Supplementary Materials for Paper "InfGen: A Resolution-Agnostic Paradigm for Scalable Image Synthesis"

Tao Han[1,2], Wanghan Xu[2], Junchao Gong[2], Xiaoyu Yue[2,3],
Song Guo[1], Luping Zhou[3], Lei Bai[2]
[1]Hong Kong University of Science and Technology [2]Shanghai Artificial Intelligence Laboratory
[3]The University of Sydney

The supplementary file for the paper entitled "**InfGen: A Resolution-Agnostic Paradigm for Scalable Image Synthesis**" provides additional details on various aspects related to the paper, which are as follows:

1) **Additional Preliminary**
2) **Detailed Network Architectures**
3) **Dataset Source**
4) **Additional Results**

I. Additional Preliminary

I.1. Optimization of VAE

Eq.2 in the main body formulates a transformation from the latent back to the inputs. To ensure the encoder and decoder work optimally, they are jointly optimized using the Evidence Lower Bound (ELBO). This objective function balances two critical aspects:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z)) \tag{1}$$

The ELBO ensures precise reconstruction, while the Kullback-Leibler divergence term encourages a well-structured latent space by aligning it with a prior, typically a standard Gaussian. In practice, the sampling is replaced with a reparametrization strategy [4], allowing gradient descent possible during the training,

$$z = \boldsymbol{\mu}_x + \boldsymbol{\sigma}_x \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I), \tag{2}$$

where mean $\boldsymbol{\mu}_x$ and covariance $\boldsymbol{\sigma}_x$ is produced by the encoder $\mu_\phi(x)$ and $\sigma^2_\phi(x)$, respectively.

Finally, the theoretical expectancy in Eq. 1 is replaced by a more or less accurate Monte-Carlo approximation [1]. The pretraining objective in this section consists of a reconstruction term and a regularisation term,

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + \frac{1}{2}\sum_{i=1}^{d}\left(-\log\sigma_i^2 + \mu_i^2 + \sigma_i^2 - 1\right). \tag{3}$$

I.2. Latent Diffusion Models

LDMs [11] are generative models that leverage diffusion processes in a latent space to efficiently generate high-quality images. They build on the principles of Denoising Diffusion Probabilistic Models (DDPMs) [3] but operate in a more compact space (i.e., the latent representation y introduced above), which reduces computational complexity while maintaining output fidelity.

The forward process in LDMs begins by gradually transforming a data point x_0 into a latent variable z_T. This transformation is achieved through a series of Gaussian noise additions, defined by a Markov chain:

$$z_t = \sqrt{\alpha_t} z_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \tag{4}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ is Gaussian noise, and α_t are predefined variance schedules controlling the noise level at each step t. This process ensures the data is encoded into a noise-like variable in the latent space.

The reverse process is employed to reconstruct the original data point x_0 from the noisy latent variable z_T. This involves learned denoising steps that sequentially remove the noise:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t) \right) + \sqrt{1-\alpha_t}\epsilon_t, \tag{5}$$

where ϵ_θ is a neural network trained to predict the noise added at each step. This reverse diffusion process gradually reconstructs the data, leveraging the learned noise patterns.

To accurately reconstruct data from noise, DDPMs are used to model the complex distributions, which is solved by optimizing the model parameters θ to minimize the variational lower bound (VLB):

$$\mathcal{L} = \mathbb{E}_{z_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2 \right]. \tag{6}$$

This objective encourages the model to accurately predict the noise ϵ at each step, effectively learning the reverse diffusion process. By refining the noise predictions, the model can achieve high-quality reconstructions.

II. Dataset Source

For our training, we utilized high-resolution image datasets from the LAION dataset, specifically sourced from `improved_aesthetics_4.5plus-ultra-hr` and `Laion_aesthetics_5plus_1024_33M`. These datasets provide high-quality images, making them ideal for our task. The image IDs used in our training process can be found in the final open-sourced code repository, ensuring full transparency and reproducibility.

III. Network Architectures

Tab. 1 elaborates the detailed network architecture of InfGen, including a frozen encoder and a trainable decoder.

