Null-text Guidance in Diffusion Models is Secretly a Cartoon-style Creator

Anonymous author

Figure 1: Image cartoonization with the proposed Image Disturbance (Image-D) strategy. In each pair, left $\overset{cartoonize}{\longrightarrow}$ right.

ABSTRACT

Classifier-free guidance is an effective sampling technique in diffusion models that has been widely adopted. The main idea is to extrapolate the model in the direction of text guidance and away from null-text guidance. In this paper, we demonstrate that null-text guidance in diffusion models is secretly a cartoon-style creator, i.e., the generated images can be efficiently transformed into cartoons by simply perturbing the null-text guidance. Specifically, we proposed two disturbance methods, i.e., Rollback disturbance (Back-D) and Image disturbance (Image-D), to construct misalignment between the noisy images used for predicting null-text guidance and text guidance (subsequently referred to as **null-text noisy image** and **text noisy image** respectively) in the sampling process. Back-D achieves cartoonization by altering the noise level of null-text noisy image via replacing x_t with $x_{t+\Delta t}$. Image-D, alternatively, produces high-fidelity, diverse cartoons by defining x_t as a clean input image, which further improves the incorporation of finer image details. Through comprehensive experiments, we delved into the principle of noise disturbing for null-text and uncovered that the efficacy of disturbance depends on the correlation between the null-text noisy image and the source image. Moreover, our proposed techniques, which can generate cartoon images and cartoonize specific ones, are training-free and easily integrated as a plug-and-play component in any classifier-free guided diffusion model. Project page is available at https://nulltextforcartoon.github.io/.

KEYWORDS

Classifier-free guidance, Cartoonization, Null-text guidance,Diffusion models

1 INTRODUCTION

Diffusion models [4, 6, 12, 18, 22, 27, 29, 34] have recently emerged as a compelling topic in computer vision and have shown remarkable results in the field of generative modeling [12, 28, 30], which generate samples by gradually removing noise from a signal with the training objective expressed as a reweighted variational lower-bound [12]. Based on diffusion models [12, 18, 31], Dhariwal et al. proposed classifier guidance [6], which boosts sample quality but requires extra trained classifier. Classifier-free guidance [13] is an alternative technique that modifies predict noise without a classifier by extrapolating the model in the direction of text guidance and away from null-text guidance. This technique has significantly improved the image generation capability in diffusion model such that it has gained wide adoption in recent works [22, 25].

To investigate the influence of null-text guidance in classifier-free guidance, we introduced a misalignment between the noisy images used for predicting null-text guidance and text guidance (subsequently referred to as **null-text noisy image** and **text noisy image** respectively) by employing noise disturbance anchored by text guidance. Our findings indicates that null-text guidance in diffusion models is secretly a cartoon-style creator. Specifically, we propose two noise disturbance methods, i.e., Rollback disturbance

(Back-D) and Image disturbance (Image-D), to achieve the misalignment of noisy images. Back-D involves modifying the noise level of null-text noisy image by replacing x_t with $x_{t+\Delta t}$, resulting in a final output resembling a cartoon. On the other hand, Image-D uses a given clean image as the null-text noisy image, enabling the capture of finer image details and producing high-fidelity cartoon images with increased diversity.

We systematically investigated and analyzed the impact of various hyper-parameters on the proposed methods, elucidating the appropriate conditions for effective cartoonization. Our exploration discovered the following insights: 1) The null-text noisy image that is more closely related to the input image can lead to improved cartoonization outcomes. 2) Text guidance has the potential to enhance the diversity and creativity of generated cartoons, as a result of the image-generating capacity inherent in the diffusion model. 3) For effective cartoonization, the noise disturbance needs to form a stable direction for image generation, requiring the number of sampling steps empirically set to over 100.

The experimental findings demonstrate that the proposed methods facilitate the generation of cartoon depictions encompassing portraits, animals, landscapes and architectures, among other entities, whilst also enabling creative cartoonization of specific images. Remarkably, our approach obviates the necessity for training, thereby facilitating simplified implementation in classifier-free guided diffusion model. In summary, this work makes pioneering contributions in the following ways:

- We conducted an in-depth exploratory analysis of null-text guidance and discovered that null-text guidance is secretly a cartoon-style creator.
- we put forth plug-and-play cartoonization components, including Rollback disturbance (Back-D) and Image disturbance (Image-D), that enable free generation of cartoons as well as the cartoonization of specific input images.
- By means of experiments, we have extracted three primary insights as described above regarding null-text guidance, thereby contributing further understanding of the principles and potential applications of classifier-free guidance.

