MegaFusion: Extend Diffusion Models towards Higher-resolution Image Generation without Further Tuning - Supplementary Material

Haoning Wu*, Shaocheng Shen*, Qiang Hu†, Xiaoyun Zhang†, Ya Zhang, Yanfeng Wang
Shanghai Jiao Tong University, China

In this appendix, we start by giving more details on the implementation details of our proposed MegaFusion in Section A. Then, we provide extra quantitative comparisons to further demonstrate the universality and effectiveness of our method in Section B. Next, we offer additional qualitative results across various experimental settings and methods to illustrate the superiority of our proposed MegaFusion in Section C. Finally, we discuss the limitations of our method and future work in Section D.

A. Implementation Details

More Details on Floyd-MegaFusion. We have evaluated the higher-resolution image generation performance of Floyd [2] at resolutions of 128 × 128 and 512 × 512. For 128 × 128 resolution, we directly apply MegaFusion to the first stage of Floyd. As for the comparison at 512 × 512 resolution, we utilize the first two stages of Floyd. Considering that the quality of the results from the first stage generation would significantly affect the second generation stage, we opt for using the 64 × 64 images generated by the original first stage model as inputs of both the baseline and our boosted Floyd-MegaFusion. That is, higher-resolution image generation is only performed under the second generation stage. Ultimately, the experimental results presented in Tab.1 of our submitted manuscript effectively demonstrate the universality and effectiveness of our proposed MegaFusion. Furthermore, we also conduct experiments where 128 × 128 out-of-distribution images are generated in the first stage, followed by 512 × 512 resolution images in the second stage. This further demonstrates that MegaFusion maintains semantic accuracy across all stages of generation.

Details on Human Evaluation. To more effectively reflect the performance of different models in generating high-resolution images, we have recruited 10 volunteers with a background in image generation research for human evaluation. Specifically, the evaluators are asked to follow these rules: (i) Rate unknown source images on a score from 1 to 5 for both image quality and semantic accuracy, with higher scores indicating better quality; and (ii) Observe the results generated by different models with the same input conditions and select their favourite one based on overall quality and semantic accuracy.

B. Additional Quantitative Results

B.1. Comparison on crop FID/KID

Following previous work [3], we also evaluate crop FID and crop KID metrics on the generated results of various models to reflect the quality of local patches in the images. As depicted in Table 1, previous methods are often limited to specific latent-space models, whereas our MegaFusion consistently improves the quality of high-resolution image generation across both latent-space and pixel-space models.

Method	SDM-1024	SDXL-2048	Floyd-128	Floyd-512
Original	41.21/0.0139	42.29/0.0125	70.16/0.0224	40.65/0.0171
ScaleCrafter	**32.24**/0.0085	26.58/0.0062	inapplicable	inapplicable
DemoFusion	inapplicable	25.91/0.0061	inapplicable	inapplicable
MegaFusion	39.42/0.0137	27.38/0.0063	57.24/0.0243	32.36/0.0122
MegaFusion++	33.39/**0.0084**	**25.64/0.0049**	**41.22/0.0188**	**29.18/0.0077**

Table 1. Comparison of $\mathrm{FID_{crop}}/\mathrm{KID_{crop}}$ on MS-COCO dataset.

B.2. Comparison on CUB-200 Dataset

To demonstrate the universality of our proposed MegaFusion, in addition to the MS-COCO [7] dataset, we also conduct quantitative evaluations on the CUB-200 [10] dataset, which is also commonly used in previous works. The CUB-200 dataset consists of over 10K images of 200 categories of birds, each accompanied by 10 textual descriptions. Considering computational costs and time expenditure, similar to the experimental settings on the MS-COCO dataset in our manuscript, we randomly select 1K images from the CUB-200 dataset. Each image is assigned a fixed caption, and the same random seed is used across different methods to eliminate the effects of randomness among models. As depicted in Table 2, our proposed MegaFusion can also be universally applied to both latent-space and pixel-space diffusion models on the CUB-200 dataset, achieving high-quality higher-resolution image generation.

Methods	resolution	FID$_r$ ↓	FID$_b$ ↓	KID$_r$ ↓	KID$_b$ ↓	CLIP-T↑	CIDEr↑	Meteor↑	ROUGE↑	GFlops	Inference time
SDM [9]	1024 × 1024	77.92	46.34	0.0363	0.0220	0.2952	8.12	7.48	7.09	135.0K	15.17s
SDM-MegaFusion	1024 × 1024	71.78	36.21	0.0303	0.0189	0.3060	24.46	11.98	12.62	**48.2K**	**7.56s**
SDM-MegaFusion++	1024 × 1024	**68.92**	**34.94**	**0.0251**	**0.0182**	**0.3115**	**28.52**	**12.32**	**13.29**	**48.2K**	**7.56s**
SDXL [8]	2048 × 2048	73.49	48.78	0.0308	0.0274	0.2994	16.43	9.90	10.35	540.2K	79.66s
SDXL-MegaFusion	2048 × 2048	72.62	13.72	0.0296	0.0039	0.3113	25.98	13.23	13.33	**216.1K**	**30.94s**
SDXL-MegaFusion++	2048 × 2048	**65.10**	**11.55**	**0.0225**	**0.0026**	**0.3122**	**26.35**	**13.98**	**14.92**	**216.1K**	**30.94s**
Floyd-Stage1 [2]	128 × 128	87.04	105.59	0.0341	0.0658	0.2866	9.95	8.28	9.07	111.7K	77.08s
Floyd-MegaFusion	128 × 128	77.82	**36.49**	0.0413	**0.0281**	0.3080	22.12	**17.06**	**20.62**	**44.9K**	**32.19s**
Floyd-MegaFusion++	128 × 128	**73.54**	45.76	**0.0334**	0.0388	**0.3086**	**22.52**	16.93	20.05	**44.9K**	**32.19s**
Floyd-Stage2 [2]	512 × 512	80.34	41.65	0.0401	0.0215	0.3013	23.59	12.28	11.67	60.7K	48.58s
Floyd-MegaFusion	512 × 512	77.66	39.34	0.0348	0.0141	0.3110	24.63	**15.74**	15.29	**24.3K**	**21.72s**
Floyd-MegaFusion++	512 × 512	**62.91**	**34.40**	**0.0232**	**0.0115**	**0.3141**	**25.44**	13.90	**18.51**	**24.3K**	**21.72s**

Table 2. **Quantitative comparison** on CUB-200 [10] dataset. **RED**: best performance, BLUE: second best performance.

Methods	resolution	FID$_r$ ↓	FID$_b$ ↓	KID$_r$ ↓	KID$_b$ ↓	CLIP-T↑	CIDEr↑	Meteor↑	ROUGE↑	GFlops	Inference time
Floyd-Stage1 [2]	128 × 128	66.27	81.65	0.0262	0.0454	0.2818	14.69	18.22	25.06	111.7K	77.08s
Floyd-MegaFusion	128 × 128	53.09	**39.73**	0.0273	**0.0334**	0.3024	**25.01**	25.00	31.35	**44.9K**	**32.19s**
Floyd-MegaFusion++	128 × 128	**43.43**	50.08	**0.0213**	0.0437	**0.3046**	20.28	**25.01**	**31.64**	**44.9K**	**32.19s**
Floyd-Stage2 [2]	64 → 512	46.64	38.15	0.0254	0.0166	0.3098	23.85	21.47	26.26	60.7K	48.58s
Floyd-MegaFusion	64 → 512	39.80	24.87	0.0164	0.0078	0.3106	23.22	23.51	29.30	**24.3K**	**21.72s**
Floyd-MegaFusion++	64 → 512	**26.34**	**24.55**	**0.0063**	**0.0077**	**0.3110**	**24.01**	**23.58**	**29.52**	**24.3K**	**21.72s**
Floyd-Stage2 [2]	128 → 512	61.24	108.01	0.0253	0.0734	0.2779	15.16	14.76	19.75	60.7K	48.58s
Floyd-MegaFusion	128 → 512	58.19	**88.56**	0.0187	**0.0379**	0.2821	16.28	**15.65**	20.02	**24.3K**	**21.72s**
Floyd-MegaFusion++	128 → 512	**57.92**	94.93	**0.0181**	0.0417	**0.2835**	**16.36**	15.47	**21.34**	**24.3K**	**21.72s**

Table 3. **More comparison results** on Floyd model and its MegaFusion boosted counterparts under different settings. Within each unit, we denote the best performance in **RED** and the second-best performance in BLUE.

B.3. More Results of Floyd-MegaFusion

As mentioned above, we also conduct experiments that first generate 128×128 out-of-distribution images, followed by 512×512 high-resolution images on the Floyd model. As depicted in Table 3, MegaFusion consistently improves the high-resolution generation capability of Floyd under both settings. This demonstrates that MegaFusion can improve the semantic accuracy of high-resolution images at any stage of the generation process.

B.4. Ablation Study of Classifier-free Guidance

As detailed in the implementation details, to ensure a fair comparison and eliminate the impact of classifier-free guidance (CFG) on generation quality and efficiency, we use the default CFG weights from official implementations for all methods and their corresponding MegaFusion-boosted counterparts. To further investigate the impact of CFG on MegaFusion at higher resolutions, we generate 100 images from the MS-COCO dataset using SDM-MegaFusion and SDXL-MegaFusion with varying CFG values, using the same text prompt and random seed as inputs, and evaluate the FID scores against our testset. The results in Figure 1 indicate that classifier-free guidance does affect our high-resolution generation quality, with preliminary findings indicating that $w = 7.0$ is a relatively good choice for SDM-MegaFusion and SDXL-MegaFusion.

