SnapGen: Taming High-Resolution Text-to-Image Models for Mobile Devices with Efficient Architectures and Training

Supplementary Material

A. Author Contribution Statement

- Jierun Chen developed the efficient UNet and AE decoder, enabling 1K resolution image generation on mobile devices *for the first time*. On ImageNet class-conditional generation, the UNet achieves an FID score on par with the recent SiT-X model while reducing parameters by 45% and compute resources by 68%. The tiny decoder is $36\times$ *smaller* and $54\times$ *faster* than conventional decoders (*e.g.*, those from SDXL and SD3) for high-resolution mobile generation. He also initiated the early T2I diffusion training and contributed to the on-device deployment.

- Dongting Hu designed and implemented the training pipeline, integrating various text encoders and incorporating a flow-matching objective to enable knowledge distillation from scalable DiT-based models (SD3.5). He prepared high-quality training data and trained the T2I base model, achieving an initial GenEval score of *0.61*. He also built the distillation pipeline and contributed to multi-level knowledge distillation, significantly enhancing the model's GenEval score to *0.66* and improving generation quality based on human evaluations. His work on step distillation further enabled efficient high-quality generation, achieving a GenEval score of *0.63* with 8-step generation and *0.61* with 4-step generation. Additionally, he managed latent decoder training, ensuring close reconstruction quality to SD3 decoder, and facilitated on-device deployment. He developed the mobile app using the Core ML Diffusers framework, which achieved 1K resolution image generation on-device in approximately *1.4* seconds, as demonstrated in the demo.

- Xijie Huang proposed the multi-level knowledge distillation scheme to improve the generation quality of our model, achieving comparable performance to the DiT-based teacher (SD3.5) across various quantitative benchmarks (*e.g.*, boosting GenEval performance from 0.61 to 0.66) and human evaluation. Specifically, he analyzed the scale difference between the distillation loss and task loss across different timesteps and proposed a timestep-aware scaling operation. He also worked on adversarial step distillation to enable efficient and effective 4/8-step generation, leading to optimal latency on mobile devices. He also conducted evaluations of our model across various benchmarks including DPG-Bench and ImageReward.

B. Demo on Mobile Devices

We present an on-device demo showcasing the capabilities of our efficient text-to-image model in generating high-resolution images (1024×1024 pixels) directly on mobile phones. The application is implemented based on the open-source Swift Core ML Diffusers framework[1]. Upon launching the application, users can input textual prompts and generate corresponding images by clicking the "Generate" button. A screenshot of the deployed application is shown in Fig. 1, and a more detailed demonstration can be found in the webpage.

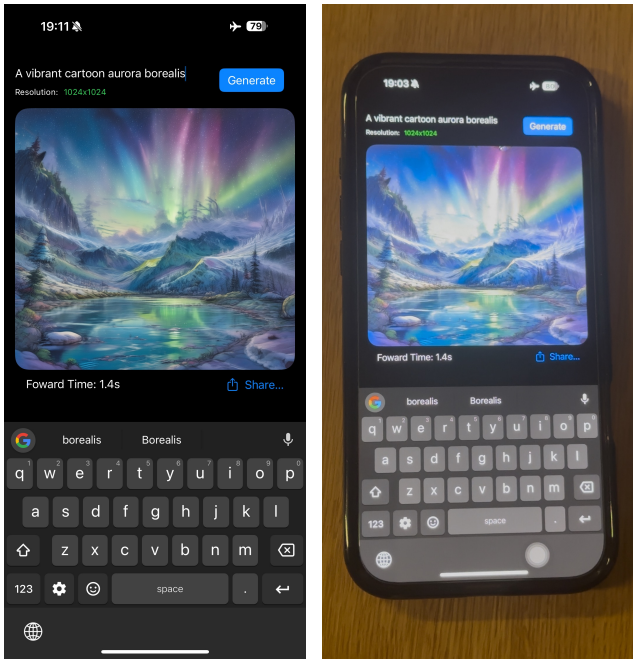

Figure 1. **Demo on iPhone 16 Pro-Max.** We report the forward time for each 4-step generation, excluding the model loading time.

C. Detailed Results on Benchmarks

We present detailed results for GenEval and DPG-Bench in Tab. 1 and Tab. 2, respectively. On GenEval, our model demonstrates exceptional performance in capturing color and positional attributes, with the counting subcategory showing significant improvement due to our proposed knowledge distillation (KD) scheme.

