Beyond Textual Constraints: Learning Novel Diffusion Conditions with Fewer Examples –Supplementary Materials–

Yuyang Yu [1*] Bangzhen Liu [1*] Chenxi Zheng [1]
Xuemiao Xu[1,2,3,4†] Huaidong Zhang [1†] Shengfeng He [5]
[1]South China University of Technology [2]State Key Laboratory of Subtropical Building Science
[3]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[4]Ministry of Education Key Laboratory of Big Data and Intelligent Robot [5]Singapore Management University
{yyyoung0611, liubz.scut, chansey0529}@gmail.com, {xuemx, huaidongz}@scut.edu.cn, shengfenghe@smu.edu.sg

This supplementary file contains the following six sections: Sec. 1 further validates and explains the motivation of our proposed method through two key experiments. Sec. 2 provides additional quantitative experimental results. Sec. 3 demonstrates the application of our method in scenarios where large-scale datasets do not exist. Sec. 4 includes more ablation experiments and their analyses. Sec. 5 presents additional visual results. Sec. 6 summarizes the limitations of our method and future directions for development.

1. Intuition and explanations of our proposal

In this section, we present two key experiments that demonstrate the intuition behind our proposal and provide detailed explanations. Regarding the first key experiment, we investigated sketch-conditioned generation using ControlNet with frozen cross-attention layers (the parameters of the cross-attention layers in ControlNet remain unchanged, which are the same as those of the corresponding cross-attention layers in Stable Diffusion.). Regarding the second key experiment, we compare the influence of feeding the correct prompt to Stable Diffusion and a generated incorrect prompt to ControlNet during inference.

The visualization results of Fig. 1 validated our approach and highlighted how text prompts negatively impact learning new conditions in few-shot scenarios. As shown in the Fig. 1b, there exists an obvious misalignment between the generated results and the given sketch when we trained ControlNet with frozen cross-attention layers on 100 pairs of training samples. These results indicate the negative impacts of text condition on learning new conditions, especially when we only have very little training data. Comparing the results from Fig. 1c to Fig. 1f, the inference results demonstrate that even when updating all parameters of ControlNet during the training process, the problem of the cross-modal gap between the text and the new conditions still exists, which is even more severe under the few-shot scenarios.

2. More Quantitative Results

In this section, we provide additional quantitative results, including an evaluation of computational cost, an evaluation of text-image alignment, and a user study.

Computational Cost. In Table 1, we report the value of trainable parameters, GPU memory, FLOPs, and inference time. Compared to ControlNet, our method can achieve better results with lower computational cost.

Methods	Params.	Mem.	FLOPs	T_{infer}
ControlNet [14]	1378.17 MB	**8.8 G**	455.2 G	10s
T2iAdapter [11]	**295.14 MB**	10.5 G	**338.7 G**	**5s**
Ours	964.57 MB	9.5 G	441.4 G	10s

Table 1. The results of the Computational Cost. The best result is indicated in **bold**, while the second-best result is underlined.

Evaluation for Text-Image Alignment. We evaluate text-image alignment scores using standard CLIP metrics for sketch-conditioned (COCO17) and segmentation-conditioned (COCO-stuff) generation. We used the "ViT-L/14" CLIP model for training and "ViT-B/16" for evaluation. Results in Table 2 demonstrate our method's effectiveness in improving alignment between conditions and image contents, while maintaining text-image correspondence.

Conditions \ Methods	ControlNet-100	T2IAdapter-100	Ours wo. CNR	Ours
Sketch	28.86	29.89	30.64	**30.84**
Segmentation	28.23	29.92	29.87	**30.26**

Table 2. The results of the Text-Image Alignment scores. The best result is indicated in **bold**, while the second-best result is underlined.

*The first two authors contributed equally.
†Corresponding authors (xuemx@scut.edu.cn, huaidongz@scut.edu.cn).

