SANA: EFFICIENT HIGH-RESOLUTION IMAGE SYNTHESIS WITH LINEAR DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

1 FULL RELATED WORK

Efficient Diffusion Transformers. The introduction of Diffusion Transformers (DiT) (Peebles & Xie, 2023) marked a significant shift in image generation models, replacing the traditional U-Net architecture with a transformer-based approach. This innovation paved the way for more efficient and scalable diffusion models. Building upon DiT, PixArt-α (Chen et al., 2024b) extended the concept to text-to-image generation, demonstrating the versatility of transformer-based diffusion models. Stable Diffusion 3 (SD3) (Esser et al., 2024) further advanced the field by proposing the Multi-modal Diffusion Transformer (MM-DiT), which effectively integrates text and image modalities. More recently, Flux (Labs, 2024) showcased the potential of DiT architectures in high-resolution image generation by scaling up to 12B parameters. In addition, earlier works like CAN (Cai et al., 2024) explored linear attention mechanisms in class-condition image generation.

Text Encoders in Image Generation. The evolution of text encoders in image generation models has significantly impacted the field's progress. Initially, Latent Diffusion Models (LDM) (Rombach et al., 2022) adopted OpenAI's CLIP as the text encoder, leveraging its pre-trained visual-linguistic representations. A paradigm shift occurred with the introduction of Imagen (Saharia et al., 2022), which employed the T5-XXL language model as its text encoder, demonstrating superior text understanding and generation capabilities. Subsequently, eDiff-I (Balaji et al., 2022) proposed a hybrid approach, ensemble T5-XXL and CLIP encoders to combine their respective strengths in language comprehension and visual-textual alignment. Recent advancements, such as Playground v3 (Li et al., 2024a), have explored the use of decoder-only Large Language Models (LLMs) as text encoders, potentially offering more nuanced text understanding and generation. This trend towards more sophisticated text encoders reflects the ongoing pursuit of improved text-to-image alignment and generation quality in the field.

On Device Deployment. Several studies have explored post-training quantization (PTQ) techniques to optimize diffusion model inference for edge devices. Research in this area has focused on calibration objectives and data acquisition methods. BRECQ (Li et al., 2021) incorporates Fisher information into the objective function. ZeroQ (Cai et al., 2020) uses distillation to generate proxy input images for PTQ. SQuant (Guo et al., 2022) employs random samples with objectives based on Hessian spectrum sensitivity. Recent work such as Q-Diffusion (Li et al., 2023) has achieved high-quality generation using only 4-bit weights. In our work, we choose W8A8 to reduce peak memory usage.

2 MORE IMPLEMENTATION DETAILS

Rectified-Flow vs. DDPM. In our theoretical analysis, we investigate the reasons behind the fast convergence of flow-matching methods, demonstrating that both 1st flow-matching and EDM models rely on similar formulations. Unlike DDPMs, which use noise prediction, flow-matching and EDM focus on data or velocity prediction, resulting in improved performance and faster convergence. This shift from noise prediction to data prediction is particularly critical at $t = T$, where noise prediction tends to be unstable and leads to cumulative errors. As noted by Balaji et al. (2022), attention activation near $t = T$ grow stronger, highlighting the necessity of accurate predictions at this key moment in the sampling process.

As discussed in Lu (2023), the behavior of diffusion models near $t = T$ reveals that when t $\approx$ T, the data distribution resembles noise, and noise prediction approaches randomness. The challenge arises because the errors made at $t = T$ propagate through all subsequent sampling steps, making it

crucial for the sampler to be particularly precise near this time step. Based on Tweedie's formula, the gradient of the log density at time t, $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$, is approximated by:

$$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) = -\frac{\mathbf{x}_t - \alpha_t \mathbb{E}_{q_{0t}(\mathbf{x}_0|\mathbf{x}_t)}[\mathbf{x}_0]}{\sigma_t^2}. \tag{1}$$

When $t \approx T$, $\mathbf{x}_0$ and $\mathbf{x}_t$ become conditionally independent, leading to $q_{0t}(\mathbf{x}_0 \mid \mathbf{x}_t) \approx q_0(\mathbf{x}_0)$. Consequently, the noise prediction model's optimal solution becomes:

$$\epsilon_\theta(\mathbf{x}_t, t) \approx -\sigma_t \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) \approx \frac{\mathbf{x}_t - \alpha_t \mathbb{E}_{q_0(\mathbf{x}_0)}[\mathbf{x}_0]}{\sigma_t}. \tag{2}$$

Since $\mathbb{E}_{q_0(\mathbf{x}_0)}[\mathbf{x}_0]$ is independent of $\mathbf{x}_t$, the noise prediction model simplifies to a linear function of $\mathbf{x}_t$. However, as discussed in Section 5.2.1, this additional linearity can result in more accumulated errors during sampling, explaining why the original DPM-Solver struggles with guided sampling in such cases.