IV. Additional Results

Comparison with the Commercial Product DALL-E-3. We provide additional visual comparisons between our InfGen + SDXL [9] and the closed-source commercial text-to-image (T2I) product DALL-E-3 [7] in Figure 1. Our method demonstrates the capability to generate high-quality images comparable to those produced by this leading commercial solution.

Additional Visual Results Varying in Resolution and Aspect Ratio from `Gen²` + `SDXL` We present additional visual results from our method in Figures 2, 3, 4, 5, and 6. Our approach demonstrates the ability to generate high-quality images across diverse resolutions and aspect ratios, handling a wide range of scenarios, including close-up portraits, creative content, and photorealistic scenes.

Comparison of Reconstructed and Original Images at Different Resolutions. Figure 7 illustrates the comparison between the original images and their reconstructed counterparts at various resolutions. Although the original images are displayed at different resolutions in the figure, it is important to note that the actual input size for the original images during processing is fixed, such as 512×512. Our method demonstrates strong reconstruction performance across resolutions, effectively preserving key details and overall visual fidelity. Achieving good reconstruction quality at different resolutions can help improve the generation performance of various models across diverse resolution requirements.

Comparison with High-Resolution Image Generation Methods. In Figure 8 and 9, we show more visual comparisons with state-of-the-art methods for generating high-resolution images. What sets our **InfGen** apart is its ability to generate images at any resolution. No matter the resolution, our results consistently look better, with sharper details, more realistic textures, and a clear preservation of small features. This shows that our method isn't just flexible—it also produces high-quality, visually appealing images across a wide range of resolutions. Additionally, the variety and consistency in the generated images demonstrate that our model is reliable and works well in different scenarios, making it a powerful tool for creating realistic images at any scale.

Improving Generation Quality Across Different Models. We applied our reconstruction model, **InfGen**, to various latent space-based generative models, including DiT [8], SiT [6], FiT [5], and MaskDiT [12]. Our approach enables these models to decode fixed latent sizes, such as 32×32 or 64×64, into images of arbitrary resolutions. Figure 10, Figure 11, Figure 12, and

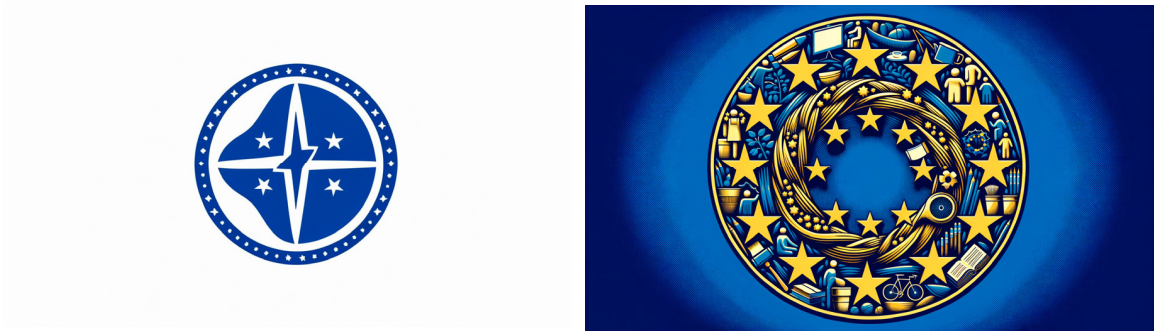
Stage	Latent-Reconstruction Encoder Architecture	Output Sizes
input data	B×3×512×512	
encoder	VAE Encoder	
quan_conv	Conv 8×512×3×3	μ_x: B×4×64×64
	μ_x σ_x	σ_x:B×4×64×64
latent $\mathbf{y}$	$\mathbf{y} \leftarrow \mu_x + \sigma_x \odot \epsilon \sim \mathcal{N}(0,1)$	B×4 ×64 ×64
Decoder Architecture		
post_quant_conv	Conv 512×4×3×3	B×512×64×64
conv_in	Conv 512×512×3×3	B×512×64×64
token generation	Latent token: B×4096×512 + Latent Position Embedding	B×4096×512
	Mask token: B×H^l $*$ W^l×512 + Mask Token Position Embedding	B×H^l $*$ W^l×512
cross_attn_blocks ×N	[Q Linear(512); K Linear(512); V Linear(512)] →Q is mask tokens, K and V are latent tokens Multi Head Cross-Attention Linear(512) MLP(512)	B×H^l $*$ W^l×512 B×H^l $*$ W^l×512
reshape	*from* B×H^l $*$ W^l×512 *to* B×512×H^l×W^l	B×512×H^l×W^l
mid_block	UNetMidBlock2D ×1	
up_blocks	get_up_block ×3	B×128×H^l $*$ 8×W^l $*$ 8
conv_norm_out	GroupNorm(512)	
conv_act	SiLU	
conv_out	Conv 128×3×3×3	B×3×H^l $*$ 8×W^l $*$ 8