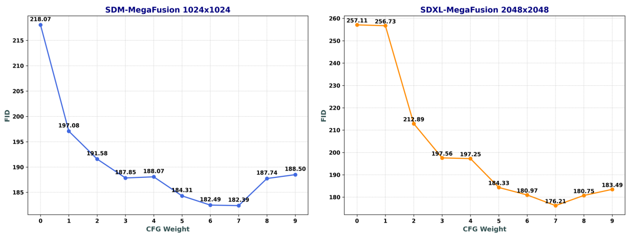


Figure 1. **Ablation study** of classifier-free guidance (CFG) weight on SDM-MegaFusion and SDXL-MegaFusion.

C. Additional Qualitative Results

C.1. Evidence Behind the Core idea & intuition

As stated in eDiff-I [1], diffusion models synthesize semantics during early denoising stages and refine image details in later stages. As depicted in Figure 2, we also observe that semantic deviations and object repetitions commonly encountered at higher resolutions primarily stem from incorrect semantics generated during early denoising, leading to irreparable errors. Thus, our **intuition and insight** here are: perform early denoising at the original resolution to generate accurate semantic information, followed by *truncate* and *relay* to continue denoising at higher resolutions, thereby enriching texture details. This enables MegaFusion to produce high-quality, semantically accurate higher-resolution images with lower computational costs, while supporting arbitrary aspect ratios.

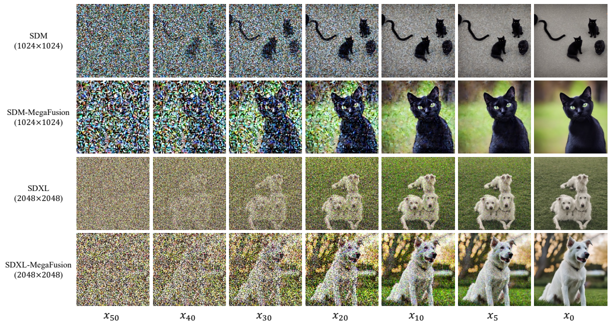


Figure 2. **Evidence behind our core idea and intuition**. For $T = 50$ steps of DDIM sampling, we visualize the key stages of the image generation process. For SDM and SDXL, incorrect semantics are generated during the early denoising stages of high-resolution generation, leading to irreparable errors. In contrast, MegaFusion generates accurate semantics and further enriches texture details at higher resolutions. The input text prompts are "*A cute black cat*" and *"A white dog sits on the grass."* For ease of visualization, the images are scaled to the same size.

C.2. Disadvantages of Direct Upsampling

Compared to our MegaFusion for higher-resolution image generation, a more straightforward approach is to directly apply upsampling to images generated by diffusion models. Although simple, this will introduce three potential issues: (i) Direct super-resolution may lead to unrealistic texture details, such as blurring and artifacts, especially at high upsampling factors; (ii) While diffusion-based SR methods can produce more realistic textures via iterative denoising, they often involve significantly higher computational costs and may not support arbitrary aspect ratios; (iii) Most critically, as shown in Figure 3, directly upsampling low-resolution images can stretch and distort content, particularly when generating under non-standard aspect ratios (e.g. $1:4$), diminishing the natural aesthetic of images.

In contrast, MegaFusion seamlessly bridges coarse-to-fine generation processes, efficiently producing accurate semantics at low resolutions and enriching texture details at high resolutions. Leveraging iterative denoising at higher resolutions, it can synthesize aesthetically pleasing high-resolution images even with non-standard aspect ratios.

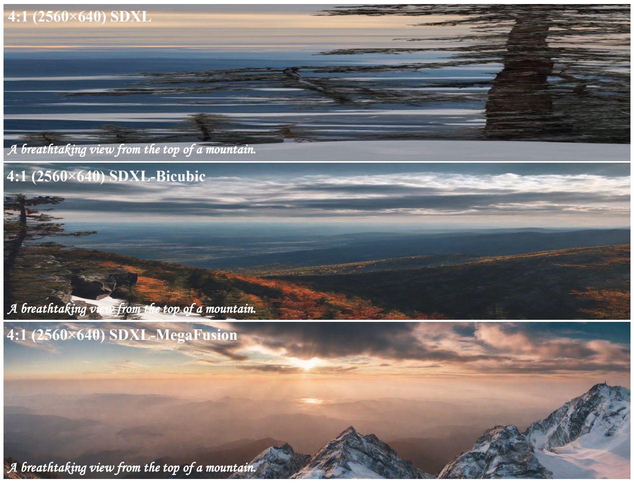

4:1 (2560×640) SDXL
A breathtaking view from the top of a mountain.
4:1 (2560×640) SDXL-Bicubic
A breathtaking view from the top of a mountain.
4:1 (2560×640) SDXL-MegaFusion
A breathtaking view from the top of a mountain.

Figure 3. **Analysis of direct upsampling**. Using diffusion models to generate images with non-standard aspect ratios directly or via upsampling, may lead to stretching and distortion (e.g., trees on both sides), while MegaFusion effectively mitigates this issue.

C.3. Effects of hyperparameters δ and γ

For denoising at the original size, we do not employ dilation. In qualitative experiments for high-resolution generation, we test various δ values and find that $\delta = 2$ is a stable choice under our experimental settings, which will not introduce blurriness or semantic deviations. As described in our manuscript, we draw inspiration from simple diffusion [6], which derives the SNR relationship between images of different resolution based on the mean and variance of pixel distributions. Substituting this into our derived relationship, we obtain that $\gamma = 4$. Qualitative experiments also confirm that this is an appropriate choice. Some visualization examples are shown in Figure 4.

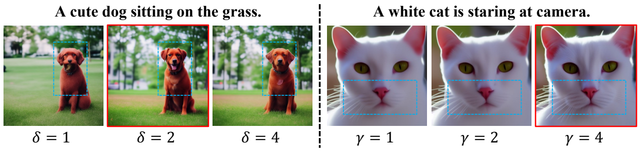


Figure 4. **Qualitative comparisons** of applying different hyperparameters δ and γ.

C.4. Ablation Study of Truncation Steps

In the *truncate and relay* strategy, the number of denoising steps at each stage may also affect generation quality. Our intuition and experience suggest that more denoising steps at lower resolutions improve generation efficiency, while additional steps at higher resolutions enhance texture details. However, conducting a comprehensive evaluation to determine the optimal truncation steps would incur significant computational costs. Therefore, in our implementation, we empirically select truncation steps for each model based on experience, and validate the above conclusions through qualitative experiments, as shown in Figure 5. Considering the trade-off between generation quality and efficiency, we choose denoising steps of $T_1 = 40$, $T_2 = 5$, and $T_3 = 5$ as the default configuration for SDM-MegaFusion.

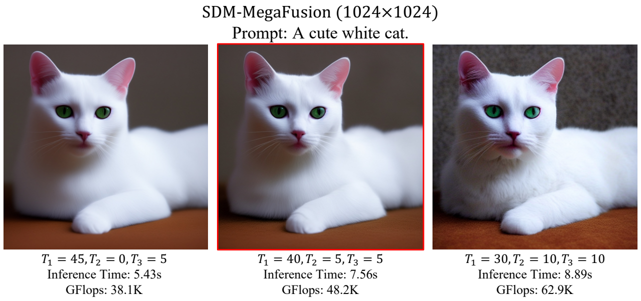

SDM-MegaFusion (1024×1024)
Prompt: A cute white cat.
$T_1 = 45, T_2 = 0, T_3 = 5$
Inference Time: 5.43s
GFlops: 38.1K
$T_1 = 40, T_2 = 5, T_3 = 5$
Inference Time: 7.56s
GFlops: 48.2K
$T_1 = 30, T_2 = 10, T_3 = 10$
Inference Time: 8.89s
GFlops: 62.9K

Figure 5. **Qualitative ablation study** of truncation steps.

C.5. Text-to-Image Foundation Models

We present more visualizations of higher-resolution image generation using both latent-space and pixel-space text-to-image models in Figure 6 and 7, respectively, to demonstrate the universality and robustness of our proposed method. The visual outcomes explicitly confirm that when pre-trained models fail to scale to higher resolutions, our approach can be universally integrated into existing latent-space and pixel-space diffusion models, improving their capability to synthesize higher-resolution images of megapixels with accurate semantics. Moreover, our further enhanced MegaFusion++ significantly boosts the quality of the generated images, producing sharper and clearer details.

C.6. Compatibility with Transformer-based Models

To further demonstrate the versatility and effectiveness of MegaFusion, we also apply it to the transformer-based (DiT) SD3 [4] model. Since DiT-based methods do not involve convolutions, we boost the model via only the *truncate and relay* strategy. As presented in Figure 8, SD3 also encounters quality degradation when generating higher-resolution images directly, while our MegaFusion effectively improves its high-resolution generation capabilities.

C.7. Comparison to state-of-the-art

To further evaluate the quality of MegaFusion, we compare it with existing state-of-the-art high-resolution generation methods. Given that these methods (ScaleCrafter [5] and DemoFusion [3]) are typically restricted to specific models, we conduct comparisons on models based on SDXL. The results in Figure 9 indicate that existing methods still face quality degradation and object repetition when generating high-resolution images. In contrast, MegaFusion produces high-quality, semantically accurate high-resolution images, and is much more efficient than existing approaches, as shown in Tab.1 of our manuscript.

