

Supplementary Materials: When ControlNet Meets Inexplicit Masks

Anonymous Authors

APPENDIX

This appendix is organized as follows:

- Additional visualized examples with detailed illustrations are presented to supplement the main paper. (§A)
- We provide additional results on other kinds of deteriorated conditions, *i.e.*, quantitative results on degraded edges. (§B)
- More discussions are provided, including the difference between our work and Layout-to-Image generation. (§C)
- We give a brief discussion about future works. (§D)

ID	Reference (main paper)	Brief Illustration
Fig.S1	§1, Line 147	More examples of ControlNet and T2I-adaptor to illustrate their contour-following ability.
Fig.S2	§4.2, Line 385	Visualization of the inductive bias of dilation radius r .
Fig.S3	§4.3, Line 431, 492	Visual results for tuning hyperparameters.
Fig.S4	§6.2, Line 677	More comparison examples of our method and ControlNet.
Fig.S5	§6.3, Line 695	Error analysis for the deterioration estimator.
Fig.S6	§6.5, Line 848	Generation with TikZ sketches.
Fig.S7	§6.5, Line 848	Generation with user scribbles.
Fig.S8	§6.5, Line 848	Examples for modifying shape priors.
Fig.S9	§6.5, Line 848	Examples for composable shape-controllable generation.
Tab.S2	§6.4, Line 822	Quantitative results of our method on degraded edges.

Table S1: Quick overview of figures and tables in the Appendix.

A ADDITIONAL QUALITATIVE EXAMPLES

We provide additional qualitative results to supplement the main paper. Tab. S1 gives a quick overview of all figures presented in the Appendix and indicates where they are referenced in the main paper for supplementary. Details are as follows.

Implication of dilation radius r . We further visualize examples of ControlNet- \mathbf{m}_r with conditional masks \mathbf{m}_r of varying deterioration degrees in Fig. S2. Note that ControlNet- \mathbf{m}_r is trained with dilated masks \mathbf{m}_r . As depicted in Fig. S2, ControlNet- \mathbf{m}_r implicitly assumes the presence of dilation in the provided masks, even for precise masks. For instance, the zebra generated by ControlNet- \mathbf{m}_r with conditional mask \mathbf{m}_r (in red boxes) shares similar size and shape

with the precise mask \mathbf{m}_0 , which explains the high CR scores in Fig.3 in our main paper. In contrast, our Shape-aware ControlNet robustly interprets all deteriorated conditions thanks to additional control over the shape priors.

Visualized examples of tuning hyperparameters. Fig. S3 presents visualized results of the vanilla ControlNet under different hyperparameter settings, including the CFG scale ω , conditioning scale λ , and condition injection strategy in our main paper. The CFG scale ω exhibits minimal impact on the contour-following effect but enhances image fidelity with higher saturation. Lowering the conditioning scale λ and omitting conditions in the early reverse sampling stages alleviate the strong contour instructions. However, because of the significant performance gaps between explicit and inexplicit conditions, bridging this gap and obtaining satisfactory results solely through hyperparameter tuning is challenging. Moreover, sudden appearance changes usually occur when we adjust hyperparameters, making it tricky to find the optimal setting. Therefore, it is necessary to develop a method to robustly interpret inexplicit masks.

More visualized comparison with the vanilla ControlNet. In Fig. S4, we provide more examples of our Shape-aware ControlNet compared with the vanilla ControlNet. Our method not only achieves competitive results with ControlNet on precise masks but also demonstrates an enhanced capability to interpret object shape and pose from inexplicit masks of varying deterioration degrees.

Full results of error analysis for the deterioration estimator. Fig. S5 reports the full results of the estimation error for the proposed deterioration estimator. The overall averaged L1 error under different dilation radius r is 5.47%. We confirm that this error has little impact on the performance including CLIP-score and FID due to the robustness of the shape-prior modulation block, as discussed in the main paper § 6.3.