[1] https://github.com/huggingface/swift-coreml-diffusers

Table 1. **Detailed Results of GenEval Bench Comparisons.**

Model	Param	Overall ↑	Single Object	Two Objects	Counting	Colors	Position	Color Attribution
PixArt-α [4]	0.6B	0.48	0.98	0.50	0.44	0.80	0.08	0.07
PixArt-Σ [5]	0.6B	0.53	0.99	0.65	0.46	0.82	0.12	0.12
SD-1.5 [1]	0.9B	0.43	0.97	0.38	0.38	0.76	0.04	0.06
SD-2.1 [2]	0.9B	0.50	0.98	0.51	0.44	0.85	0.07	0.17
KOALA [15]	1.0B	0.52	0.99	0.65	0.45	0.80	0.10	0.15
MicroDiT [21]	1.2B	0.46	0.97	0.46	0.33	0.78	0.09	0.20
Sana [22]	1.6B	0.66	0.99	0.77	0.62	0.88	0.21	0.47
LUMINA-Next [24]	2.0B	0.46	0.92	0.46	0.48	0.70	0.09	0.13
SDXL [19]	2.6B	0.55	0.98	0.74	0.39	0.85	0.15	0.23
PlayGroundv2 [17]	2.6B	0.59	0.98	0.73	0.67	0.82	0.14	0.22
PlayGroundv2.5 [17]	2.6B	0.56	0.98	0.77	0.52	0.84	0.11	0.17
IF-XL [6]	5.5B	0.61	0.97	0.74	0.66	0.81	0.13	0.35
Ours w/o KD	0.38B	0.61	0.98	0.77	0.43	0.89	0.18	0.38
SnapGen (ours)	0.38B	0.66	1.00	0.84	0.60	0.88	0.18	0.45

Table 2. **Detailed Results of DPG-Bench Comparisons.**

Model	Param	Overall ↑	Global	Entity	Attribute	Relation	Other
PixArt-α [4]	0.6B	71.1	75.0	79.3	78.6	82.6	77.0
PixArt-Σ [5]	0.6B	80.5	86.9	82.9	88.9	86.6	87.7
SDv1.5 [1]	0.9B	63.2	74.6	74.2	75.4	73.5	67.8
SDv2.1 [2]	0.9B	64.2	72.7	72.8	75.8	82.2	76.5
KOALA [15]	1.0B	74.3	83.9	81.8	81.2	87.0	61.6
MicroDiT [21]	1.2B	75.1	80.2	82.0	80.8	86.0	60.8
Sana [22]	1.6B	84.8	86.0	91.5	88.9	91.9	90.7
LUMINA-Next [24]	2.0B	74.6	82.8	88.7	86.4	80.5	81.8
SDXL [19]	2.6B	74.7	83.3	82.4	80.9	86.8	80.4
PlayGroundv2[17]	2.6B	74.5	83.6	79.9	82.7	80.6	81.2
PlayGroundv2.5[17]	2.6B	75.5	83.1	82.6	81.2	84.1	83.5
IF-XL [6]	5.5B	75.6	77.7	81.2	83.3	81.8	82.9
Ours w/o KD	0.38B	76.3	77.8	83.7	84.3	86.7	77.4
SnapGen (ours)	0.38B	81.1	88.3	85.1	87.0	87.3	87.6

D. Reconstruction by VAE Decoders

We compare reconstruction results between the SD3 VAE decoder (49.55M parameters) and our tiny decoder (1.38M parameters) in Fig. 2. Despite being 36× smaller, our tiny decoder achieves competitively high visual quality across images with intricate textures, text, and human faces.

E. Mixtures of Text Encoders

We use CLIP-L and CLIP-G for mobile inference given their small sizes (123M, 302M) and low latency (4ms, 23ms on iPhone16). For cloud inference, we incorporate Gemma-2-2b to handle long prompts.

F. Visualization of Few-step Generation

In Fig. 3, we present qualitative comparisons of the 4- and 8-step T2I generation quality of our model, both with and without step distillation. The results demonstrate that the 4- and 8-step generations, following step distillation, not only significantly outperform the baseline model but also deliver quality comparable to the 28-step generation. Remarkably, the few-step generation captures finer details and mitigates the over-saturation issue commonly observed in the 28-step generation. Additionally, it exhibits superior prompt-following fidelity, making it a more efficient and effective approach for high-quality T2I generation.

Figure 2. **Comparisons of Decoder Reconstruction** between SD3 decoder and our tiny decoder. Zoom in for better viewing.

G. Additional T2I Comparison and Examples

We present additional qualitative visualizations comparing 1024×1024 generations across various T2I models in Figs. 4 and 5. Furthermore, we showcase additional T2I examples generated by our model in Figs. 6 and 7, with corresponding prompts detailed in Tab. 4. These results demonstrate the exceptional prompt adherence and realistic generation quality achieved by our model.