(a) sketch (b) FreezeCA (c) Wrong Pr.-100 (d) ControlNet-100 (e) Wrong Pr. (f) ControlNet

Figure 1. Visualization results of two key experiments. "FreezeCA" means using ControlNet with frozen cross-attention layers. "Wrong Pr" means inputting the correct prompt to Stable Diffusion and inputting an incorrect prompt to ControlNet during inference. "-100" means that the model was trained on a training set consisting of 100 pairs of training samples.

User Study. In our user study, we assess the generation quality (qual.) and condition-content consistency (cons.) across five conditional-generation tasks. Participants were asked to rank the compared methods on 10 examples per task, evaluating them from best to worst in terms of quality and consistency. As shown in Table 3, We received 118 effective responses, which collectively highlighted the superiority of our method.

Methods \ Conditions	Sketch		Seg. map		Depth		Edge		Pose	
	qual.	cons.	qual.	cons.	qual.	cons.	qual.	cons.	qual.	cons.
ControlNet-100	2.58	1.91	2.56	2.31	2.47	2.19	3.03	2.64	2.81	2.66
T2iAdapter-100	2.23	2.04	2.78	3.13	2.64	2.93	2.58	3.09	2.75	3.31
PromptDiff-100	—	—	3.41	3.34	2.86	3.66	2.95	2.81	—	—
HumanSD-100	—	—	—	—	—	—	—	—	**2.14**	2.83
Ours	**1.16**	**1.36**	**1.25**	**1.27**	**2.03**	**1.29**	**1.44**	**1.41**	2.30	**1.28**

Table 3. The results of Human Evaluation. The best result is indicated in **bold**, while the second-best result is underlined.

3. Challenge Applications

Generative applications, especially in fields with data constraints like astronomy, healthcare, and physical simulations, greatly benefit from few-shot generation. We demonstrate its utility in generating new data with limited condition samples through tasks like face-to-thermal, thermal-to-face, and hed-to-old-photo generation. For the face-to-thermal and thermal-to-face tasks, we used 100 pairs of training samples from the Thermal Faces dataset proposed in [8] as our training set. For the hed-to-old-photo generation task, we employed 100 pairs of training samples from the dataset proposed in [3, 4] as our dataset. The prompts for all the samples in the training set were obtained through BLIP2 [9]. As shown in Fig. 2, compared with ControlNet, our method effectively learns novel conditions and possesses a better generative quality, exemplifying its efficiency with limited training data. Applications include security checks and criminal investigations for thermal-based generation and training restoration models with generated old photos.

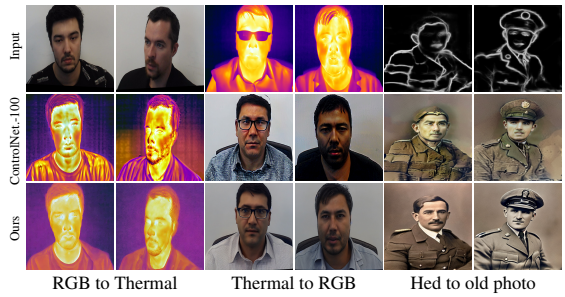


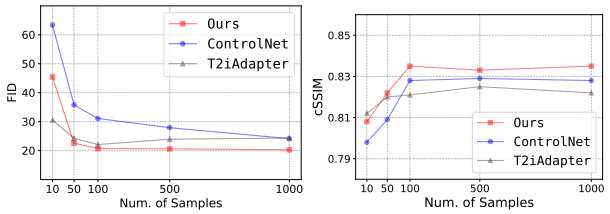
Figure 2. Visualization results for the three tasks: face-to-thermal, thermal-to-face, and hed-to-old-photo generation.

4. More Analysis on Ablation Studies

In this section, we discuss the influence of different scales of training samples on generation quality and consistency. A simple experiment for validating the robustness of our method is also conducted with different batches of training samples. We also discussed the different effects brought about by using ϵ-based prediction and x_0-based prediction during the second stage of training (**Condition-specific Negative Rectification** in the main paper).

