NEURAL INFORMATION
PROCESSING SYSTEMS

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

中国科学技术大学
University of Science and Technology of China

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

HiFlow: Training-free High-Resolution Image Generation with Flow-Aligned Guidance

Slide Talk

Jiazi Bu*, Pengyang Ling*, Yujie Zhou*, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, Jiaqi Wang

CONTENTS

1. Introduction and Motivation
2. Methodology
3. Experiments and Results

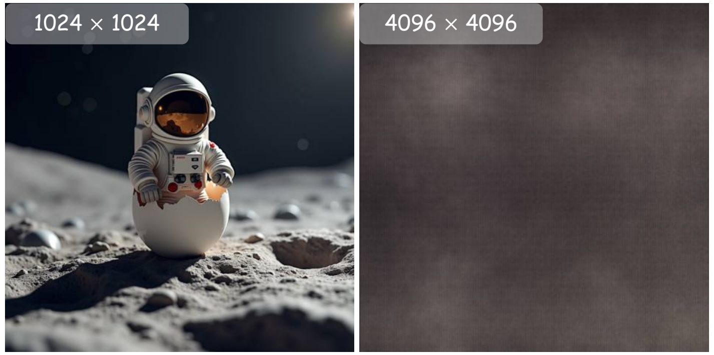
01.

Introduction and Motivation

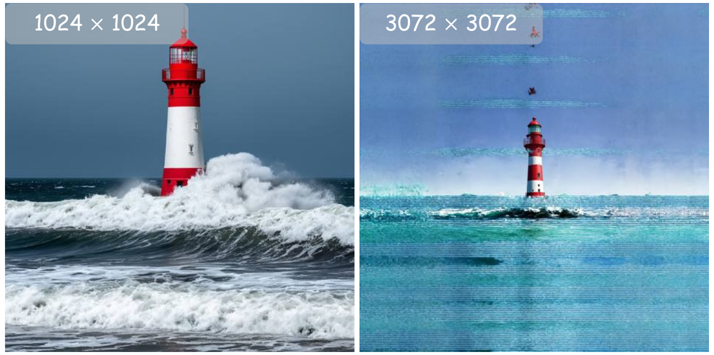
■ Background

- □ **T2I diffusion models (e.g. *Stable Diffusion*, *Flux*) have achieved a landmark advancement in the realm of visual synthesis. Despite their success, existing T2I models are typically confined to a restricted resolution (e.g., 1024×1024, 1K).**

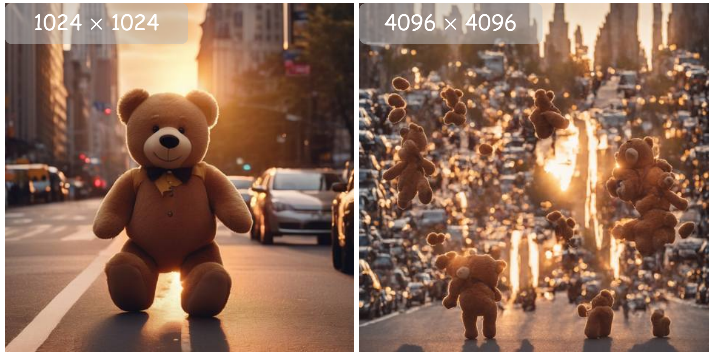
- □ **These models experience notable quality decline and even structural breakdown when attempting to generate higher-resolution images.**



FLUX (DiT/Flow)



SD3 Medium (DiT/Flow)



SDXL (U-Net/Diffusion)

■ Related Works

- □ Initial efforts (e.g., *SimpleDiffusion*, *UltraPixel*) suggest fine-tuning T2I models on higher-resolution samples to enhance the adaptability to large-scale images. However, they entail significant costs, primarily the burden of high-resolution image collection and the necessity for model-specific fine-tuning.

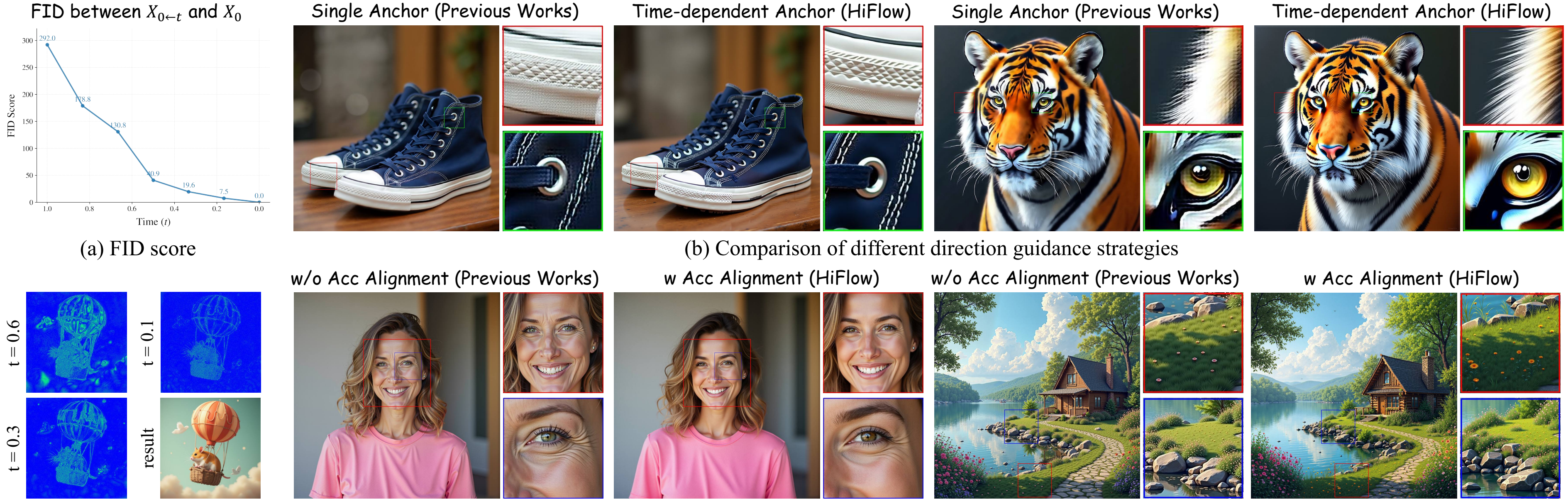
- □ Recent studies have focused on have training-free strategies to unlock the resolution potential of pretrained T2I models in high-resolution image synthesis.

