-background transitions clean? - Visual quality: Is the final image free of distortions, texture melting, blur, noise, and generation artifacts?
Perceptual image enhancement	- Instruction following: Does the output perform the requested enhancement, restoration, or super-resolution operation? - Content preservation: Does it preserve the original content, identity, geometry, colors, and layout unless the instruction asks otherwise? - Detail recovery: Are details sharper and cleaner without hallucinated or over-smoothed structures? - Artifact control: Does it avoid ringing, oversharpening, waxy texture, noise amplification, and compression artifacts? - Visual quality: Is the enhanced output natural, high-quality, and technically clean?
Object-centric manipulation	- Instruction following: Is the requested object addition, deletion, replacement, or attribute modification correctly completed? - Target localization: Is the edit applied to the intended object/region without unintended changes elsewhere? - Source preservation: Are unrelated source content, identity, background, and composition preserved? - Physical integration: Do object scale, pose, lighting, shadows, contact, occlusion, and perspective fit the scene? - Visual quality: Are there no obvious seams, broken structure, duplicated parts, blur, or artifacts?
Textual content editing	- Instruction following: Does the output place, replace, remove, or modify the requested text exactly as instructed? - Text accuracy: Are spelling, characters, capitalization, punctuation, and requested wording correct? - Text readability: Is the text legible, sharp, correctly oriented, and not garbled or pseudo-text? - Layout style match: Does the text match the original layout, font style, perspective, material, lighting, and surface geometry? - Source preservation: Are non-text regions and unrelated details preserved? - Visual quality: Is the final image free of artifacts, blur, warped letters, and unnatural overlays?
Identity-preserving editing	- Instruction following: Does the output perform the requested clothing, pose, expression, hair, makeup, age, or face-related edit? - Identity preservation: Does the person keep the same recognizable identity and key facial/body characteristics where required? - Attribute accuracy: Are the requested changed attributes accurate and complete? - Anatomy and face quality: Are face, hands, limbs, pose, gaze, expression, and body proportions natural and undistorted? - Source preservation: Are unrelated clothing, background, accessories, and composition preserved unless the instruction changes them? - Visual quality: Is the final portrait technically clean, coherent, and artifact-free?

Continued on next page

Table E (continued)

Category	<sub-score-criteria>
Stylistic transfer	- Instruction following: Does the output apply the requested style, color tone, filter, or artistic transformation? - Content preservation: Are the source subject, structure, layout, identity, and important details preserved? - Style consistency: Is the style/tone applied consistently across the image without patchy or conflicting regions? - Naturalness: Are lighting, contrast, saturation, texture, and material appearance coherent after the transfer? - Visual quality: Is the output free of over-processing, color banding, texture collapse, blur, and artifacts?

B Public Benchmark Evaluation

This section evaluates **Qwen-Image-Flash** on Qwen-Image-Bench (Li et al., 2026a) for T2I generation and GEditBench v2 (Jiang et al., 2026) for image editing. These public benchmarks complement our diagnostic evaluations by assessing the final unified model under independent evaluation protocols and contextualizing its performance against established models.

Results on Qwen-Image-Bench. Qwen-Image-Bench assesses five dimensions: image quality, aesthetics, prompt alignment, real-world fidelity, and creative generation. Table F compares **Qwen-Image-Flash** with representative generative models, with particular attention to Qwen-Image-2.0-Base and Qwen-Image-2.0-RL (Xu et al., 2026) as references for examining capability retention within the same model family.

With only 4 NFEs, Qwen-Image-Flash achieves an overall score of 56.18, exceeding the 80-NFE Base model by 0.95 points and outperforming FLUX 2 Max (55.33) and GPT Image 1 (54.07). Its gains over Base extend across all five dimensions, with the largest improvements in real-world fidelity ($+2.09$), prompt alignment ($+1.20$), and image quality ($+1.17$). This pattern indicates that the overall improvement is accompanied by gains in both semantic adherence and visual fidelity, without an observed regression in the remaining dimensions.

Relative to the RL reference, Qwen-Image-Flash trails by 1.66 points overall, but the differences vary substantially across dimensions. Prompt alignment approaches the RL score (58.84 versus 59.28), whereas creative generation exhibits the largest remaining gap (59.23 versus 64.94). This uneven profile suggests that different aspects of post-trained performance are retained to different degrees under the current recipe, with creative generation leaving the greatest room for improvement. Overall, the public evaluation supports the effectiveness of the final model beyond our diagnostic benchmarks, while revealing capability-specific differences that are obscured by the aggregate score.