2 RELATED WORKS

2.1 Classifier-free guidance

Classifier-free guidance [13] is a powerful sampling technique as it directs the model towards text guidance and away from null-text guidance by introducing a null-text guidance term. Compared to the previous study, classifier guidance [6], which utilizes a separate classifier to trades off Inception Score (IS) and Fréchet Inception Distance (FID) via truncation or low temperature sampling, classifier-free guidance can be easily implemented and applied. Specifically, classifier-free guidance trains an unconditional denoising diffusion model together with the conditional model and updates the prediction of noise by increasing the distance between target noise and null-text noise.

The utilization of classifier-free guidance has greatly enhanced the caliber of generated images and has become ubiquitous in subsequent studies. [7, 8, 10, 11, 15, 16, 19, 24, 25, 32, 38]. Based on classifier-free guidance, stable diffusion [22] has facilitated the training of diffusion models on restricted computational resources, whilst preserving their quality and adaptability by deploying them within the potent pretrained autoencoder's latent space. This has resulted in marking new milestones in the realm of image inpainting and class-conditional image synthesis whilst exhibiting exceptionally competitive performance across multiple tasks. In light of this, our study endeavors to further explore the potency of classifier-free guidance in conjunction with stable diffusion acting as the cornerstone of our investigation.

2.2 Image cartoonization

The art form of cartoon has gained immense popularity and has been widely utilized across diverse domains. In the field of image synthesis, Generative Adversarial Network (GAN) [9, 35, 37] is a potent technique that enables the generation of data with the same distribution as that of input data by solving a min-max problem between a generator network and a discriminator network. This approach holds considerable potential in generating images that are seamlessly indistinguishable from real images [2, 5, 14, 17, 20, 21, 26, 39]. White-box Cartoon [33] also employs a GAN architecture to generate cartoonized images. However, it differs from CartoonGAN by allowing each extracted representation to have its own learning objectives. This makes the framework adjustable and controllable. AnimeGAN [3] proposes a unique method for converting real-world scene photographs into anime-style imagery. The approach fuses neural style transfer and GAN to achieve this transformation rapidly while upholding high-quality standards.

Unlike these GAN-based methods [1, 3, 5, 33, 40], This work presents the incorporation of noise interference occurring in the sampling phase of diffusion models to create cartoon images, utilizing not only the input image but also supplementary textual cues. **To the best of our knowledge, this is the first instance of utilizing a diffusion model for implementing image cartoonization without model training.**

3 METHOD

3.1 Preliminaries

The success of classifier-free guidance lies in introducing null-text guidance to re-predict the noise output by extrapolating the model in the direction of text guidance and away from null-text guidance. Specifically, the noise output in the sampling process of classifier-free guidance is computed as follows:

$$\epsilon_\theta(x_t|p) = \epsilon_\theta(x_t|\oslash) + \gamma(\epsilon_\theta(x_t|p) - \epsilon_\theta(x_t|\oslash)), \tag{1}$$

where $\epsilon_\theta(x_t|\cdot)$ is a simplified notation for $\epsilon_\theta(x_t, \cdot, t)$, $\epsilon_\theta(\cdot)$ represents the prediction noise of the U-Net model [23] parameterized by θ. t indicates the time step and p is the prompt used to guide the generated content. $\oslash$ represents the null-text and γ is the guidance scale used to control the strength of the classifier-free guidance. Taking the DDIM sampling for example, a denoising step with $\epsilon_t = \epsilon_\theta(x_t|p)$ can be denoted as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t, \tag{2}$$

where x_t is the noisy image in step t. $\bar{\alpha}$ is related to a pre-defined variance schedule.

