LookFlow: Training-Free and Efficient High-Resolution Image Synthesis via Dynamic Lookahead Guidance Flow

Yuan Zhou[1,2], **Yan Zhang**[3], **Jianlong Chang**[4], **Xin Gu**[5], **Ying Wang**[1*], **Kun Ding**[1], **Guangwen Yang**[3], **Shiming Xiang**[1, 2]

[1] MAIS, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, UCAS
[3] Department of Computer Science and Technology, Tsinghua University
[4] Huawei
[5] Research and Development Department of China Academy of Launch Vehicle Technology

Abstract

Rectification flow Transformers (RFTs) have shown promising performance in diffusion-based image synthesis but are typically confined to lower-resolution scenarios, limiting their ability to generate high-resolution images. Existing resolution extrapolation approaches often suffer from excessive computational overhead, resulting in prolonged inference times. We propose LookFlow, a training-free high-resolution synthesis framework that accelerates inference while preserving visual quality. Building on pretrained text-to-image RFTs, LookFlow employs a dynamic lookahead guidance flow mechanism to refine high-resolution velocity predictions by leveraging multi-timestep lookahead information extracted from a low-resolution flow. Additionally, reusing temporally similar features across consecutive timesteps drastically reduces computation and significantly decreases inference time overhead. Extensive experiments on COCO demonstrate that LookFlow robustly scales resolutions from 4× to 25×, achieving up to a maximum speedup of 2.01× while maintaining competitive visual fidelity.

Code — https://github.com/yz413/LookFlow

1 Introduction

Diffusion models have garnered widespread attention in generative tasks due to their exceptional capability to produce high-quality and diverse outputs (Rombach et al. 2022; Podell et al. 2023; Zhang, Rao et al. 2023; Ho, Jain et al. 2020). Early diffusion models, built upon UNet architectures, have achieved remarkable results in image and video synthesis (Chen et al. 2023; Zheng et al. 2024b; Guo et al. 2023). To address the scalability challenges of earlier architectures, Diffusion Transformers (DiTs) (Peebles and Xie 2023) were introduced. Recent advancements, such as Rectification Flow Transformers (RFTs) (Lipman et al. 2022; Liu, Gong, and Liu 2022; Albergo and Vanden-Eijnden 2022) incorporating flow matching techniques, have further pushed the state-of-the-art in DiT-based generation (Li et al. 2024; Chen et al. 2023, 2024b), with notable models like Stable Diffusion 3 (Esser et al. 2024) and Flux (Labs 2024) demonstrating significant improvements in generation quality, aesthetic appeal, and text-image alignment. However, these models are often limited to native resolutions in the 1024^2-2048^2 range, posing a challenge for their direct use in high-resolution applications.

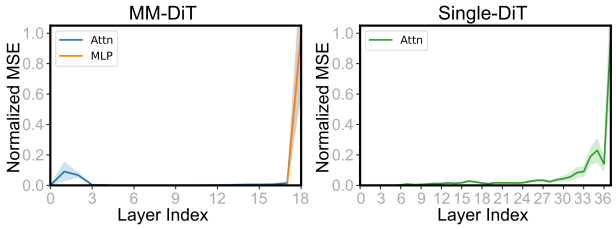

Figure 1: Normalized MSE across layers for MM-DiT (left) and Single-DiT (right). Between the input and output of each block in Flux, both attention and MLP layers in MM-DiT exhibit minimal MSE in the intermediate stages, with higher MSE appearing only in the initial and final layers. By contrast, Single-DiT shows stable early layers but significant deviations in later layers.

While conceptually viable, direct training of generation models on high-resolution data remains constrained by prohibitively expensive computational requirements and prohibitively long training durations (Zheng et al. 2024a; Liu et al. 2024b; Teng et al. 2023; Guo et al. 2024; Hoogeboom, Heek, and Salimans 2023). This fundamental limitation has driven a shift toward *training-free* resolution extrapolation strategies . (Bar-Tal et al. 2023; Du et al. 2023, 2024; Wu et al. 2024). However, high-resolution visual generation still faces two fundamental challenges: (1) Cross-scale semantic guidance. Existing methods often fail to effectively propagate hierarchical semantic cues from low- to high-resolution stages, leading to outputs lacking fine details and coherence. While methods like DemoFusion (Du et al. 2023) and I-Max (Du et al. 2024) show promise, they often introduce artifacts. (2) Computational efficiency. Higher resolution significantly increases inference time, yet few studies focus on accelerat-