C.8. Models with additional conditions

We have confirmed that our method is equally applicable to diffusion models with additional input conditions, such as ControlNet [12] with depth maps and IP-Adapter [11] with reference images as extra inputs. As depicted in Figure 10, we further discover that ControlNet with canny edges or human poses as conditional inputs also struggle with synthesizing higher-resolution images, and often produce images that are not fidelity to input conditions, with confusing semantics and poor image quality. In contrast, with the assistance of our proposed MegaFusion, our boosted model, ControlNet-MegaFusion consistently generates high-quality images of higher resolutions with accurate semantics, that are fidelity to conditions.

C.9. Generation with Arbitrary Aspect Ratios

As previously stated, our MegaFusion also enables existing pre-trained diffusion models to generate images at arbitrary aspect ratios. Figure 11, 12 and 13 showcase more qualitative results from SDXL-MegaFusion across various aspect ratios and resolutions, including $1:1$ (2048×2048), $16:9$ (1920×1080), $3:4$ (1536×2048), and $4:3$ (2048×1536). Moreover, as presented in Figure 14, 15, and 16, we also include visualizations with **non-standard** aspect ratios, such as $1:4$ (640×2560), $4:1$ (2560×640), $1:2$ (1024×2048), $2:1$ (2048×1024), $21:9$ (2016×864), and $9:21$ (864×2016). These impressive outcomes further demonstrate the scalability and superiority of our approach.

C.10. Compatibility with LoRA

To further illustrate the versatility and broad applicability of MegaFusion, we apply it to SDM and SDXL models using LoRA from the open-source community for personalized higher-resolution image generation. As depicted in Figure 17, MegaFusion can seamlessly integrate with various LoRAs of SDM and SDXL, demonstrating significant potential for artistic and commercial applications.

D. Limitations & Future Work

D.1. Limitations

Since our proposed MegaFusion is a tuning-free approach built on existing latent-space and pixel-space image generation models, it inevitably inherits some limitations of current diffusion-based generative models. For example, when handling complex textual conditions, the generated content often struggles to accurately reflect input prompts, particularly in aspects such as attribute binding and positional control. This may lead to degraded synthesis quality during high-resolution generation with MegaFusion. However, more powerful backbone models are expected to mitigate this issue, and when combined with MegaFusion, they are likely to produce higher-quality images at higher resolutions with low computational costs.

D.2. Future Work

The striking quantitative results produced by MegaFusion have confirmed its potential to overcome the limitations of existing diffusion-based generative models and to improve their capabilities to synthesize high-resolution outcomes. Additionally, we have observed that existing video generation models encounter significant semantic deviations and quality degradation when generating content beyond their pre-trained spatial resolution and temporal length. Therefore, we anticipate further applying MegaFusion to current video generation models towards efficient, low-cost, higher-resolution, and longer video content generation. Similarly, MegaFusion also holds the potential for extension to 3D generation models and models for image and video editing, which are also left for future exploration.

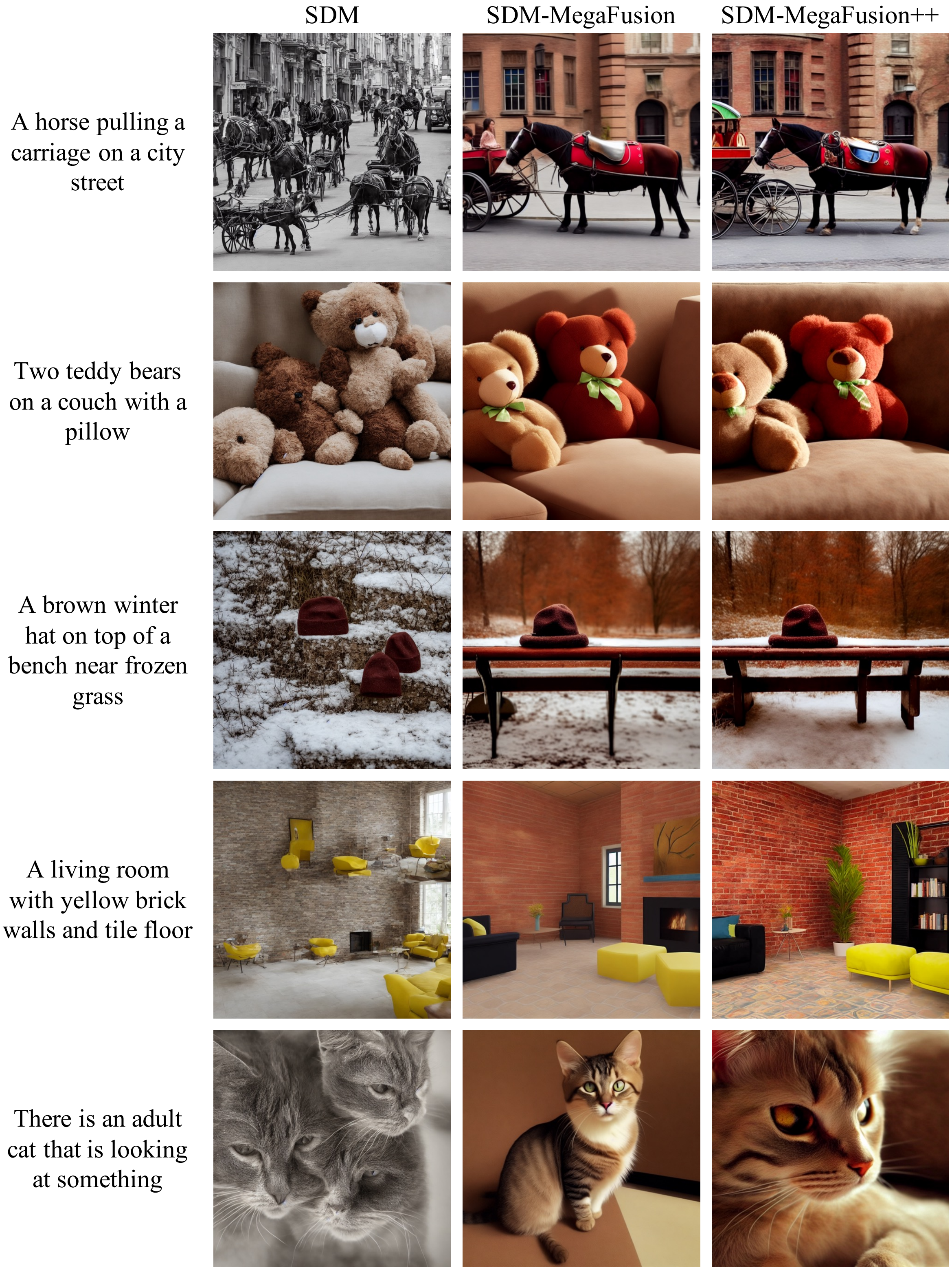

SDM
SDM-MegaFusion
SDM-MegaFusion++
A horse pulling a carriage on a city street
Two teddy bears on a couch with a pillow
A brown winter hat on top of a bench near frozen grass
A living room with yellow brick walls and tile floor
There is an adult cat that is looking at something

Figure 6. **More qualitative results** of applying our MegaFusion to latent-space diffusion model (SDM [9]) for higher-resolution (1024 × 1024) image generation on MS-COCO [7] and commonly used prompts from the Internet.