To address this issue and improve stability, DPM-Solver (Lu et al., 2022a) proposes modifying the noise prediction model to a more stable parameterization. By subtracting all linear terms inspired by equation 2, the remaining term is proportional to $\mathbb{E}_{q_0(\mathbf{x}_0)}[\mathbf{x}_0]$, corresponding to the data prediction model. Specifically, when $t \approx T$, the data prediction model approximates a constant:

$$\mathbf{x}_\theta(\mathbf{x}_t, t) \approx \frac{\mathbf{x}_t + \sigma_t^2 \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)}{\alpha_t} \approx \mathbb{E}_{q_0(\mathbf{x}_0)}[\mathbf{x}_0]. \tag{3}$$

Thus, for $t \approx T$, the data prediction model becomes approximately constant, and the discretization error for integrating this constant is significantly smaller than for the linear noise prediction model. This insight guides our development of an improved Flow-DPM-Solver based on DPM-Solver++ (Lu et al., 2022b), which adapts a velocity prediction model SANA to a data prediction one, enhancing performance for guided sampling.

Flow-based DPM-Solver Algorithm. We present the rectified flow-based DPM-Solver sampling process in Algorithm 1. This modified algorithm incorporates several key changes: hyper-parameter and time-step transformations, as well as model output transformations. These adjustments are highlighted in blue to differentiate them from the original solver.

In addition to improvements in FID and CLIP-Score, which are shown in Figure 8 of the main paper, our Flow-DPM-Solver also demonstrates superior convergence speed and stability compared to the Flow-Euler sampler. As illustrated in Figure 1, Flow-DPM-Solver retains the strengths of the original DPM-Solver, converging in only 10-20 steps to produce stable, high-quality images. By comparison, the Flow-Euler sampler typically requires 30-50 steps to reach a stable result.

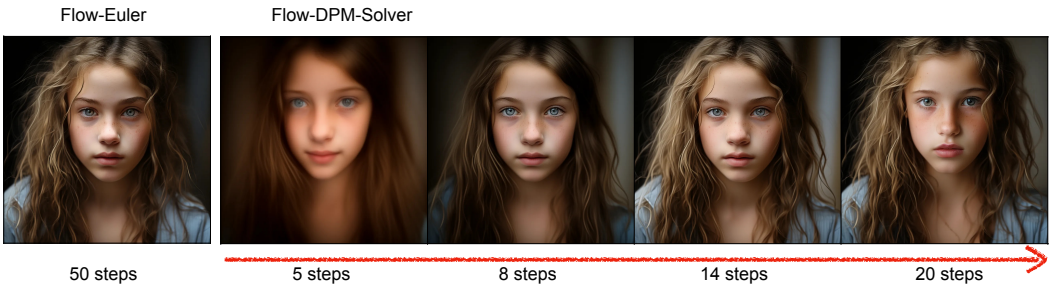

Figure 1: Visual comparison of Flow-Euler Sampler with 50 steps and Flow-DPM-Solver with 5/8/14/20 steps.

Multi-Caption Auto-labelling Pipeline. In Figure 2, we present the results of our CLIP-Score-based multi-caption auto-labelling pipeline, where each image is paired with its original prompt and 4 captions generated by different powerful VLMs. These captions complement each other, enhancing semantic alignment through their variations.