*Corresponding author. *ywang@nlpr.ia.ac.cn*

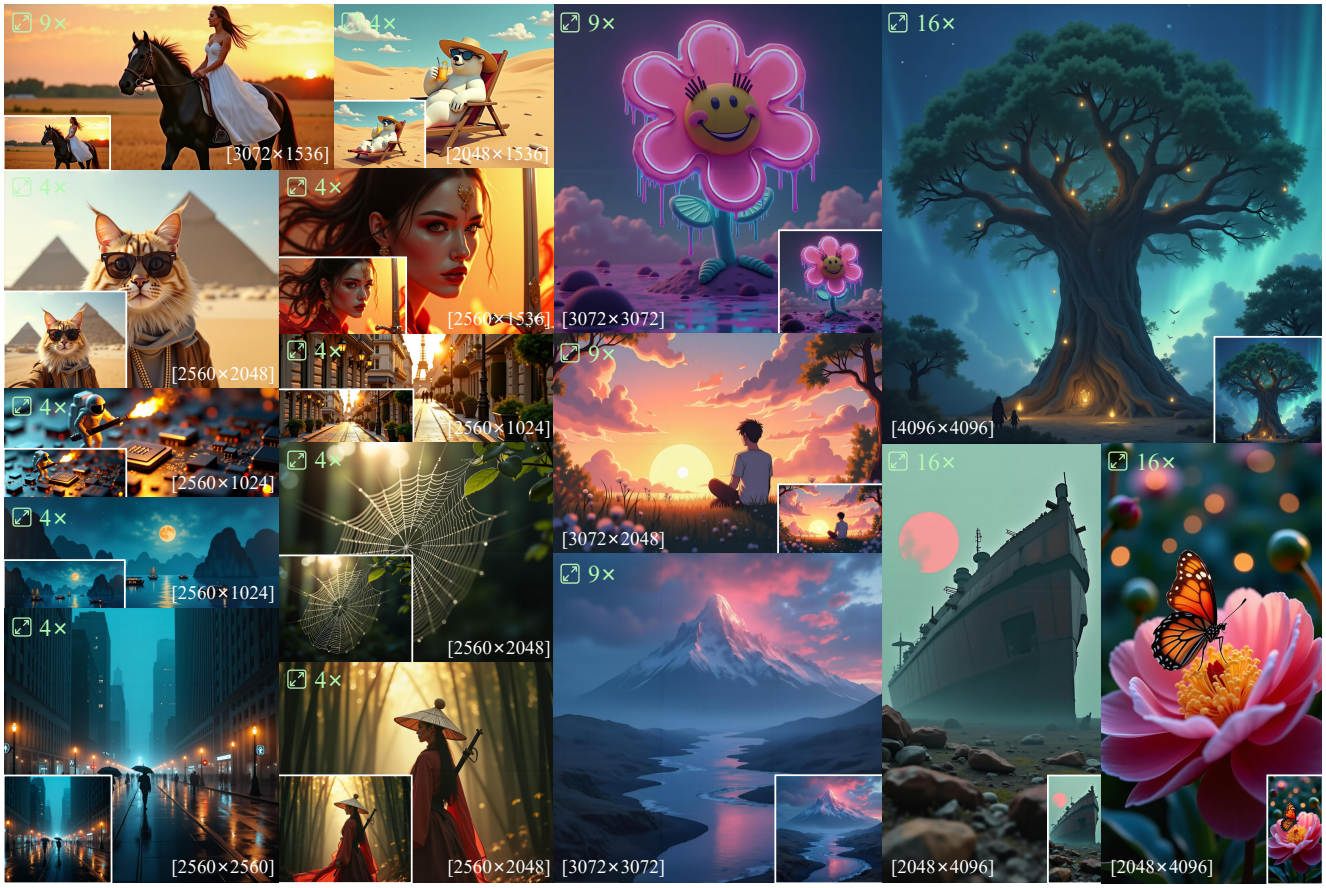

Figure 2: Representative high-resolution images generated by LookFlow. Without any additional training or fine-tuning, LookFlow can scale images to $4\times$, $9\times$, $16\times$, or even higher resolutions. It preserves semantic consistency and stability relative to the original low-resolution images, while significantly accelerating inference. All images are displayed at their actual sizes.

ing generation without compromising quality.

To address these challenges, we propose LookFlow, a *training-free* framework for efficient high-resolution image synthesis. Leveraging pretrained rectification flow Transformers, LookFlow achieves high-quality generation with significantly reduced inference time through two key components: (1) a *dynamic lookahead guidance flow* (DLG) mechanism, which refines high-resolution velocity estimates by aggregating multi-timestep information from the low-resolution denoising trajectory, thereby constructing a locally rectified velocity field that enhances cross-scale semantic consistency and yields sharper, more coherent outputs; and (2) a *sparse recomputation and caching* (SRC) strategy, which capitalizes on the high temporal coherence of intermediate representations. As shown in Fig. 1, the mean squared error between each layer's input and output remains low across layers 3–16 in MM-DiT, indicating feature stability that makes them suitable for caching. In contrast, significant deviations in Single-DiT are observed only beyond layer 30, allowing selective caching in both branches. Empirical evaluations on COCO in Fig. 3, SRC retains baseline FID while reducing inference latency by 2.01×, and caching either branch (MM-DiT, Single-DiT, or both) incurs

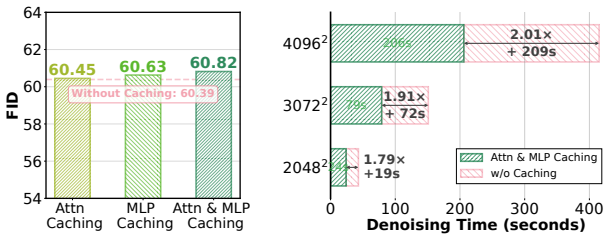

Figure 3: Evaluation of caching strategies on COCO. FID scores (left) and denoising duration (right) across multiple output resolutions under different caching strategies.

no appreciable degradation in image quality.

In summary, our contributions are threefold:

- We introduce a training-free high-resolution synthesis framework based on pretrained RFT that employs dynamic lookahead flow guidance to integrate semantic cues across resolutions, thereby ensuring semantic consistency between high- and low-resolution images.
- By leveraging the MSE stability of key blocks in MM-DiT and Single-DiT, we propose a feature sparse recom-

putation and caching mechanism that reduces redundant computation, achieving $2.01\times$ faster inference without compromising image quality.

- Extensive experiments on COCO demonstrate our method's effectiveness, supporting high-fidelity stable synthesis across scaling factors from $4\times$ to $25\times$.

