MULTI-RESOLUTION FLOW MATCHING: TRAINING-FREE DIFFUSION ACCELERATION VIA STAGED SAMPLING

Anonymous authors
Paper under double-blind review

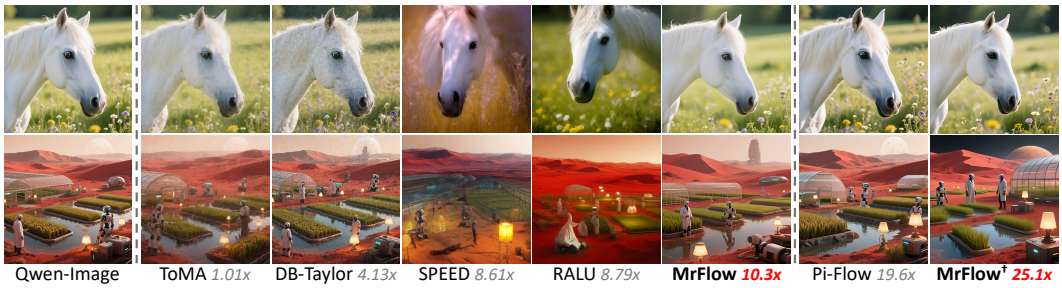

Figure 1: Qualitative comparison between MrFlow and various methods on Qwen-Image. The dashed lines separate the pretrained model, training-free strategies, and strategies that rely on or exploit timestep distillation.

ABSTRACT

Hardware-agnostic strategies for accelerating text-to-image diffusion, such as timestep distillation and feature caching, can reduce inference time without custom kernels or system-level optimization. Among them, multi-resolution generation strategies have recently received broad attention, achieving more than $5\times$ speedup without training. However, the design of upsampling in the latent space, together with the selective modification of partial regions, causes these methods to exhibit noticeable blurring or artifacts. To this end, we propose **MrFlow**, a training-free multi-resolution acceleration strategy for pretrained flow-matching models built upon a staged low-to-high-resolution pipeline. MrFlow first rapidly generates the main structure at low resolution, then performs super-resolution in the pixel space using a lightweight pretrained GAN-based model, subsequently injects low-strength noise to enable high-frequency resampling, and finally refines the details at high resolution in one step while preserving structure. Quantitative and qualitative results on FLUX.1-dev and Qwen-Image show that MrFlow exploits quadratic token reduction and reduced step requirement of low-resolution sampling to achieve $10\times$ end-to-end acceleration, while keeping the OneIG-En gap on Qwen-Image within 1%, significantly exceeding other training-free acceleration strategies, and requiring no training or runtime dynamic identification. MrFlow can also be directly combined orthogonally with pre-trained timestep distillation strategies, achieving even higher generation acceleration of up to $25\times$.

1 INTRODUCTION

Diffusion built upon Transformers as the main architecture (Peebles & Xie, 2023) and flow matching (Lipman et al., 2022; Liu et al., 2022) as the paradigm has become the mainstream in image generation tasks. However, accompanying the scaling of generation quality is the scaling of the model's computational cost. Taking Qwen-Image-20B (Wu et al., 2025a) as an example, it requires up to 47s on diffusion sampling when performing a 1024×1024 text-to-image task on an Nvidia A100. As a result, numerous works seek to accelerate the sampling-based generation of diffusion. Apart from

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

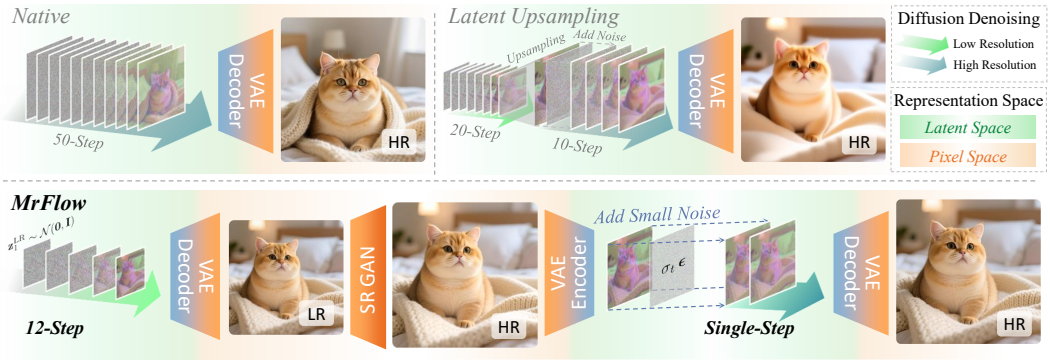

Figure 2: The framework of MrFlow. The compared strategies include the native inference scheme and methods that perform upsampling in the latent space such as LSSGen, RALU and SPEED.

strategies that rely on hardware or systems to implement acceleration, such as quantization (Li et al., 2023) and efficient attention (Dao et al., 2022), researchers have found that on diffusion one can reduce the number of computed tokens to achieve direct acceleration. Timestep distillation (Salimans & Ho, 2022; Luo et al., 2023) is one of the most effective acceleration strategies, capable of reducing the Number of Function Evaluations (NFEs) from 50-100 down to 1-4. Feature caching (Liu et al., 2025a) aims to exploit the similarity between adjacent timesteps, reusing historical features to skip the computation of partial layers or structures. However, these methods either depend heavily on training, or can only attain limited speedup of less than $4\times$. Some works that compose multiple resolution levels (Tang et al., 2025; Jeong et al., 2026) exploit the spatial characteristics of the image modality for acceleration, with training-free methods such as RALU achieving more than $5\times$ speedup, but most still rely on runtime dynamic identification of image regions, and the operation of performing upsampling in the latent space leads to unavoidable artifacts or blurring.

In this paper, we design **MrFlow**, a multi-resolution sampling strategy oriented to modern diffusion. After generating the main structure starting from random noise at low resolution (LR), it enlarges the image in the original pixel space via a super-resolution (SR) network, then injects low-strength noise into the VAE-reencoded latent to enable high-frequency resampling, and finally performs fast sampling in the high-resolution (HR) latent space to refine image details. MrFlow exploits the structure preservation and detail discrepancy of spatial information that images naturally possess at different resolutions, achieving efficient coarse-to-fine generation. This method: (1) fully exploits the fast structure-generation capability of the pretrained diffusion at low resolution, where not only the per-step execution time can be accelerated by about $4\times$, but also fewer timesteps are required; (2) adopts a pretrained lightweight GAN-based super-resolution network in the pixel space, preserving the low-resolution structural information while adding high-frequency signals; (3) the low-strength noise weakens potentially erroneous SR-imposed high-frequency details, allowing the high-resolution flow prior to resample and correct them; (4) uses a single high-resolution step to refine details that super-resolution alone cannot fully recover, completing the final refinement over the short trajectory following low-strength noising. The framework of MrFlow is illustrated in Figure 2.

Extensive experiments on evaluation metrics such as OneIG-Bench, alongside qualitative comparisons of generated images, show that MrFlow can not only achieve a speedup of more than $10\times$ while keeping the OneIG-En gap on Qwen-Image within 1%, but also require no training or runtime dynamic statistics and identification, significantly surpassing various other training-free acceleration strategies in both generation quality and speedup. MrFlow can further be directly combined orthogonally with pre-trained timestep distillation strategies without any additional training or model modification, achieving even higher generation acceleration of up to more than $25\times$.

In summary, our contributions are as follows:

- **A training-free multi-resolution pipeline.** We propose MrFlow, which combines few-step low-resolution generation, pretrained GAN-based pixel-space super-resolution, low-strength latent noising, and single-step high-resolution refinement. This design exploits the complementary roles of low-resolution structure generation and high-resolution detail

refinement, accelerating pretrained flow-matching models by concentrating most sampling at low resolution without runtime dynamic statistics or region identification.

- **Stage-wise analysis.** We empirically characterize few-step low-resolution generation, the compatibility of pixel-space SR with downstream refinement, and single-step high-resolution refinement. A local Gaussian analysis further relates noise strength to direction-wise information retention, providing a reference lower-bound estimate for high-frequency resampling to balance detail modification with coarse structure preservation.
- **Efficient and flexible acceleration.** Experiments on FLUX.1-dev and Qwen-Image demonstrate favorable quality-speed trade-offs. On Qwen-Image, MrFlow achieves over $10\times$ end-to-end speedup while keeping the absolute OneIG-En gap from native inference within 1%. It also combines directly with pretrained timestep-distilled models to achieve over $25\times$ end-to-end generation speedup without any additional model training.

2 Related Work

Flow matching. Flow matching (Lipman et al., 2022) and rectified flow (Liu et al., 2022) recast diffusion training as learning a velocity field along linear interpolants between data and noise. For a data sample x_0 and noise $\epsilon \sim \mathcal{N}(0, I)$, the forward process is:

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

where $t \in [0, 1]$ denotes normalized flow time. The model learns:

$$v_\theta(x_t, t) \approx \mathbb{E}[\epsilon - x_0 \mid x_t], \tag{2}$$

sampling integrates $\dot{x}_t = v_\theta$ from t=1 to t=0. Diffusion models since SD3 (Esser et al., 2024), including typical advanced models such as FLUX (Black Forest Labs, 2024) and Qwen-Image (Wu et al., 2025a), almost all adopt this formulation.

Acceleration of diffusion generation. Quantization (Li et al., 2023) and efficient attention (Dao et al., 2022) are general compression and acceleration strategies that rely on hardware or systems to implement, and their acceleration gains on diffusion are relatively limited. The diffusion-specific acceleration methods are mainly timestep reduction (Salimans & Ho, 2022; Song et al., 2023), feature caching (Liu et al., 2025a; DefTruth, 2025), and token pruning (Lu et al., 2025), all of which can improve sampling efficiency without hardware dependence. Accompanying a speedup of $10\times$ and above, strategies represented by timestep distillation (Luo et al., 2023; Ge et al., 2025; Chen et al., 2025) usually rely on fine-tuning training of non-negligible cost. Feature caching can typically attain around $4\times$ inference speedup without training. Existing works on token compression can usually only attain less than $1.5\times$ inference speedup.

Multi-resolution generation. Leveraging the characteristic information distribution that images naturally possess at different resolutions, many works have explored multi-stage generation from low resolution to high resolution (Saharia et al., 2022b; Ho et al., 2022; Saharia et al., 2022a). One class targets extending the attainable resolution (Du et al., 2024; Wu et al., 2025b). The goal of this class is to generate larger images beyond the native resolution, and they do not concern themselves with acceleration gains at the native resolution. The second class treats multi-resolution as an acceleration means (Tang et al., 2025; Jeong et al., 2026; Tian et al., 2025; Xiao et al., 2026); they aim to place the generation focus on the low resolution and only infer a few steps at high resolution to achieve acceleration, and they all complete the upsampling within the latent or frequency domain.

Super-resolution and image repainting. Image super-resolution in the pixel space has long been divided into three classes: regression-based methods represented by SwinIR (Liang et al., 2021), generative-adversarial-based methods represented by Real-ESRGAN (Wang et al., 2021), and diffusion-based methods represented by OSEDiff (Wu et al., 2024). Image repainting is another line of work related to ours. Methods such as SDEdit (Meng et al., 2021) add noise on an existing image and then perform reverse denoising, in order to accomplish image editing while preserving semantics; this class of methods treats noising-denoising as the basic tool of image editing.