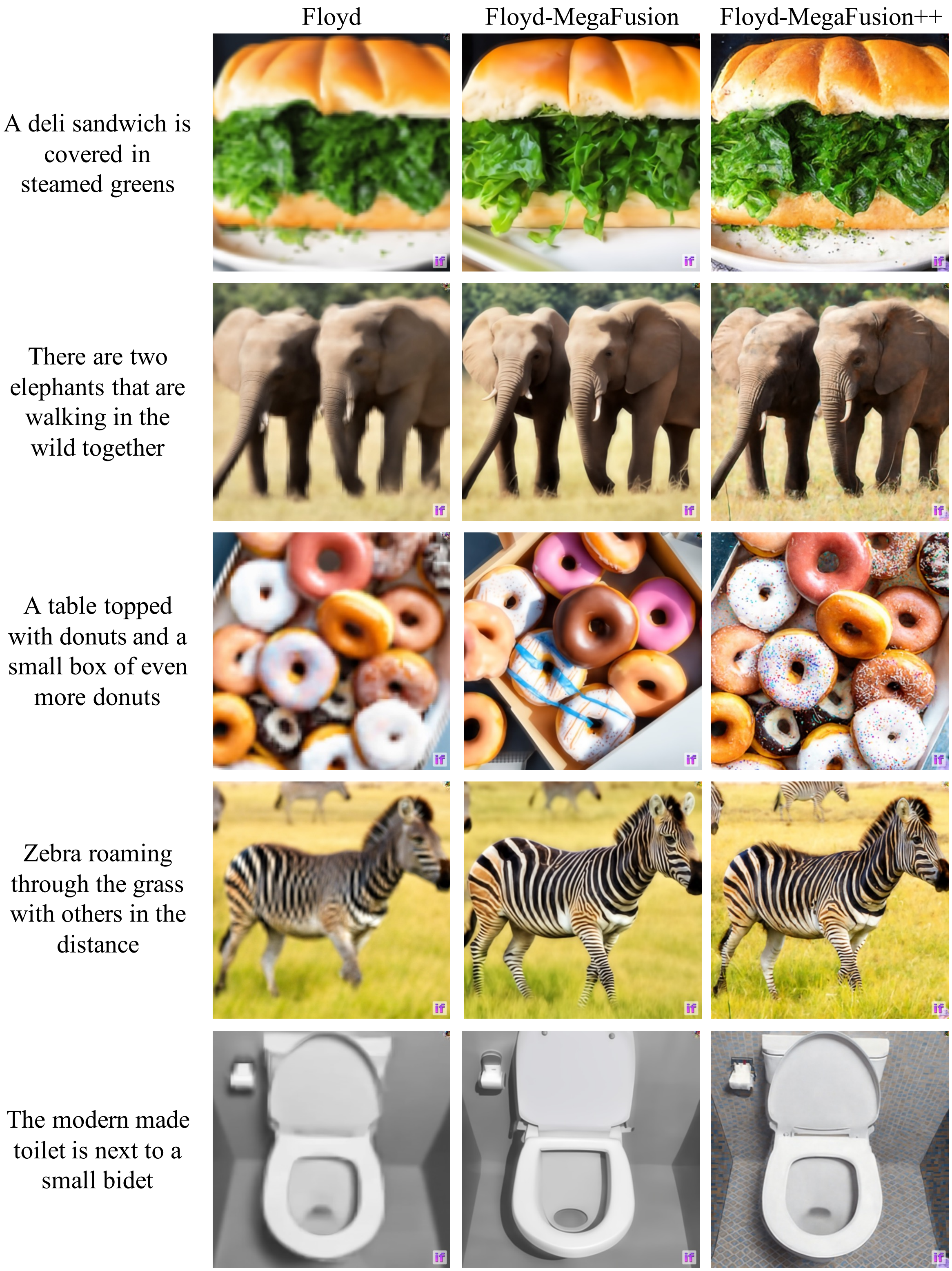
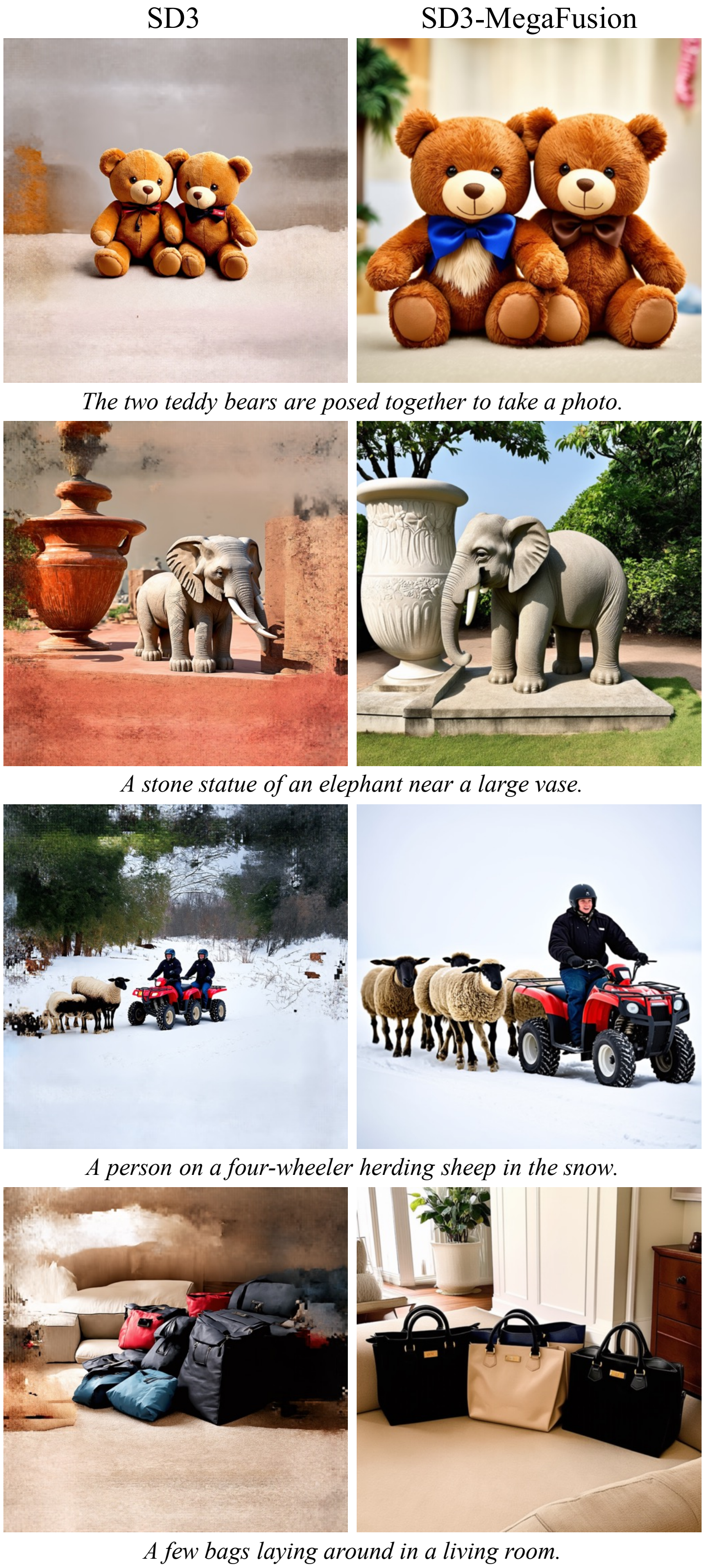

Floyd
Floyd-MegaFusion
Floyd-MegaFusion++
A deli sandwich is covered in steamed greens
There are two elephants that are walking in the wild together
A table topped with donuts and a small box of even more donuts
Zebra roaming through the grass with others in the distance
The modern made toilet is next to a small bidet

Figure 7. **More qualitative results** of applying our MegaFusion to pixel-space diffusion model (Floyd [2]) for higher-resolution (512×512) image generation on MS-COCO [7] and commonly used prompts from the Internet.



The two teddy bears are posed together to take a photo.

A stone statue of an elephant near a large vase.

A person on a four-wheeler herding sheep in the snow.

A few bags laying around in a living room.

Figure 8. **Qualitative results** of applying our MegaFusion to latent-space diffusion model (SD3 [4]) for higher-resolution (2048 × 2048) image generation on MS-COCO [7] and commonly used prompts from the Internet.

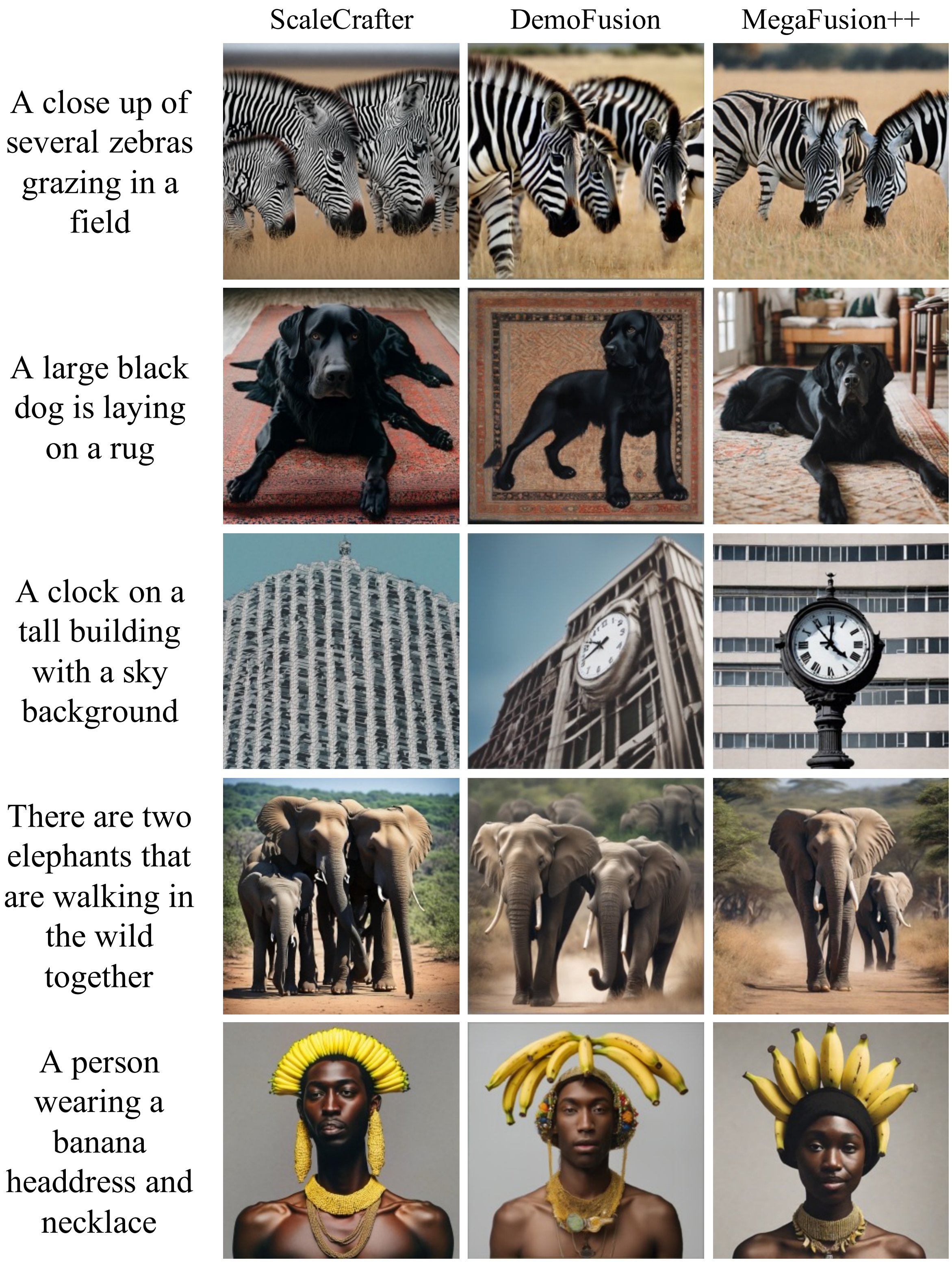

ScaleCrafter
DemoFusion
MegaFusion++
A close up of several zebras grazing in a field
A large black dog is laying on a rug
A clock on a tall building with a sky background
There are two elephants that are walking in the wild together
A person wearing a banana headdress and necklace