2 Related Work

Transformers in Diffusion Models. Diffusion models (Ho, Jain et al. 2020; Sohl-Dickstein et al. 2015) have emerged as powerful generative methods, producing high-quality and diverse outputs (Podell et al. 2023; Zhang, Rao et al. 2023; Zheng et al. 2024b). Early work like DDPM (Ho, Jain et al. 2020) and DDIM (Song, Meng, and Ermon 2021) showcased the iterative noising–denoising process, while latent diffusion models (Rombach et al. 2022) introduced diffusion in compressed latent space for more efficient training and inference. Despite these advances, early diffusion models primarily rely on UNet backbones (Ronneberger, Fischer, and Brox 2015), whose capacity can be limiting. DiT (Peebles and Xie 2023) addresses this by stacking Transformer blocks, surpassing UNet models in class-conditional image generation. Drawing on DiT's scalability, the PixArt series (Chen et al. 2023, 2024b,a) incorporates cross-attention into DiT blocks for text-to-image tasks, while the CogVideo (Yang et al. 2024b) extends attention structures for high-fidelity video generation.

High-Resolution Image Generation. Generating high-resolution images remains a significant challenge in the generative modeling domain, driven by the limited availability of large-scale high-resolution datasets and the prohibitive computational costs of training. Although training larger models on higher-resolution data appears to be a straightforward solution (Zheng et al. 2024a; Liu et al. 2024b; Teng et al. 2023; Guo et al. 2024; Hoogeboom, Heek, and Salimans 2023), the associated expenses often hinder practical deployment. Consequently, tuning-free methods requiring no additional data have gained prominence as a more feasible alternative (Bar-Tal et al. 2023; He et al. 2023; Jin et al. 2023; Du et al. 2023; Liu et al. 2024c; Du et al. 2024; Shi et al. 2024; Wu et al. 2024). For example, MultiDiffusion (Bar-Tal et al. 2023) fuses overlapped denoising paths in latent diffusion models to seamlessly generate panoramas. Similarly, ScaleCrafter (He et al. 2023) introduces a convolution kernel dilation strategy to adapt diffusion models to higher resolutions without further tuning—though, as observed in our experiments, this can impair local details. DemoFusion (Du et al. 2023) adds global perception to improve overall layout coherence, at times leading to undesirable object repetition. Recently, I-Max (Du et al. 2024) proposed a resolution extrapolation framework for RFT that leverages a novel projected flow strategy for low-resolution guidance, significantly reducing inference complexity at higher resolutions. Along the same line, FreeScale (Qiu et al. 2024) adopts a multi-scale fusion and selective frequency extraction paradigm to enhance image fidelity while avoiding common high-resolution pitfalls such as repetitive artifacts. Taken together, these approaches highlight the growing importance of training-free methodologies for high-resolution synthesis, offering a promising path toward scalable and flexible image generation.

Acceleration of Diffusion Models. Two main categories of acceleration methods have emerged for diffusion models: (1) sampling acceleration by reducing the number of iterations, and (2) computation caching to eliminate redundant calculations. DDIM (Song, Meng, and Ermon 2021) employs a deterministic, non-Markovian reverse process to shorten sampling without sacrificing quality, whereas DPM-Solver (Lu et al. 2022) analytically refines this reverse step to further reduce iteration counts. For computation caching, Faster Diffusion (Li et al. 2023) stores encoder outputs at selected timesteps, while DeepCache (Ma, Fang, and Wang 2024) buffers intermediate features via UNet skip connections. To address DiT-specific needs, Δ-DiT (Chen et al. 2024c) constructs residual connections between intermediate layer outputs, PAB (Zhao et al. 2024) enhances attention mechanisms within DiT blocks, and SmoothCache (Liu et al. 2024a) introduces a universal training-free framework that identifies reusable features through calibration passes across modalities. Collectively, these innovations optimize computational reuse to advance practical deployment of diffusion models.

3 Method

Preliminaries

Flow matching. Flow matching (Lipman et al. 2022) is a class of generative modeling frameworks based on ordinary differential equations (ODEs), which aim to learn continuous-time dynamics that transport one probability distribution into another. Formally, given empirical observations sampled from source and target distributions, $\mathbf{X}_0 \sim \pi_0$ and $\mathbf{X}_1 \sim \pi_1$, flow matching seeks to construct a time-dependent trajectory $\mathbf{X}_t$ governed by an ODE of the form:

$$\frac{\mathrm{d}\mathbf{X}_t}{\mathrm{d}t} = v_\theta(\mathbf{X}_t, t, c), \quad t \in [0, 1], \tag{1}$$

where $v_\theta(\cdot)$ is a neural network parameterized by θ, and c denotes optional conditioning variables such as class labels or text prompts. The velocity field v_θ is optimized to align with a reference vector field that deterministically transports the data from the initial distribution π_0 to the target π_1.

In rectified flow, this reference vector field is defined by a linear interpolation path between $\mathbf{X}_0$ and $\mathbf{X}_1$, leading to an analytically tractable trajectory:

$$\mathbf{X}_t = (1 - t)\mathbf{X}_0 + t\mathbf{X}_1. \tag{2}$$

The corresponding drift velocity is then given by the constant vector $\mathbf{X}_1 - \mathbf{X}_0$, and the governing ODE becomes:

$$\mathrm{d}\mathbf{X}_t = (\mathbf{X}_1 - \mathbf{X}_0)\,\mathrm{d}t. \tag{3}$$

This construction embodies the core principle of flow matching: the model is trained to approximate the deterministic drift that transports samples along linear paths from noise to data. The generation process proceeds deterministically by iteratively solving the learned ODE, which predicts a continuous velocity field $v_\theta(\mathbf{X}_t, t, c)$ that approximates the ideal linear transport from the source distribution to the target.