3 Method

We propose MrFlow, a multi-resolution generation strategy designed from an acceleration perspective. Its tenet is to first rapidly generate the overall image structure at the low-resolution stage, then

map back to the pixel space for upsampling, subsequently add low-strength noise for high-frequency resampling, and finally make a small amount of detail refinement on the image at the high-resolution stage. Specifically, MrFlow comprises the following procedures: low-resolution latent-space sampling, VAE decoding, pixel-space super-resolution, VAE encoding, noise injection, high-resolution latent-space sampling, and VAE decoding. As a typical configuration, MrFlow requires only 12 sampling steps at the low-resolution stage, the noise strength injected after super-resolution is usually only 0.1, and then one denoising step at the high-resolution stage suffices. The overall design of the pipeline is built upon one pervasive observation: the low-resolution stage determines the global structure and semantic layout of the image, while the high-resolution stage only needs to refine the details of the super-resolved image from the aforementioned stage. This chapter elaborates the design of each stage; quantitative analyses and ablations appear in Section 4.2 and Appendix C.

3.1 Low-Resolution Structure Generation

Formulation. This stage first samples initial noise in the low-resolution latent space:

$$\mathbf{z}_1^{\mathrm{LR}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \mathbf{z}_1^{\mathrm{LR}} \in \mathbb{R}^{C \times H_L \times W_L}, \tag{3}$$

where C is the number of latent channels, and $H_L \times W_L$ is determined by the target output size through the downsampling factor of the pretrained VAE.

It then performs ODE sampling in the flow matching form to obtain the clean latent:

$$\mathbf{z}_0^{\mathrm{LR}} \;=\; \Phi^{K_L}_{\mathbf{v}_\theta, \mathbf{c}}\big(\mathbf{z}_1^{\mathrm{LR}}\big), \tag{4}$$

where $\mathbf{v}_\theta$ is the pretrained flow matching velocity network, $\mathbf{c}$ is the text-condition embedding, and $\Phi^{K}_{\mathbf{v}_\theta, \mathbf{c}}$ denotes the rectified flow ODE flow map obtained by K-step Euler discretization, with the iteration rule $\mathbf{z}_{t-h} = \mathbf{z}_t - h\, \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})$. This work typically takes $K_L = 12$.

Finally, the clean latent is mapped back to the pixel space through the pretrained VAE decoder $\mathcal{D}$:

$$\mathbf{x}_{\mathrm{LR}} \;=\; \mathcal{D}\big(\mathbf{z}_0^{\mathrm{LR}}\big), \qquad \mathbf{x}_{\mathrm{LR}} \in \mathbb{R}^{3 \times H_L^{\mathrm{px}} \times W_L^{\mathrm{px}}}, \tag{5}$$

where $H_L^{\mathrm{px}} \times W_L^{\mathrm{px}}$ is the low-resolution pixel-space size.

The low-resolution image generated at this stage already almost possesses the complete global structure and the semantic features that match the prompt. However, at the same display size, it usually appears blurrier compared with conventional images generated directly at high resolution, and local fine-grained semantics such as text and symbols are difficult to generate stably.

Analysis. The acceleration of the low-resolution stage comes from two complementary factors. The first is that the overall cost of single-step inference is measured to scale almost linearly with the number of image tokens, yielding about $4\times$ comprehensive speedup when each side is shrunk by $2\times$. The second is effective structure generation with few low-resolution steps: under an equal step-count budget, a two-stage schedule using 10 low-resolution steps plus 1 high-resolution step attains higher CLIP similarity to a high-resolution reference than direct generation with 11 high-resolution steps, while reducing latency by more than half and maintaining comparable text-image alignment scores. Attention measurements further characterize text-conditioning interactions during few-step generation; the measurements and discussion are given in Appendix C.1. Appendix C.5 further characterizes the role of this stage in establishing the global structure of the generated image.

3.2 Pixel-Space Super-Resolution

Formulation. For the image obtained after low-resolution sampling and VAE decoding, we apply super-resolution in the pixel space:

$$\mathbf{x}_{\mathrm{SR}} \;\leftarrow\; U\big(\mathbf{x}_{\mathrm{LR}}\big), \qquad \mathbf{x}_{\mathrm{SR}} \in \mathbb{R}^{3 \times H_H^{\mathrm{px}} \times W_H^{\mathrm{px}}}, \tag{6}$$

where U is the pretrained super-resolution network, upsampling the pixel-space image from the low-resolution size $H_L^{\mathrm{px}} \times W_L^{\mathrm{px}}$ to the high-resolution size $H_H^{\mathrm{px}} \times W_H^{\mathrm{px}}$, with $\times 2$ upsampling typically chosen in this work. Specifically, we directly adopt the pretrained Real-ESRGAN (Wang et al., 2021) as U to supply high-frequency details for subsequent high-resolution refinement, rather than simple interpolation or pixel-regression SR models such as SwinIR (Liang et al., 2021).



Figure 3: Example comparison of MrFlow adopting different super-resolution strategies on Qwen-Image, where Low Resolution denotes the image after low-resolution generation followed by VAE decoding to the pixel space, with a resolution of 512×512. The rows of SR and High Resolution Refine are respectively the super-resolved image and the final image after high-resolution refinement, with a resolution of 1024×1024.

While enlarging the resolution, the super-resolution network both largely preserves the overall image structure generated at the low-resolution stage and supplements part of the high-frequency information. However, since the pretrained super-resolution network does not carry the semantic information in the prompt and originates only from pretrained knowledge of distribution matching between different resolutions in general scenes, the high-resolution image at this stage may still exhibit local structural confusion such as artifacts and character-stroke shifts.

Analysis. This stage involves two key design choices: performing upsampling in the pixel space, and adopting a GAN-based pretrained super-resolution model. First, directly interpolating latent features can disrupt the local statistics expected by the VAE decoder and introduce grid-like artifacts. Pixel-space upsampling instead reuses pretrained SR models and their learned edge and texture priors, then maps the enlarged image back through the VAE encoder with little additional cost. Second, the SR model should provide a useful detail initialization for subsequent refinement. In our comparisons, Real-ESRGAN supplies sharp details while retaining the coarse layout, allowing light refinement to adjust local artifacts and character strokes. Interpolation and SwinIR instead leave more diffuse blur, which persists after the same refinement step. These observations support choosing an SR model whose added details can be effectively adjusted by the downstream refinement stage. Figure 3 and Appendix C.2 provide the corresponding empirical comparisons.

3.3 Low-Strength Noise for High-Frequency Resampling

Formulation. After the pixel-space super-resolution, re-encoding through the pretrained VAE encoder $\mathcal{E}$ yields the high-resolution latent:

$$\mathbf{z}_0^{\mathrm{SR}} \;=\; \mathcal{E}\big(\mathbf{x}_{\mathrm{SR}}\big), \qquad \mathbf{z}_0^{\mathrm{SR}} \in \mathbb{R}^{C\times H_H \times W_H}. \tag{7}$$

Unlike the clean latent obtained by direct high-resolution sampling from noise, $\mathbf{z}_0^{\mathrm{SR}}$ may contain artifacts and character-stroke shifts introduced by GAN super-resolution. These errors are mostly localized in high-frequency directions, while the low-frequency structure is already determined by the low-resolution stage and should be preserved.

To this end, before entering the high-resolution refine, we inject low-strength noise in the flow matching form to obtain the noised SR latent:

$$\mathbf{z}_t^{\mathrm{HR}} \;=\; (1-\sigma_t)\,\mathbf{z}_0^{\mathrm{SR}} \;+\; \sigma_t\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{8}$$

where $\sigma_t \in (0, 1)$ controls the injected noise strength, while t denotes the corresponding refinement timestep. We typically use $\sigma_t \in [0.1, 0.15]$. This step keeps the low-frequency structure at high

SNR, while lowering the SNR of high-frequency directions so that the following denoising step can resample them according to the high-resolution flow prior to correct SR artifacts.

Analysis. The SR output already carries the global structure, but its added high-frequency details may contain local artifacts or semantic inconsistencies. Too little noise leaves these details largely intact, whereas too much can disrupt the established structure. A local Gaussian analysis shows that denoising retains less of the SR input along low-variance directions under the same noise level. We therefore compare the injected noise amplitude with the high-frequency signal scale to estimate a suitable lower bound. Let λ_{hf} denote a representative clean-latent variance along high-frequency directions; the estimate is:

$$\sigma_t \gtrsim \sigma_t^\star = \frac{\sqrt{\lambda_{\mathrm{hf}}}}{1+\sqrt{\lambda_{\mathrm{hf}}}}. \tag{9}$$

The estimate uses clean-latent variability to identify a reference noise scale. Around this scale, denoising places less weight on the SR input in high-frequency directions, while directions with larger variance retain more of the input. This frequency-dependent response supports adjusting local details, including potential artifacts introduced by the super-resolution network, without restarting global structure generation. In practice, we use $\sigma_t \in [0.1, 0.15]$ to balance detail refinement and structure preservation. Broader blur or low-frequency deviations may require stronger noising, which can alter useful SR details and increase refinement cost. Appendix C.3 provides the posterior-mean interpretation, spectral measurements, and frequency-dependent perturbation analysis.

3.4 High-Resolution Detail Refinement

Formulation. After completing the low-strength noising on the high-resolution latent, we take the noised high-resolution latent $\mathbf{z}_t^{\mathrm{HR}}$ as the initial value and apply the flow map on the high resolution along the same velocity-estimation network:

$$\mathbf{z}_0^{\mathrm{HR}} = \Phi_{\mathbf{v}_\theta,\mathbf{c}}^{K_H}\big(\mathbf{z}_t^{\mathrm{HR}}\big), \tag{10}$$

where $\mathbf{c}$ is the text-condition embedding completely identical to that of the low-resolution stage, and K_H is the number of Euler steps at this stage. In practice, refinement starts at the timestep corresponding to σ_t under the backbone's noise schedule and integrates to 0, with $K_H = 1$ by default; additional high-resolution steps bring only marginal changes under this low-strength regime. Finally, the clean latent is sent back to the pixel space through the VAE decoder $\mathcal{D}$ to obtain the final output:

$$\mathbf{x}_{\mathrm{HR}} = \mathcal{D}\big(\mathbf{z}_0^{\mathrm{HR}}\big), \qquad \mathbf{x}_{\mathrm{HR}} \in \mathbb{R}^{3\times H_H^{\mathrm{px}}\times W_H^{\mathrm{px}}}. \tag{11}$$

This stage performs the final one-step detail refinement on the local artifacts introduced by the SR output, and the entire pipeline at this point outputs a high-quality image at the target resolution.

Analysis. Low-strength noising confines high-resolution refinement to a short integration interval near the clean endpoint, helping reduce Euler discretization error. Comparisons with multi-step refinement show that a single step already captures the main visual improvements, including corrections to distorted character strokes, with additional steps producing only minor changes. These observations support $K_H = 1$ as an efficient default for detail refinement. Appendix C.4 characterizes this behavior through step-count comparisons and discrete prediction statistics, while Appendix C.5 examines the complementary roles of low-resolution structure generation and high-resolution detail refinement.



Figure 4: Images generated by MrFlow under different high-resolution step configurations on Qwen-Image.

4 Empirical Results

We mainly conduct experiments on text-to-image tasks. Section 4.1 reviews the experimental setup, Section 4.2 mainly presents extensive comparisons with various acceleration strategies on quantitative metrics, and Section 4.3 is the overall ablation experiment of MrFlow.