Different Scales of Training Samples. Taking the segmentation-guided generations task on COCO-stuff [2] as an example, we randomly select 10, 50, 100, 500, and 1000 images as the training batches and compare between ControlNet [14], T2IAdapter [11], and our method. The values of FID [5] and cSSIM [12] for each method are reported in Fig. 3. A common trend is that using more training

samples can benefit the learning of new conditions. Yet our method is more effective and efficient than the other two methods in terms of both generation quality and structural consistency. Notably, our method could achieve competitive performance compared to ControlNet and T2iAdapter which are trained with 500 or even larger amounts of training samples, with only 100 training samples, further demonstrating the necessity of eliminating the textual influence.



(a) FID(↓). (b) cSSIM(↑).

Figure 3. The curves of (a)FID score(↓), and (b)cSSIM score(↑) over a varying number of training samples for segmentation-guided generations on COCO-stuff [2].

We also visualize the generated images of the three approaches using different training samples in Fig. 4. Our method could efficiently learn the structural information of the provided conditions when the number of training samples is limited to 50, while ControlNet requires 100 training examples for a roughly aligned layout and T2iAdapter requests even more training samples. The content of our generated images is also faithful and more diverse, demonstrating the efficiency and effectiveness of our approach.

Figure 4. Visual comparison over varying numbers of training samples for segmentation-guided generations on COCO-stuff [2]. The first row shows the exemplar image and the corresponding text description.

Note that all the methods failed to obtain convincible generation results when only 10 training examples were available, which is the major limitation of these works. We will leave it to our future work to explore the potential of adapting novel conditions under the guidance of extremely low-shot exemplars.

Impact of Varing Training Data Batches. To affirm the training robustness of our methodology across distinct sets of training samples, we randomly selected 10 batches of data within the COCO training dataset, each of them consisting of 100 training pairs. The subsequent training was conducted with sketch inputs as a conditioning factor. The results are reported in Fig. 5. The average FID over the 10 generation tasks is 20.891, with a standard deviation of 0.854. Additionally, the average value of the cSSIM is 0.692, accompanied by a standard deviation of 0.006.

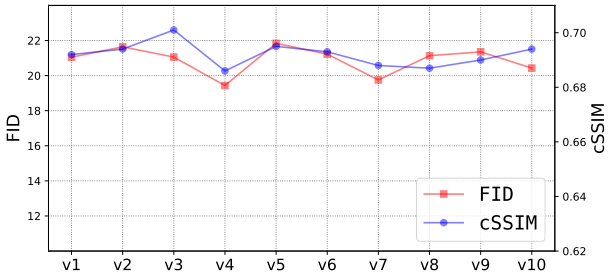


Figure 5. The curves of FID and cSSIM over varying contents of 100 training samples for sketch-guided generations on COCO [10].

Impact of the ϵ-based and x_0-based Prediction. Here we compare the visual quality of the generated results by inferencing with the two general diffusion score functions. As shown in Eq. 2 in the main paper, the ground truth of $\tilde{\varepsilon}_\theta$ is not standard Gaussian noise ε. Utilizing standard diffusion loss to bias $\tilde{\varepsilon}_\theta$ towards ε may result in decreased image quality. As shown in the Fig. 6, we performed additional experiments using the standard ϵ-based loss, which leads to deteriorated generation quality. Consequently, we adopt the x_0-based loss as the optimizing objective of our framework.

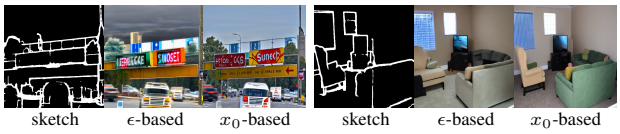


Figure 6. Visual comparison between the ϵ-based prediction and x_0-based prediction

5. Additional Visualizations

We provide more comparison of text-to-image generation results conditioned on sketch (Fig. 7), segmentation (Fig. 8), depth (Fig. 9), pose (Fig. 11), and edge (Fig. 10). The compared methods including ControlNet-1.0 [14], T2IAdapter [11], HumanSD [6], and PromptDiff [13].

More diverse generation results of sketch, segmentation, depth, and edge, are shown in Fig. 12, Fig. 13, Fig. 14, and Fig. 15, respectively. The contents of the generated images are realistic and visually pleasing, while maintaining a highly

structural consistency with the provided condition exemplar and semantic consilience with the given text description.