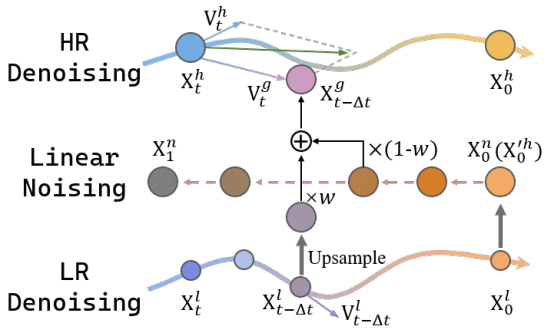

Figure 4: Overview of the LookFlow pipeline. LookFlow first denoises and upsamples a low-resolution image, then deterministically refines it at high resolution using look-ahead cues from the low-resolution trajectory to correct the velocity field. High-resolution detail preservation and sparse feature caching are applied concurrently, maintaining fidelity while boosting inference efficiency.

Feature Caching. Feature caching exploits temporal feature similarity by reusing computations from previous timesteps to avoid redundant processing. Specifically, let $b_l(\cdot)$ denote the computation performed by the l-th block. In a caching cycle spanning s timesteps, the full computation is executed at the initial, or refreshing, timestep t, and the resulting features are stored as $\mathcal{C}_t[l] := b_l(x_t)$, where $\mathcal{C}_t[l]$ represents the cached feature from block l at timestep t. For each subsequent timestep t to $t+s$ (the caching steps), these cached features are reused:

$$b_l(x_{t+i}) := \mathcal{C}_{t+i}[l], \quad i = 1, \ldots, s-1. \tag{4}$$

This approach enables the direct reuse of cached features at designated timesteps, bypassing complex computations and significantly increasing efficiency.

Dynamic Lookahead Guidance Flow

Accurate velocity estimates are essential for high-resolution diffusion to retain global structure and fine detail. As illustrated in Fig. 4, we propose a dynamic lookahead guidance mechanism that continuously refines the high-resolution velocity prediction by referencing a multi-timestep lookahead information from the low-resolution flow. DLG does not rely directly on the terminal state; instead, it periodically analyzes the low-resolution flow over subsequent timesteps to perform local velocity rectification, thereby capturing a more accurate and comprehensive estimate of the high-resolution velocity direction. This strategy not only alleviates issues such as semantic inconsistency and loss of detail between high-resolution and low-resolution images but also facilitates direct, one-step high-resolution image generation, circumventing the need for a multi-stage denoising process.

Resolution Alignment. Given a sequence of low-resolution ODE states $\mathbf{X}_t^l$, our goal is to extract a meaningful velocity field that aligns with high-resolution inference. Due to dimensional inconsistency between low and high-resolution spaces, we first upsample it to obtain $\mathbf{X}_t'^h$, which aligns the spatial dimensions with the high-resolution state $\mathbf{X}_t^h$. This transformation can be viewed as a soft projection operator that preserves structural coherence across scales.

Linear Noising. Direct reliance on $\mathbf{X}_t'^h$ may be insufficient, as simply using the upsampled low-resolution results for denoising can degrade high-resolution details or risk semantic inconsistency. For simplicity, we directly adopt a linear noise addition strategy to generate a sequence of controlled noisy variants. Specifically, following the linear trajectory formulation in Eq. (2), we directly interpolate between $\mathbf{X}_0$ and $\mathbf{X}_1$ to obtain the noisy state $\mathbf{X}_t^n$:

$$\mathbf{X}_t^n = (1-t)\,\mathbf{X}_0^l + t\,\mathbf{n}, \tag{5}$$

where $\mathbf{n}$ represents Gaussian noise. This operation provides a robust noisy reference that encapsulates both the deterministic low-resolution structure and the stochastic elements of the diffusion process.

Constructing Guidance State. Next, to synthesize these insights, we construct a composite reference state by fusing the directly upsampled state with its noisy counterpart:

$$\mathbf{X}_t^g = w \cdot \mathbf{X}_t'^h + (1-w) \cdot \mathbf{X}_t^n, \tag{6}$$

where the weight $w \in [0,1]$ balances structural fidelity against stochastic regularization. This weighted combination leverages the complementary strengths of both components, yielding a more reliable reference for guiding high-resolution updates.

Lookahead Velocity Rectification. The guidance velocity is then derived as:

$$\mathbf{V}_t^g = (\mathbf{X}_{t-s}^g - \mathbf{X}_t^h)/s, \tag{7}$$

where $\mathbf{X}_t^h$ is the latent state at time step t in the high-resolution denoising process. Unlike conventional stepwise correction approaches, which typically consider only immediate discrepancies between adjacent states, our method integrates information from farther timesteps through a dynamic lookahead information guidance. The extent of this lookahead is determined by the selection of a timestep interval s, referred to as the lookahead step, where a larger s signifies the incorporation of information from more distant states. That is, historical states derived from the established low-resolution process provide localized yet forward-looking cues to guide the refinement of velocity fields at higher resolutions, thereby producing more stable and globally coherent updates for superior image synthesis.

