Tuning Real-World Image Restoration at Inference: A Test-Time Scaling Paradigm for Flow Matching Models

Purui Bai[1], Junxian Duan[1], Pin Wang[1], Jinhua Hao[2], Ming Sun[2], Chao Zhou[2], and Huaibo Huang[1] *

[1] MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Kuaishou Technology, Beijing, China
{purui.bai,junxian.duan,huaibo.huang}@cripac.ia.ac.cn
wangpin2026@ia.ac.cn
{haojinhua,sunming03,zhouchao}@kuaishou.com

Abstract. Although diffusion-based real-world image restoration (Real-IR) has achieved remarkable progress, efficiently leveraging ultra-large-scale pre-trained text-to-image (T2I) models and fully exploiting their potential remain significant challenges. To address this issue, we propose **ResFlow-Tuner**, an image restoration framework based on the state-of-the-art flow matching model, FLUX.1-dev, which integrates unified multi-modal fusion (UMMF) with test-time scaling (TTS) to achieve unprecedented restoration performance. Our approach fully leverages the advantages of the Multi-Modal Diffusion Transformer (MM-DiT) architecture by encoding multi-modal conditions into a unified sequence that guides the synthesis of high-quality images. Furthermore, we introduce a training-free test-time scaling paradigm tailored for image restoration. During inference, this technique dynamically steers the denoising direction through feedback from a reward model (RM), thereby achieving significant performance gains with controllable computational overhead. Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple standard benchmarks. This work not only validates the powerful capabilities of the flow matching model in low-level vision tasks but, more importantly, proposes a novel and efficient inference-time scaling paradigm suitable for large pre-trained models. The source code is publicly available at `https://github.com/Rorschach-1010/ResFlow-Tuner`.

Keywords: Image Restoration · Test-Time Scaling · Flow Matching Model

* Corresponding author.

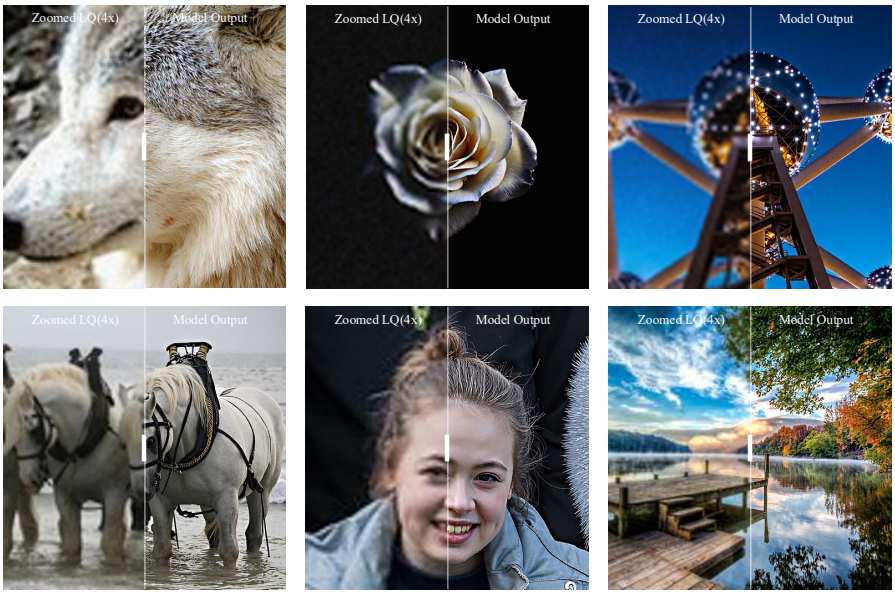

Fig. 1: ResFlow-Tuner delivers superior performance on both synthetic (the first row) and real-world (the second row) benchmarks, excelling in terms of perceptual quality and objective image quality assessment.

1 Introduction

Image restoration (IR), a vital field in computer vision, targets transforming degraded low-quality (LQ) images into high-quality (HQ) counterparts. While IR has achieved significant advancements under predefined conditions, such as super-resolution [10, 39] and denoising [45] tasks, real-world IR remains a formidable challenge due to the diversity and complexity of degradation types.

The field of image restoration has been revolutionized by generative models [12]. Numerous studies have shown that, in terms of restored image quality, SD-based methods outperform GAN-based ones. While diffusion models, which learn to reverse a stochastic degradation process, have set impressive benchmarks, a new paradigm is emerging: flow-based generative models [22], which leverage Ordinary Differential Equations (ODEs) to learn a deterministic and often more efficient mapping from noise to data. Previous research [20,30,31]has shown that this ODE formulation captures a more coherent trajectory in the latent space, leading to superior sample quality and finer granularity in generated details compared to their Stochastic Differential Equation (SDE)-based diffusion counterparts.

However, harnessing such powerful flow-based foundations for image restoration (IR) presents unique challenges. First, the scale of models like FLUX makes full fine-tuning computationally intractable. While parameter-efficient methods

(e.g., LoRA [18]) are necessary, they alone may not fully exploit the model's capabilities. Second, mainstream conditional mechanisms for IR typically employ a relatively straightforward method [2, 29, 52]of direct conditional injection, limiting the utilization of priors in pre-trained models. In this paper, we introduce **ResFlow-Tuner**, a novel framework not only adapts the flow model for IR but also introduces a pioneering test-time scaling (TTS) strategy specifically designed for its ODE dynamics. Our framework features a Unified Multi-Modal Fusion (UMMF) mechanism [13]. This design enables all semantic tags to interact directly within the multimodal attention mechanism, thereby ensuring highly faithful reconstruction that is accurate to both the pixel-level input and the high-level semantic guidance. Besides, we redesign the training-free TTS by introducing perturbations to intermediate states along the denoising ODE trajectory. The quality of these perturbed states is assessed by an ensemble of reward models, with each model providing a unique perspective on image restoration quality to guide the navigation process of the deterministic ODE trajectory. This creates a new paradigm for image restoration frameworks: dynamically trading computation for enhanced quality by exploiting the ODE structure.

We summarize the contributions of this work as follows:

- **Flow-Based IR Framework:** We present the in-depth exploration of the FLUX flow model and its multimodal processing mechanism for image restoration, augmented by VLM-derived textual conditioning for pixel-aware and semantically-aware restoration.
- **ODE-adapted Test-Time Scaling:** We pioneer the adaptation of test-time scaling for ODE-based flow models image restoration framework, introducing a novel, training-free method to dynamically steer the deterministic sampling trajectory during the restoration process, establishing a new cost-performance trade-off.
- **State-of-the-Art Performance:** We validate through rigorous experimentation that our method surpasses existing diffusion-based and other approaches, providing a powerful new baseline for future research.