Table 1: Comparison of various training-free acceleration methods on FLUX.1-dev and Qwen-Image, generated at 1024×1024 resolution.

Method	Training-free	NFEs	Speedup ↑	Geneval ↑	DPG ↑	OneIG-En ↑	OneIG-Zh ↑
FLUX.1-dev	-	50	1x	0.66	84.07	0.44	-
ToMA	√	50	1.05x	0.66	83.99	0.42	-
ToMA	√	50	1.13x	0.66	83.78	0.21	-
Teacache	√	50	4.47x	0.63	82.62	0.36	-
DB-Taylor	√	50	4.63x	0.64	**83.78**	**0.40**	-
RALU	√	4, 5, 6	4.33x	0.62	82.83	0.32	-
SPEED	√	3, 2, 7	5.71x	0.62	83.55	0.37	-
MrFlow	√	20, 1	**5.78x**	**0.65**	82.19	0.39	-
Teacache	√	50	7.57x	0.45	70.79	0.28	-
DB-Taylor	√	50	7.86x	0.46	74.23	0.28	-
RALU	√	1, 2, 3	8.21x	0.53	77.11	0.30	-
SPEED	√	2, 1, 5	8.01x	0.62	**83.43**	0.36	-
MrFlow	√	12, 1	**8.25x**	**0.63**	81.65	**0.36**	-
Qwen-Image	-	50x2	1x	0.88	88.67	0.53	0.53
Teacache	√	50x2	4.61x	0.81	84.46	0.42	0.45
DB-Taylor	√	50x2	4.13x	0.86	86.81	0.50	0.50
RALU	√	(4, 5, 5)x2	5.85x	0.85	87.07	0.46	0.48
SPEED	√	(2, 1, 7)x2	6.29x	0.79	87.50	0.47	0.51
MrFlow	√	(20, 1)x2	**6.98x**	**0.87**	**88.00**	**0.54**	**0.52**
Teacache	√	50x2	8.93x	0.26	32.51	0.16	0.19
DB-Taylor	√	50x2	9.34x	0.09	17.43	0.09	0.12
RALU	√	(1, 2, 4)x2	9.51x	0.84	82.60	0.44	0.46
SPEED	√	(1, 1, 5)x2	8.61x	0.74	84.93	0.44	0.46
MrFlow	√	(12, 1)x2	**10.3x**	**0.86**	**87.10**	**0.52**	**0.51**

4.1 Implementation Details

Our main acceleration experiments are conducted on FLUX.1-dev (Black Forest Labs, 2024) and Qwen-Image-20B (Wu et al., 2025a), with quantitative evaluation uniformly set at 1024×1024 resolution, and the evaluation metrics adopt three classes: Geneval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and OneIG-Bench (Chang et al., 2026). The selected comparison methods include typical training-free acceleration strategies of various classes, comprising the feature-cache-based Teacache (Liu et al., 2025a) and DB-Taylor (the comprehensive version fusing DBCache and TaylorSeer (Liu et al., 2025b) in Cache-Dit (DefTruth, 2025), which we abbreviate as DB-Taylor), the token-fusion-based ToMA (Lu et al., 2025), and RALU (Jeong et al., 2026) and SPEED (Xiao et al., 2026) which are also based on multi-resolution. For extremely high speedup scenarios greater than $8\times$, given that the above training-free methods essentially all face collapse, we additionally compare the state-of-the-art training-dependent timestep distillation methods, SenseFlow (Ge et al., 2025) and Pi-Flow (Chen et al., 2025), uniformly set to the 4-step inference version. For the training-required strategies, including all timestep distillation methods and LSSGen (Tang et al., 2025), we directly load the parameter weights provided in their official repositories without retraining. When MrFlow is combined with the weights pre-trained by timestep distillation, we denote it as MrFlow†; this strategy directly loads the distilled model at both the low-resolution and high-resolution stages, without any additional training in combination with the multi-resolution pipeline. Each experiment is run on a single Nvidia A100 GPU. The reported speedups are all end-to-end actual acceleration, including text encoding, random noise generation, VAE encoding/decoding, SR, and all diffusion forward passes. All detailed configurations are given in Appendix B.

4.2 Main Results

Training-free acceleration. All methods natively support FLUX.1-dev, while the native version of Qwen-Image does not adopt the distilled single-branch CFG, so every step needs two forward

Table 2: Comparison with training-dependent acceleration methods on FLUX.1-dev and Qwen-Image, generated at 1024×1024 resolution.

Method	Training-free	NFEs	Speedup ↑	Geneval ↑	DPG ↑	OneIG-En ↑	OneIG-Zh ↑
FLUX.1-dev	-	50	1x	0.66	84.07	0.44	-
LSSGen	×	25, 25	1.49x	0.65	83.51	0.40	-
LSSGen	×	25, 4	3.93x	0.64	82.77	0.40	-
MrFlow	√	12, 1	8.25x	0.63	81.65	0.36	-
SenseFlow	×	4	11.1x	0.63	82.42	0.37	-
Pi-Flow	×	4	9.49x	0.68	**84.40**	**0.41**	-
MrFlow†	†	4, 1	**11.3x**	**0.69**	83.26	0.39	-
FLUX-schnell	×	2	**20.4x**	0.69	85.03	**0.41**	-
MrFlow‡	‡	1, 1	19.5x	**0.71**	**85.10**	0.35	-
Qwen-Image	-	50x2	1x	0.88	88.67	0.53	0.53
Pi-Flow	×	4	19.6x	**0.85**	**88.10**	**0.53**	**0.52**
MrFlow†	†	4, 1	**25.1x**	**0.85**	87.52	0.52	0.51

passes. We mark whether each method requires training, and use dashed lines to separate different levels of speedup. Table 1 lists the results of various training-free methods. ToMA, as a representative method based on token pruning and merging, can only attain a very limited speedup of less than $1.2\times$. And although its GenEval drop is not noticeable, the generated images have actually collapsed significantly. The other multi-resolution-based method, RALU, already exhibits a certain accuracy degradation at a speedup of around $5\times$ on FLUX, and when the speedup is further raised to around $8\times$ the generation quality drops even more noticeably; although RALU can better preserve accuracy at high speedup on Qwen-Image, its metrics also drop noticeably when the speedup approaches $10\times$, and the actual image quality is even worse under visual inspection. SPEED provides a stronger frequency-domain multi-resolution baseline than earlier latent-upsampling strategies, but its accuracy drops sharply on Qwen-Image when pushed to the same speed range as MrFlow. Teacache and DB-Taylor likewise preserve good generation results at a speedup of around $4\times$, but when the speedup is further raised to $8\times$, they both exhibit significant collapse; on the newer, larger-scale, and more capable Qwen-Image, the feature-cache-class methods exhibit even larger losses, with a gap of more than 3% already at a speedup of 4-$5\times$, and direct collapse at the more aggressive speedup of $9\times$. MrFlow, however, demonstrates a greater advantage: it preserves good accuracy under the configuration of 20 low-resolution steps and 1 high-resolution step, with a gap of only about 2% on Geneval relative to the native 50 steps. When set to 12 low-resolution steps and 1 high-resolution step, the speedup is raised to $8.25\times$ on FLUX and $10.3\times$ on Qwen-Image, while Geneval remains at 0.63 and 0.86, respectively, significantly surpassing feature-cache-based methods such as DB-Taylor and multi-resolution-based strategies such as RALU and SPEED under aggressive configurations.

Training-dependent methods. Table 2 further compares MrFlow with training-dependent acceleration methods on the two backbones, including state-of-the-art timestep distillation methods as well as the multi-resolution method LSSGen, which requires training a lightweight latent upsampler and is therefore grouped with the training-dependent strategies rather than the training-free ones. LSSGen attains a $1.5\times$ speedup while preserving accuracy under the officially suggested default configuration with a noise strength of 0.75; however, when the noising strength is lowered to 0.2 for a $3.9\times$ speedup, the quality of the actually generated images has already dropped substantially, as detailed in Appendix F.1. Here † or ‡ both indicate that the combination of MrFlow with the distillation method directly utilizes the pre-trained distillation weights for both low-resolution generation and high-resolution detail refinement without additional training in combination with MrFlow, and the two respectively denote the combination of MrFlow with Pi-Flow and with FLUX-schnell (Black Forest Labs, 2024). On FLUX, MrFlow under the training-free $12{+}1$-step setting already attains accuracy close to the 4-step SenseFlow, while MrFlow+Pi-Flow attains a higher speedup than the native 4-step Pi-Flow while maintaining comparable Geneval. When MrFlow is combined with FLUX-schnell, although the speedup is on par with the native 2-step FLUX-schnell due to extra overhead such as VAE, DPG is also slightly higher than the native FLUX-schnell. Since timestep

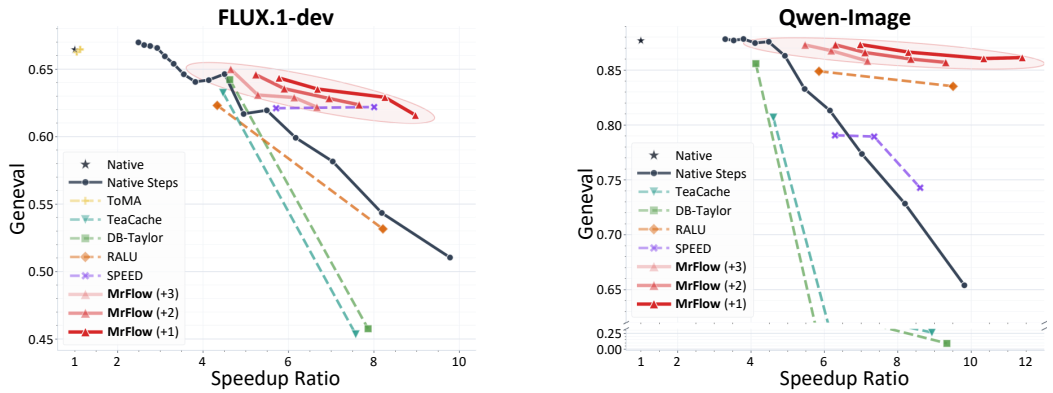

Figure 5: Trade-off curves between GenEval score and speedup for various training-free methods. Different shades of MrFlow correspond to different numbers of steps in the high-resolution stage.

distillation-class methods can directly fuse the dual-branch CFG mechanism of Qwen-Image into a single forward, they themselves naturally possess a higher speedup; combining MrFlow with Pi-Flow on Qwen-Image reaches up to $25\times$ speedup over native Qwen-Image inference while keeping the absolute OneIG-En gap within 1%. These results indicate that MrFlow under the training-free setting can already attain accuracy and efficiency comparable to training-dependent timestep distillation strategies, and has good combinability with timestep distillation strategies.

4.3 Ablation Study

Different step configurations. On FLUX.1-dev and Qwen-Image, we conduct combinatorial experiments over all training-free methods and over MrFlow configurations with the low-resolution stage set to 10, 12, 16, 20 steps and the high-resolution stage set to 1, 2, 3 steps; the trade-off result between Geneval and speedup is shown in Figure 5. Evidently, MrFlow is the only method that consistently outperforms across different models the naive acceleration strategy of directly applying different timestep configurations to the native model, and it can flexibly be configured toward either better accuracy or a more aggressive speedup through separate control of low-resolution and high-resolution budgets. Note that MrFlow's performance on test metrics improves significantly as the number of low-resolution steps increases, but has no obvious connection with the high-resolution step configuration. The visual comparison leads to the same conclusion. As shown in Figure 4, adding one high-resolution refinement step already makes the text strokes clear and removes the blur observed in the no-refinement output, while using two refinement steps produces a highly similar result. Together with the short refinement interval discussed in Section 3.4, these results support a single high-resolution step as an efficient default for detail refinement after low-strength noising.