Table F: **Comparison on Qwen-Image-Bench.** Results cover five complementary evaluation dimensions and the overall score. With only 4 NFEs, Qwen-Image-Flash consistently surpasses the Base reference across all dimensions and achieves an overall score between the Base and RL references, demonstrating competitive generation performance under aggressive sampling compression. Higher scores indicate better performance.

Model	Dimension Score					Overall ↑
	Quality ↑	Aesth. ↑	Align. ↑	Fidelity ↑	Creative ↑	
Nano Banana Pro	55.67	60.26	61.25	54.07	66.23	59.45
FLUX 2 Max	53.64	56.85	57.35	49.35	56.50	55.33
GPT Image 1	52.34	55.09	56.28	48.14	55.78	54.07
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
Qwen-Image-Flash (Ours)	**53.46**	**57.49**	**58.84**	**49.63**	**59.23**	**56.18**

Table G: **Comparison on GEditBench v2.** Models are ranked by Overall Elo from deduplicated GPT-5.5 pairwise judgments, with 95% confidence intervals (CI), preference rates (Pref.), and match counts reported. Higher Elo and preference rates indicate better performance.

Rank	Model	Source	Samples	Overall Elo ↑	95% CI	Pref. ↑	Matches
1	Nano Banana Pro	Closed	1156	1218	$-12/+14$	0.7703	20686
3	GPT Image 1.5	Closed	1081	1095	$-12/+13$	0.6246	19413
4	FLUX.2 [klein] 9B	Open	1200	1088	$-11/+11$	0.6167	21427
7	FLUX.2 [dev] Turbo	Open	1200	1079	$-11/+11$	0.6041	21427
8	FLUX.2 [klein] 4B	Open	1200	1054	$-10/+11$	0.5710	21427
9	Qwen-Image-Edit-2511	Open	1200	1042	$-9/+10$	0.5548	21427
10	Seedream 4.5	Closed	1190	1036	$-13/+13$	0.5473	21259
11	FLUX.2 [dev]	Open	1200	1030	$-11/+12$	0.5382	21427
12	LongCat-Image-Edit	Closed	1200	1019	$-10/+10$	0.5238	21427
13	Qwen-Image-Edit-2509	Open	1200	986	$-9/+9$	0.4782	21427
14	Qwen-Image-Edit	Open	1200	960	$-10/+10$	0.4439	21427
15	Step1X-Edit-v1p2	Open	1200	890	$-14/+13$	0.3517	21427
16	GLM-Image	Open	1200	868	$-12/+13$	0.3263	21427
17	BAGEL	Open	1200	810	$-13/+12$	0.2596	21427
18	FLUX.1 Kontext [dev]	Open	1200	785	$-12/+11$	0.2327	21427
19	OmniGen2	Open	1200	760	$-13/+13$	0.2088	21427
6	Qwen-Image-2.0-Base	Teacher	1200	1081	$-10/+10$	0.6069	21427
2	Qwen-Image-2.0-RL	Teacher	1200	1112	$-11/+11$	0.6476	21427
5	**Qwen-Image-Flash**	**Ours**	1200	**1087**	$-9/+10$	**0.6152**	21427

Results on GEditBench v2. GEditBench v2 (Jiang et al., 2026) comprises 1,200 real-world editing requests spanning 23 tasks, extending evaluation to a broad range of practical editing scenarios. We assess instruction-guided editing through pairwise judgments from GPT-5.5, reporting Overall Elo, 95% confidence intervals, and preference rates in Table G.

With only 4 NFEs, **Qwen-Image-Flash** achieves an Overall Elo of 1087, closely matching FLUX.2 [klein] 9B (1088). Their similar preference rates (0.6152 and 0.6167, respectively) further indicate comparable aggregate performance under this evaluation protocol. Qwen-Image-Flash also scores above Qwen-Image-2.0-Base (1081) and FLUX.2 [dev] Turbo (1079), although the overlapping confidence intervals warrant interpreting these differences as numerical advantages rather than conclusive improvements.

Within the Qwen-Image-2.0 family, the student falls between the Base and RL references, with a remaining gap of 25 Elo points to the latter (1112). This positioning suggests that the current recipe sustains competitive editing performance under a restricted sampling budget, while leaving room to better retain the stronger teacher's capabilities. The same relative ordering on Qwen-Image-Bench provides complementary evidence that the final unified model remains competitive across both generation and editing under independent public evaluations.