More application examples. We showcase additional examples of our method in three application scenarios, including robust generation with TikZ sketches (Fig. S6) following Control-GPT [7], human scribbles (Fig. S7) converted from Sketchy [5] dataset, modification of shape priors (Fig. S8), and composable shape-controllable generation (Fig. S9). These examples verify that our method helps extend ControlNet to creative applications with more flexible control signals, owing to controllable shape priors.

B QUANTITATIVE RESULTS WITH DETERIORATED EDGES

Though we mainly focus on one representative condition, *i.e.*, object masks, in the main paper for illustration, our method can be applied to other kinds of degraded conditions with little effort as discussed in the main paper § 6.4.

Here we provide quantitative results for supplementary to further prove the effectiveness of our method on more condition types,

Metric	Method	r						AVG
		0	20	40	80	100	∞	
CLIP-Score	ControlNet	26.77	26.22	25.77	25.53	25.49	25.49	25.88
	Ours	26.83	26.85	26.88	26.85	26.87	26.87	26.86
FID	ControlNet	13.51	16.54	19.45	21.84	22.50	22.97	19.48
	Ours	13.80	14.52	15.05	15.84	15.92	16.01	15.19
LC	ControlNet	0.521	0.429	0.361	0.310	0.299	0.279	0.367
	Ours	0.510	0.494	0.476	0.456	0.452	0.434	0.509
SR	ControlNet	0.607	0.533	0.496	0.477	0.474	0.469	0.470
	Ours	0.599	0.582	0.572	0.563	0.563	0.556	0.572

Table S2: Performance comparison of our method and the vanilla ControlNet, under the condition of degraded edges of different dilation radius r . The results prove the effectiveness of our method in other kinds of degraded conditions besides masks.

i.e., deteriorated edges. Tab. S2 reports the CLIP-score, FID, Layout Consistency (LC), and Semantic Retrieval (SR) following the same setting as the main paper § 6.1. Compared with the vanilla ControlNet, our method not only obtains competitive results with accurate edges but also exhibits robust performance with degraded edges of varying degrees. These results demonstrate the advance of our method in handling diverse kinds of inexplicit conditions. More types of conditions are left for future works.

C MORE DISCUSSIONS

Difference with the Layout-to-Image Generation. One may notice that when the dilation radius $r = \infty$, our Shape-aware ControlNet addresses a similar problem to Layout-to-Image (L2I) generation. But we claim that there exist subtle differences between L2I and our work. For a better understanding, explanations are as follows.

Layout-to-Image generation works, such as GLIGEN [2], LayoutDiffuse [1], LayoutDiffusion [8], *etc.*, aim to generate objects conforming to the pre-defined layouts. Generally, L2I assigns fine-grained object positions in a tuple format of $(x, y, width, height, label)$, thereby associating each object to the specific bounding box. However, in our setting, only prompts and binary masks are available, which makes it hard to achieve fine-grained position control over each object like L2I. Fig. S10 compares our method with LayoutDiffuse [1] for illustrations. While our method adheres to the global layout provided by bounding-box masks, the exact positions of each object are not fixed. A recent work [3] explores adapting ControlNet to L2I generation, thus it is promising to extend our method to fine-grained layout control. However, as this is not the main topic of this paper, we leave it for future work. Our main purpose is to visit the contour-following ability of ControlNet, and we reveal a solution to adapt ControlNet to deteriorated masks at varying precision levels.

D FUTURE WORKS

By far, we have demonstrated the effectiveness of our method in improving ControlNet in dealing with inexplicit conditions, which is an essential topic in applying ControlNet to practical usage. Moreover, we have noted that the dramatic performance degradation caused by inexplicit conditions is a common issue among methods that inject

spatially aligned control signals for spatially controllable T2I generation. The results of T2I-Adapter [4] are shown in Fig. S1, where large deterioration in the control masks would pose challenges in correctly understanding the spatial control signals. Since our method takes no assumption about the network that injects control signals into the generation process, it has the potential to extend to other adapter-based methods like T2I-Adapter to avoid deviation of spatial control brought by inexplicit conditions of varying deterioration degrees. We leave this as our future work.