H. Ablation Study on Knowledge Distillation

To demonstrate how the proposed knowledge distillation scheme improves the T2I generation quality of our model, we provide additional ablation studies into different distillation loss terms and the timestep-aware scaling operations (t-scaling) in Tab. 3. As listed in the results, distillation at all levels consistently improves our model in both GenEval and ImageReward scores.

Table 3. **Abalation Study on KD Components.**

$\mathcal{L}_{\text{task}}$	$\mathcal{L}_{\text{kd}}$	$\mathcal{L}_{\text{featkd}}$	t-scaling	GenEval ↑	ImageReward ↑
✓				0.61	1.20
✓	✓			0.62	1.23
✓	✓	✓		0.64	1.26
✓	✓	✓	✓	0.66	1.32

I. Large-scale T2I Training Details

Our training is conducted across 8 nodes, each equipped with 8 NVIDIA A100 80G GPUs, resulting in a total of 64 GPUs. The training batch size per GPU is set to 128 for 512^2 resolution, 48 for 1024^2 without knowledge distillation, and 32 for 1024^2 with knowledge distillation. Gradient checkpointing is employed to accommodate larger batch sizes. For step distillation, we use Fully Sharded Data Parallel (FSDP) and set gradient accumulation steps to 4, achieving an effective batch size of 16 per GPU. We optimize the model using the AdamW optimizer with a weight decay of 0.01 and $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate remains constant during training and is scaled based on the batch size for the current training stage. For a total batch size of 1024, we use a learning rate of 5×10^{-5}. The data collection and filtering pipeline for the T2I training dataset follows the approach described by Kag et al. [12]. For unconditional diffusion guidance [9], we set the outputs of each of the three text encoders independently to zero with a probability of 46.4%, such that we roughly train an unconditional model in 10% of all steps. Additionally, we apply the Exponential Moving Average (EMA) with a decay rate of 0.999. We use CLIP-L and CLIP-G for mobile inference given their small sizes (123M, 302M) and low latency (4ms, 23ms on iPhone16). For cloud inference, we incorporate Gemma-2-2b to handle long prompts.

J. ImageNet-1K Class-conditional Generation

We provide the experimental settings when examining each design choice in developing our efficient UNet. We train and evaluate them on the ImageNet-1K class-conditional generation task at a resolution of 256×256. To incorporate class conditions, we use a text template of the form "a photo of <class name>", which can be seamlessly fused by the cross-attention layer. As the dataset provides multiple names per class, we select one at random during both training and inference to enrich text mappings. This text is encoded by a compact CLIP-ViT/L14 [20] text encoder. For latent diffusion, we convert input images into latent using an 8-channel AE. We pre-compute both the image latent and the text embeddings, which reduces the GPU memory and non-trivial computation time during training. We adopt DDPM [10] as our training objective, applying a linear noise scheduler over 1000 time steps. Models are trained for 120 epochs (and 1000 epochs for the final model) with a batch size of 1024. The AdamW optimizer is used with a learning rate of $3e$-4, weight decay of 0.01, and $(\beta_1, \beta_2) = (0.9, 0.99)$. For inference, we utilize the Heun [13] discrete scheduler with 30 sampling steps. We report the lowest FID score for each model variant, using classifier-free guidance (cfg) [8] within a scale range of $[1.3, 2.0]$. For the final model, we implement varied cfg [3, 14] in steps $[10, 30]$ with cfg scaling from 1.1 to 5.4.

K. Evaluation Metrics Details

GenEval [7] is an object-focused evaluation framework for T2I models based on object detection and color classification to verify the fine-grained object properties in the generated images. Concretely, 6 tasks with different difficulties are focused in GenEval: single object, two object, counting, colors, position, and attribute binding. The prompts in the GenEval benchmark are generated from task-specific templates filled with randomly sampled object names (from 80 MS COCO [18] class names), colors, numbers, and relative positions. There are a total of 553 prompts in GenEval and these prompts are usually concise (less than 20 tokens).

DPG-Bench [11] is a benchmark mainly focused on dense prompts that describe multiple objects characterized by various attributes and relationships. The average number of tokens calculated by the CLIP tokenizer is 83.91, significantly longer than previous benchmarks such as 12.65 for T2I-CompBench and 12.20 for PartiPrompts. There are a total of 4286 prompts, spanning five categories: entity, global, attribute, relation, and other. 4 images are generated for each prompt and mPLUG-large [16] is used as the adjudicator to evaluate the generated images according to the designated questions.