For each picture, we recommend zooming in for more detailed visualization.

6. Limitations and Future Work

In this section, we briefly discuss the potential limitations and future influence of our work.

Limitations. While we have significantly reduced the data sample requirements during fine-tuning, our framework still demands up to 100 condition-image pairs, which might remain challenging for certain scenarios. Additionally, as our non-unified framework requires separate training for each condition, its scalability to multiple modalities is limited.

Future Works. We will continue investigating lightweight conditioning techniques to reduce data costs for "handcrafting" text-to-image models and enhance their control and scalability.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In CVPR, pages 18392–18402, 2023. 6

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In CVPR, pages 1209–1218, 2018. 2, 3, 5

[3] Weiwei Cai, Xuemiao Xu, Jiajia Xu, Huaidong Zhang, Haoxin Yang, Kun Zhang, and Shengfeng He. Hierarchical damage correlations for old photo restoration. Information Fusion, page 102340, 2024. 2

[4] Weiwei Cai, Huaidong Zhang, Xuemiao Xu, Shengfeng He, Kun Zhang, and Jing Qin. Contextual-assisted scratched photo restoration. IEEE Transactions on Circuits and Systems for Video Technology, 2023. 2

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, volume 30, 2017. 2

[6] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In ICCV, 2023. 3, 7

[7] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In CVPR, pages 618–629, 2023. 7

[8] Askat Kuzdeuov, Dana Aubakirova, Darina Koishigarina, and Huseyin Atakan Varol. Tfw: Annotated thermal faces in the wild dataset. IEEE Transactions on Information Forensics and Security, 17:2084–2094, 2022. 2

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023. 2

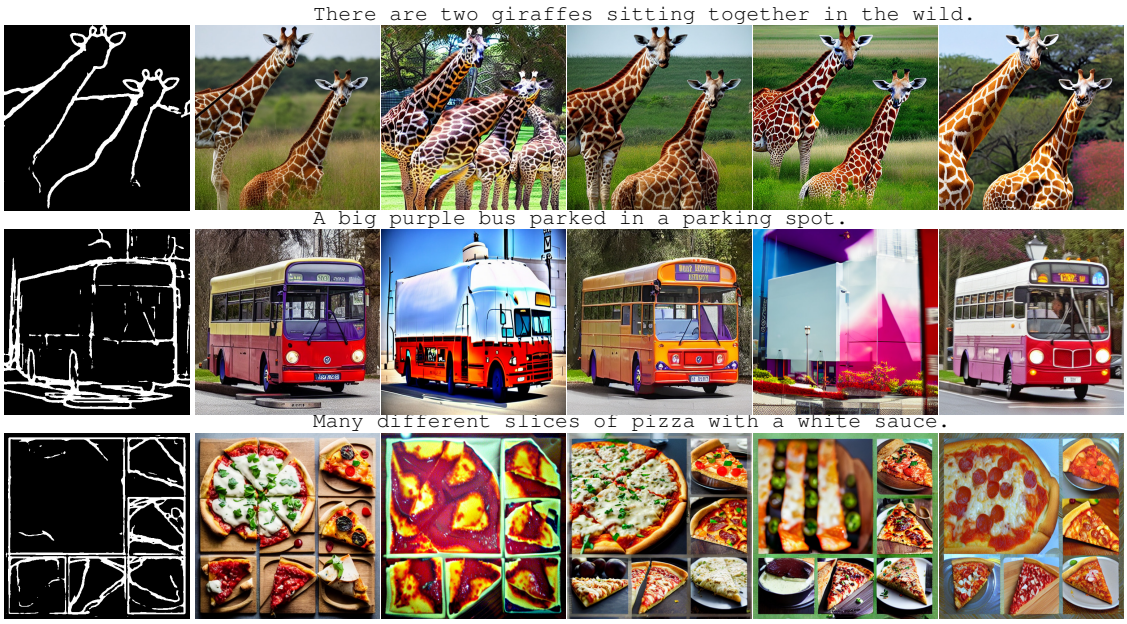
[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014. 3, 5