Super-resolution network. On Qwen-Image, we test several SR upsampling strategies, including direct interpolation, SwinIR, OSEDiff, Real-ESRGAN, and SR-only output without HR refinement, and find that their differences on quantitative metrics such as GenEval are no more than 2%; nevertheless, their perceptual quality differs substantially, as shown in Figure 3, generated with 12 LR steps and 1 HR step. Overall, Real-ESRGAN provides the most reliable perceptual quality.

5 Conclusion

We propose MrFlow, a training-free multi-resolution strategy for accelerating pretrained flow-matching models. MrFlow combines fast low-resolution structure generation, pixel-space super-resolution with a lightweight GAN-based network, low-strength latent noising, and one-step high-resolution refinement. Without any training or runtime dynamic statistics, this staged pipeline achieves more than $10\times$ end-to-end speedup while keeping the OneIG-En gap on Qwen-Image within 1%, and it offers better quality, efficiency, flexibility, and simplicity than other training-free diffusion acceleration strategies. MrFlow can also be directly combined with existing timestep distillation techniques without additional training, yielding a further compounded speedup of $25\times$.

AI USE STATEMENT

We used generative AI tools to aid manuscript writing and revision, including language polishing and structural editing. The authors reviewed the AI-assisted revisions and take full responsibility for the final content of this work.

REFERENCES

AUTOMATIC1111. Stable diffusion web UI. https://github.com/AUTOMATIC1111/stable-diffusion-webui, 2022. High-resolution fix (Hires.fix) feature.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. pi-flow: Policy-based few-step generation via imitation distillation. *arXiv preprint arXiv:2510.14974*, 2025.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

etc. DefTruth, vipshop.com. Cache-dit: A pytorch-native inference engine with cache, parallelism, quantization and cpu offload for dits., 2025. URL https://github.com/vipshop/cache-dit.git. Open-source software available at https://github.com/vipshop/cache-dit.git.

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6159–6168, 2024.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

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

Weilun Feng, Guoxin Fan, Haotong Qin, Chuanguang Yang, Mingqiang Wu, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Dingrui Wang, et al. Worldcache: Accelerating world models for free via heterogeneous token caching. *arXiv preprint arXiv:2603.06331*, 2026.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. Senseflow: Scaling distribution matching for flow-based text-to-image distillation. *arXiv preprint arXiv:2506.00523*, 2025.

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Wongi Jeong, Kyungryeol Lee, Hoigi Seo, and Se Young Chun. Training-free mixed-resolution latent upsampling for spatially accelerated diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18673–18682, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6007–6017, 2023.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Junqing Lin, Xingyu Zheng, Pei Cheng, Bin Fu, Jingwei Sun, and Guangzhong Sun. Token pruning for in-context generation in diffusion transformers. *arXiv preprint arXiv:2602.01609*, 2026.

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

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7353–7363, 2025a.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15853–15863, 2025b.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. Toma: Token merge with attention for diffusion models. *arXiv preprint arXiv:2509.10918*, 2025.

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.

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. *arXiv preprint arXiv:2108.01073*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022b.

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

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252. PMLR, 2023.

Jyun-Ze Tang, Chih-Fan Hsu, Jeng-Lin Li, Ming-Ching Chang, and Wei-Chao Chen. Lssgen: Leveraging latent space scaling in flow and diffusion for efficient text to image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5048–5057, 2025.

Ye Tian, Xin Xia, Yuxi Ren, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Yunhai Tong, Ling Yang, and Bin Cui. Training-free diffusion acceleration with bottleneck sampling. *arXiv preprint arXiv:2503.18940*, 2025.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

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

Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3944–3953. IEEE, 2025b.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024.

Howard Xiao, Brian Chao, Lior Yariv, and Gordon Wetzstein. Spectral progressive diffusion for efficient image and video generation. *arXiv preprint arXiv:2605.18736*, 2026.

Shikang Zheng, Guantao Chen, Landis He, Jiacheng Liu, Yuqi Lin, Chang Zou, and Linfeng Zhang. From sketch to fresco: Efficient diffusion transformer with progressive resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18714–18723, 2026a.

Xingyu Zheng, Xianglong Liu, Haotong Qin, Xudong Ma, Mingyuan Zhang, Haojie Hao, Jiakai Wang, Zixiang Zhao, Jinyang Guo, and Michele Magno. Binarydm: Accurate weight binarization for efficient diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 70892–70913, 2025.

Xingyu Zheng, Haotong Qin, Yuye Li, Haoran Chu, Jiakai Wang, Jinyang Guo, Michele Magno, and Xianglong Liu. First-order error matters: Accurate compensation for quantized large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28883–28891, 2026b.

A BACKGROUND

Acceleration of diffusion generation. Quantization (Li et al., 2023; Zheng et al., 2025; 2026b) and efficient attention (Dao et al., 2022) are compression and acceleration strategies oriented to general structures, and their acceleration usually relies on hardware or systems to implement. The diffusion-specific acceleration methods are mainly timestep reduction, feature caching, and token pruning, all of which can improve sampling efficiency without hardware dependence. A major direction for accelerating diffusion models is reducing the number of denoising steps. Existing methods use improved samplers (Song et al., 2020; Lu et al., 2022) or distillation techniques (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023) to compress a multi-step teacher into a few-step or even one-step generator. Accompanying a speedup of $10\times$ and above, strategies represented by timestep distillation (Ge et al., 2025; Chen et al., 2025) usually rely on fine-tuning training of non-negligible cost, and the trajectories are likely to gradually converge to losing the diversity across different random seeds. Another line of work accelerates diffusion models through feature caching (Liu et al., 2025a; DefTruth, 2025; Liu et al., 2025b; Feng et al., 2026). These methods reuse intermediate features across adjacent denoising steps based on the observation that diffusion features change slowly over time, thereby avoiding repeated computation in the U-Net or DiT backbone. These approaches typically achieve around $2\times$ to $4\times$ inference speedup with relatively small quality degradation. Token compression (Lu et al., 2025; Lin et al., 2026) reduces computation by dynamically removing or fusing spatial tokens based on their estimated importance during denoising. Existing approaches typically achieve only less than $1.5\times$ inference speedup. These diffusion acceleration strategies either rely on fine-tuning training of non-negligible cost, or can only attain limited speedup.

Multi-resolution generation. Leveraging the characteristic information distribution that images naturally possess at different resolutions, many works have explored multi-stage generation from low resolution to high resolution. This line can be roughly divided into two classes. The first class targets extending the attainable resolution. Early works such as SR3 (Saharia et al., 2022b), Cascaded Diffusion Models (Ho et al., 2022), and Imagen (Saharia et al., 2022a) adopt cascaded diffusion or diffusion-based super-resolution, generating the main structure at low resolution and then progressively enlarging and supplementing details; Hires. fix (AUTOMATIC1111, 2022) further validated in community practice the feasibility of "low-resolution generation + high-resolution repainting"; works such as DemoFusion (Du et al., 2024) and MegaFusion (Wu et al., 2025b) break through the pretrained resolution ceiling through progressive denoising and multi-scale latent fusion. The goal of this class of work is to generate larger images beyond the native resolution, and they do not concern themselves with acceleration gains at the native resolution. The second class treats multi-resolution as an acceleration means. LSSGen (Tang et al., 2025) trains a lightweight latent upsampler and advocates direct upsampling in the latent space to avoid the artifacts introduced by VAE re-encoding after pixel-space upsampling. Bottleneck Sampling (Tian et al., 2025) adopts interpolation operators such as Lanczos in the latent space, and uses a "high-low-high" U-shaped schedule together with timestep rescheduling to achieve training-free acceleration. RALU (Jeong et al., 2026) likewise does nearest-neighbor interpolation in the latent space, but suppresses aliasing artifacts through region-adaptive early-edge upsampling and noise-timestep distribution matching. Fresco (Zheng et al., 2026a) uses a Hadamard orthogonal transform to expand a single parent token into four child tokens, and maintains consistency across resolution stages with a global noise field. SPEED (Xiao et al., 2026) changes the upsampling space from the spatial domain to the frequency domain, progressively expanding by frequency band along the denoising trajectory. The above acceleration-class works all complete the upsampling within the latent or frequency domain.

Super-resolution and image repainting. Image super-resolution in the pixel space has long been divided into three classes: regression-based methods, generative-adversarial-based methods, and diffusion-based methods. Methods such as SwinIR (Liang et al., 2021), based on a Swin Transformer (Liu et al., 2021) or convolutional backbone, use pixel-wise reconstruction losses to recover the high-resolution image from the low-resolution input. Real-ESRGAN (Wang et al., 2021), building upon ESRGAN (Wang et al., 2018), introduces second-order degradation modeling and replaces pure pixel regression with discriminator loss and perceptual loss, aligning the output distribution with the natural-image distribution in the discriminative sense. Diffusion-based super-resolution is represented by OSEDiff (Wu et al., 2024), which performs single-step distillation on a pretrained diffusion and introduces the super-resolution input as a condition. Image repainting is another related line. SDEdit (Meng et al., 2021) proposes adding noise on an existing image and then performing reverse denoising, in order to accomplish local editing while preserving semantics; subsequent

works such as InstructPix2Pix (Brooks et al., 2023), Imagic (Kawar et al., 2023), and Prompt-to-Prompt (Hertz et al., 2022) extend the conditional guidance to various degrees. This class of methods treats noising-denoising as the basic tool of image editing. Unlike these editing methods, MrFlow does not use noising-denoising to preserve and edit an existing image; instead, it uses low-strength noise as a frequency-selective mechanism that preserves the generated low-frequency structure while enabling high-frequency resampling.

B Detailed Experimental Setup

Our main acceleration experiments are conducted on FLUX.1-dev (Black Forest Labs, 2024) and Qwen-Image-20B (Wu et al., 2025a), with some extended experiments on FLUX.2 Klein (Black Forest Labs, 2025). The comparison methods cover typical acceleration strategies of various classes: the feature-cache-based TeaCache (Liu et al., 2025a) and DB-Taylor (the comprehensive version fusing DBCache and TaylorSeer (Liu et al., 2025b) in Cache-Dit (DefTruth, 2025)), the token-fusion-based ToMA (Lu et al., 2025), and the multi-resolution LSSGen (Tang et al., 2025), RALU (Jeong et al., 2026), and SPEED (Xiao et al., 2026). For the extreme regime beyond $8\times$, where the training-free methods essentially all collapse, we additionally compare the state-of-the-art timestep distillation methods SenseFlow (Ge et al., 2025) and Pi-Flow (Chen et al., 2025), both at their released 4-step setting. For all training-dependent strategies, i.e. the timestep distillation methods and LSSGen, we directly load the official pre-trained weights without retraining. The native models are run with their default parameters, e.g. a CFG of 3.5 for FLUX.1-dev and 4.0 for Qwen-Image. For each baseline we report two operating points, chosen so that TeaCache, DB-Taylor, RALU, and SPEED cover comparable speed regimes; the per-method hyper-parameters at each point are listed in Tables 3 and 4. When MrFlow is combined with timestep-distillation weights, denoted MrFlow†, it directly loads the distilled model at both stages without any additional training.