REFERENCES

- [1] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. 2023. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908* (2023).
- [2] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.
- [3] Denis Lukovnikov and Asja Fischer. 2024. Layout-to-Image Generation with Localized Descriptions using ControlNet with Cross-Attention Control. *arXiv preprint arXiv:2402.13404* (2024).
- [4] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.
- [5] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.
- [6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
- [7] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. 2023. Controllable Text-to-Image Generation with GPT-4. *arXiv preprint arXiv:2305.18583* (2023).
- [8] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22490–22499.

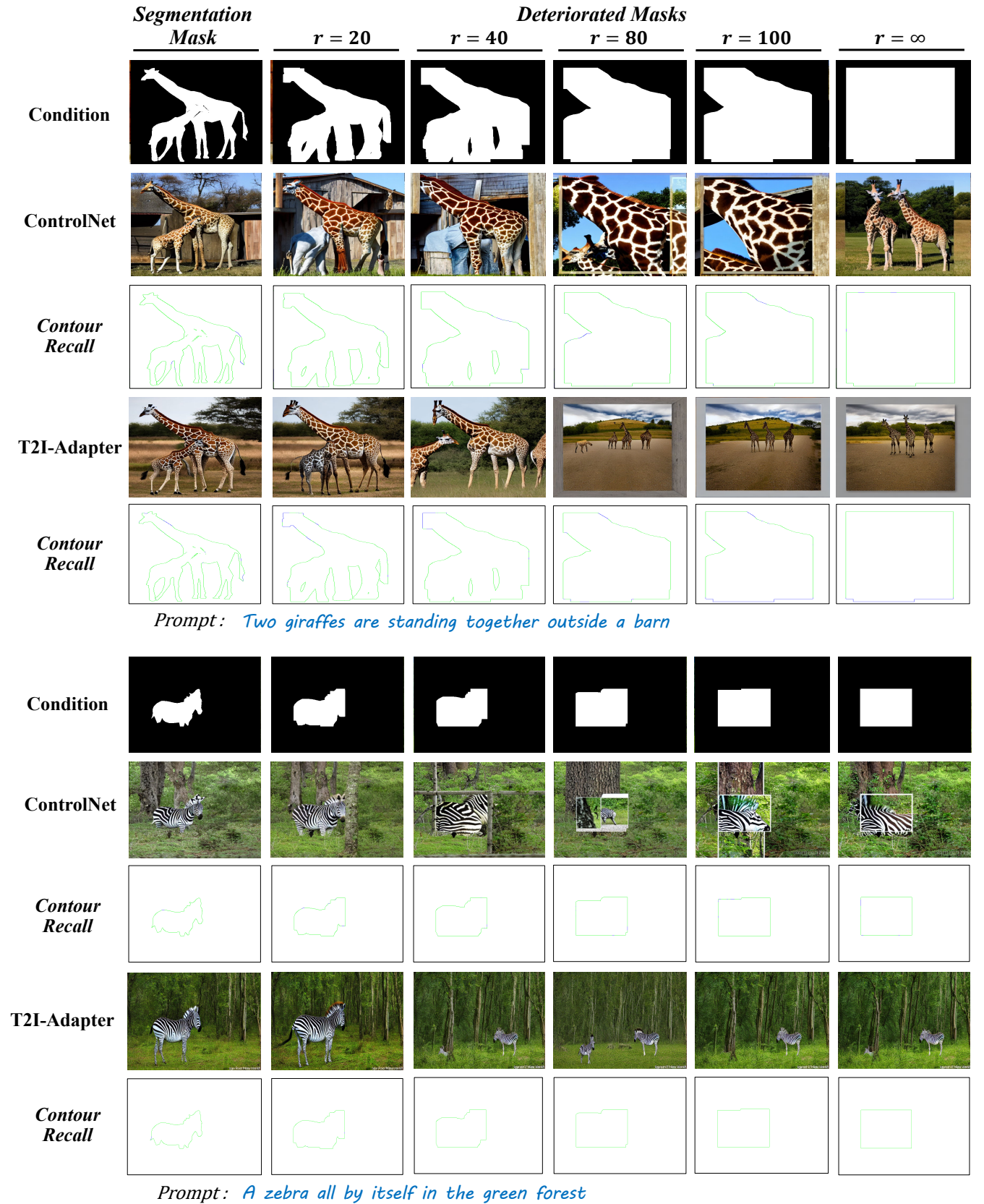
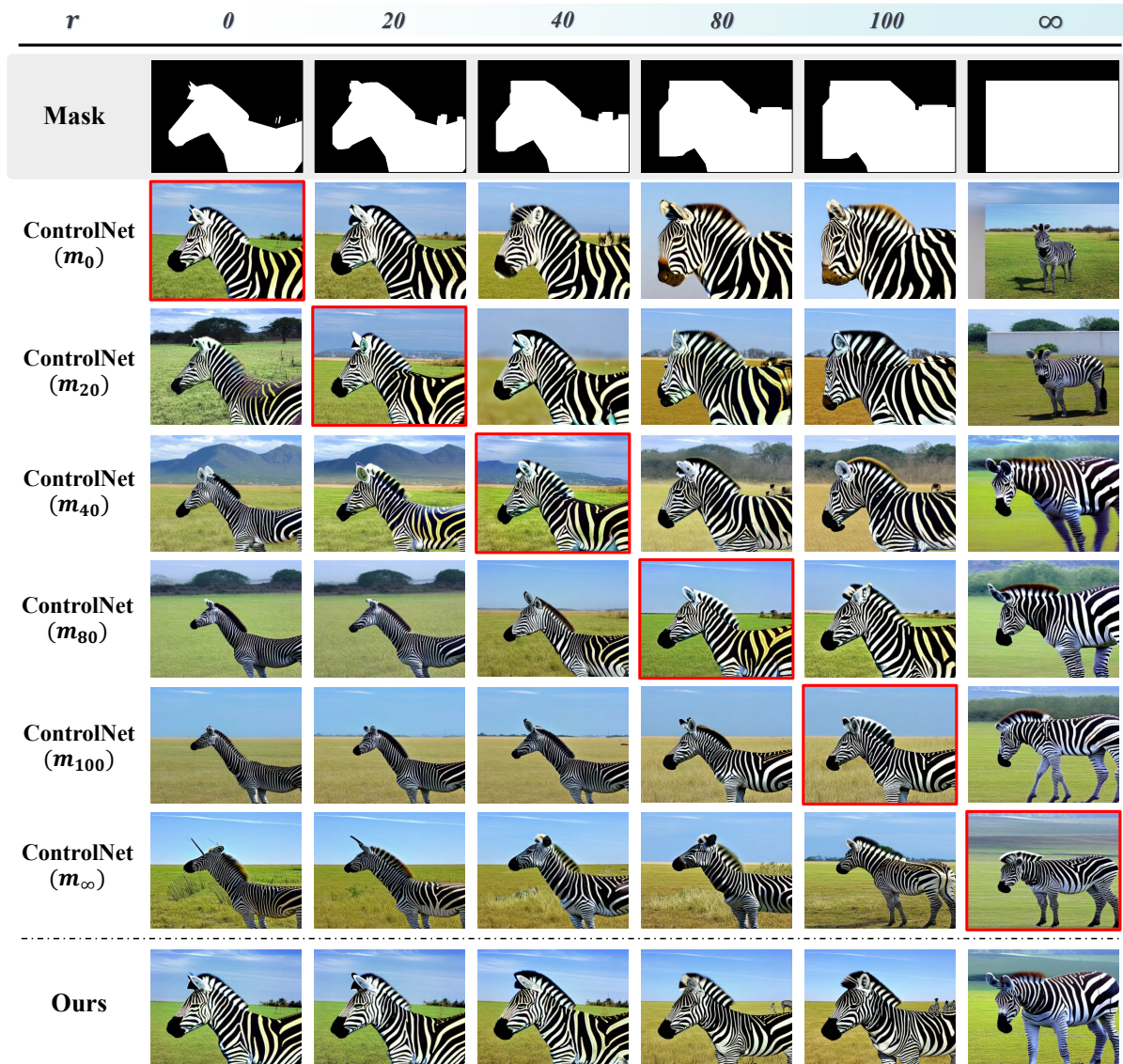


Figure S1: More examples of ControlNet and T2I-adapter illustrating spatially controllable generation with deteriorated masks of varying degrees. Both methods exhibit strong preservation of contours during the generation process. **Green** denotes the recalled contours and **blue** denotes the missing ones. Moreover, inexplicit masks with inaccurate contours would cause severe degradation of image fidelity and realism.