[11] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453, 2023. 1, 2, 3, 5, 6, 7

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE TIP, 13(4):600–612, 2004. 2

[13] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. arXiv preprint arXiv:2305.01115, 2023. 3, 5, 6

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, pages 3836–3847, 2023. 1, 2, 3, 5, 6, 7

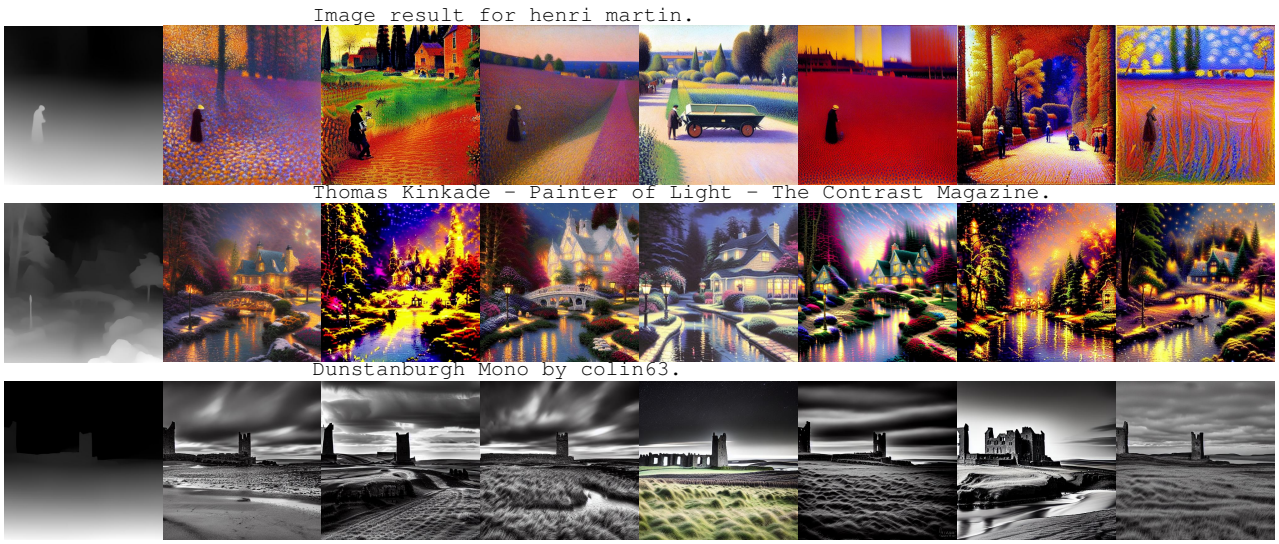


Conditions (a) ControlNet [14] (b) ControlNet-100 (c) T2IAdapter [11] (d) T2IAdapter-100 (e) Ours

Figure 7. Comparisons of sketch-guided text-to-image generations on COCO [10].

Conditions (a) ControlNet [14] (b) ControlNet-100 (c) T2IAdapter [11] (d) T2IAdapter-100 (e) PromptDiff [13] (f) PromptDiff-100 (g) Ours

Figure 8. Comparisons of mask-guided text-to-image generations on COCO-stuff [2].



Conditions (a) ControlNet [14] (b) ControlNet-100 (c) T2IAdapter [11] (d) T2IAdapter-100 (e) PromptDiff [13] (f) PromptDiff-100 (g) Ours

Figure 9. Comparisons of depth-guided text-to-image generations on InstructPix2Pix [1].

Conditions (a) ControlNet [14] (b) ControlNet-100 (c) T2IAdapter [11] (d) T2IAdapter-100 (e) PromptDiff [13] (f) PromptDiff-100 (g) Ours

Figure 10. Comparisons of edge-guided text-to-image generations on InstructPix2Pix [1].

Figure 11. Comparisons of pose-guided text-to-image generations on HumanArt [7].