Finally, the previously introduced lookahead velocity is incorporated as a correction term to refine the original high-resolution velocity. The final velocity prediction is

$$v_\theta(\mathbf{X}_t^h, t, c) = v_\theta(\mathbf{X}_t^h, t, c) + \alpha_t \mathbf{V}_t^g, \tag{8}$$

where α_t is a hyperparameter that controls the influence of the low-resolution guidance. As illustrated in Fig. 4, the blue arrow represents the original high-resolution denoising velocity prediction, while the purple arrow denotes the correction term containing the forward-looking guidance. The sum of these, depicted by the green arrow, provides the corrected velocity estimate. This final velocity is then used in the subsequent Euler iteration to obtain the high-resolution denoising result.

Sparse Recomputation and Caching

Despite the effectiveness of using low-resolution feature upsampling to guide high-resolution generation, increasing the resolution inevitably results in greater computational demands and longer processing times. Our re-examination of DiT-based model reveals several key insights:

- **Temporal Consistency.** The denoising process shows a high degree of similarity between outputs at consecutive timesteps. This visual consistency is supported by the similarity of features across layers, indicating that many computed features remain essentially unchanged over successive layers, as illustrated in Fig. 1.
- **Cache Insensitivity.** Regardless of the caching strategy used, there is no significant performance degradation compared to the non-caching case, while inference speed is improved. Fig. 3 supports this, demonstrating that suitable caching strategies accelerate inference without noticeable quality loss.
- **Computational Bottlenecks.** Most of the computational cost is concentrated in the self-attention and MLP layers. While these layers are vital for performance, they also restrict computational efficiency.

As illustrated in Fig. 5, SRC focuses on caching and reusing features from computationally intensive layers, such as the self-attention and MLP layers, thereby significantly reducing redundant computations while maintaining the integrity of the generation process. Specifically, features are recomputed and cached at fixed intervals and then reused over a predetermined number of subsequent timesteps. It involves three key hyperparameters: the cache start step t_{start}, the cache end step t_{end}, and the cache stride t_{stride}. The cache stride, an integer within the range $[t_{\text{start}}, t_{\text{end}}]$, determines the number of consecutive timesteps over which the cached features are reused. The above process is detailed as follows:

- **Initialization.** At the start of the denoising process, the model performs a full forward pass and caches the features for each layer.
- **Cache Trigger.** At any given timestep t, the model checks if the caching condition is met. When this condition holds, it indicates that certain layers are eligible for recomputation and caching.
- **Computation and Caching.** Upon triggering, the model executes a complete forward pass within each DiT block. For the l-th block that satisfies the caching criterion, the outputs from its attention and MLP layers are cached, denoted as C^l_{attn} and C^l_{MLP}, respectively.
- **Feature Reuse.** In the following timesteps, until the next caching trigger, the model bypasses redundant computation by reusing the cached features in the corresponding layers. That is, each block retrieves the previously stored features instead of performing new computations.
- **Cyclic Process.** The cycle of recomputation, caching, and feature reuse repeats continuously until the entire denoising process is complete, thereby significantly accelerating the inference.

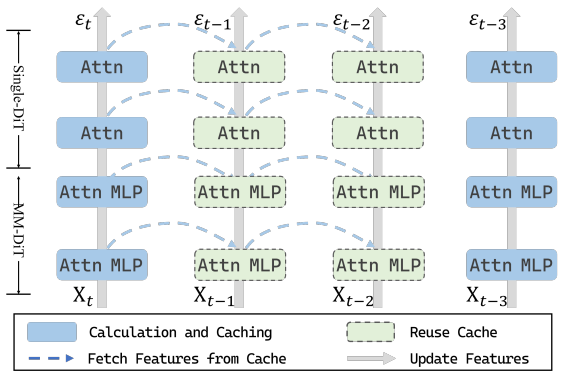

Figure 5: Overview of SRC. In the first time step of each caching period, SRC computes all features and stores them in the cache as initialization. Then, in the subsequent time steps (two are shown), SRC reuses these cached features to avoid redundant computations, alternating in this manner to form a complete caching cycle. Residual connections are omitted for clarity.

4 Experiments

Implementation Details

Experimental Settings. Our experiments employ FLUX.1-dev. We conduct experiments at various resolution scales (4×, 9×, and 16×) on a single H100 GPU, using official implementations and reality-focused textual prompts for all methods to ensure fair assessment of semantic consistency. Following established practice (Du et al. 2023; Huang et al. 2024; He et al. 2023), we randomly sample 10,000 images from the COCO-2017 dataset to form our real image set (Lin et al. 2014). We additionally select 1,000 captions from the same dataset as text prompts for image generation and subsequent evaluation.

Evaluation Metrics. We employ the Frechet Inception Distance (FID) (Heusel et al. 2017) and the Kernel Inception Distance (KID) (Salimans et al. 2016b). Because FID computation requires resizing images to 299×299, which can degrade the evaluation, we follow previous works (Du et al. 2023; Zhang et al. 2024; Yang et al. 2024a) by extracting 10 random crops from each image to compute patch-level FID_c. We also measure the Inception Score (IS) (Salimans et al. 2016a) and adopt a similar patch-level variant denoted by IS_c. Furthermore, the CLIP score (Radford et al. 2021) is reported to evaluate the semantic alignment.

Baseline. We compared the proposed LookFlow with the following baseline approaches: (1) SDXL direct inference (SDXL DI) (Podell et al. 2023), (2) FreeScale (Qiu et al. 2024), (3) FouriScale (Huang et al. 2024), (4) ScaleCrafter (He et al. 2023), (5) DemoFusion (Du et al. 2023), (6) Flux direct inference (Flux DI)(Labs 2024), and (7) I-Max (Du et al. 2024). Among these methods, the first five approaches are based on the SDXL model, whereas the last three methods are built upon the Flux.1-dev model.