2 Related Work

Image Restoration with Pre-trained Generative Models. Leveraging large-scale pre-trained generative models as priors to solve the ill-posed problem of image restoration has become a significant research trend. These methods aim to harness the powerful generative capabilities of such models to recover plausible and detailed high-quality images. A substantial body of research has been built upon Stable Diffusion (SD) models, such as StableSR [52], DiffBIR [29], and SeeSR [56]. Building on this foundation, later studies have migrated to more powerful diffusion backbones [8, 38], including SUPIR [63], FaithDiff [9], and DreamClear [3]. Recently, methods based on flow models [33, 50, 57] have shown greater potential. Flow models learn a deterministic mapping from noise to data through Ordinary Differential Equations (ODEs), resulting in more coherent trajectories and often generating higher-quality samples with finer details. Building

upon this frontier, we propose a novel image restoration framework based on the FLUX architecture. By integrating multimodal conditional guidance from both low-resolution images and semantic text, our framework delivers enriched contextual information throughout the restoration process.

Test-Time Scaling for Generative Models. The core idea of Test-Time Scaling (TTS) techniques is to dynamically mine and enhance the potential performance of pre-trained models during inference by investing additional computational resources, without modifying the model weights. This set of methods has achieved remarkable success in large language models [65]and is rapidly expanding into the visual generation domain. In the visual domain, the most straightforward TTS method is Best-of-N sampling [44, 47, 49], which directly selects the sample with the highest reward signal from a large number of randomly generated candidates. Subsequently developed Particle Sampling methods [21,24,25,41] retain high-performing sample branches and prune poorer ones during the denoising process, akin to beam search. However, their deterministic nature, due to the lack of stochasticity in the generative process, limits their applicability within more advanced diffusion and flow-based generative models. Furthermore, existing search methods heavily rely on the quality of the initial candidate particles and fail to actively explore new regions of the solution space. To address these limitations, we introduce a novel test-time extension paradigm, specifically designed for the ODE trajectories of flow models in image restoration. This training-free approach dynamically enhances output quality during inference, establishing a new computational-quality Pareto frontier for large models in low-level vision tasks.

3 Method

3.1 Flow-based Image Restoration Framework

Our goal is to leverage the powerful generative prior of the FLUX model for image restoration, guided by both the LQ image and semantic text cues, As illustrated in Fig. 2, our framework consists of two main components: (1) a condition preparation stage, comprising a lightweight degradation removal network and a multimodal large language model for generating textual descriptions; and (2) a conditional injection mechanism, specifically designed for our MM-DiT-based flow model to guide the image restoration process.

Stage-I: Multi-Modal Condition Preparation

– **Reference Image Generation via SwinIR:** Directly processing the heavily degraded I_{LR} with a Vision-Language Model (VLM) can lead to inaccurate semantic interpretations. To mitigate this, we first employ a pre-trained SwinIR model $\mathcal{G}_{SwinIR}$ to obtain a preliminary enhanced image I_{Ref}:

$$I_{Ref} = \mathcal{G}_{SwinIR}(I_{LR}). \tag{1}$$

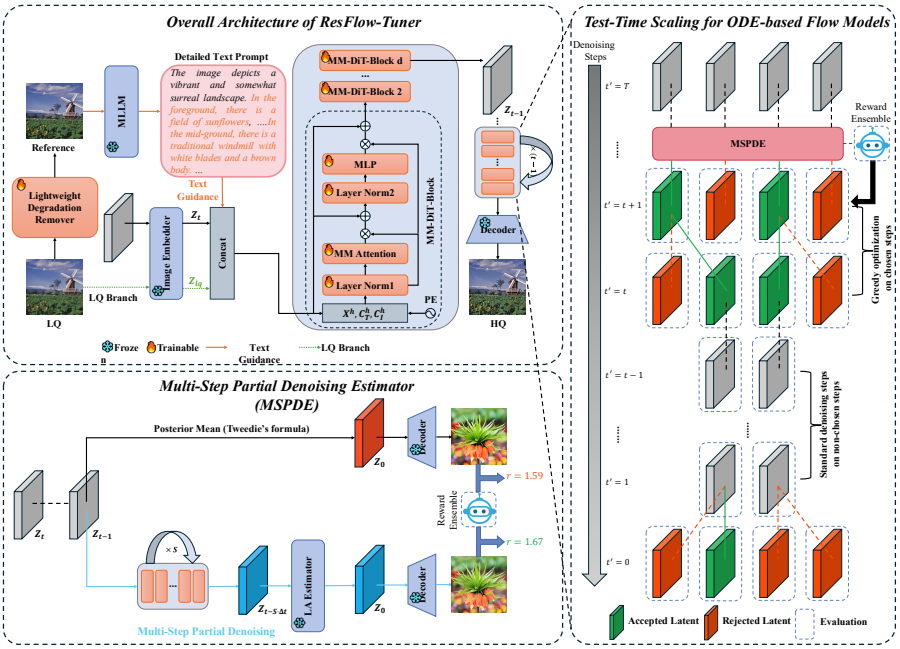

Fig. 2: Architecture of the proposed **ResFlow-Tuner**. ResFlow-Tuner enhances training performance through the seamless integration of multi-modal guidance. During inference, it adopts a greedy optimization strategy for path selection, augmented by our Multi-Step Partial Denoising Estimator (MSPDE) for more accurate path evaluation.

– **Semantic Context Extraction via VLM:** The reference image I_{Ref} is fed into a powerful VLM $\mathcal{F}_{VLM}$ (e.g., Qwen-2.5-VL) to generate a descriptive textual caption T This caption encapsulates the high-level semantic content, serving as a weak textual condition to guide the generative process towards semantically coherent content synthesis.

$$T = \mathcal{F}_{VLM}(I_{Ref}). \tag{2}$$

Stage-II: Unified Multi-Modal Fusion (UMMF) and ODE Solving

– **Unified Sequence Construction and Position-Aware Embedding:** Drawing from and adapting the multi-modal processing paradigm of advanced architectures like MM-DiT [46], we employ a unified sequence concatenation approach for condition fusion, rather than using cross-attention. Specifically, we map the latent variable $z_t \in \mathbb{R}^{L_z \times d}$ representing the generative process into a token sequence. Concurrently, the low-quality image I_{LQ} and the text caption T are encoded into a visual token sequence $C_{img} = \mathcal{E}_{img}(I_{LQ}) \in \mathbb{R}^{L_{img} \times d}$ and a text token sequence $C_{txt} = \mathcal{E}_{txt}(T) \in \mathbb{R}^{L_{txt} \times d}$, respectively.