Table 1. **Architectures details.** {B×D×H×W} denotes a feature shaped as batch size, channel, width, and height, respectively. {B×N×D} denotes a sequence with the shape of batch size, sequence length, and dimension.

Figure 13 provide a comparison between using our method to change the resolution and directly performing interpolation on the original generated results. The visual comparisons clearly demonstrate that our method produces images with significantly richer details, sharper textures, and more realistic structures compared to simple interpolation techniques.

By leveraging our reconstruction model, the generative models not only achieve higher resolution outputs but also enhance the overall visual quality, especially in preserving fine-grained details that are often lost or blurred with interpolation. Furthermore, our method introduces flexibility, allowing these generative models to adapt seamlessly to different resolution requirements without additional modifications to their architectures. This highlights the effectiveness of our approach in bridging the gap between fixed latent spaces and high-quality image generation at arbitrary resolutions, making it a valuable tool for improving the versatility and performance of existing generative models.

Editor's Choice, pumpkin bread



Think and Create your own Hobbies - EU Logo

Thai Style Soup with Beef by graficallyminded

Parkland hospital emergency room

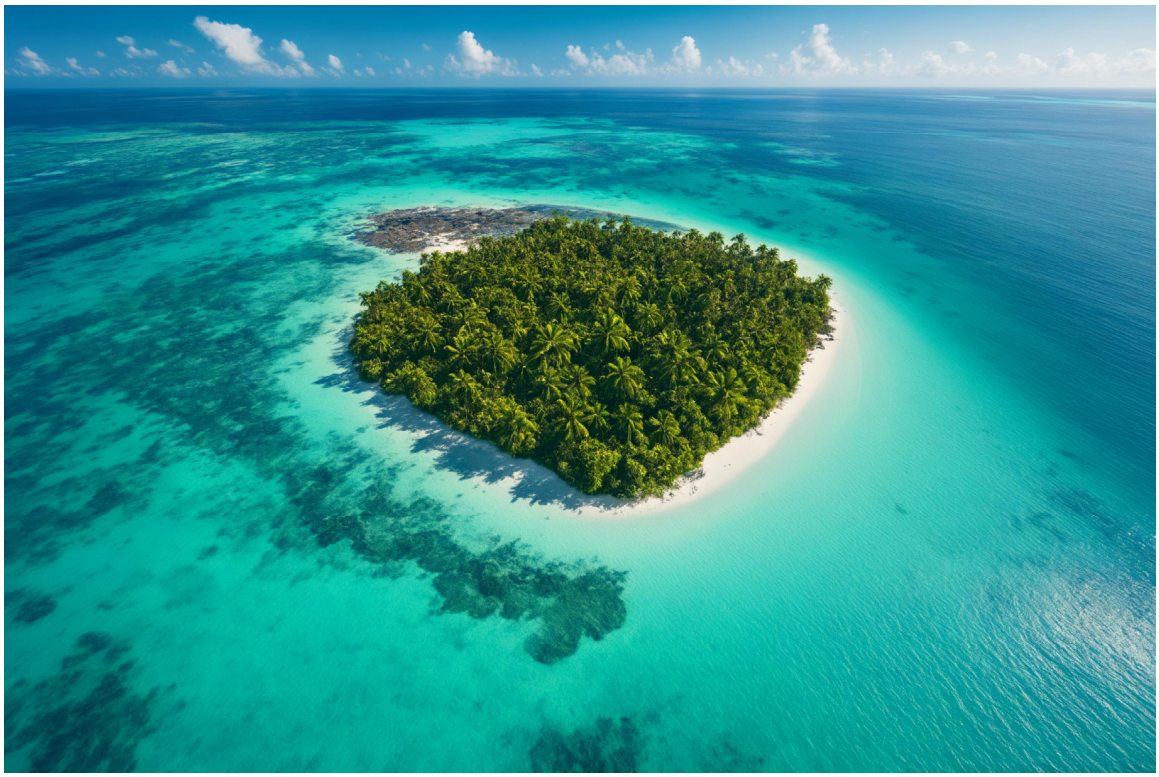
Figure 1. Visual comparison with DALL-E-3 at 1024 × 1792 resolution.