For the quantitative evaluation, we uniformly set it at 1024×1024 resolution, as this is the most common resolution setting in the literature. The evaluation metrics adopt three classes: Geneval, DPG-Bench, and OneIG-Bench, which characterize text-to-image quality from three angles respectively: compositional semantics, dense alignment under long prompts, and multi-dimensional attributes under both Chinese and English bilinguals, including character rendering and style consistency. For some analytical experiments, we additionally generate at officially suggested resolutions of the model such as 1328×1328 and 1584×1056. The experiments are conducted on Nvidia A100 GPUs, with each experiment run on a single GPU. The reported speedups are all end-to-end actual acceleration, including text encoding, random noise generation, VAE encoding/decoding, SR, and all diffusion forward passes. For multi-resolution baselines, timing also includes transition overheads such as region selection, token rearrangement, and stage-wise re-noising in RALU, or spectral expansion with timestep realignment in SPEED.

C Stage-wise Analysis

This appendix follows the four stages of MrFlow, providing empirical analyses and a local model to support the design choices in the main text. Sections C.1–C.4 examine low-resolution generation, pixel-space super-resolution, low-strength noising, and high-resolution refinement, respectively. Section C.5 combines seed sensitivity and before-and-after comparisons to characterize their complementary roles in structure generation and detail refinement.

C.1 Sources of Acceleration at the Low-Resolution Stage

We examine two complementary benefits of low-resolution generation: lower per-step cost and effective generation with few steps. The following Qwen-Image comparisons characterize these benefits, while Appendix C.5 examines the low-resolution stage's role in establishing global structure.

Single-step inference acceleration. Halving each image dimension reduces the number of image tokens by $4\times$. The image-token self-attention component can therefore admit up to a $16\times$ reduction in arithmetic cost, whereas linear operators such as feed-forward layers scale more modestly. The measured overall per-step speedup is about $4\times$.

Table 3: Parameter settings on FLUX.1-dev used in the main comparison tables. N_{LR} and N_{ref} denote the low-resolution denoising and high-resolution refinement steps of MrFlow.

Method	NFEs	Parameter setting
FLUX.1-dev *- training-free acceleration*		
ToMA	50	Token-merging ratio $r \in \{0.5, 0.8\}$; destination budget over M=256 local tiles, merging every Δt=3 steps.
TeaCache	50	Triggered by the accumulated relative ℓ_1-change criterion; thresholds $\tau \in \{1.0, 2.5\}$ for the two speed regimes.
DB-Taylor	50	DBCache threshold $\delta \in \{0.30, 0.60\}$, warm-up $K \in \{4, 1\}$; TaylorSeer order p=1, (F_n, B_n)=$(1, 0)$.
RALU	$4, 5, 6$ / $1, 2, 3$	Three-stage schedule $\mathbf{N}$=$(4, 5, 6)$ or $(1, 2, 3)$; transitions $\mathbf{e}$=$(0.30, 0.45, 1.0)$ or $(0.20, 0.30, 1.0)$; HR ratio ρ=0.20 or 0.075.
SPEED	$3, 2, 7$ / $2, 1, 5$	Official s3 setting with total steps $12/8$, scales $(0.25, 0.5, 1.0)$, and δ=0.01; discrete schedules are derived from the power-spectrum transitions.
MrFlow	$20, 1$ / $12, 1$	Pixel-space staged sampling with $(N_{\mathrm{LR}}, N_{\mathrm{ref}})$=$(20, 1)$ or $(12, 1)$; Real-ESRGAN $\times 2$ before refinement.
FLUX.1-dev *- training-dependent acceleration*		
LSSGen	$25, 25$ / $25, 4$	Trained latent upsampler; LR stage 25 steps, HR stage from noise level $\sigma_s \in \{0.75, 0.2\}$, giving 25 and 4 HR steps.
SenseFlow	4	Released 4-step FLUX student model.
Pi-Flow	4	Released pi-FLUX 4-step policy model, FlowMap schedule as in the official adapter.
MrFlow + Pi-Flow	$4, 1$	LR Pi-Flow 4 steps, then Real-ESRGAN $\times 2$ and one HR refinement step.
FLUX-schnell	2	Released FLUX.1-schnell 2-step model.
MrFlow + FLUX-schnell	$1, 1$	LR FLUX-schnell 1 step, then Real-ESRGAN $\times 2$ and one HR FLUX-schnell step.

Few-step generation. We compare direct high-resolution generation with 11 steps (`HR11`) against 10 low-resolution steps followed by one high-resolution refinement step (`LR10_HR1`). Table 5 reports latency, CLIP image similarity to a 28-step high-resolution reference ($\mathrm{CLIP}_{\mathrm{ref}}$), and text-image similarity:

$$\mathrm{CLIP}_{\mathrm{txt}}(\mathbf{x}, \mathbf{c}) = \cos\big(\boldsymbol{\phi}_I(\mathbf{x}), \boldsymbol{\phi}_T(\mathbf{c})\big), \tag{12}$$

where $\boldsymbol{\phi}_I$ and $\boldsymbol{\phi}_T$ are the CLIP (Radford et al., 2021) image and text encoders. At the same total sampling-step count, the low-resolution-dominated schedule more than halves latency and improves agreement with the high-resolution reference, while text-image similarity remains comparable. This supports allocating most generation steps to low resolution.

Text-conditioning interaction. As a complementary observation, we measure the attention mass that image queries assign to text keys. For the attention matrix $\mathbf{A}^{(\ell)}$ at layer ℓ, let $\mathcal{Q}_{\mathrm{img}}$ and $\mathcal{K}_{\mathrm{txt}}$ denote the image-query and text-key sets:

$$\mathcal{T}^{(\ell)} = \frac{1}{|\mathcal{Q}_{\mathrm{img}}|} \sum_{q \in \mathcal{Q}_{\mathrm{img}}} \sum_{k \in \mathcal{K}_{\mathrm{txt}}} \mathbf{A}^{(\ell)}_{q,k}. \tag{13}$$

We average attention mass from the text-conditioned branch over 10 sampling steps and 3 prompts with one seed. Table 6 shows a higher text-attention share at low resolution at all three measured layers.

C.2 Two Design Choices of Pixel-Space Super-Resolution

This section examines the upsampling domain and the choice of SR model, focusing on the detail supplied to subsequent high-resolution refinement.

Table 4: Parameter settings on Qwen-Image used in the main comparison tables. "$\times 2$" marks true-CFG evaluation of the positive and negative branches.

Method	**NFEs**	**Parameter setting**
Qwen-Image *- training-free acceleration*		
TeaCache	50×2	Accumulated relative ℓ_1-change criterion; thresholds $\tau \in \{0.75, 12.0\}$ for the two speed regimes.
DB-Taylor	50×2	DBCache threshold $\delta \in \{0.35, 1.0\}$, warm-up $K \in \{2, 0\}$; TaylorSeer p=1, (F_n, B_n)=$(1, 0)$ or $(5, 0)$, separate CFG branches.
RALU	$(4, 5, 5) \times 2$ / $(1, 2, 4) \times 2$	Three-stage schedule $\mathbf{N}$=$(4, 5, 5)$ or $(1, 2, 4)$; transitions $\mathbf{e}$=$(0.30, 0.45, 1.0)$ or $(0.20, 0.30, 1.0)$; HR ratio ρ=0.20 or 0.10.
SPEED	$(2, 1, 7) \times 2$ / $(1, 1, 5) \times 2$	S3 setting with total steps $10/7$, scales $(0.25, 0.5, 1.0)$, and δ=0.01; discrete schedules are derived from the Qwen shifted scheduler.
MrFlow	$(20, 1) \times 2$ / $(12, 1) \times 2$	Pixel-space staged true-CFG sampling with $(N_{\mathrm{LR}}, N_{\mathrm{ref}})$=$(20, 1)$ or $(12, 1)$; Real-ESRGAN $\times 2$ before refinement.
Qwen-Image *- training-dependent acceleration*		
Pi-Flow	4	Released pi-Qwen-Image 4-step policy model, FlowMap schedule as in the official adapter.
MrFlow + Pi-Flow	4, 1	LR Pi-Flow 4 steps, then Real-ESRGAN $\times 2$ and one HR refinement step.

Table 5: CLIP evaluation of Qwen-Image under an equal total step-count budget, averaged over 3 prompts $\times$ 3 seeds.

Schedule	Latency / s	$\mathrm{CLIP}_{\mathrm{ref}} \uparrow$	$\mathrm{CLIP}_{\mathrm{txt}} \uparrow$
`HR11`	10.96	0.897	0.359
`LR10_HR1`	**5.08**	**0.935**	**0.360**

The pixel space as the upsampling domain. Pixel-space upsampling allows MrFlow to reuse pretrained SR models and their learned image priors. The SR output is then mapped to the high-resolution latent space through the backbone's VAE encoder. Directly interpolating latent features can instead alter the local statistics expected by the decoder; the examples in Appendix F.1 show grid-like artifacts in the latent-upsampling baseline LSSGen. These comparisons complement the practical motivation for our pixel-space design.

Compatibility with downstream refinement. On Qwen-Image, we compare interpolation, SwinIR, OSEDiff, Real-ESRGAN, and Real-ESRGAN output without HR refinement. Table 7 shows broadly similar automatic scores, despite the perceptual differences in Figure 3, generated with 12 LR steps and 1 HR step. In these examples, interpolation and SwinIR remain blurry, OSEDiff introduces character inaccuracies, and Real-ESRGAN supplies sharper details that can be further adjusted by refinement. The SR-only comparison likewise shows similar automatic scores despite the visible effect of refinement.

The role of SR is therefore to provide both the enlarged layout and a useful detail initialization. In Figure 3, light refinement adjusts local deviations in the Real-ESRGAN output, including character strokes, while retaining the overall composition; smoother SR inputs remain less clear after refinement. These observations support Real-ESRGAN as the SR module in MrFlow.

Frequency-dependent sensitivity provides a complementary perspective: the perturbation analysis in Appendix C.3 shows that low-frequency perturbations disrupt the refined output more strongly

Table 6: Text-attention mass of Qwen-Image, averaged over 10 sampling steps and 3 prompts at each resolution.

Layer ℓ	$\mathcal{T}^{(\ell)}_{R=512}$	$\mathcal{T}^{(\ell)}_{R=1024}$
15	0.347	0.275
30	0.384	0.369
45	0.200	0.177

Table 7: Comparison of different super-resolution method choices on Qwen-Image, generated at 1024×1024 resolution.

SR method	NFEs	Speedup ↑	Geneval	DPG ↑	OneIG-En ↑	OneIG-Zh ↑
Interpolate	(12, 1)x2	10.7x	0.87	87.00	0.52	0.51
SwinIR	(12, 1)x2	4.67x	0.86	86.97	0.52	0.51
OSEDiff	(12, 1)x2	9.45x	0.85	87.25	0.51	0.49
Real-ESRGAN	(12, 1)x2	10.3x	0.86	87.10	0.52	0.51
Real-ESRGAN	12x2	14.0x	0.87	87.45	0.52	0.51

than the tested high-frequency perturbations. This is consistent with preserving a reliable coarse layout while adjusting local details. Appendix C.5 directly compares SR inputs with refined outputs to further characterize this division of work.