Subsequently, we concatenate these three components into a unified conditional generation sequence $\mathcal{S}$:

$$\mathcal{S} = [z_t; C_{img}; C_{txt}] \in \mathbb{R}^{(L_z+L_{img}+L_{txt})\times d}. \tag{3}$$

To distinguish between different modalities and positional information within the sequence, we inject RoPE positional encoding P and add a modality type encoding M (e.g., $M_{latent}, M_{image}, M_{text}$) to each token. Thus, the final sequence input to the Transformer block is:

$$S^{(0)} = \mathcal{S} + P + M. \tag{4}$$

– **Forward Process of Concatenation-Based Multi-Modal DiT:** Condition injection no longer occurs externally to the attention layers but is naturally achieved through the self-attention mechanism within the unified sequence. Through this approach, the latent variable tokens z_t can directly and adaptively retrieve and aggregate relevant information from the visual and textual condition tokens, achieving a deeper level of multi-modal fusion. The input sequence $S^{(0)}$ is processed by N consecutive Transformer blocks. For the ℓ-th block, the computation can be formally described as:

$$S^{(\ell)'} = \text{LayerNorm}(S^{(\ell)}), \tag{5}$$

$$\mathcal{A}^{(\ell)} = S^{(\ell)'}\text{SelfAttention}(S^{(\ell)'}) + S^{(\ell)}, \tag{6}$$

$$S^{(\ell+1)} = S^{(\ell)'}\text{MLP}(\text{LayerNorm}(\mathcal{A}^{(\ell)})) + S^{(\ell)}. \tag{7}$$

3.2 Test-Time Scaling for ODE-based Flow Models

While our two-stage image restoration framework produces high-quality results, we posit that the deterministic ODE trajectory of a flow model presents a unique and under-explored opportunity for inference-time optimization [15, 20]. By fundamentally reformulated for the ODE dynamics of flow models and the perceptual demands of image restoration, we introduce a Test-Time Scaling (TTS) paradigm.

ODE-SDE Transformation Framework. To enable flow-based TTS, it is useful to revisit the theoretical connection between ODEs and SDEs. Flow models learn a velocity field vector, $u_t \in \mathbb{R}^d$, which defines an ODE trajectory. Sampling x_0 is then achieved by solving this ODE [43] in reverse time from $t = T$ to $t = 0$:

$$x_{t-1} = x_t + u_t(x_t)dt. \tag{8}$$

This results in identical samples of x_{t-1} being derived from any given x_t, implying a deterministic mapping. Consequently, the applicability of test-time extension methods to flow models is limited, as the sampling process lacks inherent stochasticity beyond the initial noise. To overcome this constraint, we

transform the deterministic Flow-ODE into an equivalent Stochastic Differential Equation (SDE) by introducing a stochastic diffusion term. Following prior work [4, 20, 32, 36, 42], the ODE sampling process in Eq. 8 is reformulated as follows:

$$dx_t = \left(u_t(x_t) - \frac{\sigma_t^2}{2} \nabla \log p_t(x_t) \right) dt + \sigma_t dw. \tag{9}$$

In this formulation, the score $\log p_t(x_t)$ can be computed from the velocity field u_t [42], while the Wiener process dw injects stochasticity at each sampling step. This SDE framework thereby introduces randomness into the trajectory. Building upon this, our method operates directly on the deterministic ODE trajectory of the FLUX model and performs its search by guiding perturbations with this injected randomness, effectively leveraging the inherently straighter paths characteristic of ODEs.

An ODE, by construction, defines a deterministic mapping: given an initial noise and conditioning input, its output is uniquely determined. Arbitrarily perturbing an intermediate state would disrupt this learned deterministic dynamics, potentially causing the resulting state to deviate from the model's learned data manifold and leading to quality degradation or semantic errors. Critically, the ODE-to-SDE theory provides the theoretical foundation for our approach. In essence, this theory establishes that for a deterministic ODE defined in the form of Eq. 8, one can construct a family of SDEs as in Eq. 9. By designing their drift coefficients (set to σ_t in our paper) and introducing stochastic noise dw, the resulting stochastic processes share the same marginal distributions as the corresponding ODE model but exhibit diverse sampling trajectories.

The "mutation" operation we introduce next is precisely designed to explore within this legitimate SDE family, rather than arbitrarily disrupting the model. Our objective is not to compromise the distribution, but rather to explore—while preserving the marginal distributions—for samples that yield superior outcomes according to our reward function (*i.e.*, image quality). The SDE theory thereby delineates the search space for our "legitimate exploration." However, since our algorithmic goal is not general stochastic sampling but rather injecting controlled stochasticity into the pre-trained, highly deterministic FLUX ODE trajectory to explore better image restoration paths, we specifically apply and discretize Eq. 9 as shown in Eq. 11 and Eq. 12 below. The principles underlying this discretization and the logical connections behind it will be detailed in the corresponding section of a later chapter.

Formalizing TTS as Trajectory Optimization in Latent Space. The core intuition is to treat the initial state z_1 of the ODE as a searchable variable within a neighborhood, and the ODE solver as a dynamical system whose trajectory can be steered. Unlike diffusion models with stochastic differential equations (SDEs), the ODE of a flow model like FLUX defines a deterministic mapping $\Phi_{1\to 0}(z_1; c)$ from initial noise z_1 to the final latent z_0, given condition c. Our goal is to find an initial state z_1^* within a constrained set $\mathcal{Z}$ that maximizes the

(a) LQ Input (b) Real-ESR (c) DiffBIR (d) SeeSR (e) SUPIR (f) FaithDiff (g) Ours

Fig. 3: Qualitative comparisons on both synthetic (the first row) and real-world (the last three rows) benchmarks. Please zoom in for a better view.

expected quality of the resulting image:

$$z_1^* = \arg\max_{z_i \in \mathcal{Z}} \mathbb{E}\left[\mathcal{R}(\mathcal{D}(\Phi_{1\to 0}(z_1; c)))\right], \tag{10}$$

where $\mathcal{R}$ is a reward function quantifying image quality, and $\mathcal{D}$ is the decoder. Directly solving this optimization over the high-dimensional z_1 is intractable. Instead, we perform an iterative, greedy optimization over the trajectory at K strategically chosen time steps $\{t_k\}_{k=1}^K$, where $1 = t_1 > t_2 > \cdots > t_K > 0$.

ODE-aware Search with Verifier Ensemble. At each intervention step t_k, let the current state be z_{t_k}. The algorithm generates N candidate perturbations of the current state, evaluates their potential future quality via a partial ODE rollout and a learned reward model, and selects the most promising path. This "perturb-rollout-evaluate-select" cycle is detailed below:

1. **Perturbation**: We generate N candidate perturbations $\{\bar{z}_{t_k}^{(i)}\}_{i=1}^N$ by injecting structured noise:

$$\bar{z}_{t_k}^{(i)} = z_{t_k} + \epsilon^{(i)}, \quad \epsilon^{(i)} \sim \mathcal{N}(0, \sigma_k^2 I). \tag{11}$$

   This is equivalent to $\bar{z}_{t_k}^{(i)} = z_{t_k} + \sigma_k^{(i)} \epsilon_k^{(i)}, \quad \epsilon_k^{(i)} \sim \mathcal{N}(0, I)$, an alternative mutation operator we propose, inspired by the reverse-time SDE, to synthesize meaningful variations while preserving the intrinsic structure of the latent state z_{t_k}.
   Here, σ_k is the diffusion coefficient defined in the reverse-time SDE (Eq. 9), which controls the level of injected stochasticity and can be scheduled to

decrease with k, encouraging coarse exploration early and fine-tuning later. This mutation operation effectively generates novel latent states, enabling exploration of an expanded state space while preserving the inherent distribution established during the denoising process. In the next generation of our TTS search, we sample $z_0^{(i)} \sim p_0(z_0^{(i)} \mid \bar{z}_{t_k}^{(i)})$ based on the new offspring $\bar{z}_{t_k}^{(i)}$ and repeat the above evolutionary search process, including evaluation, selection, and mutation. Therefore, $\epsilon^{(i)}$ here in Eq. 11 is essentially equivalent to the Wiener process $\sigma_t dw$ in Eq. 9.