  - □ Most of them (e.g., *HiDiffusion*, *FreeScale*) involve manipulating the internal features within models, exhibiting restricted transferability across architectures. (e.g. apply methods for U-Net-based models to DiT-based models)

  - □ Another line of research (e.g., *DemoFusion*, *I-Max*) suggests fusing the upsampled low-resolution images into the denoising target during high-resolution synthesis for structure guidance.

■ Observation & Motivation

- □ Distribution discrepancy exists between predicted clean sample $X_{0\leftarrow t}$ and clean sample X_0. Directly fusing a constant structural guidance anchor can cause artifacts.
  - □ Constant Anchor → Time-dependent Anchor
- □ Lack of detail-oriented guidance, leading to decline in detail fidelity and the emergence of unrealistic contents, such as repetitive patterns and abnormal textures.
  - □ No detail guidance → Detail Guidance (via Flow Acceleration Alignment)



(a) FID score

(b) Comparison of different direction guidance strategies

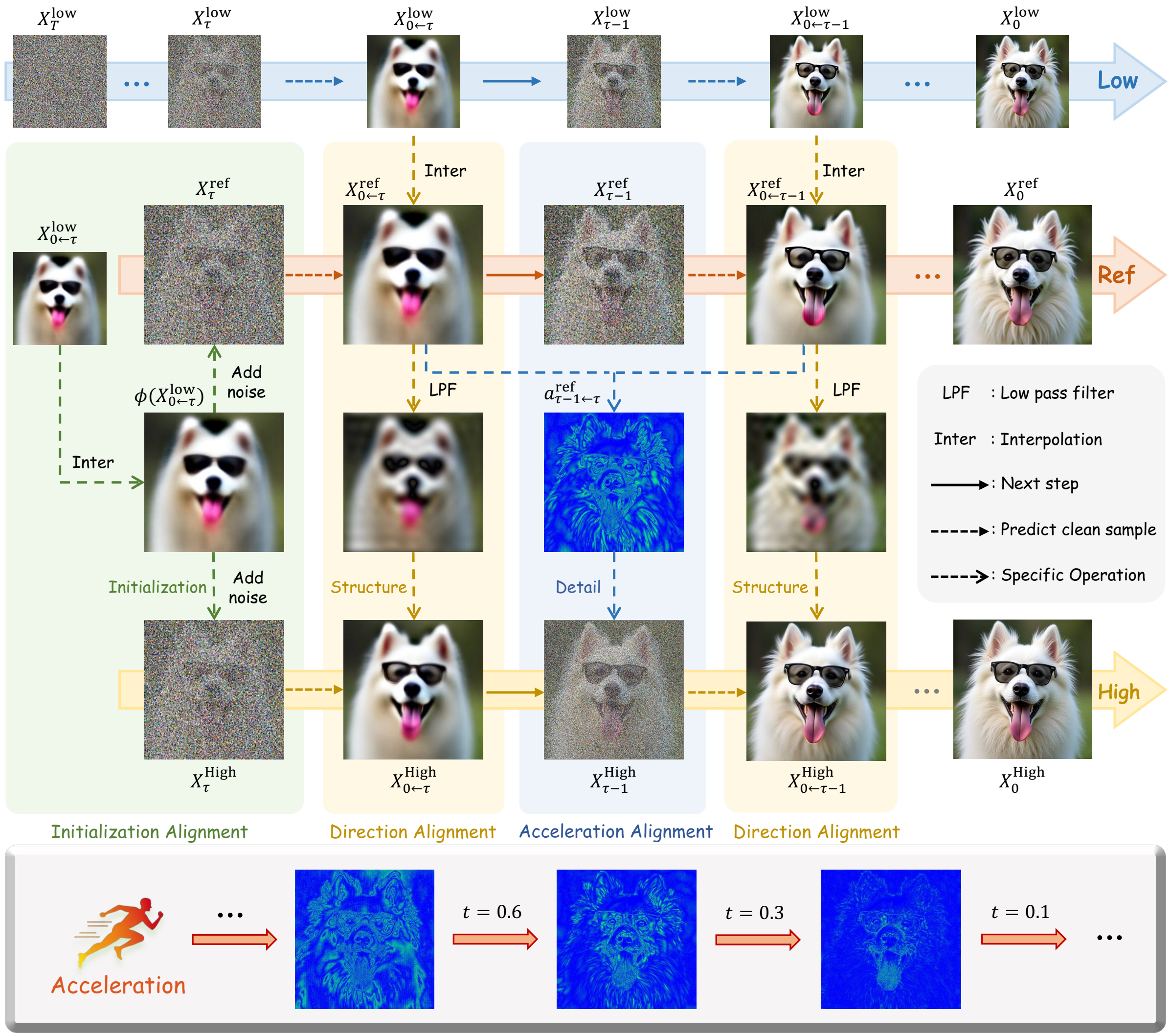
(c) Acceleration visualization

(d) Effects of acceleration alignment in detail synthesis

02.