Triton Acceleration Training/Inference Detail. This section describes how to accelerate the training inference with kernel fusion using Triton. Specifically, for the forward pass, the ReLU activation

Algorithm 1 Flow-DPM-Solver (Modified from DPM-Solver++)

Require: initial value x_T, time steps $\{t_i\}_{i=0}^{M}$, data prediction model x_θ, velocity prediction model v_θ, time-step shift factor s

Denote $h_i := \lambda_{t_i} - \lambda_{t_{i-1}}$ for $i = 1, \dots, M$

$\tilde{\sigma}_{t_i} = \frac{s \cdot \sigma_{t_i}}{1+(s-1)\cdot\sigma_{t_i}}$, $\alpha_{t_i} = 1 - \tilde{\sigma}_{t_i}$ ▷ Hyper-parameter and Time-step transformation

$x_\theta(\tilde{x}_{t_i}, t_i) = \tilde{x}_{t_i} - \tilde{\sigma}_{t_i} v_\theta(\tilde{x}_{t_i}, t_i)$ ▷ Model output transformation

$\tilde{x}_{t_0} \leftarrow x_T$. Initialize an empty buffer Q.

$Q_{\text{buffer}} \leftarrow x_\theta(\tilde{x}_{t_0}, t_0)$

$\tilde{x}_{t_1} \leftarrow \frac{\tilde{\sigma}_{t_1}}{\tilde{\sigma}_{t_0}} \tilde{x}_{t_0} - \alpha_{t_1} \left(e^{-h_1} - 1\right) x_\theta(\tilde{x}_{t_0}, t_0)$

$Q_{\text{buffer}} \leftarrow x_\theta(\tilde{x}_{t_1}, t_1)$

for $i = 2$ to M **do**

  $r_i \leftarrow \frac{h_{i-1}}{h_i}$

  $D_i \leftarrow \left(1 + \frac{1}{2r_i}\right) x_\theta(\tilde{x}_{t_{i-1}}, t_{i-1}) - \frac{1}{2r_i} x_\theta(\tilde{x}_{t_{i-2}}, t_{i-2})$

  $\tilde{x}_{t_i} \leftarrow \frac{\tilde{\sigma}_{t_i}}{\tilde{\sigma}_{t_{i-1}}} \tilde{x}_{t_{i-1}} - \alpha_{t_i} \left(e^{-h_i} - 1\right) D_i$

  if $i < M$ **then**

    $Q_{\text{buffer}} \leftarrow x_\theta(\tilde{x}_{t_i}, t_i)$

  end if

end for

return $\tilde{x}_{t_M}$

are fused to the end of QKV projection, the precision conversions and padding operations are fused to the start of KV and QKV multiplications, and the divisions are fused to the end of QKV multiplication. For the backward pass, the divisions are fused to the end of the output projection, and the precision conversions and ReLU activation are fused to the end of KV and QKV multiplications.

3 More Results

Ablation on SANA Blocks. Table 3 describes how different block designs affect performance. Directly switching from DiT's self-attention to linear attention will result in FID and Clip Score performance loss, but adding Mix-FFN can compensate for the performance loss. Adding triton kernel fusion can speed up training/inference without negatively impacting performance.

Complex Human Instruction Analysis. To observe the effectiveness of CHI, we input the user prompt with/without CHI into Gemma-2. We believe a strong positive correlation exists between LLM output and text embedding quality. As shown in Figure 3, without CHI, although Gemma-2 can understand the meaning of the input, the output is conversational and does not focus on understanding the user prompt itself. After adding CHI, Gemma-2's output is better focused on understanding and enhancing the details of the user prompt.

Detailed Results on DPG-Bench, GenEval and ImageReward. As an extension of Table. 7 in the main paper, we show all the metric details of GenEval, DPG-Bench and ImageReward for reference in Table 1 and Table 2 respectively.

Finding: Zero-shot Language Transfer Ability. As shown in Figure 4, we were surprised to find that by using Gemma-2 as the text encoder and Chinese/Emoji expressions as text prompts; our SANA can also understand and generate corresponding images. Note that we filter out all prompts other than English during training, so the zero-shot generation capability of Chinese/Emoji is brought by Gemma-2.

Detailed Speed Comparison of Text Encoder. In Table 4, we present a comparison of the latency and parameters for various T5 models alongside the Gemma models. Notably, the Gemma-2B model exhibits a similar latency to T5-large while significantly increasing the model size. This enhancement in model size is a key factor in achieving improved capabilities with greater efficiency.