2. **Accurate z_0 Estimation via Multi-Step Partial Denoising (MSPDE)**: We obtain partially denoised latent variables by performing a fixed number of ODE solving steps, then decode and evaluate them. Given a candidate perturbation $\tilde{z}_{t_k}^{(i)}$, we perform S steps of ODE solving to obtain the partially denoised latent variable:

$$\tilde{z}_{t_k - S\cdot\Delta t}^{(i)} = \text{ODESolver}(\tilde{z}_{t_k}^{(i)}, t_k, t_k - S \cdot \Delta t, \mathbf{c}) \tag{12}$$

where Δt is the single-step time interval and S is the fixed number of denoising steps. We then use a a lookahead estimator [35]to obtain a high-fidelity preview of the final image $\tilde{z}_0^{(i)}$, rather than decoding a noisy intermediate latent state. This method balances computational efficiency with high assessment accuracy, mitigating the error inherent in one-step estimators.

Eq. 12 sets the drift term in the SDE (*i.e.*, $\frac{\sigma_t^2}{2}\nabla \log p_t(x_t)dt$ in Eq. 9) to zero—a key design choice in our framework. In the continuous SDE formulation shown in Eq. 9, $\nabla \log p_t(x_t)$ serves as a score function that guides the sampling process toward the data distribution $p_t(x_t)$, representing a general, task-agnostic gradient. However, image restoration is inherently a conditional generation problem that prioritizes local optimization within the output space over exact simulation of the full data distribution. Unlike unconditional text-to-image tasks that pursue broad distribution coverage and diversity (e.g., as measured by FID), our objective is to recover a restored image that is optimal in terms of pixel fidelity, perceptual quality, and semantic consistency. This constitutes an optimization problem with a well-defined objective function. Accordingly, we employ a Verifier Ensemble to steer the search direction toward human visual preferences and semantic realism, rather than merely toward the original data distribution.

This conceptual shift—from a *data distribution score* to a *task-specific reward signal*—represents a core innovation in adapting SDE-based generative principles to discriminative, optimization-driven tasks such as image restoration. Our results demonstrate that even without strictly adhering to the continuous SDE formalism, the introduction of discrete stochasticity can effectively correct biases (e.g., variance underestimation) that arise from discretization errors in purely deterministic ODE solvers.

Our strategy can thus be interpreted as follows: at sparse yet critical time steps, we harness the stochasticity inherent in SDE theory to "jump" to the starting point of a new latent trajectory. From there, we perform a fast,

low-cost rollout evaluation along the corresponding deterministic ODE path originating from that point. Notably, this ODE path itself corresponds to a specific instance of the original SDE family, obtained by setting its drift term to zero.

3. **Verifier Ensemble for Image Restoration**: Our reward function $\mathcal{R}$ is designed tailored for perceptual image restoration. Instead of relying on a single metric, we form a Verifier Ensemble $\mathcal{R}_{ensemble}$ that synthesizes judgments from multiple perspectives (*i.e.*, Aesthetic Score Predictor [40], CLIP-Score [16] and ImageReward [58]). Since these validators operate at different scales, we adopt a rank-based ensemble strategy. Specifically, for the N candidates at a given step, we compute the ordinal rank from each validator—denoted as $r_{aes}^{(i)}$, $r_{clip}^{(i)}$ and $r_{ir}^{(i)}$ for candidate i.

$$\mathcal{R}_{ensemble}(\hat{I}^{(i)}) = -\frac{1}{3}\left(r_{aes}^{(i)} + r_{clip}^{(i)} + r_{ir}^{(i)}\right) \tag{13}$$

   This approach favors candidates that demonstrate consistent performance across all quality dimensions, making it inherently robust to scale discrepancies.
   During the TTS search loop, we specifically utilize generation metrics rather than standard IR metrics such as MANIQA [59] and MUSIQ [19]. Optimizing IR metrics directly in a TTS search loop often causes the model to inject unnatural, high-frequency noise. Because these metrics are highly sensitive to high-frequency details, this results in "reward hacking"—artificially inflating the score while degrading actual human perceptual quality. Conversely, generation metrics are trained on human preferences for holistic structural and semantic harmony. Using them as an ensemble regularizes the search, ensuring that the enhanced details remain perceptually realistic.
4. **Selection and Commitment**: Instead of a simple "winner-takes-all" strategy, we adopt a progressive filtering approach inspired by EvoSearch [15]. The N candidates are ranked by their ensemble reward. We then retain the top-M performers (where $M < N$), and these form the parent population for the next iteration, which is chosen as the starting point for the immediate next ODE segment. This strategy maintains diversity in the search space for longer, increasing the probability of finding a globally superior path.

4 Experiments

4.1 Experimental Setup

Training Dataset: Our model was trained on a combination of the DIV2K [28], Flickr2K [1], LSDIR [26], DIV8K [14]. Following common practices in Real-ESRGAN [53], we generated LQ counterparts by applying a bicubic downsampling (with a scaling factor of $\times 4$) to the HQ images, using the same degradation settings as StableSR [52] to ensure a fair comparison.