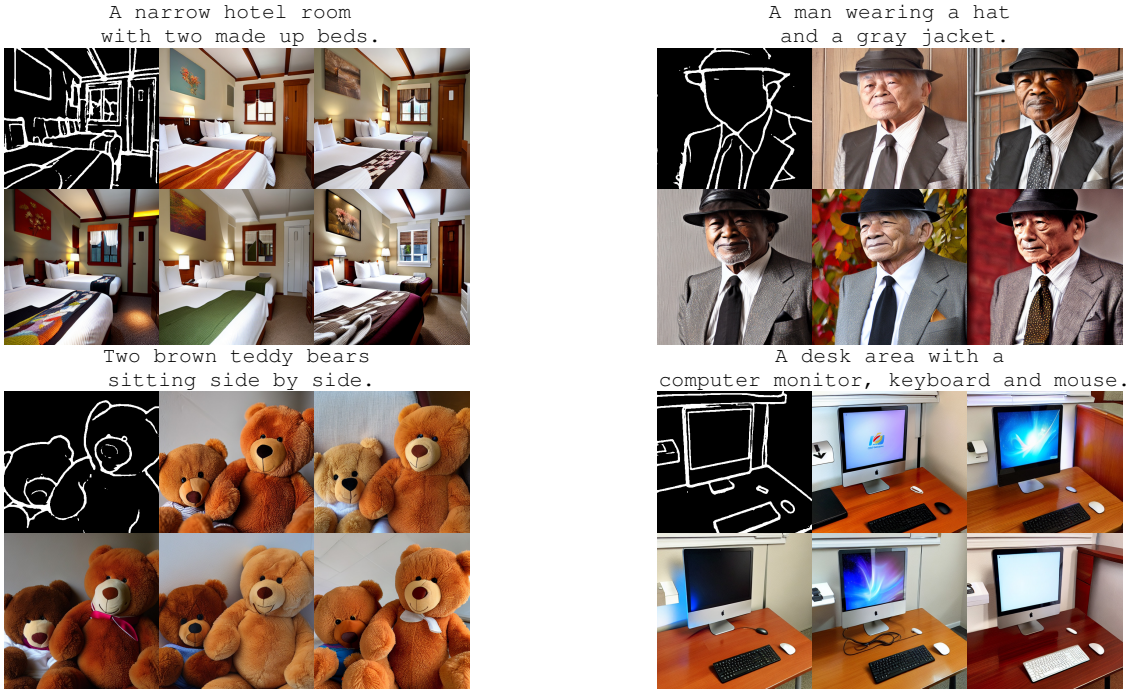


Figure 12. Our results of diverse generation conditioned on sketch.

Figure 13. Our results of diverse generation conditioned on segmentation map.