C.3 Frequency-Dependent Information Retention under Low-Strength Noising

SR supplies a high-resolution image with an established layout, but its added details can include distorted edges, character strokes, or textures inconsistent with the prompt. Too little noise leaves these errors difficult to change; too much can disturb the reliable structure. We use a local Gaussian model to characterize how the same noise level affects different information directions, then relate this characterization to measured frequency-band statistics and perturbation experiments.

Local model and direction-wise SNR. Let $\mathbf{z}_0^{\mathrm{SR}} = \mathcal{E}(\mathbf{x}_{\mathrm{SR}})$ be the SR latent. We approximate the clean-latent prior in its neighborhood by $\mathbf{Z}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \Lambda)$, where $\boldsymbol{\mu}$ is the local prior mean. In the orthonormal eigenbasis of Λ, λ_i denotes the clean-latent variance along direction i. For injected noise strength σ_t at refinement timestep t, the forward model is:

$$Z_{t,i} = (1-\sigma_t)Z_{0,i} + \sigma_t \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1), \tag{14}$$

where noise is independent of the clean latent. Centering the observation separates its signal and noise components:

$$Z_{t,i} - \mathbb{E}[Z_{t,i}] = (1-\sigma_t)(Z_{0,i} - \mu_i) + \sigma_t \epsilon_i, \tag{15}$$

with direction-wise SNR:

$$\mathrm{SNR}_i(\sigma_t) = \frac{(1-\sigma_t)^2 \lambda_i}{\sigma_t^2}. \tag{16}$$

Thus, directions with different clean-data variances have different SNRs even under isotropic noise.

Posterior-mean interpretation. The posterior mean provides an idealized MMSE view of clean-latent estimation (Efron, 2011; Vincent, 2011; Karras et al., 2022). Under the local Gaussian approximation, it has a closed form that characterizes the sensitivity of this estimate to the SR input.

Proposition 1 (Direction-wise retention of the SR input)**.** *For $0 < \sigma_t < 1$, consider the actual noised SR observation $z^{\mathrm{HR}}_{t,i} = (1-\sigma_t) z^{\mathrm{SR}}_{0,i} + \sigma_t \epsilon_i$. Under the local Gaussian model, its posterior-mean clean estimate is:*

$$\hat{z}_{0,i} = \kappa_i(\sigma_t) z^{\mathrm{SR}}_{0,i} + [1 - \kappa_i(\sigma_t)]\mu_i + \eta_i(\sigma_t)\epsilon_i, \tag{17}$$

Table 8: Measured clean-latent band scales and empirical noise-scale estimates.

Band	$\sqrt{P_B}$	Reference estimate $\tilde{\sigma}_B$
Low-freq $k < 0.10$	0.32	0.24
Mid-freq $0.10 \le k < 0.25$	0.18	0.16
High-freq $k \ge 0.25$	0.08	0.075

with coefficients:

$$\kappa_i(\sigma_t) = \frac{(1-\sigma_t)^2\lambda_i}{(1-\sigma_t)^2\lambda_i + \sigma_t^2} = \frac{\mathrm{SNR}_i(\sigma_t)}{1+\mathrm{SNR}_i(\sigma_t)}, \qquad \eta_i(\sigma_t) = \frac{\sigma_t}{1-\sigma_t}\kappa_i(\sigma_t). \tag{18}$$

Proof. For jointly Gaussian variables, the conditional mean follows from their covariance and observation variance (Bishop & Nasrabadi, 2006, Eq. (2.81)). Here, we have:

$$\begin{aligned} \mathbb{E}[Z_{t,i}] &= (1-\sigma_t)\mu_i, \\ \mathrm{Var}(Z_{t,i}) &= (1-\sigma_t)^2\lambda_i + \sigma_t^2, \\ \mathrm{Cov}(Z_{0,i}, Z_{t,i}) &= (1-\sigma_t)\lambda_i. \end{aligned} \tag{19}$$

Consequently, we obtain:

$$\mathbb{E}[Z_{0,i} \mid Z_{t,i} = z_{t,i}^{\mathrm{HR}}] = \mu_i + \frac{(1-\sigma_t)\lambda_i}{(1-\sigma_t)^2\lambda_i + \sigma_t^2}\big[z_{t,i}^{\mathrm{HR}} - (1-\sigma_t)\mu_i\big]. \tag{20}$$

Substituting the actual noised SR observation gives the stated decomposition. □

The Wiener gain $\kappa_i(\sigma_t)$ measures how strongly the estimate retains the SR input along direction i. At a fixed noise level, a larger λ_i gives greater retention, whereas a smaller λ_i allows the local prior to exert more influence. This is the mechanism captured by the model: reduced dependence on uncertain SR details while retaining directions with stronger signal.

A reference lower-bound estimate. To identify a noise scale that substantially reduces SR retention, we use equal signal and noise power as a reference. For a representative high-frequency direction with variance λ_{hf}, this yields the following model-level criterion.

Proposition 2 (Equal-power noise scale)**.** *Under Proposition 1,* $\mathrm{SNR}_{\mathrm{hf}}(\sigma_t) \le 1$*, equivalently* $\kappa_{\mathrm{hf}}(\sigma_t) \le \frac{1}{2}$*, if and only if:*

$$\frac{\sigma_t}{1-\sigma_t} \ge \sqrt{\lambda_{\mathrm{hf}}} \qquad \Longleftrightarrow \qquad \sigma_t \ge \sigma_t^\star = \frac{\sqrt{\lambda_{\mathrm{hf}}}}{1+\sqrt{\lambda_{\mathrm{hf}}}}. \tag{21}$$

This follows by rearranging $\sigma_t^2 \ge (1-\sigma_t)^2\lambda_{\mathrm{hf}}$. We use $\sigma_t^\star$ as a reference lower-bound estimate for reducing high-frequency SR retention. This scale guides noise injection to balance detail modification and layout preservation.

Frequency-band measurements. We characterize clean-latent frequency scales using band-pass FFT on 18 high-resolution latents of shape $16 \times 128 \times 128$ from the main experiments. We define k as radial spatial frequency divided by the maximum radial frequency on the latent grid. Let P_B denote the variance of the spatial-domain component retained by band B. The reported Parseval check, comparing the summed band powers with the total variance, has a ratio of 1.000.

The band power P_B aggregates contributions from multiple frequency modes. We use the measured band scales as empirical proxies, defining the reference estimate $\tilde{\sigma}_B = \sqrt{P_B}/(1+\sqrt{P_B})$.

Table 8 shows a substantially smaller characteristic scale in high frequency than in low frequency. Its high-frequency reference of approximately 0.075 is of the same order as the practical noise level $\sigma_t \in [0.1, 0.15]$. Together with the direction-wise model, this supports the design intuition that modest noising can facilitate detail changes without restarting global structure generation.

Table 9: DINOv2 nearest-neighbor distance of Qwen-Image outputs under different injected perturbations, averaged over 3 prompts $\times$ 3 seeds.

Noise component	$\sigma_t = 0.2$	$\sigma_t = 0.3$
SR input (no refine)	0.155	0.155
Gaussian	0.154	0.155
High-freq	0.150	0.187
Channel-correlated	0.164	0.180
Low-freq	0.252	0.498

Table 10: CLIP similarity between Qwen-Image high-resolution refinement and the 8-step reference, averaged over 3 prompts.

Strength σ_t	$K_H = 1$	$K_H = 2$	$K_H = 3$	$K_H = 5$
0.1	**0.9974**	0.9995	0.9997	0.9999
0.3	0.9902	0.9957	0.9962	0.9996

Frequency-dependent perturbation sensitivity. We further replace the default Gaussian injection with predominantly high-frequency, low-frequency, or channel-correlated noise and examine the refined output. Low- and high-frequency perturbations are rescaled to match the Gaussian noise variance. Table 9 reports nearest-neighbor cosine distance to native high-resolution reference images of the same prompt in DINOv2 (Oquab et al., 2023) feature space. We use $\sigma_t \in \{0.2, 0.3\}$ to compare perturbation responses at higher noise levels.

Low-frequency perturbations produce much larger feature-space deviations than Gaussian or high-frequency perturbations, especially at $\sigma_t = 0.3$. The high-frequency variant also has a greater distance to native references at the larger setting, indicating that the noise strength still matters. This frequency-dependent response supports protecting the coarse structure during lightweight refinement.

C.4 Empirical Analysis of Single-Step Refinement

This section examines the empirical behavior of high-resolution refinement under low-strength noising. Step-count comparisons assess whether additional refinement steps are needed, while sampled prediction statistics characterize the corresponding short denoising trajectories.

Comparison with multi-step refinement. We compare $K_H \in \{1, 2, 3, 5\}$ refinement steps with an 8-step reference, keeping the SR input and refinement noise seed fixed. We test noise strengths $\sigma_t \in \{0.1, 0.3\}$, with $\sigma_t = 0.3$ included as a larger-noise diagnostic setting. Agreement is measured by CLIP image similarity:

$$\mathrm{CLIP}_{\mathrm{ref}}(\hat{\mathbf{x}}_0^{(K_H)}, \hat{\mathbf{x}}_0^{(8)}) \;=\; \cos\big(\boldsymbol{\phi}_I(\hat{\mathbf{x}}_0^{(K_H)}), \boldsymbol{\phi}_I(\hat{\mathbf{x}}_0^{(8)})\big). \tag{22}$$

Table 10 shows that a single step at $\sigma_t = 0.1$ already reaches 0.9974 similarity to the reference, with additional steps yielding small gains in agreement. At $\sigma_t = 0.3$, the single-step gap is larger. The visual comparison in Figure 4 further shows that one HR step captures the main visible improvements. Together, these observations support $K_H = 1$ as an efficient default in the low-strength regime.

Discrete prediction statistics. To characterize these trajectories, we record the raw conditional predictions $\{\mathbf{q}_k\}_{k=0}^{K-1}$ of the Qwen-Image denoising transformer along $K = 6$ refinement steps. For prediction dimension D, we measure the mean per-coordinate magnitude, adjacent-prediction

Table 11: Discrete statistics of raw conditional predictions during Qwen-Image high-resolution refinement, 3 prompts $\times$ 2 seeds. Magnitudes and differences are normalized by $\sqrt{D}$.

Strength σ_t	m_q	d_q	θ_q / rad
0.05	0.924	0.095	0.087
0.1	0.970	0.119	0.107
0.2	1.002	0.140	0.126
0.3	1.017	0.151	0.136
0.5	1.025	0.170	0.157

difference, and angle:

$$m_q = \frac{1}{K}\sum_{k=0}^{K-1}\frac{\|\mathbf{q}_k\|_2}{\sqrt{D}}, \tag{23}$$

$$d_q = \frac{1}{K-1}\sum_{k=0}^{K-2}\frac{\|\mathbf{q}_{k+1}-\mathbf{q}_k\|_2}{\sqrt{D}}, \tag{24}$$

$$\theta_q = \frac{1}{K-1}\sum_{k=0}^{K-2}\arccos\frac{\langle \mathbf{q}_{k+1}, \mathbf{q}_k\rangle}{\|\mathbf{q}_{k+1}\|_2\|\mathbf{q}_k\|_2}. \tag{25}$$

Table 11 reports their averages over 3 prompts and 2 seeds. Prediction magnitudes vary modestly across the tested strengths, while adjacent differences and angles are smaller in the low-strength runs. These measurements characterize the limited prediction changes along the sampled low-strength refinement trajectories.