Prompt: *A zebra standing on top of a grass covered field*

Figure S2: Visualized examples to illustrate the inductive bias, where ControlNet- m_r implicitly learns the dilated radius r from the data. Therefore, it always assumes a dilation in the provided mask. The ControlNet- m_r conditioned on m_r (in red boxes) produces objects with similar shapes and sizes of mask m_0 . While ControlNet- m_∞ also breaks the inductive bias, our method surpasses it in terms of image fidelity and additional control over the shape priors.

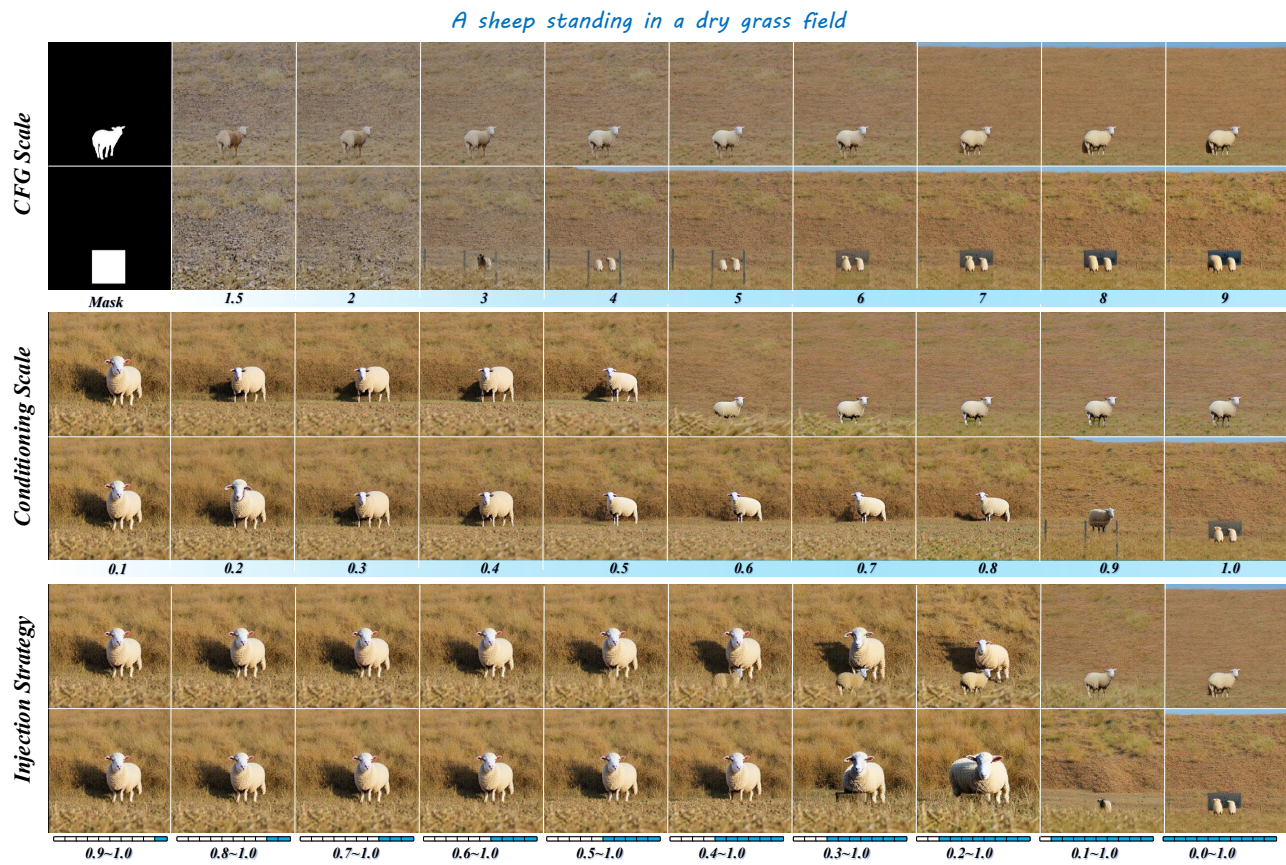


Figure S3: Visualized examples under different hyperparameter settings, *i.e.*, CFG scale, conditioning scale, and condition injection strategy. Though the CFG scale shows little impact on the contour-following ability, reducing conditioning scales and discarding conditions