Algorithm ODE-adapted Test-Time Scaling

Require: Pre-trained flow IR model Φ, condition c, initial noise z_1
Ensure: Optimized restored image $\hat{x}_0$

Initialize population $\mathcal{P} \leftarrow \{z_1\}$
for $k = 1$ **to** K **do** ▷ K intervention steps
  $\mathcal{C} \leftarrow \text{Perturb}(\mathcal{P}, \sigma_k)$ ▷ Generate N candidates with noise σ_k
  for each $z^{(i)} \in \mathcal{C}$ **do**
    $\hat{z}_0^{(i)} \leftarrow \text{MSPDE}(z^{(i)}, t_k, S)$ ▷ Multi-step partial denoising
    $r^{(i)} \leftarrow \mathcal{R}_{\text{ensemble}}(\mathcal{D}(\hat{z}_0^{(i)}))$ ▷ Rank-based reward ensemble
  end for
  $\mathcal{P} \leftarrow \text{TopM}(\mathcal{C}, r, M)$ ▷ Select top-M candidates
  $\mathcal{P} \leftarrow \text{AdvanceODE}(\mathcal{P}, t_k, t_{k+1}, c)$ ▷ Move to next time step
end for
$\hat{x}_0 \leftarrow \mathcal{D}(\text{ODESolver}(\mathcal{P}[0], t_K, 0, c))$ ▷ Final denoising
return $\hat{x}_0$

Testing Datasets: To comprehensively evaluate the generalization ability and performance of our method, we conducted tests both on synthetic test datasets and real-world test datasets. For synthetic benchmarks, we employ different levels of degradations (D-level) to synthesize degraded validation sets of DIV2K-Val, similar to DASR [27], and degrade LSDIR-Val using the same settings as training. For real-world benchmarks, we conduct experiments on commonly used RealPhoto60 [63], RealSR [7] and DRealSR [54] datasets.
Implementation Details: We build our method upon FLUX.1 [22] and Qwen2.5-VL [6]. By default, we use FLUX.1-dev to generate images based on given low-quality inputs and Qwen2.5-VL-7B to generate captions based on reference images. Our method utilizes LoRA [18] for fine-tuning the base model with a default rank of 16. The SwinIR model in DiffBIR [29] is used as the lightweight degradation remover. Our model is trained on 512×512 resolution images, with a batch size of 16 and gradient accumulation over 4 batch (effective batch size of 64). We employ the Prodigy optimizer [34] with safeguard warmup and bias correction enabled, setting the weight decay to 0.01. Models are trained for 30,000 iterations. For inference of our model, we employ the sde-dpmsolver++ sampler in FlowDPMSolverMultistepScheduler [37] in SDE process with 50 denoising steps and set guidance scale as 5.5. We set the mutation rate $\beta = 0.3$, with σ_t following the default sde-dpmsolver configurations.

4.2 Comparison with State-of-the-Art Methods

We compare our proposed method against several leading approaches, including: GAN-based methods (*i.e.*, Real-ESRGAN [53] and BSRGAN [64]), Diffusion-based methods (*i.e.*, StableSR [52], DiffBIR [29], PASD [60], SeeSR [56], DreamClear [3], OSEDiff [55], SUPIR [63], and FaithDiff [9]) and Flow-based methods (*i.e.*, PnP-Flow [33], FlowSR [57] and LP [50]). We use PSNR, SSIM, perceptual-oriented metrics (*i.e.*, LPIPS [66], DISTS [11], FID [17]) and no-reference metrics (*i.e.*, MUSIQ [19], CLIPIQA+ [51], PaQ-2-PiQ [61], CLIPIQA [51] and

Table 1: Quantitative comparison with state-of-the-art methods on synthetic benchmark. 'I', 'II', and 'III' denote the datasets with mild, medium, and severe degradations, respectively. The best and second performances are marked in **red** and **blue**, respectively.

Dataset	Metric	Real-ESRGAN [53]	BSRGAN [64]	StableSR [52]	DiffBIR [29]	PASD [60]	SeeSR [56]	DreamClear [3]	OSEDiff [55]	SUPIR [63]	FaithDiff [9]	PnP-Flow [33]	FlowSR [57]	LP [50]	Ours
DIV2K-Val-I	PSNR↑	**26.64**	**27.63**	24.71	24.60	25.31	25.08	23.76	25.13	25.09	24.29	23.82	25.02	22.98	23.11
	SSIM↑	**0.7737**	**0.7897**	0.7131	0.6595	0.6995	0.6967	0.6574	0.7098	0.7010	0.6668	0.6632	0.7219	0.6932	0.6945
	LPIPS↓	**0.1964**	**0.2038**	0.2393	0.2496	0.2370	0.2263	0.2259	0.2382	0.2139	0.2187	0.2214	0.2143	0.2198	0.2098
	PaQ-2-PiQ↑	73.85	74.17	74.19	72.81	73.12	73.92	73.86	74.07	74.24	**74.76**	73.21	74.14	73.82	**78.82**
	MANIQA↑	0.5611	0.5507	0.5743	0.5898	0.5524	0.5749	0.5750	0.5767	**0.6049**	0.5945	0.5511	0.5828	0.5842	**0.7261**
	MUSIQ↑	62.38	61.81	65.55	66.23	64.57	66.48	66.15	66.39	65.49	**66.53**	65.44	66.24	64.85	**75.54**
	CLIPIQA+↑	0.4649	0.4588	0.5156	0.5407	0.4764	0.5336	**0.5478**	0.5271	0.5202	0.5432	0.5294	0.5328	0.5422	**0.7925**
DIV2K-Val-II	PSNR↑	**25.49**	**26.42**	24.26	24.42	24.89	24.65	23.39	25.12	24.42	23.80	23.45	24.79	22.54	22.79
	SSIM↑	**0.7274**	**0.7402**	0.6771	0.6441	0.6764	0.6734	0.6330	0.6983	0.6703	0.6413	0.6421	0.6983	0.6708	0.6778
	LPIPS↓	0.2309	0.2465	0.2590	0.2708	0.2502	0.2428	0.2518	0.2499	0.2432	0.2407	0.2398	**0.2303**	0.2346	**0.2255**
	PaQ-2-PiQ↑	73.39	73.26	73.81	72.07	72.91	73.52	73.28	73.67	74.26	**74.77**	73.05	73.82	73.54	**78.94**
	MANIQA↑	0.5499	0.5261	0.5825	0.5827	0.5466	0.5691	0.5667	0.5675	**0.6012**	0.5919	0.5342	0.5682	0.5624	**0.7038**
	MUSIQ↑	61.84	60.00	64.76	64.83	64.45	66.09	64.96	65.04	65.58	**66.42**	64.31	65.08	63.98	**74.68**
	CLIPIQA+↑	0.4719	0.4463	0.5057	0.5246	0.4718	0.5226	0.5295	0.5209	0.5202	**0.5460**	0.5188	0.5204	0.5342	**0.7811**
DIV2K-Val-III	PSNR↑	22.81	**23.45**	23.34	23.42	22.58	22.58	21.82	**24.10**	21.90	21.77	20.22	21.53	19.48	19.74
	SSIM↑	**0.6288**	0.6281	0.6277	0.5992	0.5985	0.5944	0.5510	**0.6432**	0.5611	0.5662	0.5034	0.5536	0.5286	0.5263
	LPIPS↓	0.3535	0.3462	0.3559	0.3676	0.3646	0.3278	0.3336	**0.2978**	0.3172	0.3080	0.3222	0.3182	0.3228	**0.3047**
	PaQ-2-PiQ↑	71.93	72.09	70.22	74.20	72.68	73.46	72.49	73.49	74.07	**75.05**	73.12	73.48	73.31	**78.62**
	MANIQA↑	0.5057	0.4917	0.4896	0.5776	0.4999	0.5462	0.5515	0.5405	**0.5882**	0.5686	0.4976	0.5206	0.5223	**0.6605**
	MUSIQ↑	60.11	62.41	57.89	58.86	63.08	65.82	62.59	65.28	65.46	**66.28**	64.12	64.88	63.54	**74.67**
	CLIPIQA+↑	0.4637	0.4838	0.4124	0.5154	0.4815	0.5106	0.4914	0.5132	0.5134	0.5275	0.5072	0.5126	0.5225	**0.7704**
LSDIR-Val	PSNR↑	18.13	18.27	18.11	**18.42**	18.21	18.03	17.01	**19.02**	16.95	16.87	15.98	16.82	14.82	15.02
	SSIM↑	**0.4867**	0.4673	0.4508	0.4618	0.4592	0.4564	0.4236	**0.4796**	0.4080	0.4259	0.3802	0.4044	0.3907	0.3845
	LPIPS↓	0.3986	0.4378	0.4152	0.4049	0.3883	0.3759	0.3836	**0.3742**	0.4119	0.3802	0.3954	0.3822	0.3810	**0.3721**
	DISTS↓	0.2278	0.2539	0.2159	0.2439	0.1734	0.1966	0.1656	0.1619	0.1838	**0.1557**	0.1675	0.1582	0.1653	**0.1406**
	FID↓	46.46	53.25	31.26	35.91	22.65	25.91	22.06	22.32	30.03	**20.69**	22.93	22.18	22.76	**19.82**
	MANIQA↑	0.4381	0.3829	0.3098	0.4551	0.4043	0.5700	0.4811	0.4630	0.4683	0.5956	0.4878	**0.6023**	0.5912	**0.6658**
	MUSIQ↑	68.25	65.98	59.37	65.94	71.02	**73.00**	70.40	70.74	70.98	72.64	71.15	71.68	70.82	**74.21**
	CLIPIQA↑	0.6218	0.5648	0.5190	0.6592	0.5729	**0.7261**	0.6914	0.6679	0.6174	0.7224	0.7152	0.7208	0.7097	**0.7389**