C.5 Structure Generation and Detail Refinement

We examine the division of labor between low-resolution generation and high-resolution refinement through two complementary comparisons: how the final image responds to each stage's randomness, and what actually changes from the SR input to the refined output.

Measurement protocol. We use 3 prompts and 4 seeds per condition on Qwen-Image. In $\mathrm{vary_LR}$, the HR refinement seed is fixed while the LR seed varies; in $\mathrm{vary_HR}$, the LR and SR outputs are fixed while the HR seed varies. Each condition yields $\binom{4}{2} = 6$ output pairs per prompt, or 18 pairs in total. Images are resized to 256×256 with bicubic interpolation and represented as RGB values in $[0, 1]$. We obtain a low-frequency component $L(\mathbf{x})$ using a Gaussian blur of radius 16 pixels and compute the high-frequency residual $H(\mathbf{x}) = \mathbf{x} - L(\mathbf{x})$ by floating-point subtraction.

For each image pair $(\mathbf{x}, \mathbf{y})$, we compute the mean squared difference over pixels and channels for the full image, L, and H.

Sensitivity to stage-specific randomness. Table 12 shows that varying the LR seed induces about 8000 times greater low-frequency change and 1900 times greater full-image change than varying the HR seed. The HR-seed changes are much smaller overall and relatively more concentrated in the high-frequency residual, with a mean high/low ratio of 20.5 versus 7.9 for LR-seed changes. This asymmetry supports the LR stage as the primary source of global structure, with HR randomness contributing mainly small-scale variation.

Changes introduced by refinement. To complement the seed-sensitivity experiment, we directly compare each SR input with its corresponding refined output in the fixed-LR condition, giving 12 input-output pairs (3 prompts $\times$ 4 HR seeds). Applying the same decomposition, Table 13 shows small low-frequency changes for every prompt, with larger changes in the high-frequency residual.

The direct comparison complements the seed experiment: HR refinement preserves the coarse layout while changing finer-scale content. The visual examples in Figure 3 further show how refinement adjusts local artifacts and character strokes. Taken together, these observations support the intended

Table 12: Stage-specific seed sensitivity on Qwen-Image: 3 prompts $\times 6$ pairs, or 18 pairs per condition. Values are mean pairwise MSEs and mean per-pair high/low ratios.

Control variable	low-freq MSE	high-freq MSE	full-image MSE	high/low ratio
vary_LR	1.087×10^{-2}	3.282×10^{-2}	4.842×10^{-2}	7.9
vary_HR	1.298×10^{-6}	2.671×10^{-5}	2.575×10^{-5}	20.5

Table 13: Direct SR-to-refined image differences on the same saved Qwen-Image samples. Each prompt averages 4 input-output pairs; all MSE values are in units of 10^{-5}.

Prompt ID	low-freq MSE	high-freq MSE	full-image MSE
1	0.692	1.889	2.008
2	0.563	4.454	4.471
3	0.445	2.843	2.775
Mean	0.567	3.062	3.084

division of labor: the LR stage primarily establishes global structure, while HR refinement preserves that structure and improves local details.

Prompts. Both experiments use the following prompts, numbered as in Table 13:

1. A university lecture hall chalkboard covered in handwritten physics equations in white chalk, including 'E = mc^2', '$\nabla \times$ B $-$ (1/c) ∂E/∂t = (4π/c) J', 'ψ(x,t) = A e^(i(kx $-$ ωt))', and a small labeled diagram of a pendulum with angle θ and length L. A piece of chalk and a dusty eraser rest on the ledge. Soft morning light from a tall window on the left, cinematic, 50mm.
2. A vintage jazz concert poster pinned to a brick wall. The poster text, in clean art-deco typography, reads:
'MIDNIGHT BLUE QUARTET'
'Live at The Blue Room'
'Friday, March 14th - 9:00 PM'
'Tickets $25 at the door'
Gold and teal color scheme, subtle paper grain, slight corner curl, studio product photography lighting.
3. A cozy corner bookstore at dusk with a large wooden hanging sign that reads 'THE WANDERING PAGE' in hand-painted serif letters, and below it a smaller chalkboard sign saying 'Est. 1987 - Open 9am to 9pm'. Warm amber window light, rain-slick cobblestone street, autumn leaves, shallow depth of field, 35mm film photograph.

D MrFlow on FLUX.2 Klein

To further examine whether the same staged design transfers beyond the two main backbones, we additionally evaluate MrFlow on FLUX.2 Klein (Black Forest Labs, 2025). The results are summarized in Table 14. For the non-base FLUX.2 Klein variants, which already use a reduced-step generation regime, we report the speedup relative to the corresponding base model at the standard 50-step setting, so that the number reflects the practical acceleration limit against the full model in the same family.

On FLUX.2 Klein Base, MrFlow keeps the same behavior observed on FLUX.1-dev and Qwen-Image. With 20 low-resolution steps and 1 high-resolution step, it reaches $5.16\times$ and $5.40\times$ speedup on the 4B and 9B variants, respectively, while preserving the OneIG-En score within a very small gap. The more aggressive $12, 1$ configuration further improves the speedup to $8.03\times$ and $8.79\times$, with a moderate metric drop. The distilled Klein variants also remain compatible with the multi-

Table 14: Extended results on FLUX.2 Klein variants at 1024×1024 resolution. Speedups of non-base variants are measured relative to the corresponding Klein Base 50-step model with the same model scale. Here "$\times 2$" denotes the effective positive/negative CFG branches.

Model variant	Method	Training-free	NFEs	Speedup ↑	Geneval ↑	DPG ↑	OneIG-En ↑
FLUX.2-klein-base-4B	Native	-	50×2	1x	0.78	83.05	0.49
FLUX.2-klein-base-4B	MrFlow	$\surd$	$(20,1)\times2$	5.16x	0.75	83.24	0.49
FLUX.2-klein-base-4B	MrFlow	$\surd$	$(12,1)\times2$	8.03x	0.74	82.63	0.48
FLUX.2-klein-4B	Native	-	4	20.5x	0.84	85.78	0.49
FLUX.2-klein-4B	MrFlow‡	$\ddagger$	4, 1	20.1x	0.83	85.17	0.50
FLUX.2-klein-base-9B	Native	-	50×2	1x	0.83	85.84	0.54
FLUX.2-klein-base-9B	MrFlow	$\surd$	$(20,1)\times2$	5.40x	0.79	85.09	0.54
FLUX.2-klein-base-9B	MrFlow	$\surd$	$(12,1)\times2$	8.79x	0.77	84.62	0.54
FLUX.2-klein-9B	Native	-	4	22.5x	0.84	86.31	0.54
FLUX.2-klein-9B	MrFlow‡	$\ddagger$	4, 1	26.9x	0.87	85.92	0.54

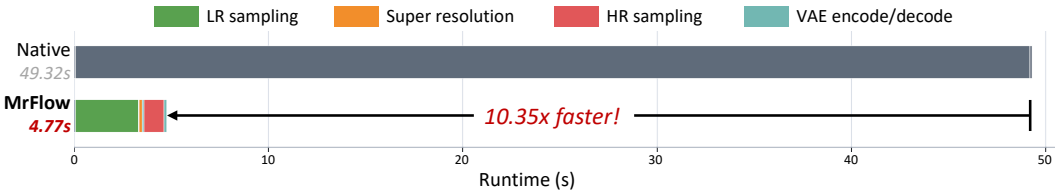

Figure 6: Stage-wise runtime breakdown of native Qwen-Image and MrFlow-$12, 1$.

resolution pipeline: adding one high-resolution refinement step to the 4-step model reaches $20.10\times$ on 4B and $26.92\times$ on 9B relative to their base counterparts, denoted as MrFlow‡ in the table.

E Efficiency Analysis

We decompose the end-to-end runtime of MrFlow to clarify where the acceleration comes from. Taking the $12{+}1$ setting on Qwen-Image as an example, we measure all stages in the inference path, including text encoding, low-resolution initial noise generation, low-resolution denoising, VAE decoding, pixel-space super-resolution, VAE encoding, high-resolution noise generation and noising, high-resolution denoising, and the final VAE decoding. The average end-to-end latency of this configuration is 4.77s, compared with 49.32s for native 50-step Qwen-Image, corresponding to a practical speedup of $10.35\times$.

Figure 6 visualizes the same measurement as a left-aligned horizontal runtime breakdown. The dominant costs of MrFlow are low-resolution sampling and high-resolution refinement, which take 3.24s and 1.03s, respectively. In contrast, the fixed overhead beyond denoising is small: the low-resolution initial noise generation, intermediate VAE decoding, Real-ESRGAN super-resolution, and VAE encoding together take about 0.30s. Therefore, the efficiency gain mainly follows from two factors. First, each low-resolution denoising step processes only about $1/4$ of the image tokens. Second, the low-resolution stage can establish the global structure with substantially fewer steps. Since pixel-space super-resolution and the additional VAE transformations introduce only a small fixed cost, MrFlow converts a large portion of expensive high-resolution sampling into cheaper low-resolution sampling while retaining one high-resolution refinement step for local details.

F More Generation Examples

Rasterized figures are compressed by over 90% to keep the PDF compact.

Table 15: Stage-wise runtime breakdown on Qwen-Image at 1024×1024 resolution. Times are reported in seconds. Blank entries indicate stages that are not present in the corresponding pipeline.

Method	NFE	Text enc.	LR noise	LR samp.	Mid VAE dec.	SR	VAE enc.	HR noise	HR samp.	Final VAE dec.	Total
Native	50×2	0.067	-	-	-	-	-	0.0025	49	0.14	49
MrFlow	$(12,1)\times2$	0.067	0.00012	3.2	0.035	0.18	0.083	0.0025	1.0	0.14	4.8



Figure 7: Comparison of the effects of various methods on FLUX.1-dev. The dashed lines separate the pretrained model, training-free strategies, and strategies that rely on or exploit timestep distillation.

F.1 Comparison with Various SOTA Strategies

Beyond the quantitative metrics in Section 4.2 of the main text, we observe that the gap between methods is even more noticeable when directly inspecting generation examples.

Benchmark samples. In addition to the advantage in the quantitative evaluation of Section 4.2, MrFlow in fact also looks better on the actually generated images. Figure 7 directly extracts generation examples from the DPG-Bench task on FLUX.1-dev. MrFlow significantly surpasses other existing training-free strategies in both generation quality and efficiency, and MrFlow†, which fuses timestep distillation without extra training, further attains a higher speedup.

Multi-resolution strategy comparison. Figure 8 further compares MrFlow with other multi-resolution acceleration strategies, including LSSGen, RALU, and SPEED. LSSGen already exhibits dense artifacts even at a moderate speedup, while RALU and SPEED can still obtain competitive quantitative scores on FLUX.1-dev at speedups above $8\times$, with SPEED even partially exceeding MrFlow on some metrics in Table 1. However, the generated samples reveal a clear gap in perceptual fidelity: these baselines exhibit severe local degradation, including blurred fine details, unstable textures, or collapsed structures. In contrast, MrFlow maintains sharp and coherent details at comparable or higher acceleration levels, suggesting that pixel-space super-resolution followed by low-strength high-resolution resampling is more reliable than latent- or frequency-domain expansion.