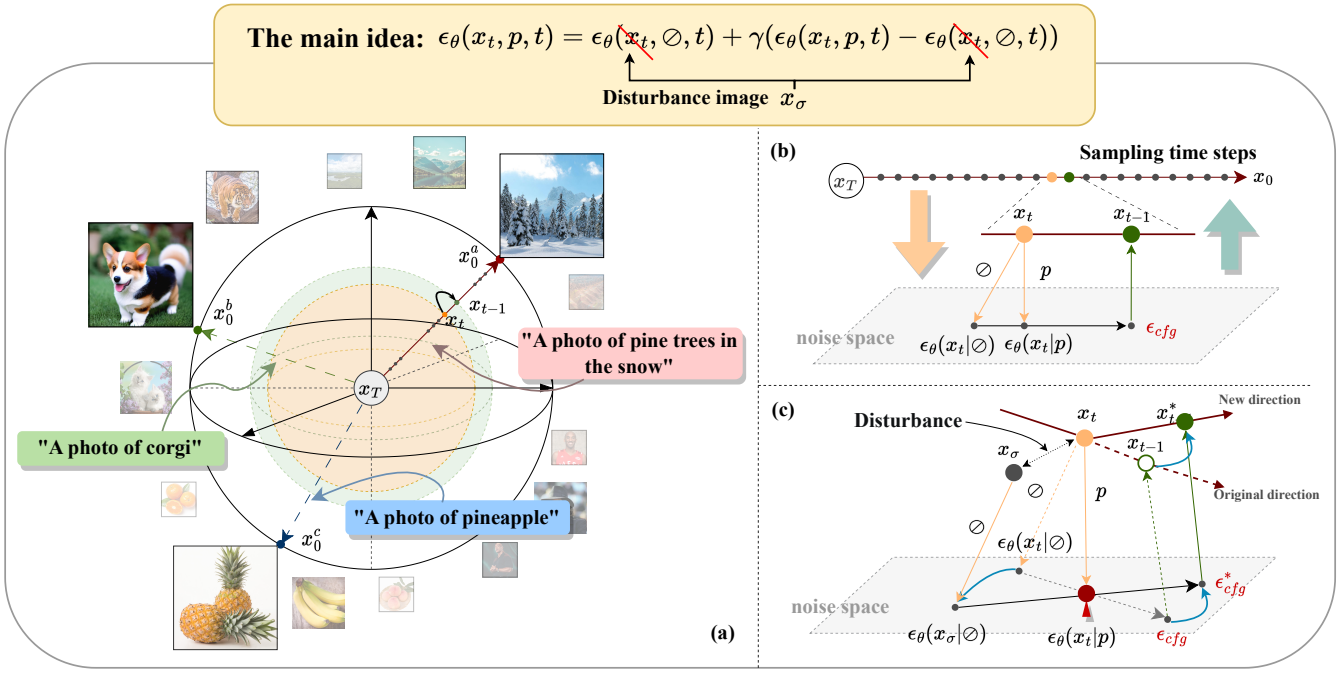

Figure 2: Conceptual diagram of the proposed methods. (a) x_T is random noise in the center of the image space, which is gradually pushed towards the specific image out by a sampling process guided by a given prompt. (b) Predicted noise ϵ_{cfg} in classifier-free guidance. (c) The proposed Noise Disturbance replaced the null-text noisy image with x_σ, resulting in the noise disturbance anchored by text guidance. ϵ^*_{cfg} represents the predicted noise after disturbance. The main idea of this work is summarized on the top of this figure.

Classifier-free guidance greatly improves the generation effect of text-to-image task based on the diffusion model. We believe that there are more potential functions in null-text guidance worth exploring.

3.2 Noise disturbances anchored by text guidance

The text-guided diffusion model is used to denoise a random noise x_T to a clean image from step T to 0 with the guidance of the given text prompt as shown in Figure 2(a). During this procedure, the starting noise x_T is viewed as the central point of the image space, traveling to different endpoints x_0 with the assistance of varied prompts following T time samplings. For the sampling process of classifier-free guidance, the model first predicts the noise guided by text p and null-text $\varnothing$ respectively, getting the noises $\epsilon_\theta(x_t|p)$ and $\epsilon_\theta(x_t|\varnothing)$, and then re-compute the noise output as shown in Eq.(1) in the noise space, which is denoted as ϵ_{cfg} in Figure 2(b). Finally, the noisy image x_{t-1} is obtained by the mapping from noise space to the image space via Eq.(2).