at early reverse sampling stages help to relieve contour instructions, resulting in better results with inexplicit conditional masks. However, it is still tricky and hard to achieve satisfactory results through hyperparameter tuning, especially for deteriorated control masks.



Figure S4: More examples of our Shape-aware ControlNet compared with the vanilla ControlNet, i.e., ControlNet- m_0 and ControlNet- m_∞ given the conditional mask m_r . Our method exhibits robust performance on inexplicit masks of varying deterioration degrees.

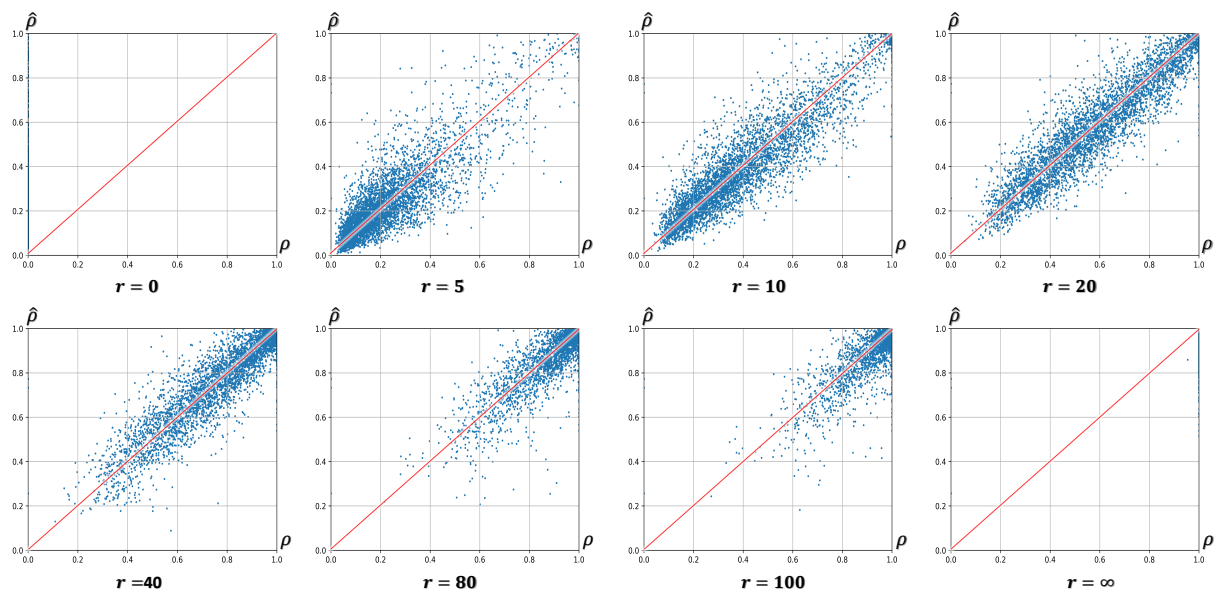


Figure S5: Error analysis of our deterioration estimator at different dilation radius r . The overall average $L1$ error is 5.47%. Notably, such errors show little impact on the performance of our Shape-aware ControlNet, referring to our main paper § 6.3.