Methodology

■ Method Overview



- □ We introduce ***HiFlow***, a novel training-free and model-agnostic framework to advance high-resolution synthesis of T2I models.

- □ HiFlow: A cascade generation paradigm:

  - □ Construct a virtual reference flow in the high-resolution space based on step-wise low-resolution predicted clean samples.

  - □ During high-resolution generation process, the above reference flow offers guidance in initialization, denoising direction, and flow acceleration, aiding in achieving consistent low-frequency patterns, preserving overall structural features, and maintaining high-fidelity details, respectively.

■ Reference Flow Construction

□ A virtual reference flow is constructed in the high-resolution space that can fully characterize the information of the low-resolution sampling trajectory.

□ Specifically, for reference flow, its predicted clean image $X_{0\leftarrow t}^{\text{ref}}$ at timestep t is defined as:

$$\left\{X_{0\leftarrow t}^{\text{ref}}\right\} = \left\{\phi(X_{0\leftarrow t}^{\text{low}})\right\}, \quad t \in [0, 1]$$

□ $\Phi(\cdot)$ is an interpolation function like bilinear interpolation or bicubic interpolation.

□ Note that the noisy image X_t^{ref} and the corresponding flow vector v_t^{ref} are both imaginary, i.e., they cannot be actually sampled by the flow model. Finally, the reference flow can produce $\Phi(X_0^{\text{low}})$, which stands for the upsampled low-resolution image output.

□ Such a reference flow acts as a bridge connecting low-resolution and high-resolution sampling trajectories , facilitating guided high-resolution synthesis with enhanced structure and fidelity.

Flow-Aligned Guidance: Initialization Guidance

- For a given virtual reference flow, the sampling of high-resolution generation starts from the noisy variant of $X_{0\leftarrow\tau}^{\mathrm{ref}}$, which can be expressed as:

$$X_{\tau}^{\mathrm{high}} = \tau X_{1}^{\mathrm{high}} + (1-\tau)X_{0\leftarrow\tau}^{\mathrm{ref}},$$

- $X_{0\leftarrow\tau}^{\mathrm{high}}$ is the sampling initialization of high-resolution generation, X_{1}^{high} is a random gaussian noise in the high-resolution space.

- Such initialization alignment allows skipping the early stage in high-resolution generation, thereby maintaining the consistency of high-resolution results and low-resolution images in low-frequency components, while also facilitating a higher inference speed.

■ Flow-Aligned Guidance: Direction Guidance

- □ Initialization alignment ensures low-frequency consistency at the beginning of high-resolution generation. However, such structural information may be destroyed at subsequent denoising steps.

- □ Direction alignment is designed for structural preservation, which is achieved by modifying the denoising direction of the high-resolution flow based on reference flow:

$$\hat{X}^{\text{high}}_{0\leftarrow t} = X^{\text{high}}_{0\leftarrow t} + \alpha_t[\widetilde{\mathcal{F}}(\mathcal{F}(X^{\text{ref}}_{0\leftarrow t}) \odot \mathcal{L}(D)) - \widetilde{\mathcal{F}}(\mathcal{F}(X^{\text{high}}_{0\leftarrow t}) \odot \mathcal{L}(D))],$$

- □ $\mathcal{F}$ and $\widetilde{\mathcal{F}}$ denote 2D Fast Fourier Transform and its inverse operation. $\mathcal{L}(\cdot)$ denotes a butterworth low-pass filter and D is the normalized cutoff frequency. α_t is the direction guidance weight.

- □ Essentially, direction alignment replaces the low-frequency component of $X^{\text{high}}_{0\leftarrow t}$ with that in $X^{\text{ref}}_{0\leftarrow t}$, thus rejecting the updating on low-frequency structures when synthesizing high-frequency details.

- □ Unlike previous methods, we utilize a time-dependent guidance anchor $X^{\text{ref}}_{0\leftarrow t}$ instead of a single constant anchor X^{low}_{0}, thereby avoiding artifacts caused by distribution discrepancy.

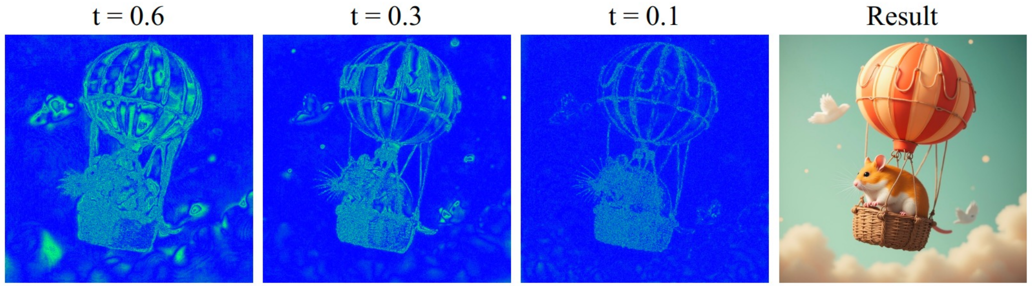
■ Flow-Aligned Guidance: Acceleration Guidance

- □ Although the above strategies allow for rich detail generation while maintaining structure, the fidelity of synthesized details risks dropping in some cases, in which unrealistic contents appear.