In this work, we seek to investigate further potential functions of null-text guidance by introducing a misalignment between null-text noisy image and text noisy image. As illustrated in Figure 2(c), a disturbance image x_σ is chosen from the image space as the null-text noisy image, while keeping the text noisy image x_t unchanged, thus forming the noise disturbance anchored by text guidance, which can be represented mathematically as shown in Eq.(3) and the summit of Figure 2 manifests this main idea. As a result, the sampling direction has been altered from $x_t \rightarrow x_{t-1}$ to $x_t \rightarrow x^*_{t-1}$ through the addition of noise disturbance.

$$\mathcal{F}_{dis}(\underbrace{\epsilon_\theta(x_\sigma|\varnothing)}_{Disturbance}, \underbrace{\epsilon_\theta(x_t|p)}_{Anchor}) = \epsilon_\theta(x_\sigma|\varnothing) + \gamma(\epsilon_\theta(x_t|p) - \epsilon_\theta(x_\sigma|\varnothing)) \tag{3}$$

3.3 Rollback Disturbance (Back-D)

To conduct the misalignment between null-text noisy image and text noisy image, resulting in the noise disturbance and leading to the change of sampling direction, we propose the Rollback disturbance (Back-D) strategy, which set:

$$x_\sigma = x_{t+b}, b \in (0, T). \tag{4}$$

where b, i.e., Δt, is a hyper-parameter to control the rollback degree. Compared to x_t, x_{t+b} contains more noise and is closer to the initial noisy image x_T.

To preserve the initial structure of the original image, the imposition of noise perturbations is limited to the concluding stage of the sampling procedure. Specifically, such disturbances are initiated solely when $t > s$, where s is a hyper-parameter. Take a cat as an example,

we assessed the effectiveness of the proposed Back-D across varying values of b and s, as demonstrated in Figure 3. Compared with the original image x_0(located at the top-left corner of the

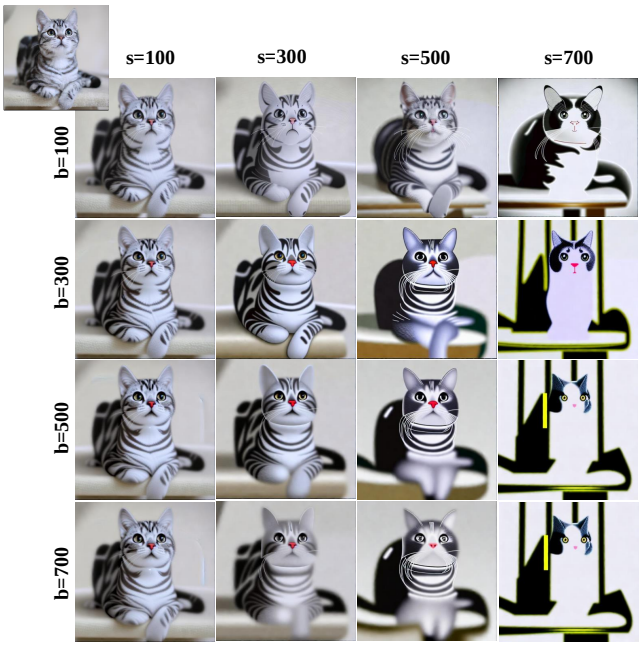

Figure 3: Exploration of Back-D with rollback step b and disturbance time s. The presence of an ideal cartoon-style version near ($b = 300, s = 300$).

figure), we observed the presence of an ideal cartoon-style version near (b=300, s=300).

Additionally, it can be observed that the degree of cartoonization in the generated images is insufficient when b and s are too small (e.g. 100). Conversely, when b and s are excessively large (e.g. 700), the directional deviation caused by noise disturbance is excessive, leading to blurry or even chaotic content in the generated images. In a word, Back-D can enable the free generation of cartoons expeditiously by utilizing a prompt as input. A concise summary of the algorithm is presented in Algorithm 1.