Detailed Speed Comparison of Diffusion Model. In Table 5, we compare the throughput and latency of the mainstream DiT-based text-to-image method and our model in detail and test them at resolutions of 512, 1024, 2048, and 4096, respectively. Our SANA is far ahead of other methods at

Table 1: **Comparison of SOTA methods on GenEval with details.** The table includes different metrics such as overall performance, single object, two objects, counting, colors, position, and color attribution.

Model	Params (B)	Overall ↑	Objects		Counting	Colors	Position	Color Attribution
			Single	Two				
512 ×512 resolution								
PixArt-α	0.6	**0.48**	0.98	0.50	0.44	0.80	0.08	0.07
PixArt-Σ	0.6	**0.52**	0.98	0.59	0.50	0.80	0.10	0.15
SANA-0.6B (Ours)	0.6	**0.64**	0.99	0.71	0.63	0.91	0.16	0.42
SANA-1.6B (Ours)	0.6	**0.66**	0.99	0.79	0.63	0.88	0.18	0.47
1024 ×1024 resolution								
LUMINA-Next (Zhuo et al., 2024)	2.0	**0.46**	0.92	0.46	0.48	0.70	0.09	0.13
SDXL (Podell et al., 2023)	2.6	**0.55**	0.98	0.74	0.39	0.85	0.15	0.23
PlayGroundv2.5 (Li et al., 2024a)	2.6	**0.56**	0.98	0.77	0.52	0.84	0.11	0.17
Hunyuan-DiT (Li et al., 2024b)	1.5	**0.63**	0.97	0.77	0.71	0.88	0.13	0.30
DALLE3 (OpenAI, 2023)	-	**0.67**	0.96	0.87	0.47	0.83	0.43	0.45
SD3-medium (Esser et al., 2024)	2.0	**0.62**	0.98	0.74	0.63	0.67	0.34	0.36
FLUX-dev (Labs, 2024)	12.0	**0.67**	0.99	0.81	0.79	0.74	0.20	0.47
FLUX-schnell (Labs, 2024)	12.0	**0.71**	0.99	0.92	0.73	0.78	0.28	0.54
SANA-0.6B (Ours)	0.6	**0.64**	0.99	0.76	0.64	0.88	0.18	0.39
SANA-1.6B (Ours)	1.6	**0.66**	0.99	0.77	0.62	0.88	0.21	0.47

Table 2: **Comparison of SOTA methods on DPG-Bench and ImageReward with details.** The table includes different metrics such as overall performance, entity, attribute, relation, and other categories.

Model	Params (B)	Overall ↑	Global	Entity	Attribute	Relation	Other	ImageReward ↑
512 ×512 resolution								
PixArt-α (Chen et al., 2024b)	0.6	**71.6**	81.7	80.1	80.4	81.7	76.5	0.92
PixArt-Σ (Chen et al., 2024a)	0.6	**79.5**	87.5	87.1	86.5	84.0	86.1	0.97
SANA-0.6B (Ours)	0.6	**84.3**	82.6	90.0	88.6	90.1	91.9	0.93
SANA-1.6B (Ours)	0.6	**85.5**	90.3	91.2	89.0	88.9	92.0	1.04
1024 ×1024 resolution								
LUMINA-Next (Zhuo et al., 2024)	2.0	**74.6**	82.8	88.7	86.4	80.5	81.8	-
SDXL (Podell et al., 2023)	2.6	**74.7**	83.3	82.4	80.9	86.8	80.4	0.69
PlayGroundv2.5 (Li et al., 2024a)	2.6	**75.5**	83.1	82.6	81.2	84.1	83.5	1.09
Hunyuan-DiT (Li et al., 2024b)	1.5	**78.9**	84.6	80.6	88.0	74.4	86.4	0.92
PixArt-Σ (Chen et al., 2024a)	0.6	**80.5**	86.9	82.9	88.9	86.6	87.7	0.87
DALLE3 (OpenAI, 2023)	-	**83.5**	91.0	89.6	88.4	90.6	89.8	-
SD3-medium (Esser et al., 2024)	2.0	**84.1**	87.9	91.0	88.8	80.7	88.7	0.86
FLUX-dev (Labs, 2024)	12.0	**84.0**	82.1	89.5	88.7	91.1	89.4	-
FLUX-schnell (Labs, 2024)	12.0	**84.8**	91.2	91.3	89.7	86.5	87.0	0.91
SANA-0.6B (Ours)	0.6	**83.6**	83.0	89.5	89.3	90.1	90.2	0.97
SANA-1.6B (Ours)	1.6	**84.8**	86.0	91.5	88.9	91.9	90.7	0.99