- □ Acceleration is defined as the second-order derivative of movement X_t, and also denotes the first-order derivative of vector v_t, can be expressed as:

$$a_{t_{i-1}\leftarrow t_i} = \frac{d^2 X_t}{dt^2} = \frac{dv_t}{dt} = \frac{v_{t_{i-1}} - v_{t_i}}{t_{i-1} - t_i}.$$

- □ By visualizing the flow acceleration, we discover that it primarily captures texture and contour information while also indicating the sequence of content synthesis at each timestep, i.e., it showcases what content the model is responsible for adding at different t.



■ Flow-Aligned Guidance: Acceleration Guidance

- Furthermore, we can simplify the acceleration as:

$$
\begin{aligned}
a_{t_{i-1}\leftarrow t_i} &= \frac{\frac{1}{t_{i-1}}(X_{t_{i-1}} - X_{0\leftarrow t_{i-1}}) - \frac{1}{t_i}(X_{t_i} - X_{0\leftarrow t_i})}{t_{i-1} - t_i} \\
&= \frac{t_i X_{t_i} + (t_{i-1} - t_i)(X_{t_i} - X_{0\leftarrow t_i}) - t_i X_{0\leftarrow t_{i-1}} - t_{i-1}(X_{t_i} - X_{0\leftarrow t_i})}{t_{i-1} t_i (t_{i-1} - t_i)} \\
&= -\frac{1}{t_{i-1}} \frac{X_{0\leftarrow t_{i-1}} - X_{0\leftarrow t_i}}{t_{i-1} - t_i}.
\end{aligned}
$$

- It can be concluded that the acceleration depicts the variation in the predicted clean sample $X_{0\leftarrow t}$ between adjacent timesteps, with time-dependent term $1/t$.

- Therefore, we propose aligning the acceleration of high-resolution generation with that of the reference flow to synchronize the model's preference for content synthesis order, enabling guided detail synthesis in both content and timing:

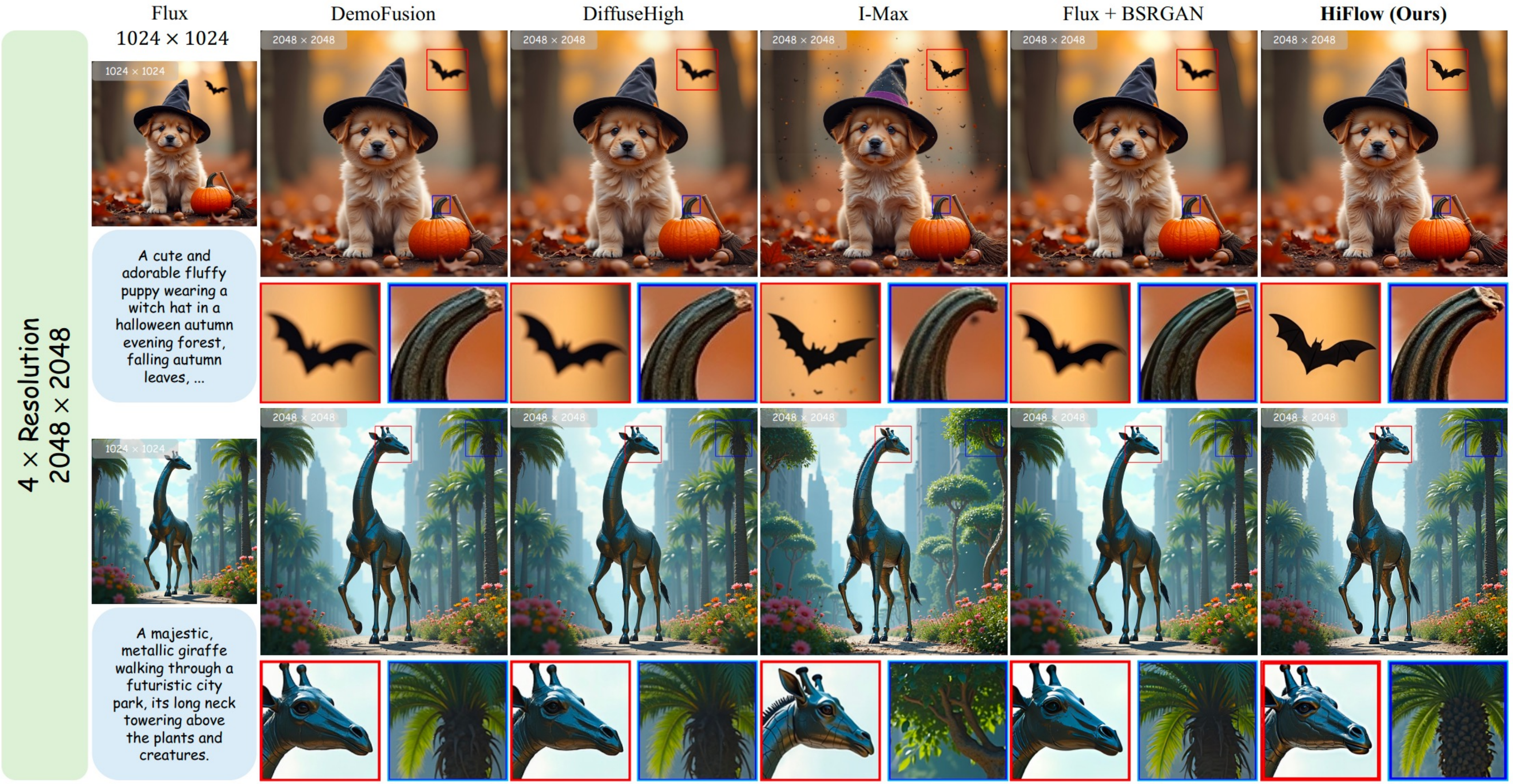
$$
\hat{a}^{\text{high}}_{t_{i-1}\leftarrow t_i} = a^{\text{high}}_{t_{i-1}\leftarrow t_i} + \beta_t (a^{\text{ref}}_{t_{i-1}\leftarrow t_i} - a^{\text{high}}_{t_{i-1}\leftarrow t_i}),
$$