3.4 Image Disturbance (Image-D)

Inspired by the free generation of cartoons with Back-D, we endeavor to achieve image cartoonization through Back-D. Unlike the free generation task that only requires a prompt for guidance, image cartoonization necessitates the use of the base structure of the designated input image, denoted as x_{ref}, as the initial noisy image. Precisely, the process involves obtaining the initial noisy image x_s by adding noise to input image x_{ref} in s steps, wherein the sampling process from $t = s$ to $t = 0$ is consistent with that used in free generation.

The first row on the right of Figure 4 shows the results of Image cartoonization with Back-D, in which the generation is guided by the input image x_{ref} and the prompt "a photo of Dwayne Johnson". Despite its ability to preserve the image structure and enable cartoonization for particular images, Back-D fails to achieve sufficient fidelity.

To enhance the fidelity of image cartoonization, we propose an Image disturbance (Image-D) strategy, which set the null-text

Algorithm 1 Free generation of cartoon with Back-D

Input: A pre-trained Diffusion model $\epsilon_\theta(\cdot)$ with classifier-free guidance, prompt p, guidance scale γ, rollback step b and disturbance time s.
Output: The cartoon-style image x_0^*.

$x_T \sim \mathcal{N}(0, I)$
for t from T to 0 **do**
  if t > s **then**
    Get ϵ_t via Eq.(1)
  else
    $\epsilon_\theta(x_\sigma|\varnothing) = \epsilon_\theta(x_{t+b}|\varnothing)$.
    $\epsilon_t = \mathcal{F}_{dis}(\epsilon_\theta(x_\sigma|\varnothing), \epsilon_\theta(x_t|p))$ via Eq.(3)
  end if
  $x_{t-1} \leftarrow \epsilon_t$ via Eq.(2).
end for
return x_0 as x_0^*.

Figure 4: Comparison between Back-D and Image-D of portrait cartoonization. Image-D can generate better fidelity.

noisy image as a clean image to extract additional details from it. Specifically, we empirically set $x_\sigma = x_{ref}$, i.e., the input image.

The results depicted in the second row on the right of Figure 4 demonstrate that utilizing Image-D leads to enhanced preservation of intricate features present in the input image, and thereby results in superior fidelity. The algorithm of image cartoonization with Back-D or Image-D is summarized in Algorithm 2.

3.5 Analysis of the null-text noisy image

The main idea of this work is to substitute the null-text noisy image with an disturbance image x_σ, which introduces misalignment between null-text noisy image and text noisy image. By utilizing x_{t+b} and x_{ref}, Back-D and Image-D modify the sampling direction, thereby inducing cartoon-style generation. Further, our analysis in Figure 4 suggests that Image-D yields greater fidelity compared to Back-D owing to the detailed information provided by x_{ref}.

To explore the function of null-text noisy image, we vary the correlation of x_σ with x_{ref}. Specifically, Figure 5 illustrates two additional settings for x_σ, viz., an unrelated image x_{irr} and an isotropic image x_{iso} that shares structural similarity with the input image. The degree of correlation between x_{ref} and x_σ in various

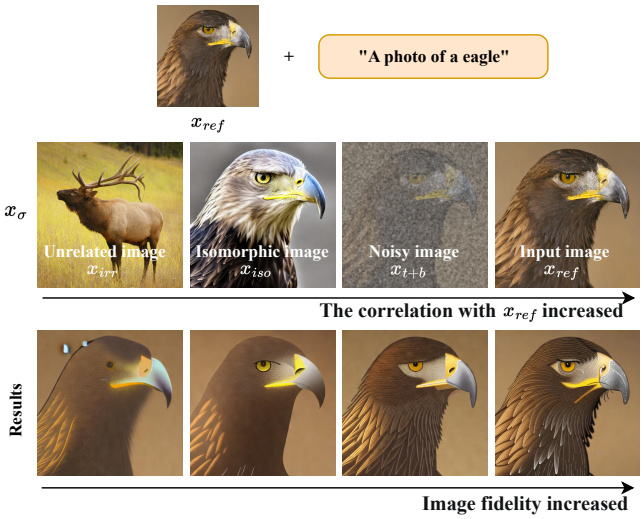

Figure 5: The cartoon effect is affected by the correlation between the noise input of null-text and the input image.