Caption_original (ClipScore: 25.67)

top view the written " HAPPY VALENTINE " on a tart chocolate cake, black wood background

Caption_VILA-13B (ClipScore: 26.33)

The image captures a delightful scene of a Valentine's Day celebration. At the center of the image is a round chocolate cake, rich and inviting. The cake is adorned with a red heart on top, symbolizing love and affection. The words "Happy Valentine's" are elegantly written in red icing on the cake, adding to the festive atmosphere. The cake is placed on a wooden table, which provides a rustic charm to the setting. Scattered around the cake are red heart-shaped confetti, further enhancing the Valentine's theme. In the background, there's a hint of a bouquet of red roses, a classic symbol of love and romance. The roses are slightly blurred, drawing focus to the cake in the foreground. Overall, the image beautifully encapsulates the spirit of Valentine's Day, with its focus on love, affection, and celebration.

Caption_VILA-3B (ClipScore: 25.46)

The image captures a delightful scene of a Valentine's Day celebration. At the center of the frame is a round chocolate cake, its surface adorned with a glossy chocolate ganache. The cake is further embellished with a scattering of chocolate shavings and a heart-shaped topper, adding to its festive appeal. The cake is placed on a wooden surface, which provides a rustic contrast to the rich, dark color of the chocolate. The background is dark, allowing the cake to stand out prominently. Adding to the Valentine's Day theme, red heart-shaped sprinkles are scattered around the cake, their vibrant color popping against the dark backdrop. The cake is positioned slightly off-center, drawing the viewer's attention to it. The overall composition of the image suggests a carefully arranged celebration, with the cake as the star of the show. The image does not contain any discernible text. The relative positions of the objects and their detailed descriptions provide a comprehensive understanding of the image content.

Caption_InternVL2-26B (ClipScore: 25.57)

The image features a beautifully decorated chocolate cake placed on a dark wooden surface. The cake is richly covered in a layer of dark chocolate ganache, giving it a smooth and glossy appearance. The top of the cake is adorned with piped chocolate swirls and heart- shaped decorations, adding a touch of elegance and romance. At the center of the cake, the words "HAPPY VALENTINE" are written in bold, red icing, creating a striking contrast against the dark chocolate background. Surrounding the cake, there are additional heart- shaped decorations in red, some of which are placed on the wooden surface, while others are scattered around the cake. The overall setting is warm and inviting, with a rustic wooden background that complements the rich, indulgent nature of the chocolate cake. The use of red hearts and the "HAPPY VALENTINE" message clearly indicate that this cake is intended for a Valentine's Day celebration

Caption_InternVL2-8B (ClipScore: 26.37)

The image features a beautifully decorated chocolate cake with a heartfelt Valentine's Day message. The cake is round and sits on a dark, rustic wooden surface. The cake is covered in a rich, dark chocolate ganache, and it is adorned with a generous amount of chocolate shavings and crushed nuts, giving it a luxurious and indulgent appearance. At the top of the cake, the words "HAPPY VALENTINE" are written in bold, red icing, creating a striking contrast against the dark chocolate. The icing is piped in a slightly uneven, handwritten style, adding a personal touch to the message. Surrounding the cake are several red heart-shaped decorations, some of which are whole and others that are broken into smaller pieces, scattered around the cake. These hearts add to the romantic and festive theme of the cake. The cake is placed on a dark, round cake board, which complements the rustic wooden background. The overall presentation is elegant and festive, making it perfect for a Valentine's Day celebration

Figure 2: Illustration of re-caption of an image with multiple VLMs.

different resolutions. As the resolution increases, the efficiency advantage of our SANA becomes more significant.

More Visualization Images. As shown in Figure 5, we can see that 4K images can directly generate more details than 1k images. In Figure 6, we show more images generated by our model with various prompts. We also provide a video mp4 demo in the supplementary material (zip file) to show SANA is deployed on a laptop.