Figure 9. **Qualitative comparison** with existing state-of-the-art methods (ScaleCrafter [5] and DemoFusion [3]). Our MegaFusion can generate images with details and accurate semantics at high resolution, whereas existing methods struggle to do so.

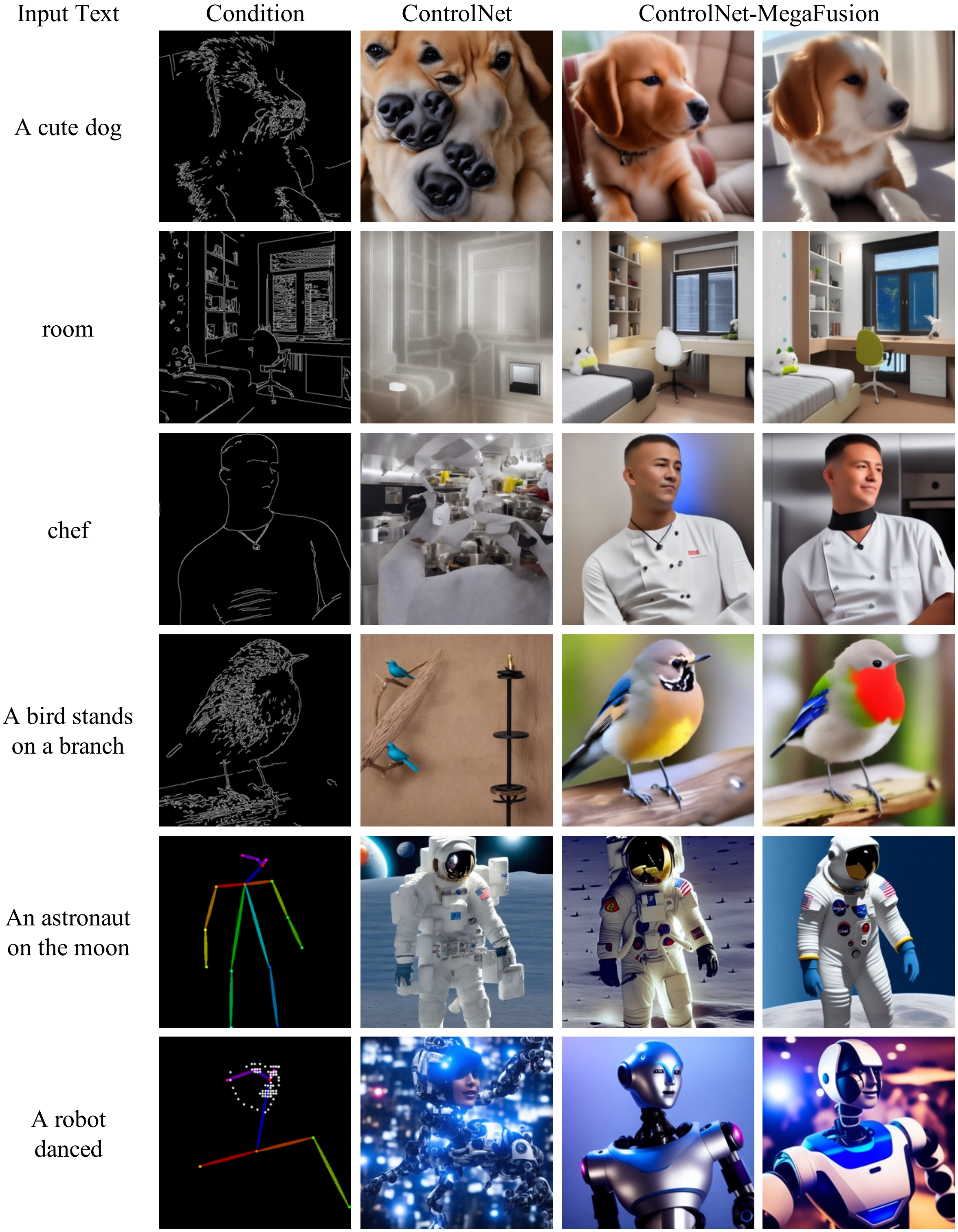


Figure 10. **Qualitative results** of applying our MegaFusion to ControlNet [12] with canny edges or human poses as extra conditions for higher-resolution (1024 × 1024) image generation with better semantics and fidelity.

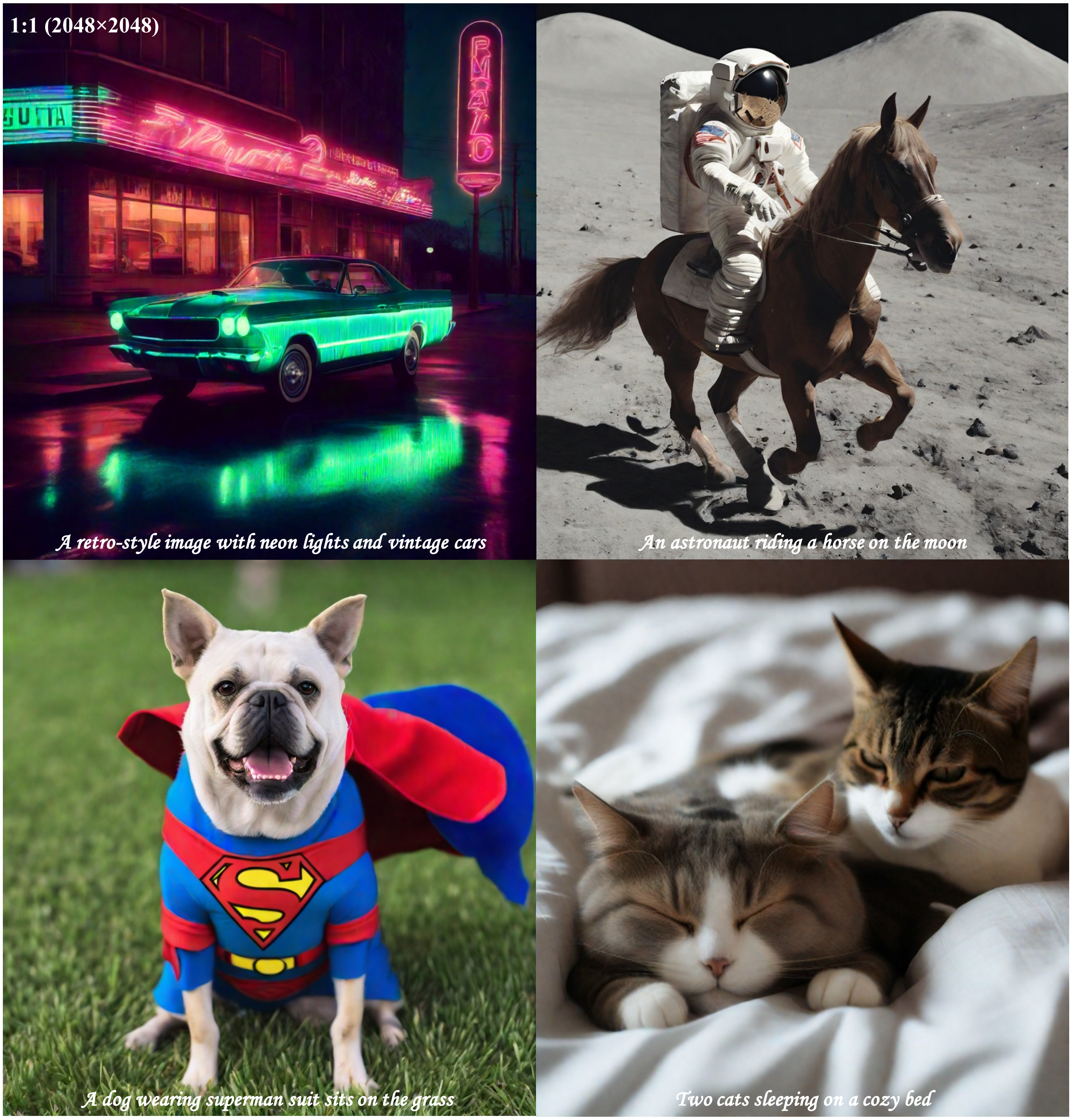

1:1 (2048×2048)
A retro-style image with neon lights and vintage cars
An astronaut riding a horse on the moon
A dog wearing superman suit sits on the grass
Two cats sleeping on a cozy bed

Figure 11. **More qualitative results** of applying our MegaFusion to SDXL [8] model for higher-resolution image generation with various aspect ratios and resolutions.