Figure 2. Visual results of InfGen + SDXL at 2888×3720 resolution.

A dense, mystical forest with sunlight streaming through the tall trees, illuminating the lush green moss and foliage, cinematic atmosphere

A tranquil beach at sunrise, gentle waves lapping the shore, with palm trees swaying in the breeze, and soft pastel colors in the sky

Figure 3. Visual results of InfGen + SDXL at 3072×2048 resolution.



An aerial view of a tropical island surrounded by crystal-clear turquoise water, coral reefs visible beneath the surface, and white sandy beaches

A peaceful Japanese garden with a koi pond, arched wooden bridge, cherry blossoms in full bloom, and lanterns softly glowing, tranquil and serene

Figure 4. Visual results of InfGen + SDXL at 3072×2048 resolution.

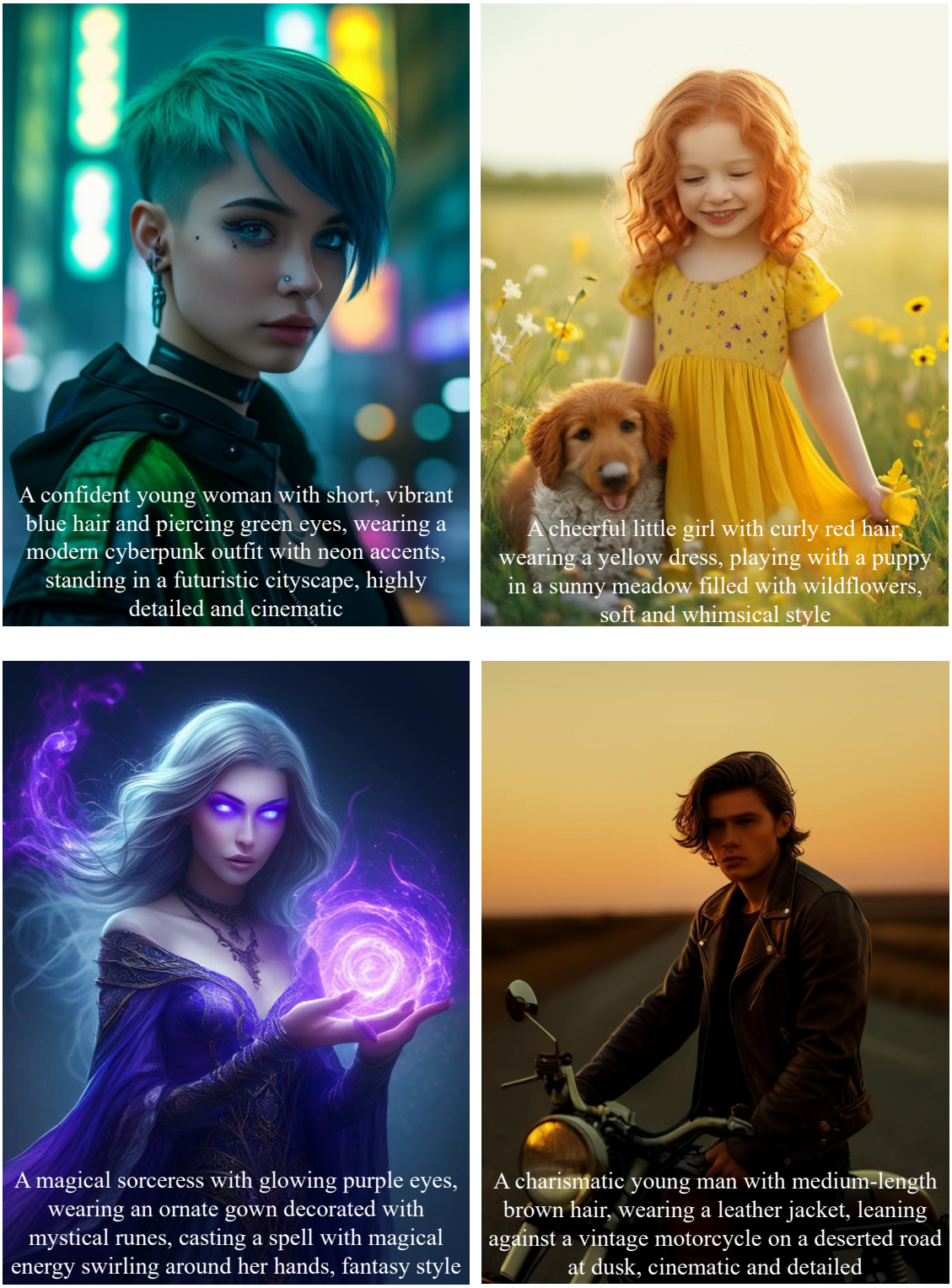


Figure 5. Visual results of InfGen + SDXL at 768×1024 resolution.

Unit 10 Kinetics Equilibrium And Thermodynamics

The Units Digital Stimulation

Hlluya Professional Sink Mixer Tap Kitchen Faucet The sink faucet basin wash basin full copper single hole hot and cold brushed nickel waterfall faucet Mixer Taps

Howard Safe & Lock Co

HAAS ST-30 with C axis & Milling 2010

Palenque

Cadillac XT4 Abogado Ley Limon

Somewhere (Saku) – LIKE A GIRL, LIKE A BOY [AAC 320 / WEB] [2018.12.23]

Figure 6. Visual results of InfGen + SDXL at 2800×1664 resolution.