Image Reward [23] is a zero-shot metric to encode the human preference on the text-to-image results. The model uses BLIP as the backbone and an MLP to obtain a scalar for preference comparison. After training on human-annotated preference data with ratings on alignment, fidelity, and harmlessness, the reward model aligns with human preference. Different from GenEval and DPG-Bench, we observe that Image Reward prefers images with detailed textures and backgrounds with rich color patterns.

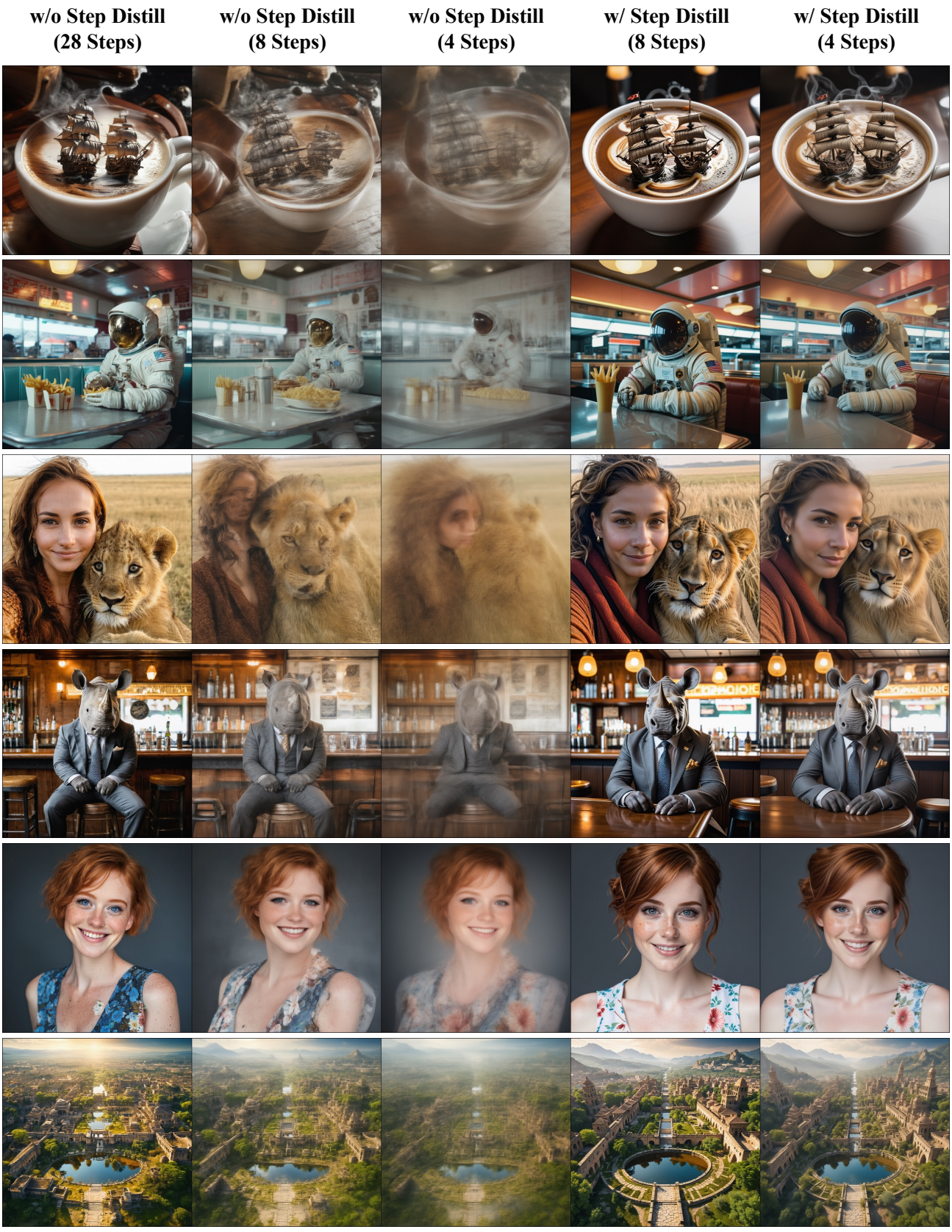

Figure 3. **Few-step generation qualitative comparison at** 1024^2 **resolution.** The prompts used in these examples are from PixArt [4].

Figure 4. **Additional Qualitative Comparison.** Our model demonstrates competitive visual quality and superior prompt-following ability. Input text prompts are shown above each image grid; all images are generated at 1024^2 resolution. Zoom in for details.