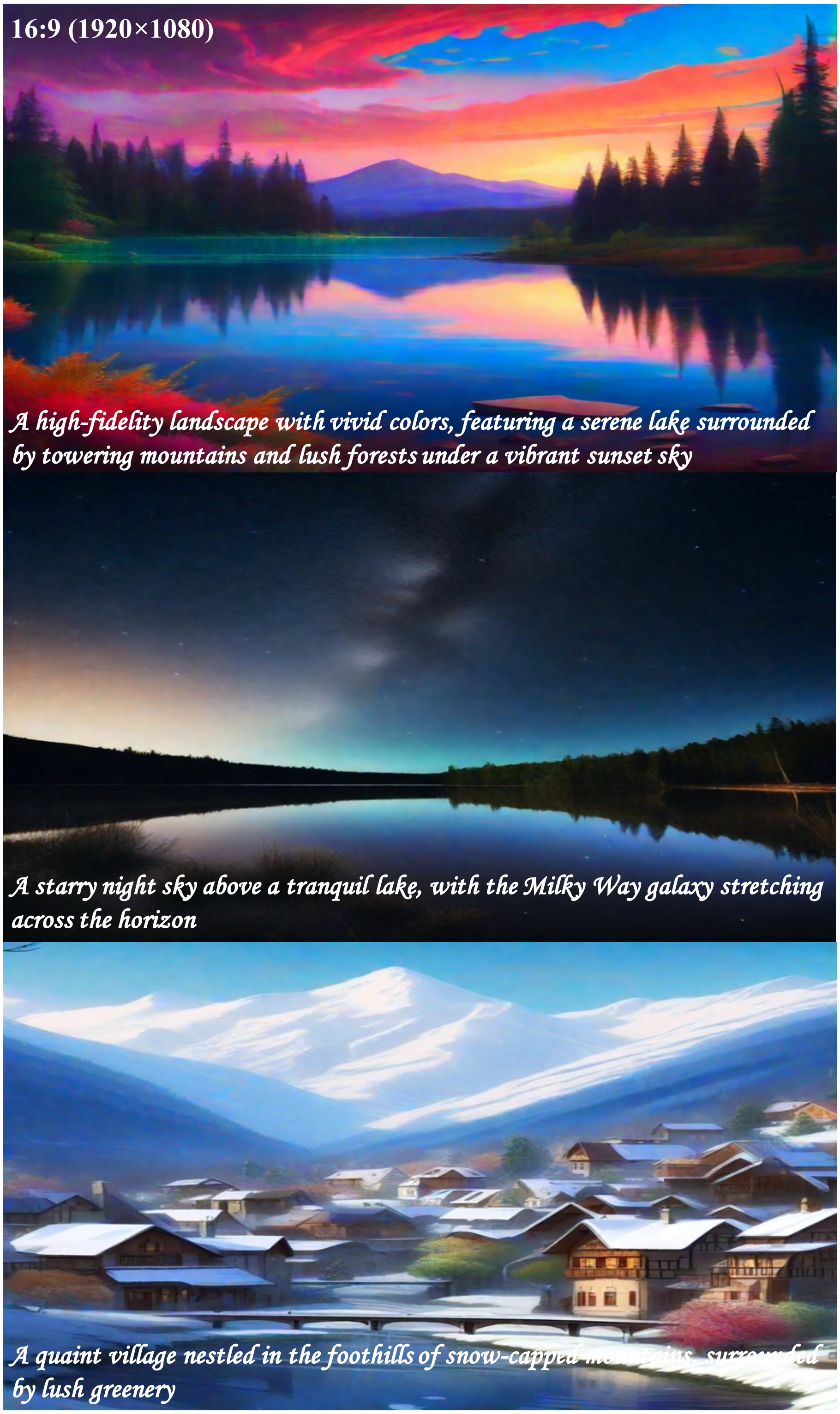

16:9 (1920×1080)
A high-fidelity landscape with vivid colors, featuring a serene lake surrounded by towering mountains and lush forests under a vibrant sunset sky
A starry night sky above a tranquil lake, with the Milky Way galaxy stretching across the horizon
A quaint village nestled in the foothills of snow-capped mountains, surrounded by lush greenery

Figure 12. **More qualitative results** of applying our MegaFusion to SDXL [8] model for higher-resolution image generation with various aspect ratios and resolutions.

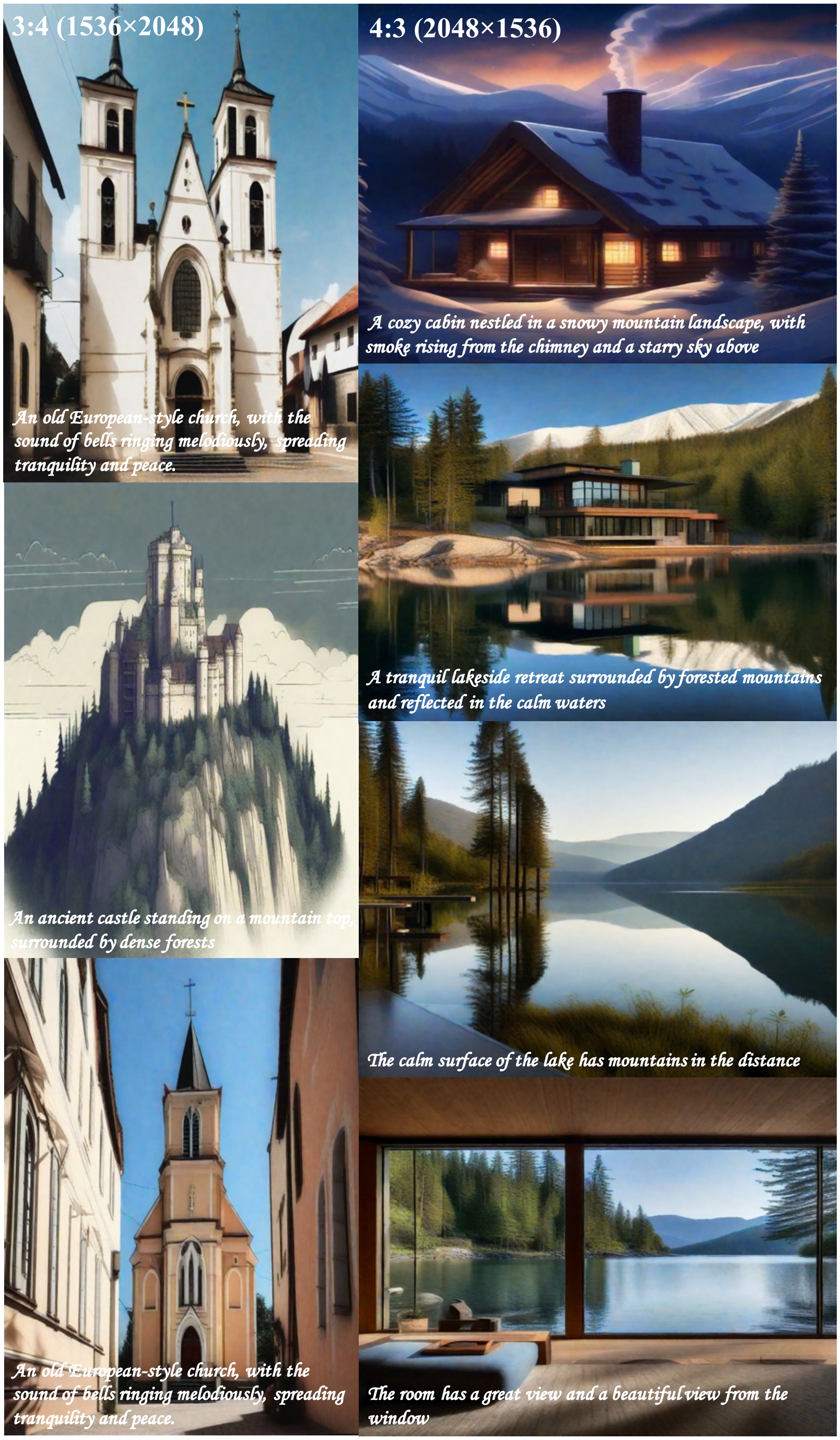

3:4 (1536×2048)
4:3 (2048×1536)
A cozy cabin nestled in a snowy mountain landscape, with smoke rising from the chimney and a starry sky above
An old European-style church, with the sound of bells ringing melodiously, spreading tranquility and peace.
A tranquil lakeside retreat surrounded by forested mountains and reflected in the calm waters
An ancient castle standing on a mountain top, surrounded by dense forests
The calm surface of the lake has mountains in the distance
An old European-style church, with the sound of bells ringing melodiously, spreading tranquility and peace.
The room has a great view and a beautiful view from the window

Figure 13. **More qualitative results** of applying our MegaFusion to SDXL [8] model for higher-resolution image generation with various aspect ratios and resolutions.