Method	2048 × 2048							3072 × 3072							4096 × 4096						
	FID↓	IS↑	KID↓	FID_c ↓	IS_c ↑	CLIP↑	T(s)	FID↓	IS↑	KID↓	FID_c ↓	IS_c ↑	CLIP↑	T(s)	FID↓	IS↑	KID↓	FID_c ↓	IS_c ↑	CLIP↑	T(s)
SDXL DI	100.642	14.769	0.025	76.551	21.154	29.413	14	176.220	8.494	0.067	110.728	16.168	26.217	42	214.927	7.171	0.095	117.076	15.057	24.610	104
FreeScale	73.586	17.544	0.012	52.410	25.443	29.688	54	105.865	12.583	0.027	58.240	**23.887**	28.374	152	119.182	10.875	0.030	105.305	**22.493**	28.104	330
FouriScale	62.382	21.071	0.008	43.832	26.707	**31.356**	90	74.037	20.572	0.015	76.970	22.355	29.351	250	90.732	19.126	0.027	118.394	18.760	28.807	483
ScaleCrafter	62.296	20.713	0.007	45.526	25.708	30.897	54	82.480	16.923	0.015	69.809	19.371	29.440	147	89.621	15.804	0.017	79.538	21.335	29.102	630
DemoFusion	65.774	16.371	0.016	**37.624**	**30.441**	29.121	38	71.353	16.663	0.018	59.334	20.137	29.233	122	73.805	17.042	0.023	90.302	13.328	29.360	281
Flux DI	**59.204**	21.367	0.011	47.357	24.639	30.881	38	203.495	4.580	0.139	281.168	1.444	24.131	135	392.112	1.924	0.348	380.194	1.641	19.298	360
I-Max	64.458	18.569	0.015	50.426	21.417	30.691	52	65.345	18.470	0.017	59.550	19.676	30.294	165	65.924	18.263	0.018	**64.589**	18.971	30.155	435
LookFlow	60.815	**23.191**	**0.005**	49.209	26.772	31.001	35	**59.597**	**24.414**	**0.005**	**54.087**	22.073	**30.830**	92	**59.433**	**24.215**	**0.005**	66.818	17.801	**30.662**	222

Table 1: Quantitative comparison results (best in **bold**, second-best underlined).

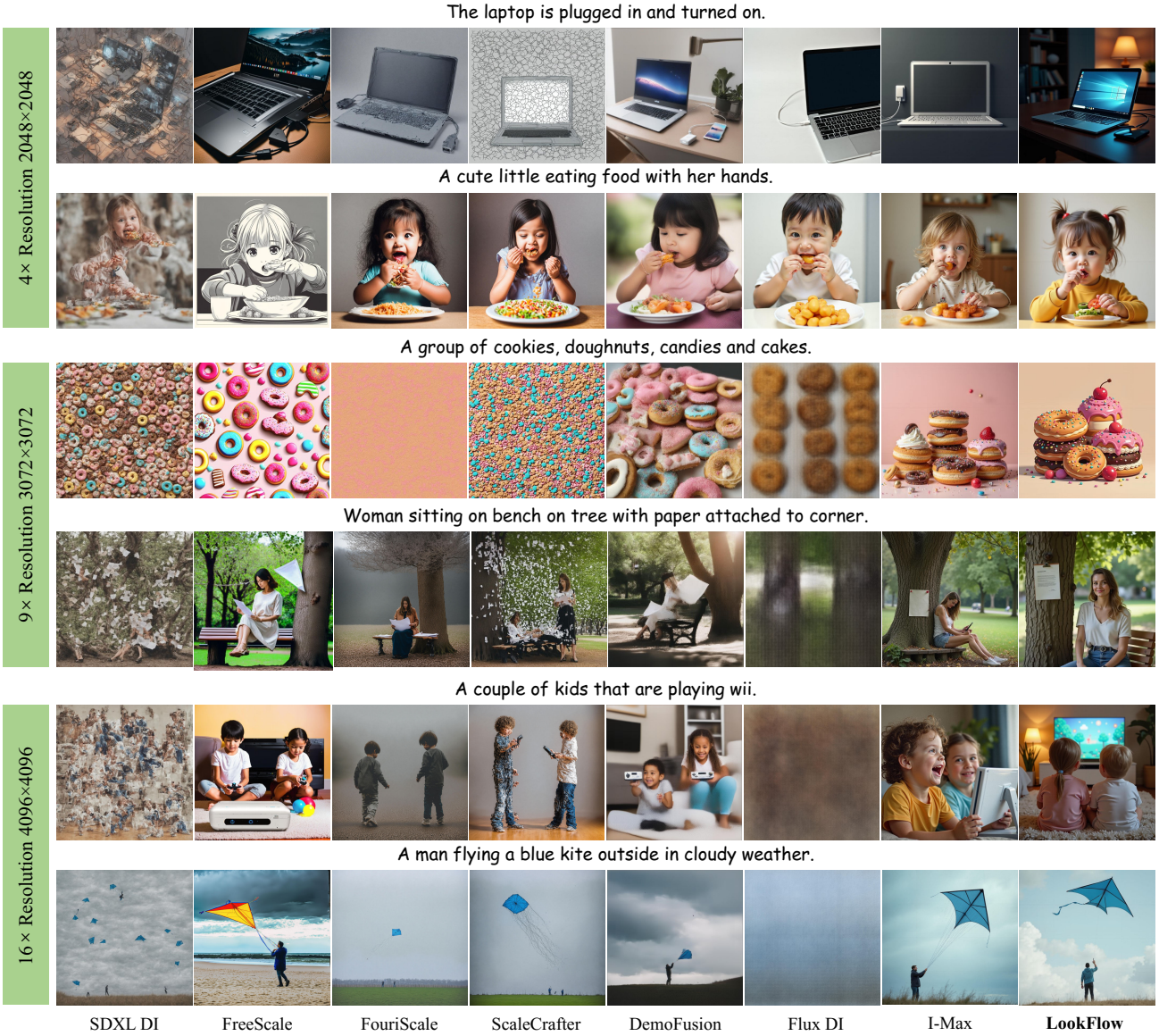

Figure 6: Qualitative comparison with other baselines. Best viewed ZOOMED-IN.