03.

Experiments and Results

■ Qualitative Comparison with Existing Baselines

- □ **Comparison in 2K resolution. Compared to previous approaches, HiFlow yields high-resolution images characterized by high-fidelity details and coherent structure.**

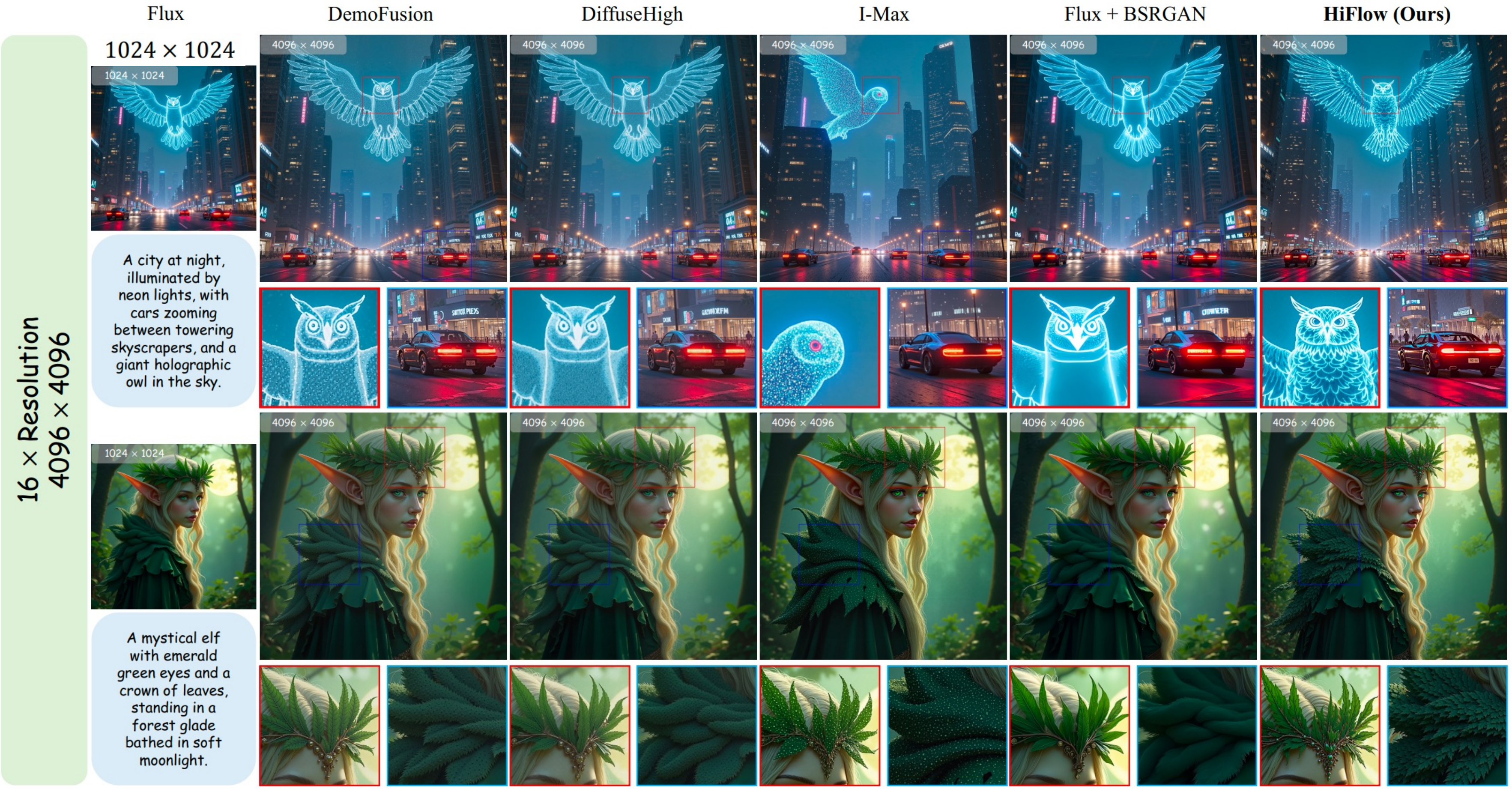


A cute and adorable fluffy puppy wearing a witch hat in a halloween autumn evening forest, falling autumn leaves, …

A majestic, metallic giraffe walking through a futuristic city park, its long neck towering above the plants and creatures.