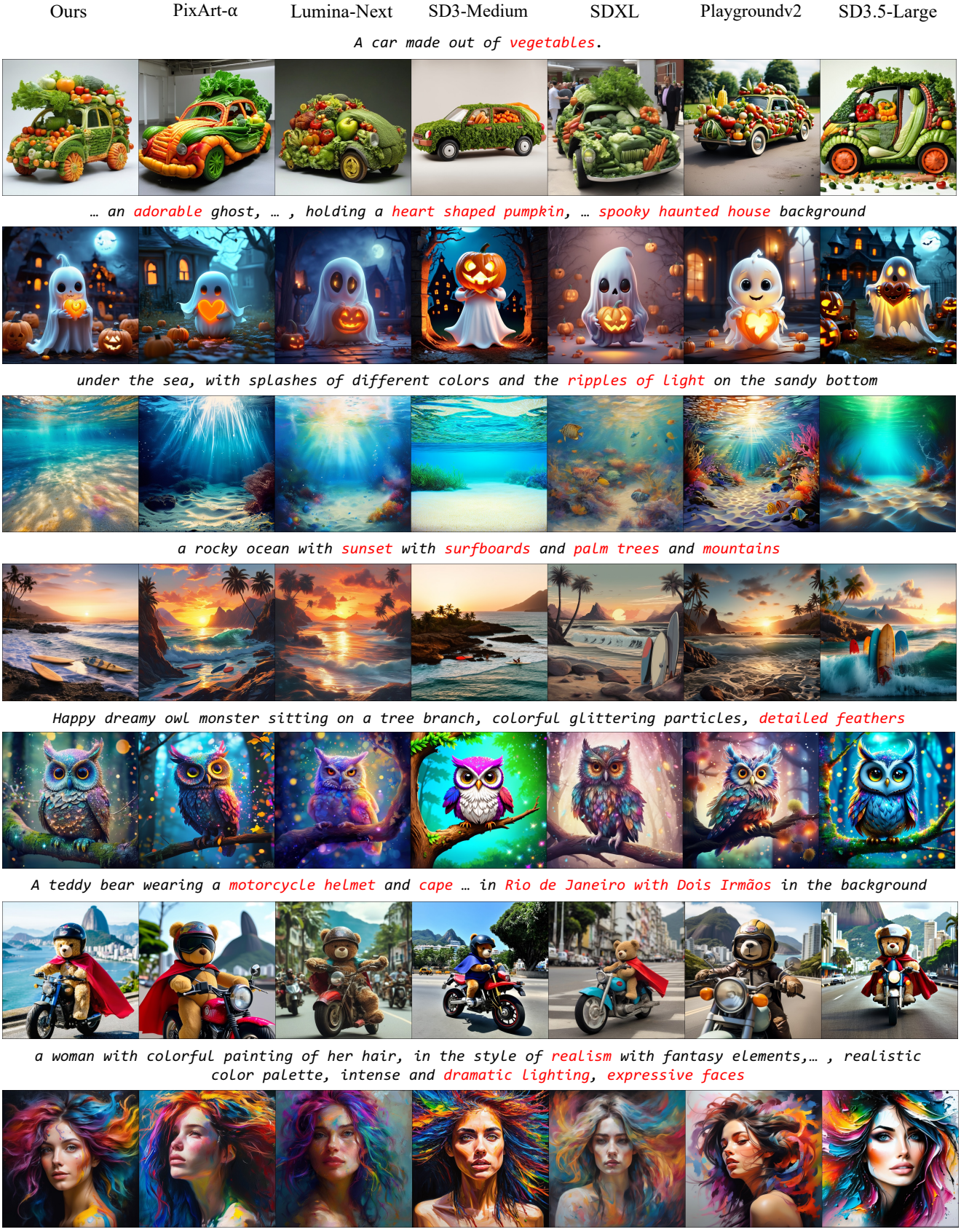

Figure 5. **Additional qualitative comparison** at 1024^2 resolution. Zoom in for details.