Main Results

Quantitative Comparison. In Tab. 1, LookFlow consistently outperforms other methods across all metrics. While Flux DI achieves the best FID at lower resolutions (e.g., 2048) due to its native generation range, its performance

Figure 7: Ablation study on DLG and SRC.

degrades sharply at higher resolutions, resulting in blurred and incoherent outputs (Fig. 6). Similarly, SDXL DI suffers from pattern repetition beyond its default resolution. Among SDXL-based approaches, DemoFusion remains competitive, and I-Max maintains quality but at the cost of higher latency. In contrast, our method demonstrates superior scalability, maintaining or improving performance as resolution increases.

Visualization. Qualitative comparisons are presented in Fig. 6. While direct high-resolution generators like SDXL DI and Flux DI often suffer from object duplication, chaotic textures, and blurriness, other specialized methods also exhibit distinct limitations. ScaleCrafter frequently introduces locally repetitive elements and fragmented distortions, whereas DemoFusion shows degraded edge definition and shadow artifacts during upscaling. FouriScale exhibits insufficient adherence to textual prompts, (e.g., missing cookies), and FreeScale, though avoiding repetition, lacks semantic coherence. Furthermore, I-Max lacks fine-grained detail and occasionally generates extraneous artifacts like floating objects or discontinuous lines. In contrast, our proposed LookFlow consistently produces high-fidelity images with superior detail richness and semantic coherence, effectively eliminating unintended repetitions.

s	FID↓	IS↑	KID↓	FID_c↓	IS_c↑	CLIP↑
1	61.138	22.302	0.010	**48.277**	26.634	30.972
2	**60.815**	**23.191**	**0.005**	49.209	**26.772**	**31.001**
3	61.812	22.933	0.009	50.501	26.135	30.994
4	62.604	22.115	0.012	50.419	24.980	30.892
5	64.207	22.364	0.013	51.963	24.039	30.914

Table 2: Performance across look-ahead steps s.

Ablations

In Fig. 7, comparative experiments clearly indicate that DLG substantially enhances image quality, ensuring robust semantic consistency between high-resolution outputs and their low-resolution counterparts. Interestingly, incorporating SRC minimally impacts image quality metrics but significantly reduces inference time, particularly notable in higher-resolution scenarios, achieving speedups up to $2.01\times$ (refer to Fig. 3). We also conduct comprehensive ablation studies to evaluate: (1) the optimal lookahead horizon for balancing semantic consistency and generation quality compared to conventional single-step guidance, (2) the optimal cache step t_{stride} for achieving the quality-efficiency trade-off. All images are generated at 2048×2048.

t_{stride}	FID↓	IS↑	KID↓	FID_c↓	IS_c↑	CLIP↑
2	61.189	22.908	0.007	49.863	25.750	30.979
3	60.994	**23.226**	0.012	**48.992**	26.441	30.996
4	60.815	23.191	**0.005**	49.209	**26.772**	**31.001**
5	**60.599**	22.861	0.010	49.706	26.321	30.916

Table 3: Performance under different caching strides (t_{stride}).

Performance at different lookahead step. To investigate the optimal look-ahead strategy, we evaluated the impact of varying look-ahead step s, with results detailed in Tab. 2. Employing adjacent-step guidance (i.e., $s = 1$), equivalent to directly utilizing the immediate next step in the low-resolution space, yields suboptimal performance. In contrast, extending the look-ahead to two steps strikes an optimal balance between metrics, enhancing both semantic alignment and quantitative quality. However, as the step size increases to 3, 4, and 5, performance gradually degrades, corroborating our hypothesis that a moderate temporal context (two steps) provides sufficient guidance while mitigating the noise accumulation observed in longer horizons.

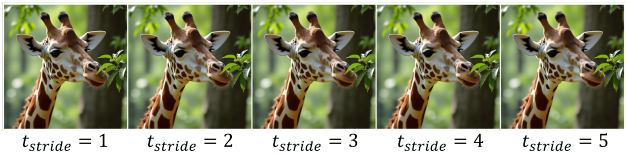

Figure 8: Results with different t_{stride}. The case where $t_{\text{stride}} = 1$ is equivalent to no SRC strategy being applied.

Quality-Efficiency trade-off. The caching step interval t_{stride} governs the frequency of feature caching and reuse, influencing both inference speed and image quality. Larger strides reuse more features, cutting computation and accelerating inference, yet may degrade fidelity. Experiments with strides 2–5 in Tab. 3. show that t_{stride}=4 delivers the best efficiency–quality balance. Notably, even with a stride of 5, the quality degradation is minimal, as shown in Fig. 8, highlighting the robustness of the proposed SRC method.

5 Conclusion

This paper presents LookFlow, a training-free framework for efficient, high-resolution image synthesis. LookFlow addresses semantic inconsistencies and inference overhead in rectified flow Transformers through two core innovations: (1) dynamic lookahead guidance, which refines high-resolution velocity predictions via multi-timestep low-resolution context, and (2) sparse recomputation and caching, which leverages temporal consistency to eliminate redundant computations. Empirical results on the COCO dataset demonstrate that LookFlow consistently maintains competitive visual quality across various scaling factors.