■ Qualitative Comparison with Existing Baselines

□ **Comparison in 4K resolution. Compared to previous approaches, HiFlow yields high-resolution images characterized by high-fidelity details and coherent structure.**


Flux
DemoFusion
DiffuseHigh
I-Max
Flux + BSRGAN
HiFlow (Ours)
16 × Resolution
4096 × 4096
1024 × 1024
1024 × 1024
4096 × 4096
A city at night, illuminated by neon lights, with cars zooming between towering skyscrapers, and a giant holographic owl in the sky.
A mystical elf with emerald green eyes and a crown of leaves, standing in a forest glade bathed in soft moonlight.

Quantitative Evaluation

- **HiFlow achieves superior performance in terms of both image quality metrics (FID, $\text{FID}_{\text{patch}}$, IS, IS_{patch}) and image-text alignment (CLIP Score) under different resolutions, validating its stable and superior performance in high-resolution image generation.**

- **HiFlow surpasses all its training-free competitors in inference speed, offering higher efficiency.**

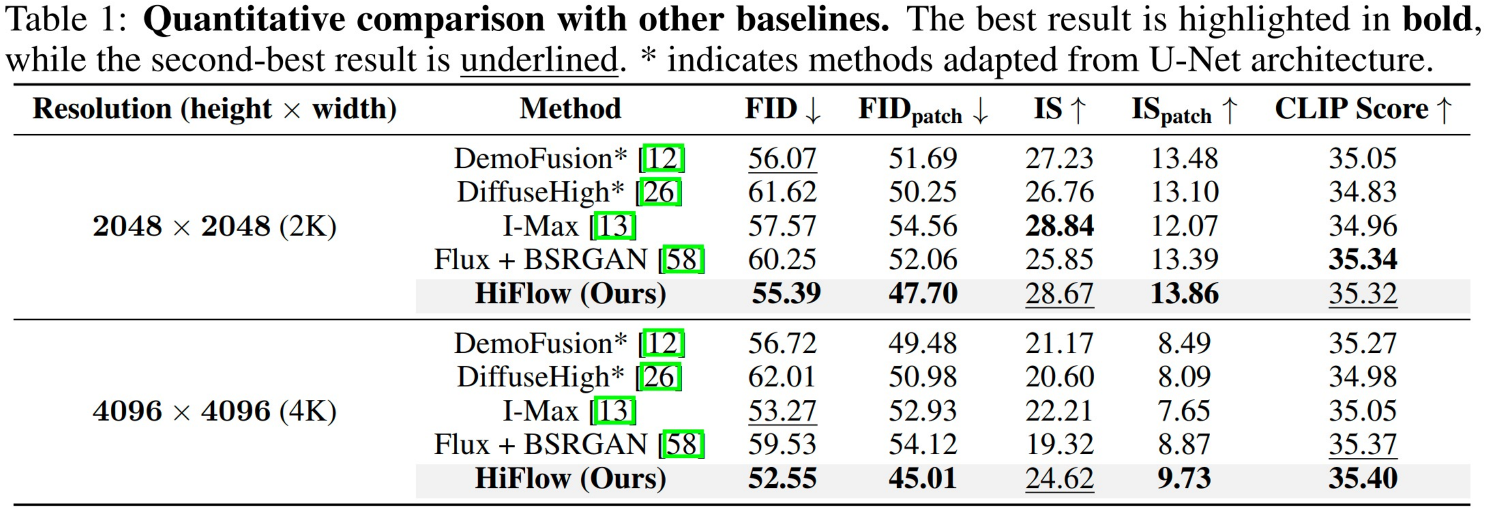


Table 1: **Quantitative comparison with other baselines.** The best result is highlighted in **bold**, while the second-best result is underlined. * indicates methods adapted from U-Net architecture.

Resolution (height × width)	Method	FID ↓	$\text{FID}_{\text{patch}}$ ↓	IS ↑	IS_{patch} ↑	CLIP Score ↑
2048 × 2048 (2K)	DemoFusion* [12]	56.07	51.69	27.23	13.48	35.05
	DiffuseHigh* [26]	61.62	50.25	26.76	13.10	34.83
	I-Max [13]	57.57	54.56	**28.84**	12.07	34.96
	Flux + BSRGAN [58]	60.25	52.06	25.85	13.39	**35.34**
	HiFlow (Ours)	**55.39**	**47.70**	28.67	**13.86**	35.32
4096 × 4096 (4K)	DemoFusion* [12]	56.72	49.48	21.17	8.49	35.27
	DiffuseHigh* [26]	62.01	50.98	20.60	8.09	34.98
	I-Max [13]	53.27	52.93	22.21	7.65	35.05
	Flux + BSRGAN [58]	59.53	54.12	19.32	8.87	35.37
	HiFlow (Ours)	**52.55**	**45.01**	24.62	**9.73**	**35.40**