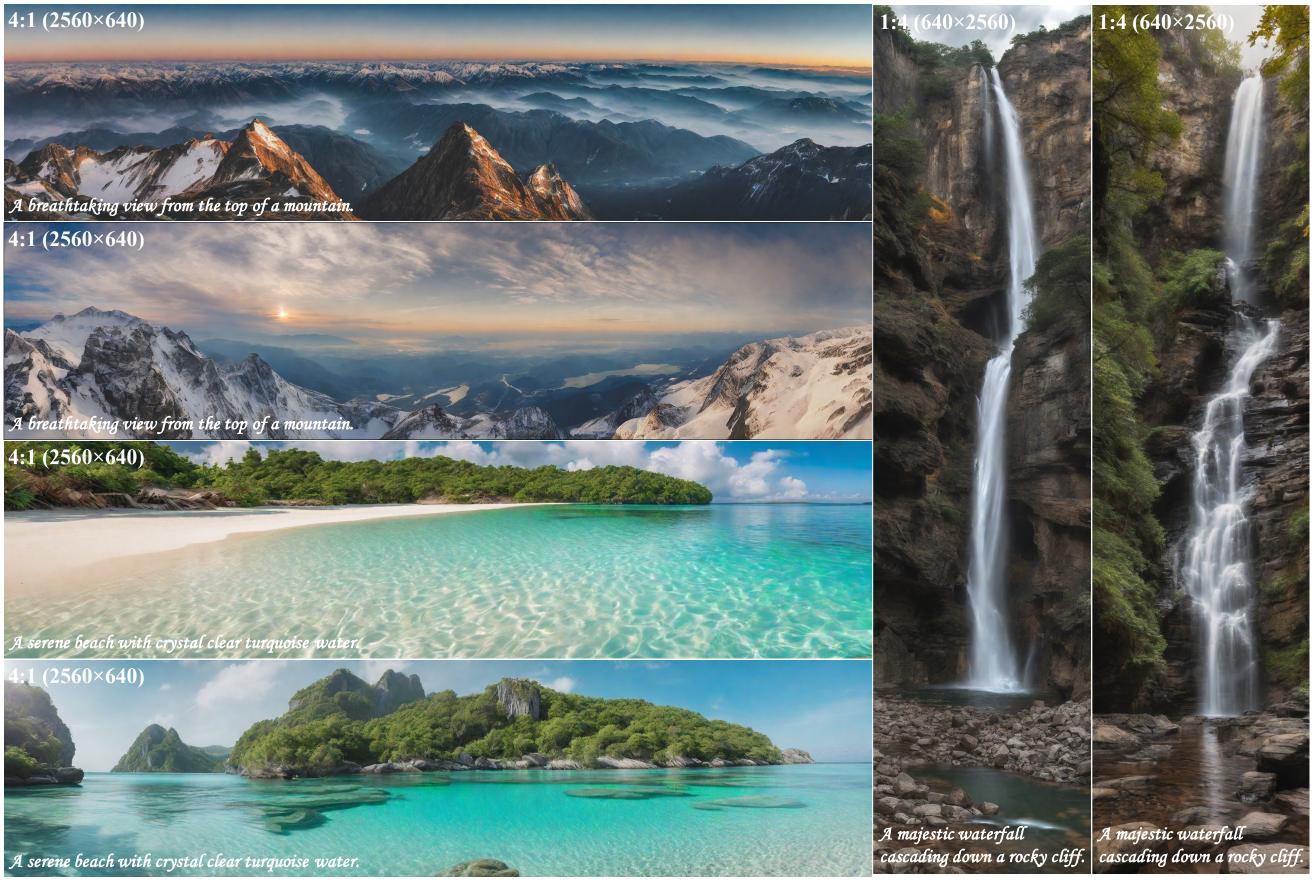

4:1 (2560×640)
A breathtaking view from the top of a mountain.
4:1 (2560×640)
A breathtaking view from the top of a mountain.
4:1 (2560×640)
A serene beach with crystal clear turquoise water.
4:1 (2560×640)
A serene beach with crystal clear turquoise water.
1:4 (640×2560)
A majestic waterfall cascading down a rocky cliff.
1:4 (640×2560)
A majestic waterfall cascading down a rocky cliff.

Figure 14. **More qualitative results** of applying our MegaFusion to SDXL [8] model for higher-resolution image generation with various **non-standard** aspect ratios and resolutions.

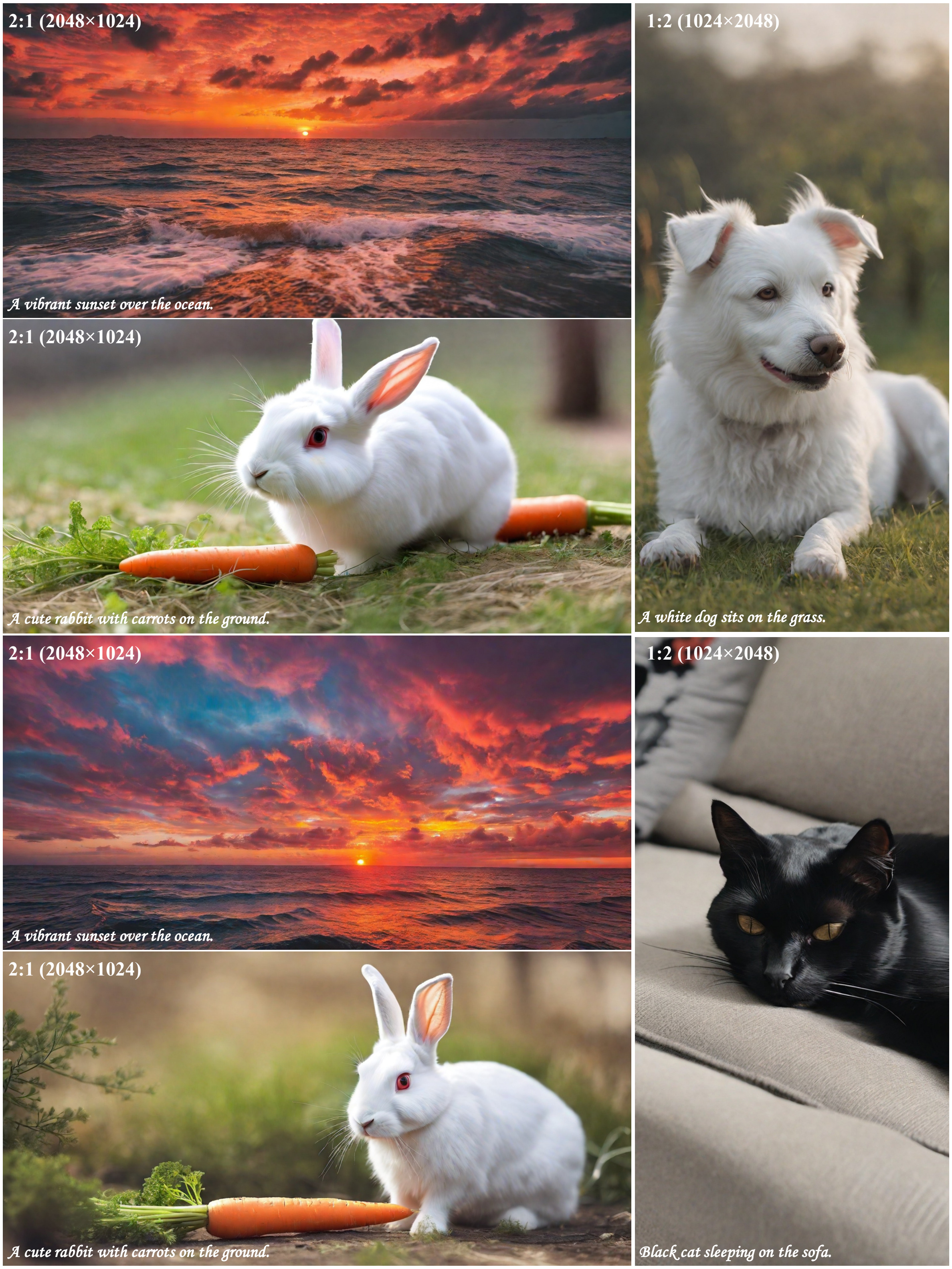

2:1 (2048×1024)
A vibrant sunset over the ocean.
1:2 (1024×2048)
2:1 (2048×1024)
A cute rabbit with carrots on the ground.
A white dog sits on the grass.
2:1 (2048×1024)
A vibrant sunset over the ocean.
1:2 (1024×2048)
2:1 (2048×1024)
A cute rabbit with carrots on the ground.
Black cat sleeping on the sofa.

Figure 15. **More qualitative results** of applying our MegaFusion to SDXL [8] model for higher-resolution image generation with various **non-standard** aspect ratios and resolutions.