Table 2: Quantitative comparison with state-of-the-art methods on real-world benchmark. The best and second performances are marked in **red** and **blue**, respectively.

Dataset	Metric	Real-ESRGAN [53]	BSRGAN [64]	StableSR [52]	DiffBIR [29]	PASD [60]	SeeSR [56]	DreamClear [3]	OSEDiff [55]	SUPIR [63]	FaithDiff [9]	PnP-Flow [33]	FlowSR [57]	LP [50]	Ours
RealPhoto60	PaQ-2-PiQ↑	69.04	63.38	64.40	71.88	69.97	73.77	**75.84**	73.41	74.42	75.27	72.78	73.82	73.44	**78.38**
	MANIQA↑	0.5094	0.3759	0.4456	0.6253	0.5290	0.6057	0.6166	0.5989	0.6165	**0.6527**	0.6176	0.6348	0.6320	**0.6425**
	MUSIQ↑	59.29	45.46	57.89	63.67	64.53	70.80	70.46	70.44	70.26	**72.74**	70.08	71.82	70.96	**74.81**
	CLIPIQA+↑	0.4389	0.3397	0.4214	0.4935	0.4786	0.5691	0.5273	0.5720	0.5528	**0.5932**	0.5795	0.5842	0.5856	**0.7861**
RealSR	MANIQA↑	0.3656	0.3660	0.3465	0.4182	0.3855	0.5189	0.4337	0.4052	0.4296	0.4893	0.4910	0.5232	**0.5308**	**0.6205**
	MUSIQ↑	62.06	64.67	61.07	61.74	62.03	69.38	65.33	64.87	66.09	**69.77**	62.27	68.82	63.96	**72.34**
	CLIPIQA↑	0.4872	0.5329	0.5139	0.6202	0.5043	0.6839	**0.6895**	0.5729	0.5371	0.6804	0.6822	0.6792	0.6842	**0.6937**
DrealSR	MANIQA↑	0.3423	0.3420	0.3171	0.3801	0.3456	0.4974	0.3676	0.3572	0.4174	**0.5647**	0.5228	0.5424	0.5386	**0.6492**
	MUSIQ↑	58.37	61.22	56.43	55.14	53.82	67.42	59.83	66.68	64.53	**68.92**	67.32	68.08	68.26	**71.82**
	CLIPIQA↑	0.4847	0.5385	0.5344	0.6005	0.5458	**0.7022**	0.6620	0.6037	0.5800	0.6848	0.6821	0.6946	0.6884	**0.7169**

MANIQA [59]) to evaluate the fidelity and quality of restored images. Our method performs strongly across all these metrics, attesting to the high quality of our restorations. Notably, ResFlow-Tuner surpasses all prior methods by a considerable margin in no-reference metrics, establishing a new state-of-the-art. Although resulting in lower PSNR/SSIM scores, recent studies [62,63] make compelling cases that these metrics fail to adequately capture perceived visual quality, which is also consistent with our experimental observations.

4.3 Evaluation on Downstream Benchmarks

We quantitatively assess our method by evaluating its performance on Optical Character Recognition (OCR) tasks. Following the established evaluation

Table 3: Recognition results on the dataset of Occluded RoadText 2024 [48] (OCR recognition). The best and second performances are highlighted in **red** and **blue**, respectively.

Metrics	GT	LQ	Real-ESRGAN [53]	BSRGAN [64]	StableSR [52]	DiffBIR [29]	SeeSR [56]	DreamClear [3]	SUPIR [63]	FaithDiff [9]	Ours
Precision	52.72%	7.54%	13.19%	12.04%	19.87%	26.21%	30.07%	22.45%	31.78%	**36.45%**	**42.23%**
Recall	56.67%	7.59%	13.68%	12.33%	20.31%	27.90%	33.09%	23.50%	41.57%	**46.74%**	**49.75%**

Table 4: Ablation results on LSDIR-Val for ResFlow-Tuner.