6 Acknowledgments

This document is the result of the research project funded by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB0500103) and the National Natural Science Foundation of China (Grant No. 62306310). The authors gratefully acknowledge this support.

References

Albergo, M. S.; and Vanden-Eijnden, E. 2022. Building Normalizing Flows with Stochastic Interpolants. *ArXiv*, abs/2209.15571.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024a. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, 74–91. Springer.

Chen, J.; Wu, Y.; Luo, S.; Xie, E.; Paul, S.; Luo, P.; Zhao, H.; and Li, Z. 2024b. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426.

Chen, P.; Shen, M.; Ye, P.; Cao, J.; Tu, C.; Bouganis, C.-S.; Zhao, Y.; and Chen, T. 2024c. Δ-DiT: A Training-Free Acceleration Method Tailored for Diffusion Transformers. arXiv:2406.01125.

Du, R.; Chang, D.; Hospedales, T. M.; Song, Y.-Z.; and Ma, Z. 2023. DemoFusion: Democratising High-Resolution Image Generation With No $$$. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6159–6168.

Du, R.; Liu, D.; Zhuo, L.; Qi, Q.; Li, H.; Ma, Z.; and Gao, P. 2024. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. *arXiv preprint arXiv:2410.07536*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Guo, L.; He, Y.; Chen, H.; Xia, M.; Cun, X.; Wang, Y.; Huang, S.; Zhang, Y.; Wang, X.; Chen, Q.; et al. 2024. Make a Cheap Scaling: A Self-Cascade Diffusion Model for Higher-Resolution Adaptation. *arXiv preprint arXiv:2402.10491*.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

He, Y.-Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models. *ArXiv*, abs/2310.07702.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Neural Information Processing Systems*.

Ho, J.; Jain; et al. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Hoogeboom, E.; Heek, J.; and Salimans, T. 2023. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, 13213–13232. PMLR.

Huang, L.; Fang, R.; Zhang, A.; Song, G.; Liu, S.; Liu, Y.; and Li, H. 2024. FouriScale: A Frequency Perspective on Training-Free High-Resolution Image Synthesis. In *European Conference on Computer Vision*.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2023. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36: 70847–70860.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Li, S.; Hu, T.; Khan, F. S.; Li, L.; Yang, S.; Wang, Y.; Cheng, M.-M.; and Yang, J. 2023. Faster Diffusion: Rethinking the Role of UNet Encoder in Diffusion Models. *ArXiv*, abs/2312.09608.

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. 2024. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, J.; Geddes, J.; Guo, Z.; Jiang, H.; and Nandwana, M. K. 2024a. SmoothCache: A Universal Inference Acceleration Technique for Diffusion Transformers. arXiv:2411.10510.

Liu, S.; Yu, W.; Tan, Z.; and Wang, X. 2024b. LinFusion: 1 GPU, 1 Minute, 16K Image.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *ArXiv*, abs/2209.03003.

Liu, X.; He, Y.; Guo, L.; Li, X.; Jin, B.; Li, P.; Li, Y.; Chan, C.-M.; Chen, Q.; Xue, W.; et al. 2024c. Hiprompt: Tuning-free higher-resolution generation with hierarchical mllm prompts. *arXiv preprint arXiv:2409.02919*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Ma, X.; Fang, G.; and Wang, X. 2024. DeepCache: Accelerating Diffusion Models for Free. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15762–15772.

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

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qiu, H.; Zhang, S.; Wei, Y.; Chu, R.; Yuan, H.; Wang, X.; Zhang, Y.; and Liu, Z. 2024. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. *arXiv preprint arXiv:2412.09626*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016a. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Salimans, T.; Goodfellow, I. J.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016b. Improved Techniques for Training GANs. *ArXiv*, abs/1606.03498.

Shi, S.; Li, W.; Zhang, Y.; He, J.; Gong, B.; and Zheng, Y. 2024. ResMaster: Mastering High-Resolution Image Generation via Structural and Fine-Grained Guidance. *arXiv preprint arXiv:2406.16476*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Teng, J.; Zheng, W.; Ding, M.; Hong, W.; Wangni, J.; Yang, Z.; and Tang, J. 2023. Relay Diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*.

Wu, H.; Shen, S.; Hu, Q.; Zhang, X.; Zhang, Y.; and Wang, Y. 2024. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. *arXiv preprint arXiv:2408.11001*.

Yang, H.; Bulat, A.; Hadji, I.; Pham, H. X.; Zhu, X.; Tzimiropoulos, G.; and Martinez, B. 2024a. FAM Diffusion: Frequency and Attention Modulation for High-Resolution Image Generation with Stable Diffusion. *arXiv preprint arXiv:2411.18552*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Zhang, L.; Rao; et al. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, S.; Chen, Z.; Zhao, Z.; Chen, Y.; Tang, Y.; and Liang, J. 2024. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, 145–161. Springer.

Zhao, X.; Jin, X.; Wang, K.; and You, Y. 2024. Real-Time Video Generation with Pyramid Attention Broadcast. arXiv:2408.12588.

Zheng, Q.; Guo, Y.; Deng, J.; Han, J.; Li, Y.; Xu, S.; and Xu, H. 2024a. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7571–7578.

Zheng, W.; Teng, J.; Yang, Z.; Wang, W.; Chen, J.; Gu, X.; Dong, Y.; Ding, M.; and Tang, J. 2024b. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*, 1–22. Springer.