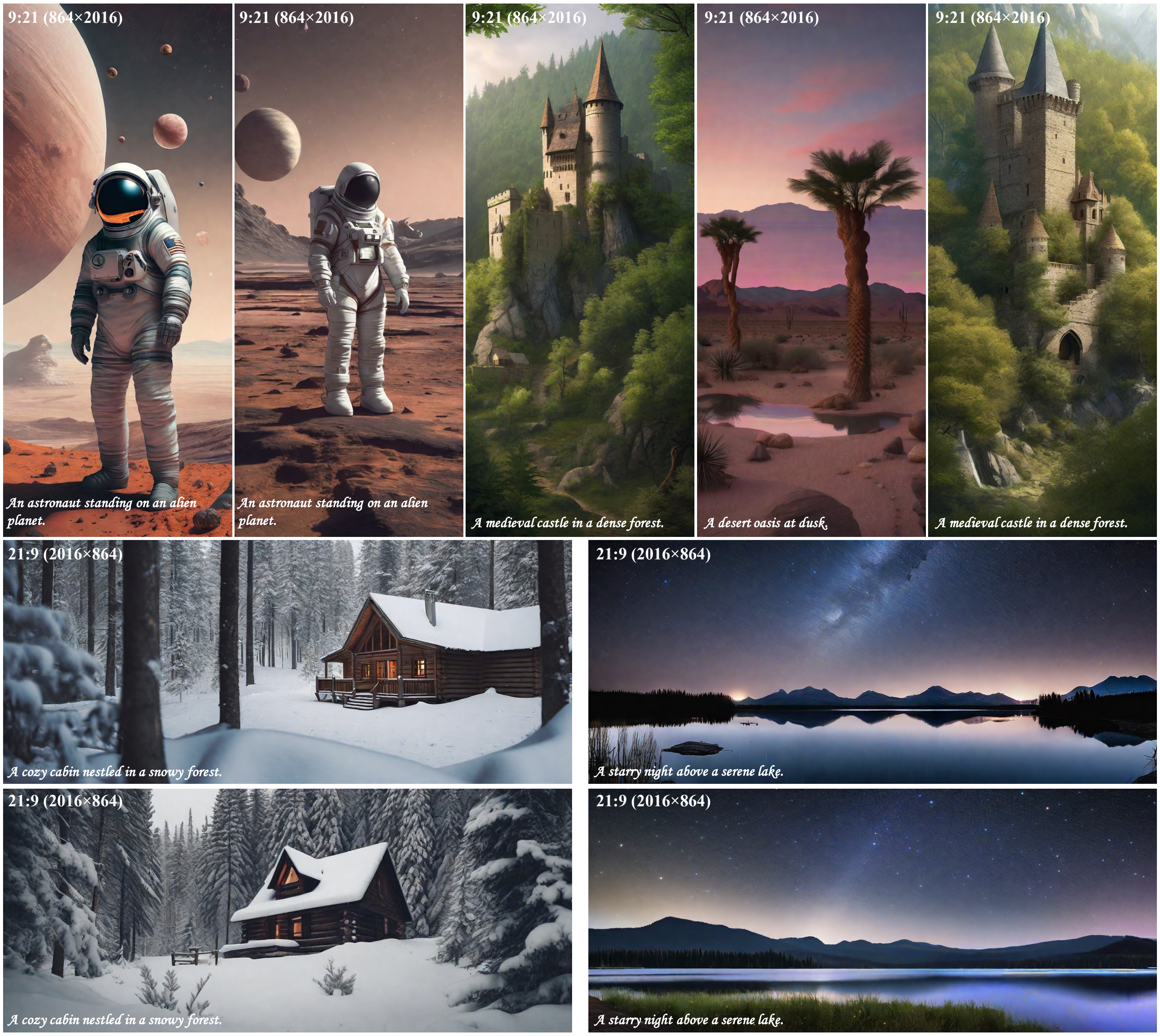

9:21 (864×2016)
An astronaut standing on an alien planet.
9:21 (864×2016)
An astronaut standing on an alien planet.
9:21 (864×2016)
A medieval castle in a dense forest.
9:21 (864×2016)
A desert oasis at dusk.
9:21 (864×2016)
A medieval castle in a dense forest.
21:9 (2016×864)
A cozy cabin nestled in a snowy forest.
21:9 (2016×864)
A starry night above a serene lake.
21:9 (2016×864)
A cozy cabin nestled in a snowy forest.
21:9 (2016×864)
A starry night above a serene lake.

Figure 16. **More qualitative results** of applying our MegaFusion to SDXL [8] model for higher-resolution image generation with various **non-standard** aspect ratios and resolutions.

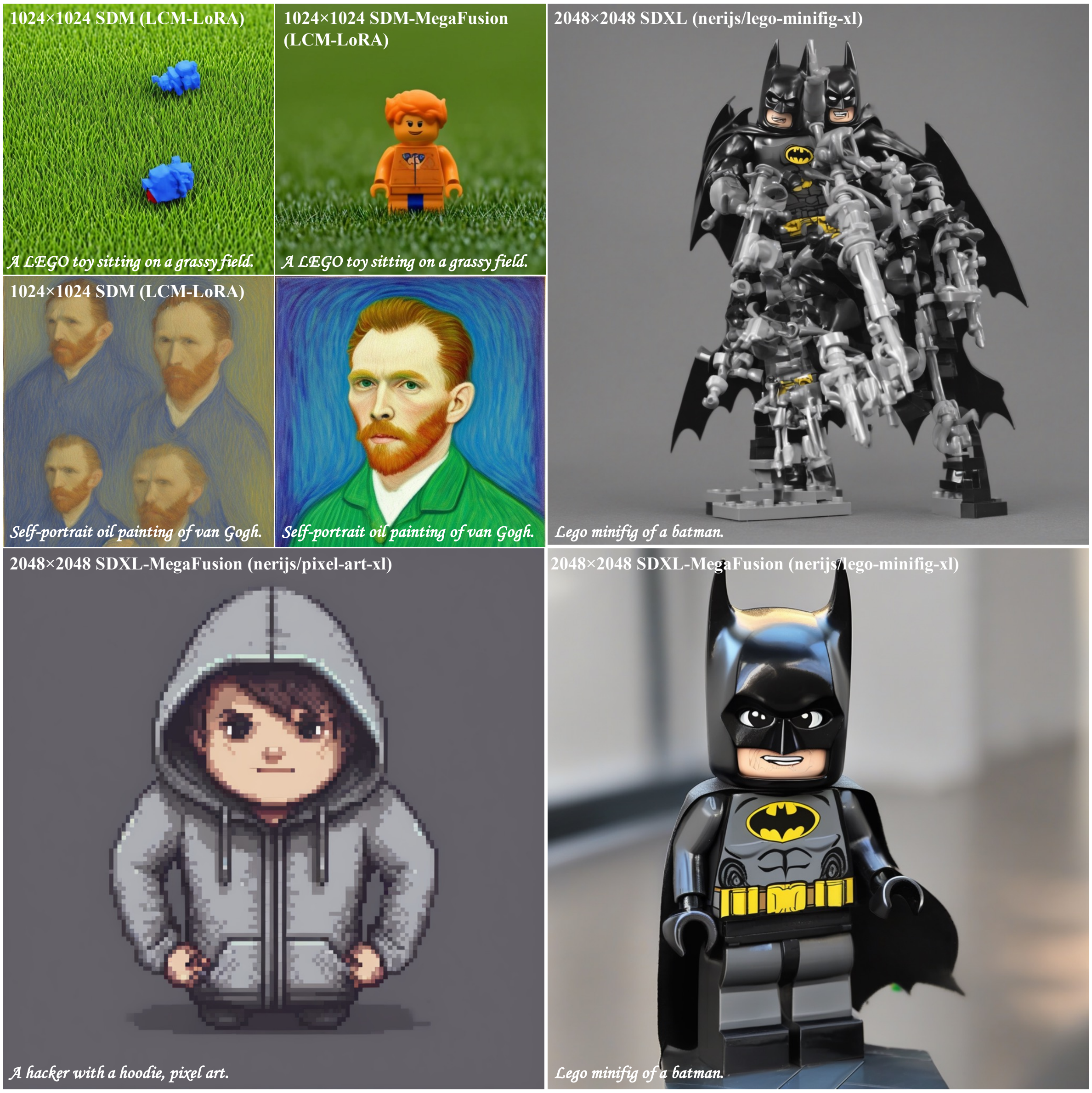

1024×1024 SDM (LCM-LoRA)
A LEGO toy sitting on a grassy field.
1024×1024 SDM-MegaFusion (LCM-LoRA)
A LEGO toy sitting on a grassy field.
2048×2048 SDXL (nerijs/lego-minifig-xl)
Lego minifig of a batman.
1024×1024 SDM (LCM-LoRA)
Self-portrait oil painting of van Gogh.
Self-portrait oil painting of van Gogh.
2048×2048 SDXL-MegaFusion (nerijs/pixel-art-xl)
A hacker with a hoodie, pixel art.
2048×2048 SDXL-MegaFusion (nerijs/lego-minifig-xl)
Lego minifig of a batman.

Figure 17. **Qualitative results** of applying MegaFusion to high-resolution image generation with LoRA-integrated SDM and SDXL. Similarly, SDM and SDXL integrated with LoRA also face common challenges like semantic deviations and object repetitions in high-resolution generation, while MegaFusion effectively addresses these challenges.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[2] Deepfloyd. Deepfloyd. *URL https://www.deepfloyd.ai/.*, 2023. 1, 2, 7

[3] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 4, 9

[4] 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 *Proceedings of the International Conference on Machine Learning*, 2024. 4, 8

[5] 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 *Proceedings of the International Conference on Learning Representations*, 2023. 4, 9

[6] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. In *Proceedings of the International Conference on Machine Learning*, 2023. 3

[7] 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 *Proceedings of the European Conference on Computer Vision*, 2014. 1, 6, 7, 8

[8] 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. In *Proceedings of the International Conference on Learning Representations*, 2024. 2, 11, 12, 13, 14, 15, 16

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 6

[10] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 1, 2

[11] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 4

[12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the International Conference on Computer Vision*, 2023. 4, 10