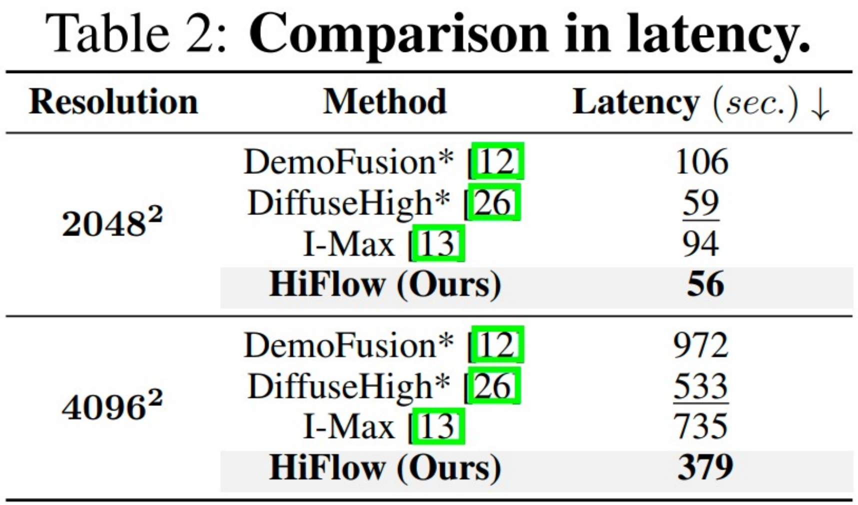


Table 2: **Comparison in latency.**

Resolution	Method	Latency (*sec.*) ↓
2048^2	DemoFusion* [12]	106
	DiffuseHigh* [26]	59
	I-Max [13]	94
	HiFlow (Ours)	**56**
4096^2	DemoFusion* [12]	972
	DiffuseHigh* [26]	533
	I-Max [13]	735
	HiFlow (Ours)	**379**

Ablation and Analysis

- HiFlow performs guided high-resolution generation by aligning its flow with the reference flow in three aspects: initialization (A_i), direction (A_d) and acceleration (A_a).
  - A_i helps avoid semantic incorrectness by facilitating low-frequency consistency with low-resolution images.
  - A_d enhances structure preservation in the generation process by suppressing the updating in the low-frequency component.
  - A_a contributes to high-fidelity detail synthesizing.

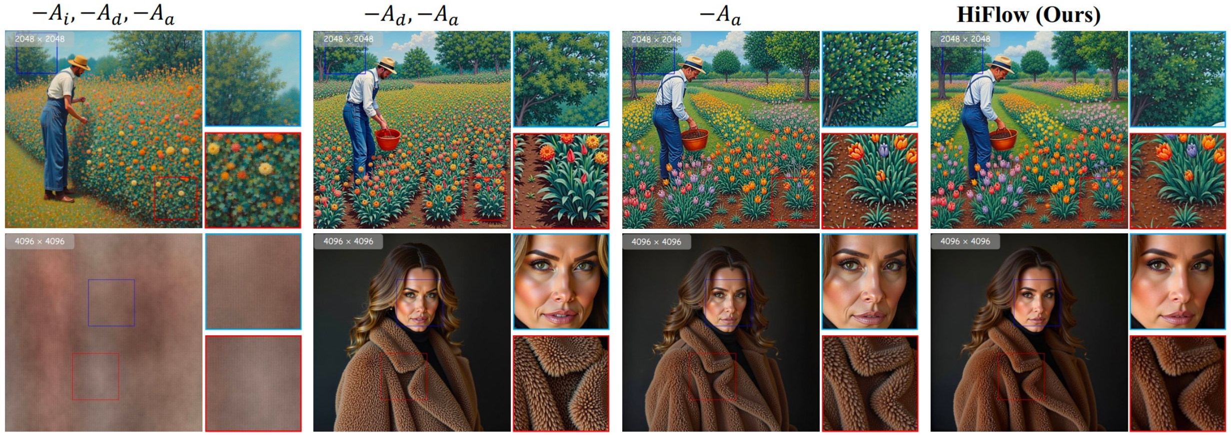
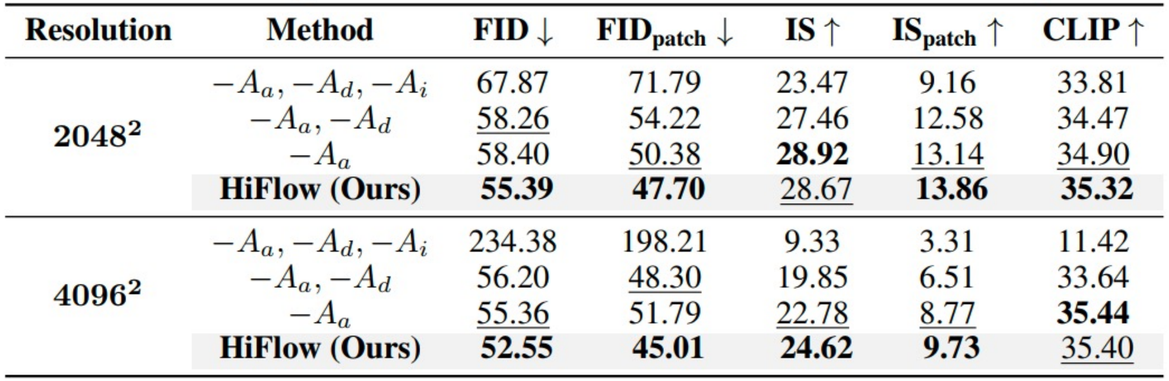