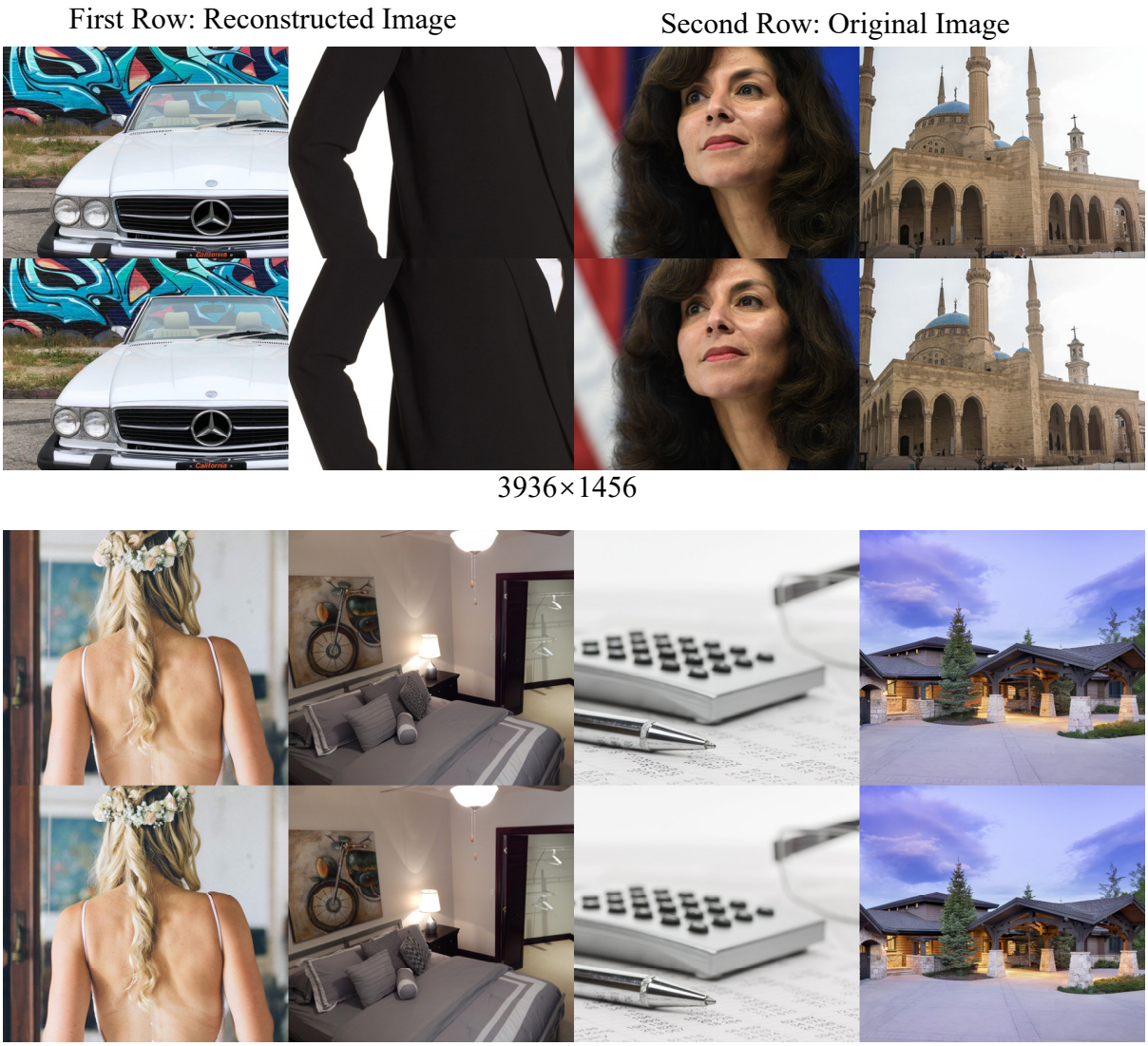


Figure 7. The reconstruction performance of **InfGen** on LAION dataset at multiple resolutions.