Algorithm 2 Image cartoonization based on noise diturance

Input: A pre-trained Diffusion model $\epsilon_\theta(\cdot)$ with classifier-free guidance, prompt p, guidance scale γ, rollback step b, disturbance time s and a specific input image x_{ref}.

Output: The cartoon-style image x_0^*.

1: Adding noise for x_{ref} with $t = s$ steps, and get the noised image x_s.
2: **for** t from s to 0 **do**
3: $\quad\epsilon_\theta(x_\sigma|\oslash) = \epsilon_\theta(x_{t+b}|\oslash)$ if use Back-D else $\epsilon_\theta(x_{ref}|\oslash)$
4: $\quad\epsilon_t = \mathcal{F}_{dis}(\epsilon_\theta(x_\sigma|\oslash), \epsilon_\theta(x_t|p))$ via Eq.(3)
5: $\quad x_{t-1} \leftarrow \epsilon_t$ via Eq.(2).
6: **end for**
7: **return** x_0 as x_0^*.

settings satisfies:

$$x_{irr} < x_{iso} < x_{t+b} < x_{ref}(100\%) \tag{5}$$

The results indicate that as the correlation degree between null-text noisy image and input images x_{ref} increases, both the quality and fidelity of generation improve.

It is noteworthy that both Back-D and Image-D yield desirable cartoonization results, with the latter providing richer details. Furthermore, additional experimental results (see supplementary material) demonstrate that Back-D and Image-D each have their respective advantages, and can be used selectively depending on the specific case.

4 EXPERIMENTS

4.1 Implementation details

Our experiments are carried out based on Stableb Diffusion Model v1.4 [22] with DDIM steps initialized to 100. During most of our experimental trials, we assigned hyper-parameters $s \in (200, 300)$ and $b \in (200, 300)$. To achieve better cartoonization of special input, we suggest fine-tuning s and b without going over 400. The isomorphic image x_{iso} in section 3.5 is provided by ControlNet [36].

4.2 Experimental results

The experimental results for three settings, i.e., free generation with Back-D, Image cartoonizaiton with Back-D, and Image cartoonization with Image-D, are showcased in Figure 6 (a), (b) and (c), which fully demonstrate the validity of our method. Check out the supplementary materials for more test results.

The diversity of image cartoonization. Exploiting the creative potential of text-guided diffusion models, our image cartoonization method based on noise disturbance yields a diverse array of results dependent upon input images. Figure 7 illustrates the richness of the output: the bird's eye texture varied in Back-D while Image-D induced lifelike variation within the eye, crown, and body. See supplementary materials for more comparison between Back-D and Image-D.

4.3 Comparison

4.3.1 Comparison in free generation. Figure 22 displays a comparison between our method and cartoon image generation model Anything v3 [22] and stable diffusion model v1.4[22]. Anything v3 is trained extensively with cartoon images but fails to accurately generate cartoons for new concepts or scenes not featured within its training data-as seen. For example, case "A photo of Robert Downey Jr." and case "The city of lights". It also suffers from scenario construction failure (case"A rabbit is eating carrot") and over-anthropomorphization of animals as illustrated in case "A koala is climbing a tree". Meanwhile, the stable diffusion model v1.4 operates by modifying guided prompts "xxx" with "xxx in cartoon style", and its resulting images lack spatial information and appear too flat, as demonstrated in the first three cases of row 2 in Figure 22. Moreover, it remains prone to cartoonization failures, as shown in the case "A koala is climbing a tree". Conversely, our method generates more accurate, vivid, and artistically textured cartoon images.

4.3.2 Comparison in image cartoonization. We compare our cartoonization method with two well-known established techniques, i.e., AnimeGANv3 [3] and white-box [33], and the experimental results are presented in Figure 23. As can be seen from the comparison, both AnimeGANv3 and White-box apply additional line textures for planarization purposes, resulting in content being divided into blocks and appearing more akin to comic book illustrations. By contrast, the results achieved through our Back-D are markedly distinct from previous efforts, and more vivid and lifelike, resulting in a style that is better suited for animation scenes.

4.4 The influence of text guidance

In comparison to other image cartoonization methods that transform an input image solely based on its visual content, our text-guided cartoonization method, which generates an image not only based on the original image but also the text prompt, offers an innovative basis for the diversity and creativity of generated cartoons.