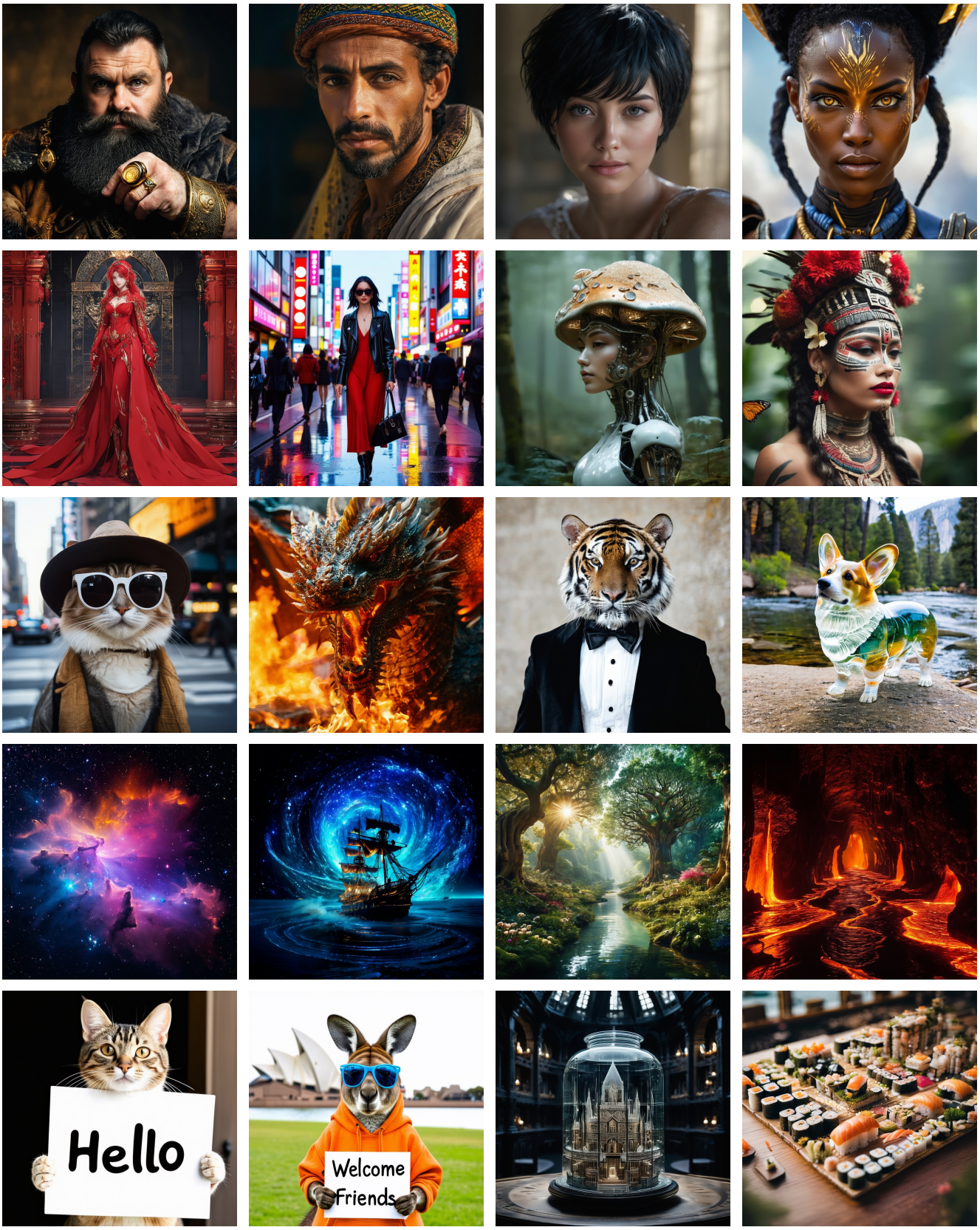

Figure 6. **Additional T2I example visualization** at 1024^2 resolution of our model. Zoom in for details.