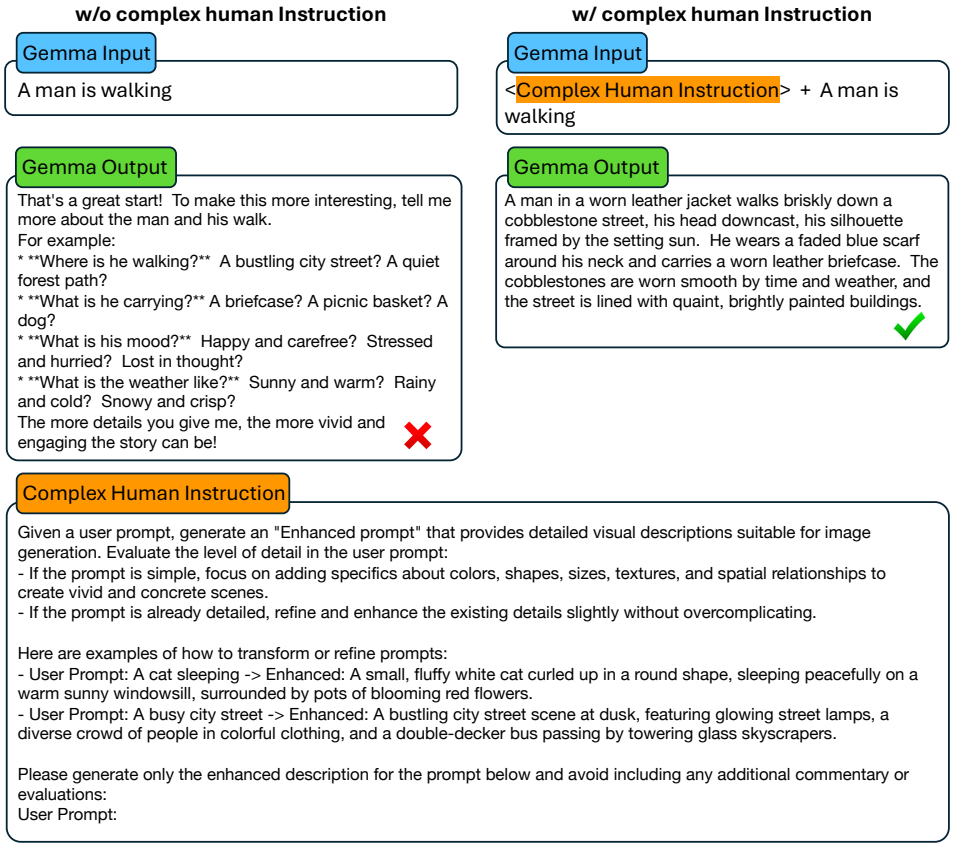

Figure 3: Illustration of Gemma-2's output with/without complex human instruction, and the full prompt of our complex human instruction.

Figure 4: Visualization of zero-shot language transfer ability. Our SANA only has English prompts during training but can understand Chinese/Emoji during inference. This benefits from the generalization brought by the powerful pre-training of Gemma-2.

Table 3: **Performance of SANA block design space.** We train all the models with the same training setting with 52K iterations.

Blocks	AE	Res.	FID ↓	CLIP ↑
FullAttn & FFN	F8C4P2	256	18.7	24.9
+ Linear	F8C4P2	256	21.5	23.3
+ MixFFN	F8C4P2	256	18.9	24.8
+ Kernel Fusion	F8C4P2	256	18.8	24.8
Linear+GLUMBConv2.5	F32C32P1	512	6.4	27.4
+ Kernel Fusion	F32C32P1	512	6.4	27.4

Table 4: Comparison of various T5 models and Gemma models based on speed and parameters. The sequence length (Seq Len) is the number of text tokens.

Text Encoder	Batch Size	Seq Len	Latency (s)	Params (M)
T5-XXL	32	300	1.6	4762
T5-XL			0.5	1224
T5-large			0.2	341
T5-base			0.1	110
T5-small			0.0	35
Gemma-2b			0.2	2506
Gemma-2-2b			0.3	2614

Table 5: Comparison of throughput and latency under different resolutions. All models tested on an A100 GPU with FP16 precision.