More scenarios. We also test some more diverse prompts on Qwen-Image to further examine the accelerated generation capability of each method. Figure 1 shows some examples. Consistent with the observation on FLUX.1-dev, MrFlow can not only significantly outperform various other advanced training-free strategies, but also be directly combined with training-dependent timestep distillation strategies to attain a higher speedup.

F.2 More Examples of MrFlow

In Figures 10 and 9, we show the generation effects of MrFlow performing acceleration under the configuration of 12 low-resolution steps and 1 high-resolution step on FLUX.1-dev and Qwen-Image respectively. At a speedup of 8-$10\times$, MrFlow demonstrates excellent generation effects at arbitrary aspect ratios and resolutions.



Figure 8: Detail comparison of multi-resolution acceleration methods on FLUX.1-dev.

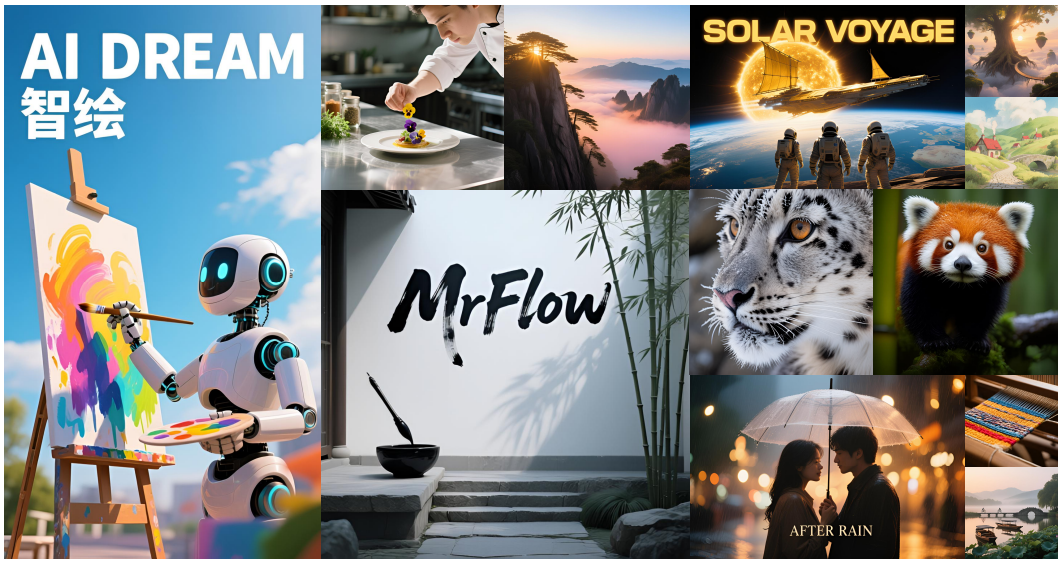

Figure 9: Generation examples of MrFlow on Qwen-Image. The configuration is 12 steps at the low-resolution stage and 1 step at the high-resolution stage, with speedups all above $10\times$.

F.3 Prompts of the Images in the Paper

For reproducibility and easier inspection of qualitative examples, we list below the text prompts used for the images shown in the paper. For figures containing multiple samples, prompts are listed in the same order as the corresponding images.

Prompts for Figure 1.

1. Close-up of a white horse in a spring pasture, sunlight on mane, clear eye reflection, wildflowers, airy realistic photography.
2. Mars rice greenhouse, shallow water beds, red dunes outside, agronomists, robots, warm lamps, detailed sci-fi realism.

Prompt for Figure 2.

1. Close-up of a chubby golden British shorthair on a knitted blanket, tiny ears, plush cheeks, bright cozy bedroom light.

Prompt for Figure 3.

1. A cozy corner bookstore at dusk with a large wooden hanging sign that reads 'THE WANDERING PAGE' in hand-painted serif letters, and below it a smaller chalkboard sign saying



Figure 10: Generation examples of MrFlow on FLUX. The configuration is 12 steps at the low-resolution stage and 1 step at the high-resolution stage, with speedups all above $8\times$.

'Est. 1987 - Open 9am to 9pm'. Warm amber window light, rain-slick cobblestone street, autumn leaves, shallow depth of field, 35mm film photograph.

Prompt for Figure 4.

1. Top-down view of a watchmaker's workbench, featuring a disassembled mechanical wristwatch with tiny gears, jewels, screws and springs laid out on a green felt mat. A loupe, tweezers, and a brass screwdriver set sit beside a leather-bound notebook with the handwritten heading 'Caliber 2824-2 - Service Log'. Warm tungsten desk lamp, hard shadows, macro photograph at f/5.6, razor-sharp focus across the scene.

Prompt for Figure 7.

1. An expansive field, blanketed by the soft light of morning, cradles a collection of eight cabbages, their green heads round and plump. These vegetables are nestled among rows of rich soil, dotted with glistening droplets of dew that cling to their crinkled leaves. As wisps of mist begin to lift, the cabbages lie poised, ready for the day's impending harvest.

Prompt for Figure 8.

1. An immaculate white vanity desk featuring an array of beauty products, among which a soft-bristled cosmetics brush and a sleek black eyeliner pencil are neatly arranged. To the

side, a stark and dissonant element, a metallic handgun, lies on the textured surface of a grey concrete floor, creating a jarring juxtaposition against the delicate makeup tools. The vanity itself is positioned against a white wall, illuminated by the natural light streaming in from a nearby window.

Prompts for Figure 9. The items marked with † were originally written in Chinese; they are shown here as English translations or English descriptions.

1. (†) Vertical AI art festival poster, a friendly AI robot is painting a glowing mural by itself in daylight, robot arm holding a brush, vivid colors, clean promotional cover composition, no human painter; only readable text: "AI DREAM" and a Chinese phrase meaning "AI painting"; strictly no other letters, no extra words, no logos, no signatures, no captions, no watermark.
2. A focused young chef with rolled-up sleeves carefully placing edible flowers on a finely plated dish in a sunlit open kitchen, gleaming stainless-steel surfaces, herbs and spice jars softly out of focus in the background, 35mm reportage style.
3. (†) The main peak of Huangshan at sunrise, bathed in golden morning glow, with ancient pine trees leaning out from steep granite cliffs, a rolling sea of clouds tinted pale orange and rose, distant mountain layers fading into cool blue, ultra-wide panoramic view, sharp details under natural light.
4. (†) A courtyard calligraphy wall with the only readable text "MrFlow", surrounded by brushes, an ink bowl, stone steps, and bamboo shadows, with no other readable text.
5. Bright sci-fi film poster, three astronauts facing a golden solar sail ship, Earth glow, heroic composition, only readable text: "SOLAR VOYAGE".
6. A colossal ancient tree with sprawling roots breaking through layers of clouds at sunrise, tiny floating islands suspended around its branches, sunlight pouring through the foliage onto a winding skybridge, painterly fantasy concept art with a warm saturated palette.
7. A cozy hillside village in Studio Ghibli illustration style, hand-painted clouds drifting over rolling green hills, small red-roofed cottages with smoking chimneys, a winding cobblestone path leading to a stone bridge, warm afternoon light, soft watercolor textures.
8. Extreme close-up of a snow leopard face, frost on whiskers, amber eye, wet nose texture, snowy rocks blurred behind, realistic wildlife photo.
9. Close-up of a red panda on a mossy branch, damp fur, tiny claws, round eyes, bamboo forest rain blur, detailed wildlife realism.
10. Romantic drama poster, two figures under a clear umbrella after rain, warm city lights, soft cinematic bokeh, only readable text: "AFTER RAIN".
11. Woven resolution tapestry, coarse threads becoming fine image detail, wooden loom, colored yarn, tactile macro scene.
12. Hangzhou lake at sunrise, wooden boats, lotus leaves, stone bridge, misty hills, cyclists on shore, soft realistic light.

Prompts for Figure 10.

1. Orbital lighthouse array, huge mirrors, maintenance craft, Earth glow, astronauts, hard sci-fi panorama.
2. Ceramic table arrangement, celadon bowls, tea, wet leaves, linen, bronze spoon, window light, macro material detail.
3. A majestic snow leopard resting on a moss-covered rock at the edge of a Himalayan cliff, soft alpine sunlight on its thick spotted fur, distant snow-capped peaks in the background, sharp golden eyes, wildlife photograph at 200mm with anamorphic depth.
4. An extreme macro photograph of a morpho butterfly's iridescent blue wing with visible scale tile structure, tiny dewdrops clinging to a few scales acting as natural lenses, soft greenery as a cool blurred background, 1:2 magnification.

5. Red alpine train crossing a stone viaduct above pine forest, waterfalls, misty peaks, tiny hikers, crisp morning light.
6. Jade palace floating above a cloud sea, cranes, waterfalls, tiled roofs, glowing bridges, bright Chinese fantasy painting.
7. River rescue training exercise, inflatable boats, ropes, helmets, fast water, command tents, realistic documentary action.
8. A focused young chef with rolled-up sleeves carefully placing edible flowers on a finely plated dish in a sunlit open kitchen, gleaming stainless-steel surfaces, herbs and spice jars softly out of focus in the background, 35mm reportage style.
9. Subway signal repair, workers in tunnel, orange lights, cables, tools, train blur, realistic infrastructure action.
10. AI chip foundry floor, wafer carriers, robotic arms, cleanroom suits, amber reflections, precise industrial photograph.
11. Alpine seed collection, botanists, tiny flowers, rock slopes, clouds, specimen boxes, crisp realistic macro landscape.
12. An elderly Japanese craftsman with white hair and round wire-rim glasses focused on carving a small wooden box in his sunlit workshop, soft sawdust floating in the warm light, neatly arranged hand tools on the workbench, traditional indigo apron with subtle wear, 50mm photograph with shallow depth of field.
13. Text-free science thriller poster, research lab on the rim of an active volcano, orange lava glow, white steam, scientists in suits, no text, no logos.
14. Promotional event cover, bright outdoor technology fair, robots, image screens, families, clean daylight, only readable text: “FUTURE FAIR”.
15. Only the word ’MrFlow’ sculpted from clouds above green hills, balloons, sunlight, no other readable text.
16. Adventure poster, huge centered title “DRIFT” in windblown white letters, small sailboat under a giant cloud, turquoise sea, sunny cinematic poster, no other text.
17. Mediterranean boat repair dock, painted hulls, ropes, gulls, buckets, sunlit water reflections, busy realistic scene.
18. Coastal drama poster, huge top title “TIDE” in white painted letters, old lighthouse, crashing waves, yellow raincoat figure, cinematic clouds, no other text.
19. A ruby-throated hummingbird hovering with rapid wing motion blur in front of a bright trumpet vine flower, droplets of nectar caught mid-air, lush greenery softly out of focus behind, 1/4000s shutter, vivid macro photograph.
20. Text-free space blockbuster poster, astronaut reaching toward a damaged satellite above Earth, golden sunlight, debris field, heroic composition, no text, no logos.
21. Hopeful sci-fi cover, huge centered title “HALO” in soft luminous letters, child and service robot under a ring-shaped sky station, clean daylight, no other text.
22. Close-up portrait of a little Chinese girl in red hanfu, soft courtyard sunlight, delicate embroidery, warm cinematic bokeh.
23. Action poster, huge centered title “BLAZE” in fiery orange letters, red sports car and sparks on a desert road, dramatic sunset, no other text.