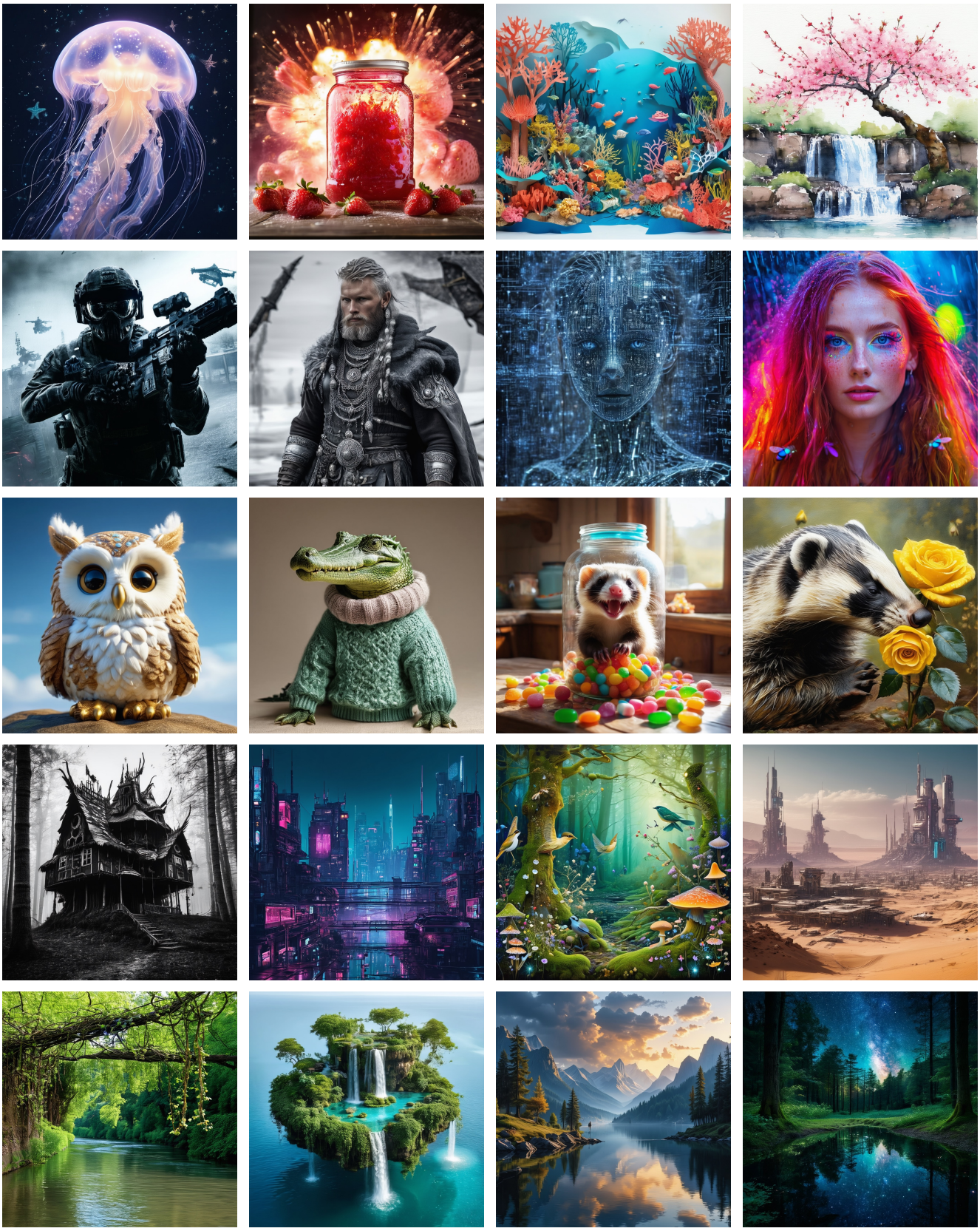

Figure 7. **Additional T2I example visualization** at 1024^2 resolution of our model. Zoom in for details.

Table 4. **Prompts used for visualization.**

Prompt used for Visualization in Fig. 6, (i,j) represents the image in i-th row and j-th column

(1,1) A black bearded dwarf with a big golden ring on his finger is looking seriously into the camera port
(1,2) Moroccan man, portrait, dynamic pose, detailed hair texture, ornate, sharp focus, in the style of National Geographic, photoportrait, Cinematic, ...
(1,3) beauty, short black hair, cinematic, photorealism, intricate ultra detail, high sharpness, 8K cinematic, photography, realistic, ...
(1,4) black African woman warrior character in the style of Avatar and Overwatch searching the path of true, extremely detailed skin, centered portrait, ...

(2,1) Lady in red, anime, cartoon, unreal engin, concept art, full body view, ornate, ultra detail, cinematic, beauty shot.
(2,2) A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.
(2,3) the mushroom on the head of a cyborg woman, in the style of atmospheric woodland imagery, hyperrealistic atmospheres, ...
(2.4) Close up out of focus blurry photo of a stunning interpretation of the most artistically and aesthetically refined representation of a Mayan tribal female models, heavy tribal makeup and facial tattoos with lush red lips, large traditional Mayan headdress, looking to the side, butterflies and flowers jungle background, minimal desaturated color palette, soft vignette, f1.4 110 sec shutter, soft light

(3,1) very realistic, detailed, cat with big hat and white sunglasses, posing, hyperrealistic, atmospheric, in city, street, new york, cinematic, dramatic lighting, photorealistic, Leica M10R 8k Leica 35mm lens, Tilt Blur, Shutter Speed 1 1000, F 5.6, Super Resolution
(3,2) a close-up of a fire spitting dragon, cinematic shot
(3,3) a tiger wearing a tuxedo
(3,4) Color photo of a corgi made of transparent glass, standing on the riverside in Yosemite National Park.