	LPIPS ↓	DISTS ↓	FID ↓	MANIQA ↑	MUSIQ ↑	CLIPIQA ↑	Infer-Compute ↓	Parameters ↓
LR w/o Text Caption	0.3856	0.1468	20.02	0.6542	72.98	0.7212	/	/
LR w/ Text Caption	**0.3721**	**0.1406**	**19.82**	**0.6658**	**74.21**	**0.7389**	/	/
Direct Adding	0.4223	0.1847	28.39	0.5796	69.14	0.6863	/	612M / 5.1%
ControlNet	0.4056	0.1663	24.73	0.6172	71.20	0.6917	/	3.3B / 27.5%
UMMF	**0.3721**	**0.1406**	**19.82**	**0.6658**	**74.21**	**0.7389**	/	**58.0M / 0.5%**
Aesthetic	0.3872	0.1662	23.25	0.6224	72.57	0.7128	/	/
CLIPScore	0.3824	0.1524	21.04	0.5774	69.88	0.6982	/	/
ImageReward	0.3789	0.1564	21.98	0.5842	70.64	0.6874	/	/
Aesthetic + CLIPScore	0.3820	0.1518	20.58	0.6302	73.24	0.7181	/	/
Aesthetic + ImageReward	0.3762	0.1558	21.76	0.6483	73.65	0.7308	/	/
CLIPScore + ImageReward	0.3752	0.1448	20.25	0.5964	70.98	0.7064	/	/
Verifier Ensemble	**0.3721**	**0.1406**	**19.82**	**0.6658**	**74.21**	**0.7389**	/	/
w/o TTS	0.4015	0.1643	22.91	0.6032	71.47	0.6946	0.025e4	/
w/ TTS ($K = 2, N = 5$)	0.3874	0.1523	20.87	0.6298	72.85	0.7149	0.15e4	/
w/ TTS ($K = 4, N = 7$)	**0.3721**	**0.1406**	**19.82**	**0.6658**	**74.21**	**0.7389**	**0.375e4**	/
w/ TTS ($K = 10, N = 15$)	0.3675	0.1332	18.17	0.6878	76.33	0.7574	1.0e4	/

benchmark [9], we conduct experiments on 200 images from the ICDAR 2024 Occluded RoadText dataset [48]. The recognition is performed with the robust PaddleOCR v3 [23] framework. As reported in Tab. 3, our approach consistently outperforms existing image restoration methods in terms of OCR accuracy. This result provides strong evidence that our method is particularly effective at recovering semantically correct and robust structural information.

5 Ablation Study

5.1 Impact of Multi-Modal Conditioning

Effect of Text Caption: As shown in Tab. 4, incorporating text captions (LR w/ Text Caption) consistently improves performance across all metrics compared to the variant without captions (LR w/o Text Caption). For instance, LPIPS decreases from 0.3856 to 0.3721, DISTS from 0.1468 to 0.1406, and FID from 20.02 to 19.82. Similarly, no-reference quality metrics also exhibit notable gains. These results demonstrate that semantic guidance from text captions effectively steers the generative process toward outputs that are not only pixel-accurate but also perceptually realistic and semantically coherent.

Modality Fusion Strategy: We compare three fusion strategies: direct feature addition, ControlNet-based conditioning, and our proposed UMMF. Our method achieves the best results across all metrics. For example, it attains an LPIPS of 0.3721, compared to 0.4223 and 0.4056 from the alternatives. Similarly, on no-reference quality metrics such as MANIQA, our method also registers an improvement, outperforming Direct Adding and ControlNet by 14.87% and 7.87%, respectively. This superiority stems from our unified sequence modeling formulation, which enables flexible cross-modal token interactions via DiT's attention mechanism, thereby capturing rich semantic relationships without imposing hard structural constraints.

Efficiency and Cost Analysis: Unlike traditional "train-time scaling" strategies that rely on heavy adapters (e.g., ControlNet) to inject low-quality image

features, our approach unleashes the model prior through a lightweight alignment module. This architectural choice results in a remarkably small number of trainable parameters—only 58M (0.5%) for the 12B FLUX.1 model. This is an order of magnitude more efficient than adapter-based counterparts. Our parameter-efficient design enables rapid adaptation to new tasks with minimal training overhead, which constitutes a significant efficiency advantage.

5.2 Study on Verifier Ensemble

Our TTS framework incorporates three distinct reward verifiers: Aesthetic Score Predictor, CLIPScore, and ImageReward. To validate the necessity of our design choices, we conduct a comprehensive ablation study by either removing or substituting individual components. Specifically, our ablations evaluate (i) the performance of each reward function in isolation and (ii) the performance of all possible pairwise combinations.

Experimental results indicate that single or pairwise reward models tend to optimize performance within a specific dimension: The Aesthetic Score Predictor is designed to estimate human ratings of synthetic image quality, thereby biasing model optimization toward no-reference evaluation metrics. In contrast, CLIPScore is a reference-free metric derived from the CLIP model. It projects visual and textual embeddings into a shared latent space by leveraging 400M manually annotated image-text pairs, and quantifies semantic consistency via cosine similarity between image and corresponding text embeddings. Consequently, CLIPScore directs optimization towards reference-based metrics that assess text-image alignment. ImageReward, on the other hand, learns broader human preferences through a meticulously designed annotation protocol that encompasses ratings and rankings of text-image alignment, aesthetic quality, and sample harmlessness.

As shown in Tab. 4, these results demonstrate that only the ensemble of all three reward models, as proposed in this paper, enables a balanced and superior performance across all evaluation dimensions.

5.3 Analysis of Test-Time Scaling (TTS)

As summarized in Tab. 4, our method exhibits monotonic performance improvement with increasing inference-time computation (As the number of candidates per round N and intervention rounds K increase), measured by the Number of Function Evaluations (NFEs). For instance, moving from no TTS to the configuration $K = 4, N = 7$ reduces LPIPS from 0.4015 to 0.3721 and FID from 22.91 to 19.82, while MUSIQ increases from 71.47 to 74.21. A critical observation is that the marginal benefit is not constant. The performance growth rate is highest until the computational cost reaches approximately 15× that of the baseline. This operating point corresponds to a sweet spot on the performance-cost Pareto frontier—beyond this point, the performance gain per unit of additional computation diminishes significantly. Guided by this principle, we set our runtime computation to 15× the baseline ($K = 4, N = 7$). As mentioned

before, when comparing the total cost of our approach (minimal training cost + controlled inference cost) against "train-time scaling" methods (typically high training cost + single inference cost), our total resource consumption is often more favorable. More importantly, our method achieves a level of performance through controlled test-time computation that remains unattainable by others, even with their substantial training investments.

6 Conclusion

In this paper, we presented a novel image restoration framework ResFlow-Tuner. Our key innovation lies in the seamless integration of multi-modal guidance and a Test-Time Scaling technique specifically redesigned for ODE-based restoration models. Extensive experiments demonstrate that our method achieves state-of-the-art performance, excelling particularly in generating perceptually realistic and semantically faithful details. This work underscores the immense potential of leveraging large-scale pre-trained flow models for low-level vision tasks and introduces a new direction for post-training optimization of such models. For future work, we plan to explore more efficient TTS strategies to reduce the computational overhead and investigate the application of our framework to other image restoration tasks, such as deblurring and inpainting. Furthermore, extending our unified multi-modal fusion paradigm from static images to complex temporal sequences, thereby bridging the gap between low-level vision enhancement and high-level multi-video perception [5]. The code and models will be made publicly available to facilitate further research.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62576342, 62550062, 62425606, 32341009), the Beijing Natural Science Foundation (Grant No. 4252054), the Beijing Nova Program (Grant Nos. 20230484276, 20240484601), and the CCF-Kuaishou Large Model Explorer Fund (No. CCF-KuaiShou 2024005).