4.4.1 Rough and precise guidance. We conducted experiments testing the influence of rough and precise text guidance on creating

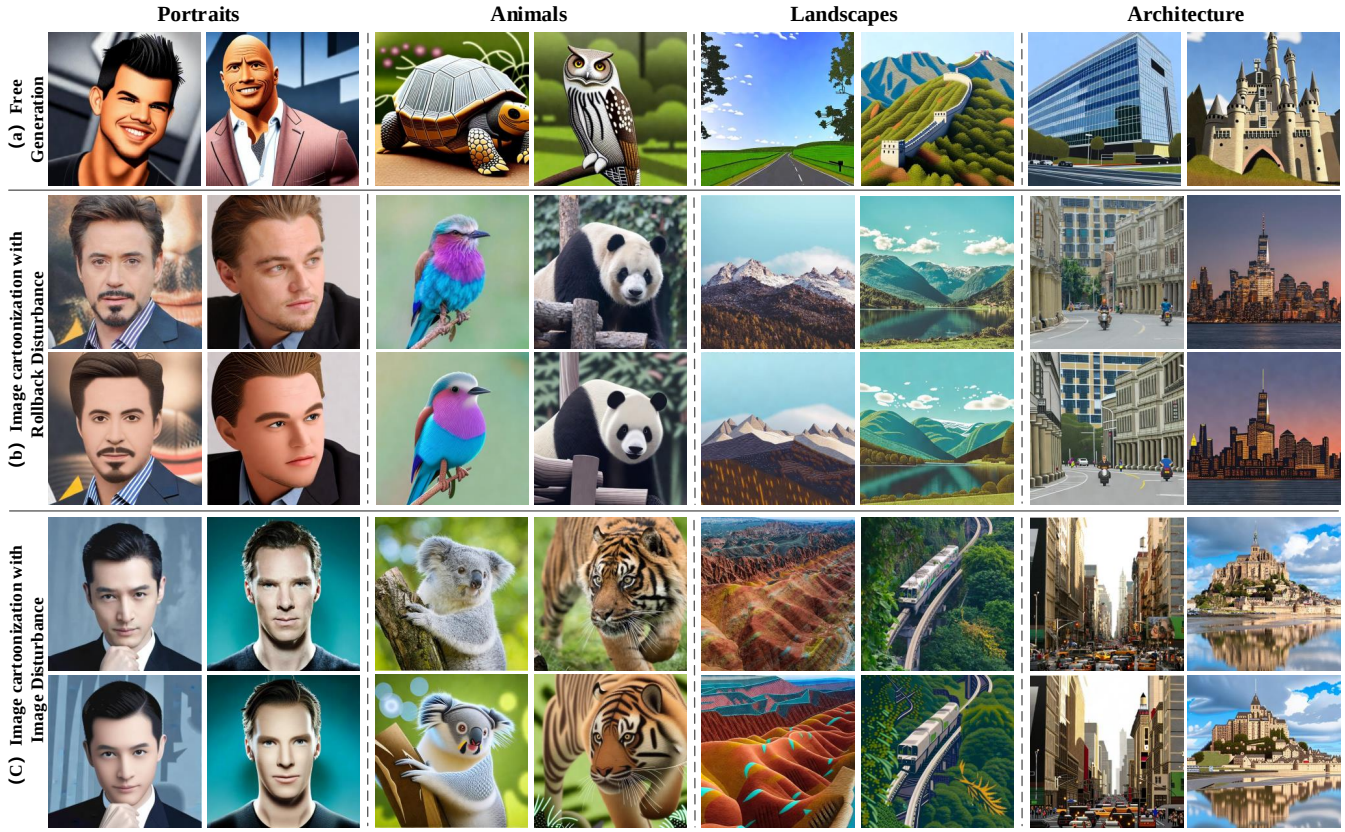

Figure 6: Results of (a) free generation using Back-D, (b) Image cartoonization using Back-D and (c) using Image-D. The results indicate that the proposed method enables free cartoon generation of portraits, animals, landscapes, and architectures, while achieving image cartoonization.