(4,1) Colorful shining nebulae
(4,2) Pirate ship trapped in a cosmic maelstrom nebula, rendered in cosmic beach whirlpool engine, volumetric lighting, spectacular, ambient lights, light pollution, cinematic atmosphere, art nouveau style, illustration art artwork by SenseiJaye, intricate detail.
(4,3) A sureal parallel world where mankind avoid extinction by preserving nature, epic trees, water streams, various flowers, intricate details, rich colors, rich vegetation, cinematic, symmetrical, beautiful lighting, V-Ray render, sun rays, magical lights, photography
(4,4) River of lava flows through a hallway of caves

(5,1) a cat holds a sign saying "Hello"
(5,2) a portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grassin front of the Sydney Opera House holding a sign on the chest that says Welcome Friends
(5,3) Spectacular Tiny World in the Transparent Jar On the Table, interior of the Great Hall, Elaborate, Carved Architecture, Anatomy, Symetrical, ...
(5,4) tilt shift aerial photo of a cute city made of sushi on a wooden table in the evening.

Prompt used for Visualization in Fig. 7, (i,j) represents the image in i-th row and j-th column

(1,1) aesthetic light colored blue jellyfish with stars
(1,2) realistic photo of a jar of strawberry jam with explosions
(1,3) A gorgeously rendered papercraft world of a coral reef, rife with colorful fish and sea creatures.
(1,4) A watercolor painting of a cherry blossom tree and waterfall

(2,1) call of duty ghosts
(2,2) black and grey, historical viking, Historic, historically accurate, black and grey clothes, with silver jewellery, full body, dark fantasy, hyperdetailed, intricate details, hyper realistic
(2,3) Create a virtual environment where the world has dissolved into a torrent of rapidly shifting code and floating in this place we see a beautiful female artificial intelligence whos skin, hair, and body are made out of patterns and sequences intertwining. The very essence of the virtual environment and the constructs within it are now exposed in this woman, revealing their true nature as complex algorithms and digital architecture. The womans algorithm poses a significant threat to her control over the virtual world
(2,4) portrait of pretty caucasian blueeyed woman with marked freckles on her cheeks, neon red flowing Hair, long hair with rainbow colors, neon Bright colors background, face makeup divided into 4 different parts with solid bright colors, colored light wasps fall like sparkles from rain in a romantic and glamorous atmosphere, the hair with an incredible movement that surrounds the whole scene, hyper realistic photography, 8k, high contrast in detail

(3,1) Pixar animation, little brown and white fluffy soft owl toy, sitting, ultra detailed, sky blue and golden details, 8k bright front lighting, fine luster, ...
(3,2) Crocodile in a sweater
(3,3) A mischievous ferret with a playful grin squeezes itself into a large glass jar, surrounded by colorful candy. The jar sits on a wooden table in a cozy kitchen, and warm sunlight filters through a nearby window
(3,4) A young badger delicately sniffing a yellow rose, richly textured oil painting.

(4,1) Monster Baba yaga house with in a forest, dark horror style, black and white.
(4,2) Midnight video game street with glitchy effects
(4,3) surreal magical fairy forest, soft brushstrokes, birds, dmt, fish, moss, wildflowers, mushrooms
(4,4) a cyberpunk city far away in a desert

(5,1) pretty river with overhanging vines
(5,2) A floating island with crystal-clear waterfalls and lush vegetation
(5,3) blue water lake reflecting clouds
(5,4) A deep forest clearing with a mirrored pond reflecting a galaxy-filled night sky

References

[1] Stability AI. Stable diffusion 1.5. *https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5*, 2022. 2

[2] Stability AI. Stable diffusion 2.1. *https://huggingface.co/stabilityai/stable-diffusion-2-1*, 2022. 2

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 4

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 2

[6] DeepFloyd. Deepfloyd. *https://github.com/deep-floyd/IF*, 2023. 2

[7] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

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

[11] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 3

[12] Anil Kag, Huseyin Coskun, Jierun Chen, Junli Cao, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, and Jian Ren. Ascan: Asymmetric convolution-attention networks for efficient recognition and generation. *arXiv preprint arXiv:2411.04967*, 2024. 3

[13] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3

[14] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024. 3

[15] Youngwan Lee, Kwanyong Park, Yoorhim Cho, Yong-Ju Lee, and Sung Ju Hwang. Koala: Empirical lessons toward memory-efficient and fast diffusion models for text-to-image synthesis, 2023. 2

[16] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259, 2022. 3

[17] 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*, 2024. 2

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[19] 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

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[21] Vikash Sehwag, Xianghao Kong, Jingtao Li, Michael Spranger, and Lingjuan Lyu. Stretching each dollar: Diffusion training from scratch on a micro-budget. *arXiv preprint arXiv:2407.15811*, 2024. 2

[22] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Yujun Lin, Zhekai Zhang, Muyang Li, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 2

[23] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[24] 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. 2