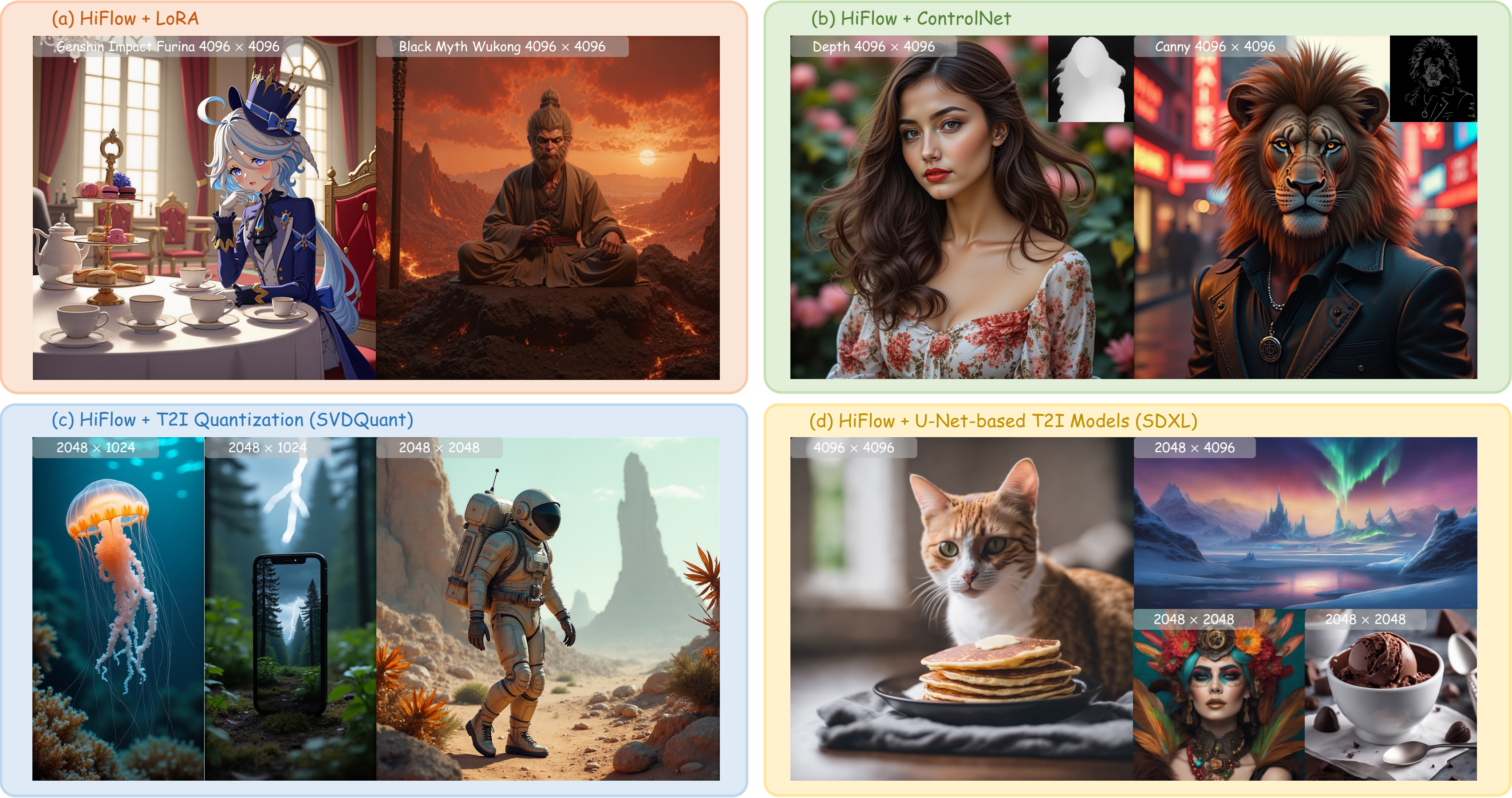
Table 3: **Quantitative results of ablation study.**



Resolution	Method	FID ↓	FID$_{\text{patch}}$ ↓	IS ↑	IS$_{\text{patch}}$ ↑	CLIP ↑
2048^2	$-A_a, -A_d, -A_i$	67.87	71.79	23.47	9.16	33.81
	$-A_a, -A_d$	58.26	54.22	27.46	12.58	34.47
	$-A_a$	58.40	50.38	**28.92**	13.14	34.90
	HiFlow (Ours)	**55.39**	**47.70**	28.67	**13.86**	**35.32**
4096^2	$-A_a, -A_d, -A_i$	234.38	198.21	9.33	3.31	11.42
	$-A_a, -A_d$	56.20	48.30	19.85	6.51	33.64
	$-A_a$	55.36	51.79	22.78	8.77	**35.44**
	HiFlow (Ours)	**52.55**	**45.01**	**24.62**	**9.73**	35.40

Versatile Applications of HiFlow

- **HiFlow can be seaminglessly integrated with customization T2I models (e.g., *LoRA*, *ControlNet*), T2I quantization (e.g., *SVDQuant*) and U-Net-based models (e.g., *SDXL*), highlighting its versatility.**



NEURAL INFORMATION
PROCESSING SYSTEMS

THANKS

Jiazi Bu*, Pengyang Ling*, Yujie Zhou*, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, Jiaqi Wang