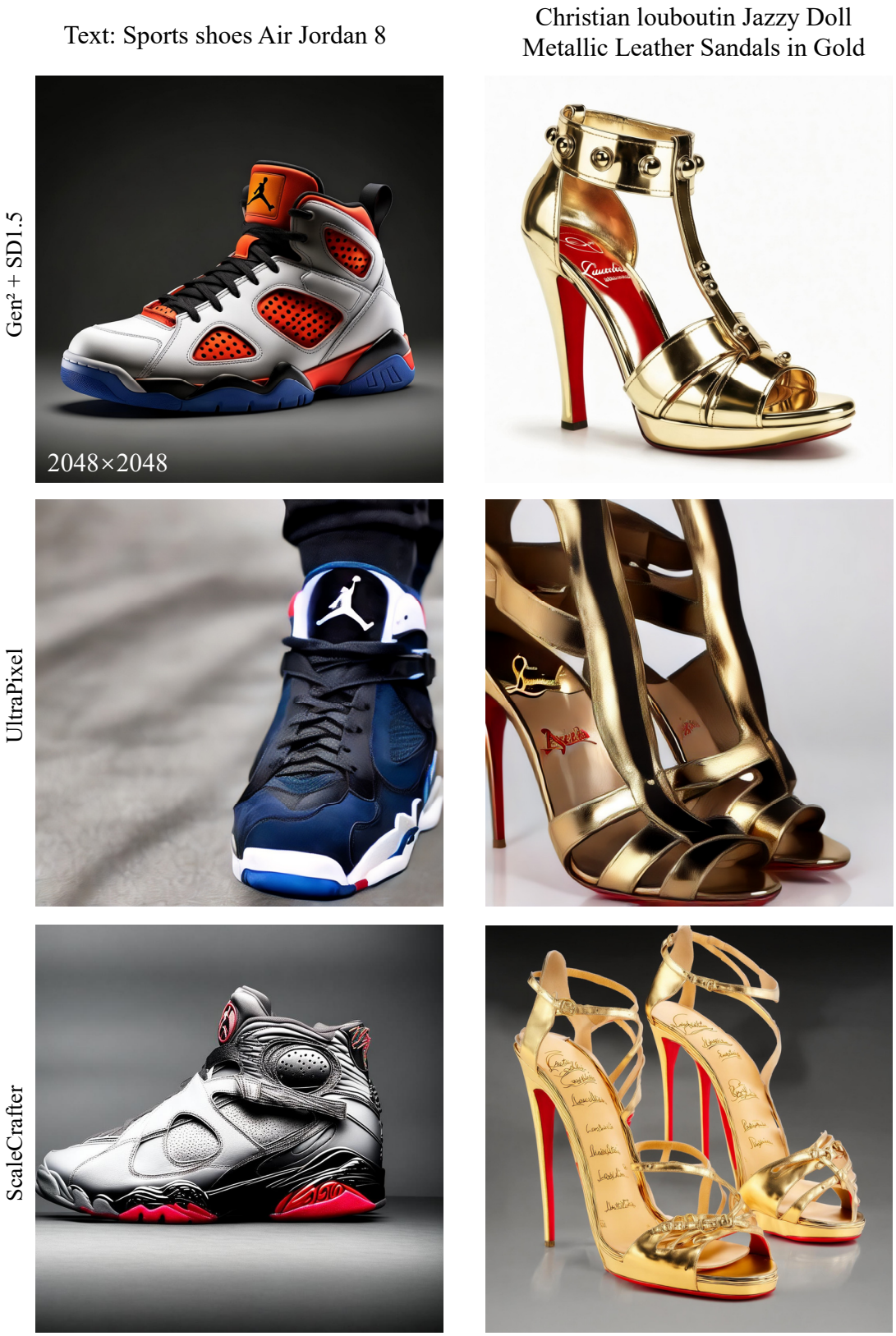


Figure 8. Visual comparison with other high-resolutio image generation methods (UltraPixel [10] and ScaleCrafter [2]). Images are at the 2048×2048 resolution.

Text: acheter tonnelle 3x3 m avec rideaux moustiquaire pas cher. Black Bedroom Furniture Sets. Home Design Ideas

Gen² + SDXL

UltraPixel

Figure 9. Visual comparison with UltraPixel [10]. Images are at the 3072×3072 resolution. .

Figure 10. Comparison of generated images from fixed latent sizes (i.e., 32×32) decoded to higher resolutions using our **InfGen**+DiT-XL/2 model versus interpolation-based upscaling. Zoom in for a better view.

Figure 11. Comparison of generated images from fixed latent sizes (i.e., 32×32) decoded to higher resolutions using our **InfGen**+FiTv1 model versus interpolation-based upscaling. Zoom in for a better view.

Figure 12. Comparison of generated images from fixed latent sizes (i.e., 32 × 32) decoded to higher resolutions using our **InfGen**+SiT model versus interpolation-based upscaling. Zoom in for a better view.