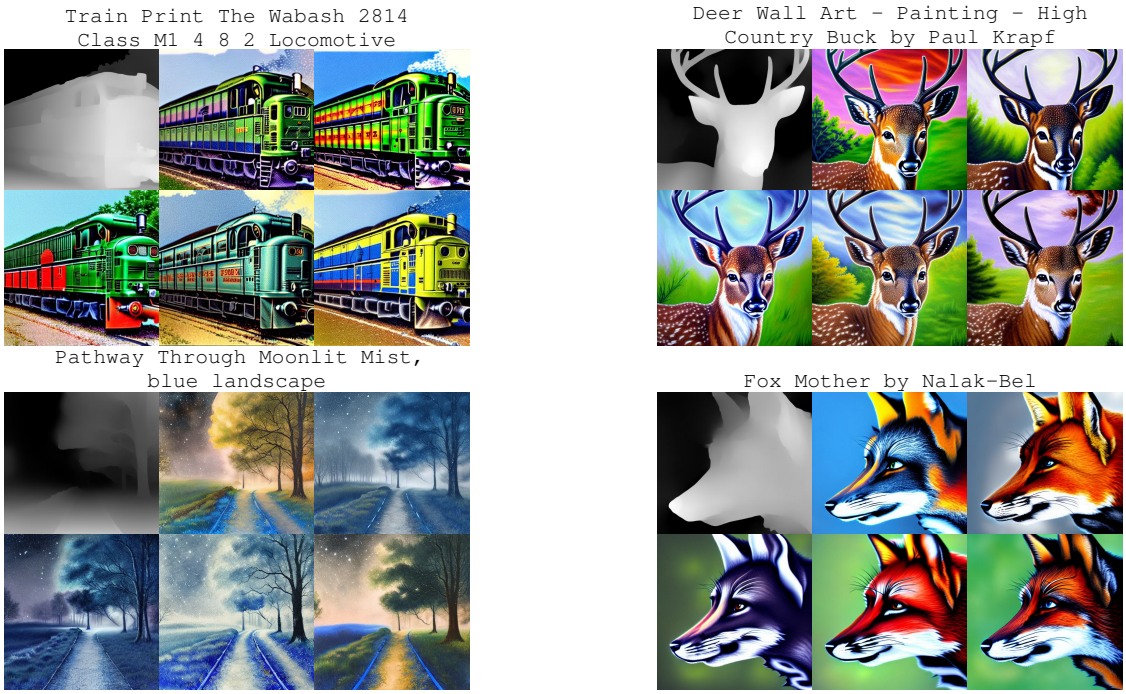


Figure 14. Our results of diverse generation conditioned on depth map.

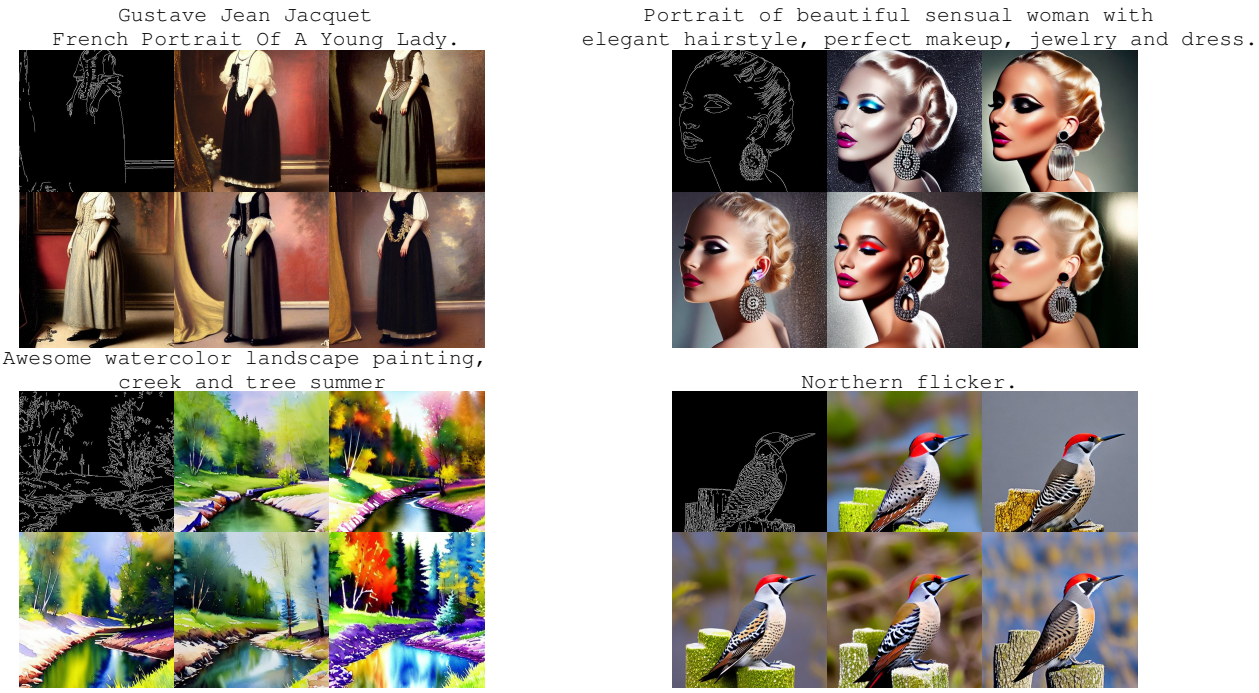


Figure 15. Our results of diverse generation conditioned on edge map.