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW. pp. 126–135 (2017)
2. Ai, Y., Huang, H., Zhou, X., Wang, J., He, R.: Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25432–25444 (2024)
3. Ai, Y., Zhou, X., Huang, H., Han, X., Chen, Z., You, Q., Yang, H.: Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. NIPS **37**, 55443–55469 (2024)

4. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)
5. Bai, P., Wu, T., Sun, J., Liu, X., Huang, H., He, R.: Mvpbench: A multi-video perception evaluation benchmark for multi-modal video understanding. arXiv preprint arXiv:2603.22756 (2026)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
7. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV. pp. 3086–3095 (2019)
8. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
9. Chen, J., Pan, J., Dong, J.: Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In: CVPR. pp. 28188–28197 (2025)
10. Chen, Z., Wu, Z., Zamfir, E., Zhang, K., Zhang, Y., Timofte, R., Yang, X., Yu, H., Wan, C., Hong, Y., et al.: Ntire 2024 challenge on image super-resolution (x4): Methods and results. In: CVPR. pp. 6108–6132 (2024)
11. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. TPAMI **44**(5), 2567–2581 (2020)
12. Duan, J., Liu, S., Hao, Y., Huang, H., He, R.: Dual frequency-guided spatiotemporal feature learning for face forgery detection. T-BIOM (2025)
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
14. Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., Timofte, R.: Div8k: Diverse 8k resolution image dataset. In: ICCVW. pp. 3512–3516. IEEE (2019)
15. He, H., Liang, J., Wang, X., Wan, P., Zhang, D., Gai, K., Pan, L.: Scaling image and video generation via test-time evolutionary search. arXiv preprint arXiv:2505.17618 (2025)
16. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP. pp. 7514–7528 (2021)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NIPS **30** (2017)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
19. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (2021)
20. Kim, J., Yoon, T., Hwang, J., Sung, M.: Inference-time scaling for flow models via stochastic generation and rollover budget forcing. arXiv preprint arXiv:2503.19385 (2025)
21. Kim, S., Kim, M., Park, D.: Test-time alignment of diffusion models without reward over-optimization. arXiv preprint arXiv:2501.05803 (2025)
22. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
23. Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., et al.: Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. arXiv preprint arXiv:2206.03001 (2022)

24. Li, X., Uehara, M., Su, X., Scalia, G., Biancalani, T., Regev, A., Levine, S., Ji, S.: Dynamic search for inference-time alignment in diffusion models. arXiv preprint arXiv:2503.02039 (2025)
25. Li, X., Zhao, Y., Wang, C., Scalia, G., Eraslan, G., Nair, S., Biancalani, T., Ji, S., Regev, A., Levine, S., et al.: Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. arXiv preprint arXiv:2408.08252 (2024)
26. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., et al.: Lsdir: A large scale dataset for image restoration. In: CVPRW. pp. 1775–1787 (2023)
27. Liang, J., Zeng, H., Zhang, L.: Efficient and degradation-adaptive network for real-world image super-resolution. In: ECCV. pp. 574–591. Springer (2022)
28. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW. pp. 136–144 (2017)
29. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: ECCV. pp. 430–448. Springer (2024)
30. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
31. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
32. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV. pp. 23–40. Springer (2024)
33. Martin, S., Gagneux, A., Hagemann, P., Steidl, G.: Pnp-flow: Plug-and-play image restoration with flow matching. In: ICLR (2025)
34. Mishchenko, K., Defazio, A.: Prodigy: An expeditiously adaptive parameter-free learner. arXiv preprint arXiv:2306.06101 (2023)
35. Oshima, Y., Suzuki, M., Matsuo, Y., Furuta, H.: Inference-time text-to-video alignment with diffusion latent beam search. arXiv preprint arXiv:2501.19252 (2025)
36. Patel, Z., DeLoye, J., Mathias, L.: Exploring diffusion and flow matching under generator matching. arXiv preprint arXiv:2412.11024 (2024)
37. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
38. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
39. Ren, B., Guo, H., Sun, L., Wu, Z., Timofte, R., Li, Y.: The tenth ntire 2025 efficient super-resolution challenge report. In: CVPR. pp. 917–966 (2025)
40. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NIPS **35**, 25278–25294 (2022)
41. Singh, A., Mukherjee, S., Beirami, A., Jamali-Rad, H.: Code: Blockwise control for denoising diffusion models. arXiv preprint arXiv:2502.00968 (2025)
42. Singh, S., Fischer, I.: Stochastic sampling from deterministic flow models. arXiv preprint arXiv:2410.02217 (2024)

43. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
44. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. NIPS **33**, 3008–3021 (2020)
45. Sun, L., Guo, H., Ren, B., Van Gool, L., Timofte, R., Li, Y.: The tenth ntire 2025 image denoising challenge report. In: CVPR. pp. 1342–1369 (2025)
46. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: ICCV. pp. 14940–14950 (2025)
47. Tang, L., Ruiz, N., Chu, Q., Li, Y., Holynski, A., Jacobs, D.E., Hariharan, B., Pritch, Y., Wadhwa, N., Aberman, K., et al.: Realfill: Reference-driven generation for authentic image completion. ACM TOG **43**(4), 1–12 (2024)
48. Tom, G., Mathew, M., Mondal, A., Karatzas, D., Jawahar, C.V., Weinman, J.: Icdar2024 challenge on occluded roadtext. In: ICDAR2024 Workshops (2024)
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. Tsao, L.Y., Lo, Y.C., Chang, C.C., Chen, H.W., Tseng, R., Feng, C., Lee, C.Y.: Boosting flow-based generative super-resolution models via learned prior. In: CVPR. pp. 26005–26015 (2024)
51. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI. vol. 37, pp. 2555–2563 (2023)
52. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. IJCV **132**(12), 5929–5949 (2024)
53. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW. pp. 1905–1914 (2021)
54. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: ECCV. pp. 101–117. Springer (2020)
55. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. NIPS **37**, 92529–92553 (2024)
56. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: CVPR. pp. 25456–25467 (2024)
57. Xu, J., Li, W., Sun, H., Li, F., Wang, Z., Peng, L., Ren, J., Yang, H., Hu, X., Pei, R., et al.: Fast image super-resolution via consistency rectified flow. In: ICCV. pp. 11755–11765 (2025)
58. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. NIPS **36**, 15903–15935 (2023)
59. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: CVPR. pp. 1191–1200 (2022)
60. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: ECCV. pp. 74–91. Springer (2024)
61. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: CVPR. pp. 3575–3585 (2020)

62. You, Z., Li, Z., Gu, J., Yin, Z., Xue, T., Dong, C.: Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In: ECCV. pp. 259–276. Springer (2024)
63. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: CVPR. pp. 25669–25680 (2024)
64. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV. pp. 4791–4800 (2021)
65. Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., et al.: A survey on test-time scaling in large language models: What, how, where, and how well? arXiv preprint arXiv:2503.24235 (2025)
66. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)