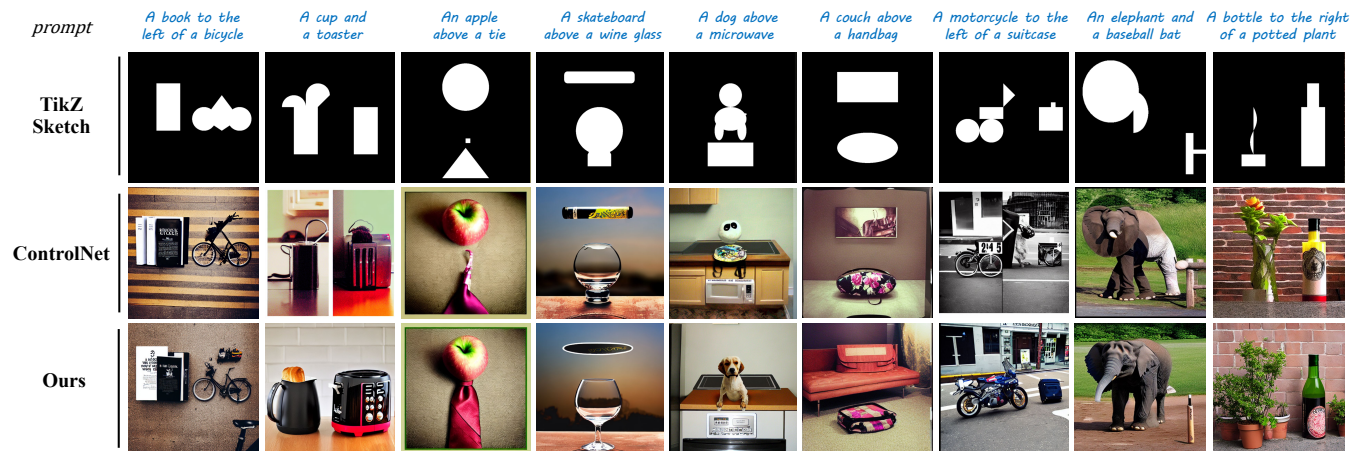


Figure S6: More examples of our method on TikZ sketches compared with the vanilla ControlNet. Though our Shape-aware ControlNet is only trained with dilated masks, it generalizes well to abstract programmatic masks with inaccurate contours and exhibits advanced performance.

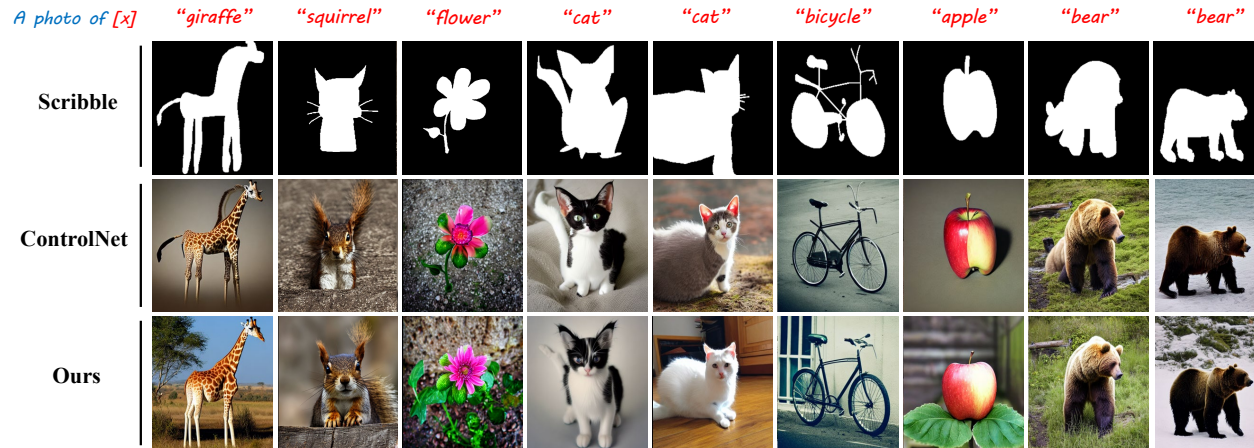


Figure S7: More examples of our method on human scribbles compared with the vanilla ControlNet. Our Shape-aware ControlNet exhibits advanced and robust performance on realistic user-provided masks with inaccurate contours.