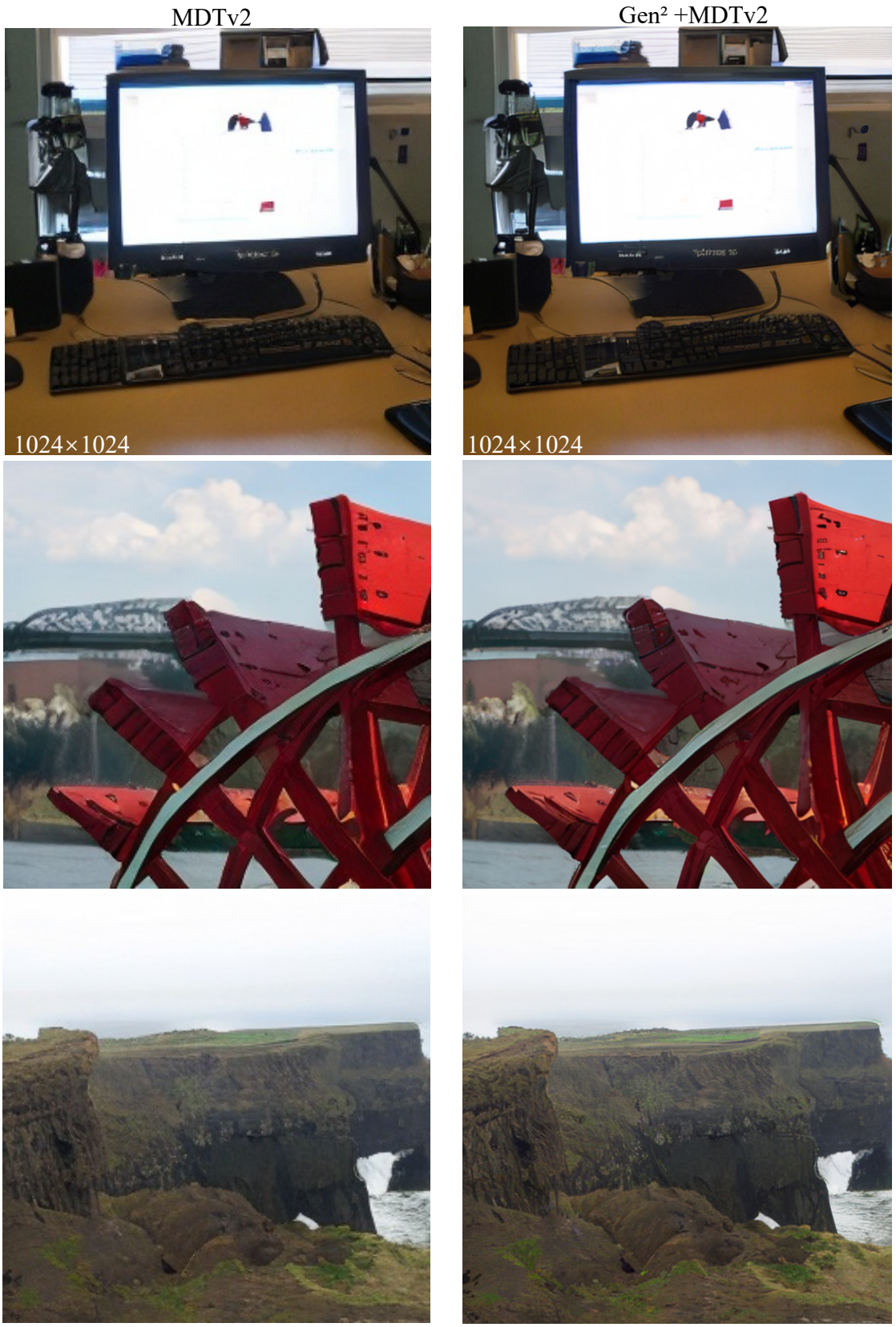


Figure 13. Comparison of generated images from fixed latent sizes (i.e., 32×32) decoded to higher resolutions using our **InfGen**+MDTv2 model versus interpolation-based upscaling. Zoom in for a better view.

References

[1] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987. 1

[2] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 11

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *In International Conference on Learning Representations*, 2014. 1

[5] Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. *ICML*, 2024. 2

[6] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024. 2

[7] OpenAI. Dall-e 3. https://openai.com/dall-e-3, 2023. Accessed: 2024-11-21. 2

[8] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[9] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[10] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. In *NeurIPs*, 2024. 11, 12

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[12] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. In *Transactions on Machine Learning Research (TMLR)*, 2024. 2