Methods	Speedup	Throughput(/s)	Latency(ms)	Methods	Speedup	Throughput(/s)	Latency(ms)
512×512 Resolution				1024×1024 Resolution			
SD3	7.6x	1.14	1.4	SD3	7.0x	0.28	4.4
FLUX-schnell	10.5x	1.58	0.7	FLUX-schnell	12.5x	0.50	2.1
FLUX-dev	1.0x	0.15	7.9	FLUX-dev	1.0x	0.04	23
PixArt-Σ	10.3x	1.54	1.2	PixArt-Σ	10.0x	0.40	2.7
HunyuanDiT	1.3x	0.20	5.1	HunyuanDiT	1.2x	0.05	18
SANA-0.6B	44.5x	6.67	0.8	**SANA-0.6B**	43.0x	1.72	0.9
SANA-1.6B	25.6x	3.84	0.6	**SANA-1.6B**	25.2x	1.01	1.2
2048×2048 Resolution				4096×4096 Resolution			
SD3	5.0x	0.04	22	SD3	4.0x	0.004	230
FLUX-schnell	11.2x	0.09	10.5	FLUX-schnell	13.0x	0.013	76
FLUX-dev	1.0x	0.008	117	FLUX-dev	1.0x	0.001	1023
PixArt-Σ	7.5x	0.06	18.1	PixArt-Σ	5.0x	0.005	186
HunyuanDiT	1.2x	0.01	96	HunyuanDiT	1.0x	0.001	861
SANA-0.6B	53.8x	0.43	2.5	**SANA-0.6B**	104.0x	0.104	9.6
SANA-1.6B	31.2x	0.25	4.1	**SANA-1.6B**	66.0x	0.066	5.9

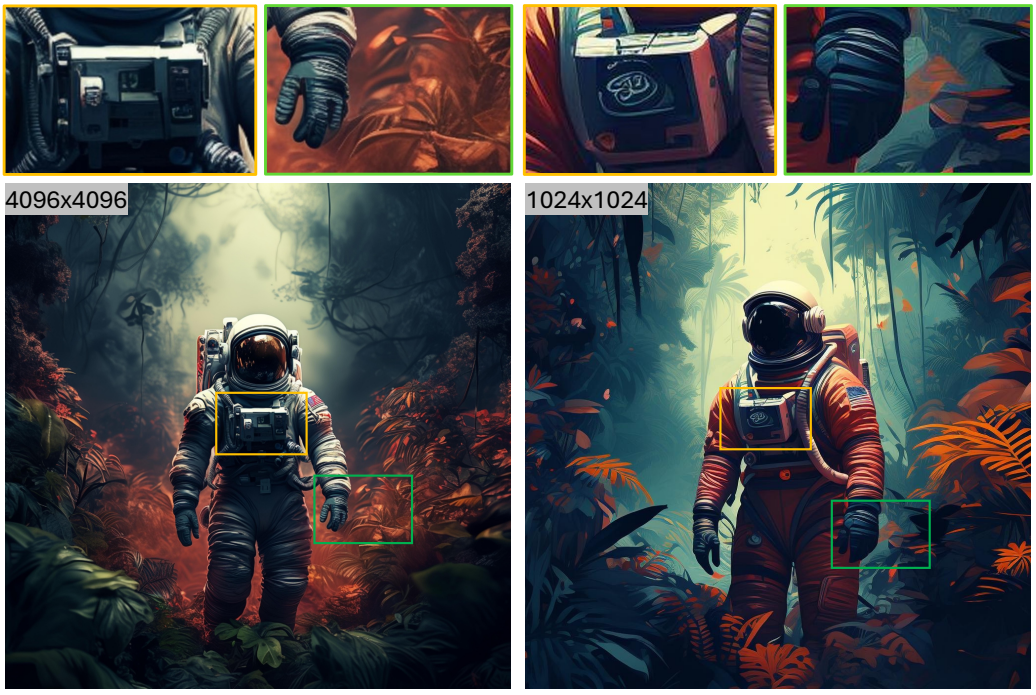

Figure 5: Comparison of 4K and 1K resolution images. We can see that the 4K image contains more details.