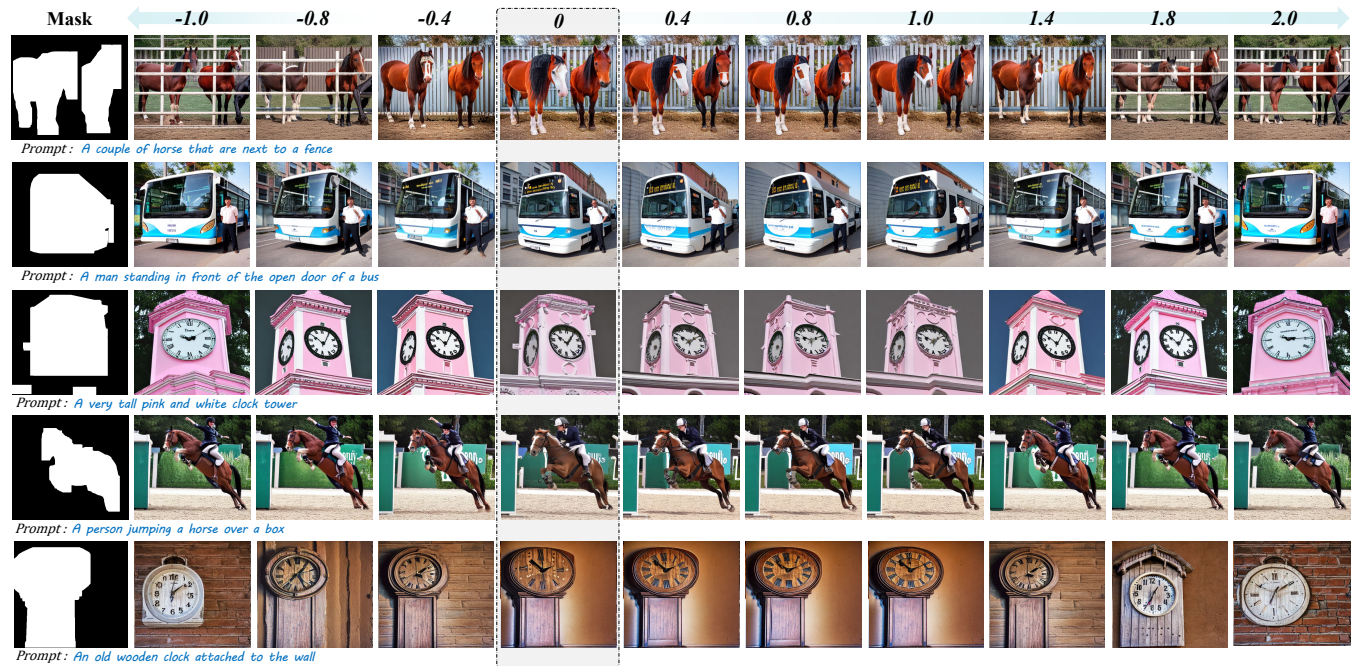


Figure S8: More examples of controlling object shape via the deterioration ratio ρ . The value of ρ is depicted on the top row. Despite training with $\rho \in [0, 1]$, we empirically find it generalizes well to other values and offers additional control over the shape of generated objects. A small $|\rho|$ encourages generated objects to adhere more tightly to the outlines, and vice versa. Moreover, note that adjusting ρ has little impact on image fidelity.



Figure S9: More examples of composable shape-controllable generation with our method. Leveraging a Multi-ControlNet structure [6], we can assign different priors to each part of the control masks. This enables strict shape control over specific masks, while simultaneously allowing T2I diffusion models to unleash creativity in imagining objects of diverse shapes within inexplicit masks.



Figure S10: Comparison with L2I method, i.e., LayoutDiffuse [1]. While our method achieves spatial control over the global layout, it may confuse the fine-grained layout for each object, which differs from the Layout-to-Image generation tasks.