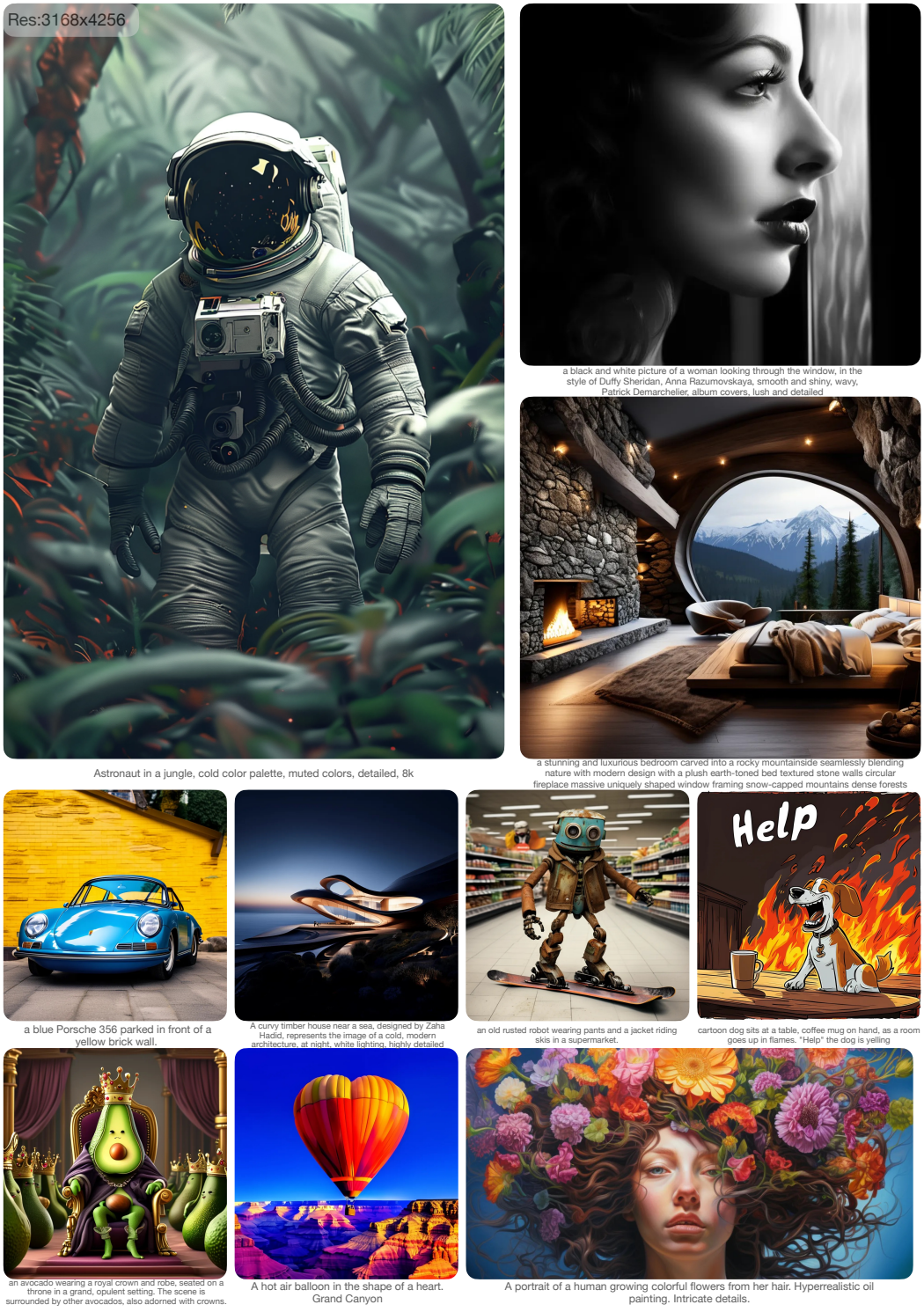

Figure 6: More samples generated from SANA.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Han Cai, Muyang Li, Qinsheng Zhang, Ming-Yu Liu, and Song Han. Condition-aware neural network for controlled image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7194–7203, 2024.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13166–13175, 2020.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024a.

Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024b.

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

Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. Squant: On-the-fly data-free quantization via diagonal hessian approximation. *ArXiv*, abs/2202.07471, 2022.

Black Forest Labs. Flux, 2024. URL https://blackforestlabs.ai/.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

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

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. {BRECQ}: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024b.

Cheng Lu. Research on reversible generative models and their efficient algorithms, 2023. URL https://luchengthu.github.io/files/chenglu_dissertation.pdf.

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

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

OpenAI. Dalle-3, 2023. URL https://openai.com/dall-e-3.

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

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.

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*, pp. 10684–10695, 2022.

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

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.