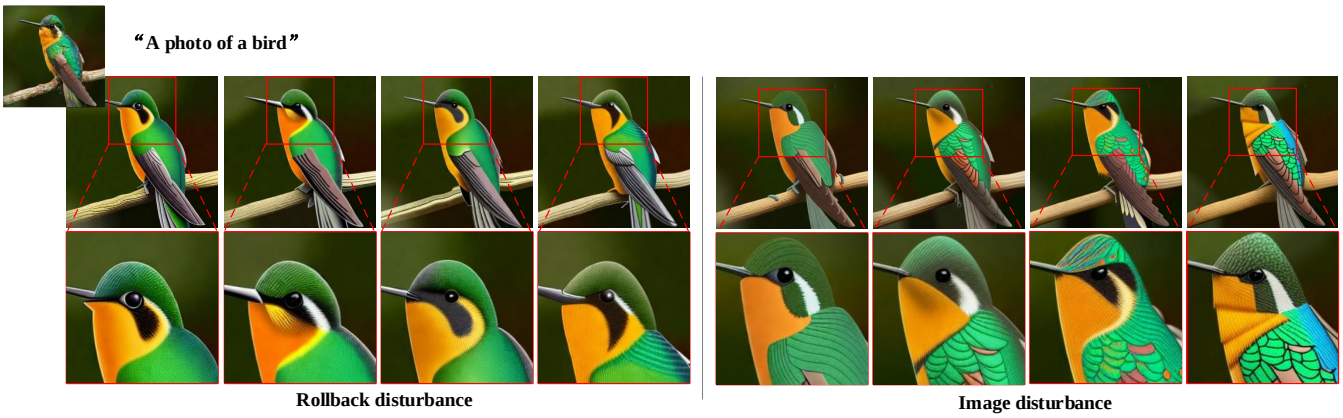

Figure 7: Image cartoonization showcases diversity. The Image disturbance (Image-D) contain richer diversity of details.

cartoon-style images. Based on the picture of a fox on the left of Figure 11, we created two cartoon version of it through rough guidance "a photo of an animal" and precise guidance "a photo of a fox" respectively. The results shown in Figure 11 (b) demonstrate that the model cannot recognize many of the animal's characteristic features under rough guidance, producing a simplified cartoonization of the original image. However, when given accurate information that the subject is indeed a fox, the model could successfully infuse fox-specific characteristics into the cartoon version, such as the slender eyes and long beard in Figure 11 (a).

It is clear that our method represents an intriguing and expressive means of generating cartoons. With appropriate text guidance, the resulting cartoon effect can be highly captivating.

Figure 8: Comparison with other cartoon generation works. Our free generation using Back-D generates more accurate, vivid, and artistically textured cartoon images.

Figure 9: Comparison with other Image cartoonization works. The images resulting from the cartoonization of AnimeGANv3 [3] and White-box [33] appear to have been flattened, resembling drawings on a two-dimensional plane. However, our method produces cartoonized images that are more vivid and lifelike, approaching the three-dimensional quality of animated scenes.

4.4.2 Mismatched guidance. To investigate the influence of text guidance more extensively, we tested our cartoonization approach under the context of mismatched text prompts. Figure 11 (c) and (d) present the results produced by setting prompts to "a photo of a tiger" and "a photo of a pig" respectively, despite using an image of a fox as input. The results demonstrate that the generated fox developed tiger markings on its faces and bodies with tiger as guides, while it changed the shape of its eyes and noses, and lost much of the colour and texture of its head with pig as guides.

Overall, the combination of input images and text prompts can serve as inspiration for creating more creative cartoon characters.

4.5 Restrictions on the sampling steps

We conducted research on number of DDIM sampling steps N, which determines the actual number of sampling in the process of $T \rightarrow 0$. The actual number of sampling k satisfies $k = \frac{s*N}{T}$, where s denotes the time step at which the execution of noise perturbation commences. Figure 10 depicts the cartoonification results of N at varying values, and furthermore, as N increases, the noise in the resulting images progressively decreases. Remarkably, when $N = 20$, the actual number of sampling k dropped to only 4 times (T usually defaults to 1000). In this instance, noise disturbance fails to establish a stable sampling direction and consequentially results in significant noise throughout the generated image. The number

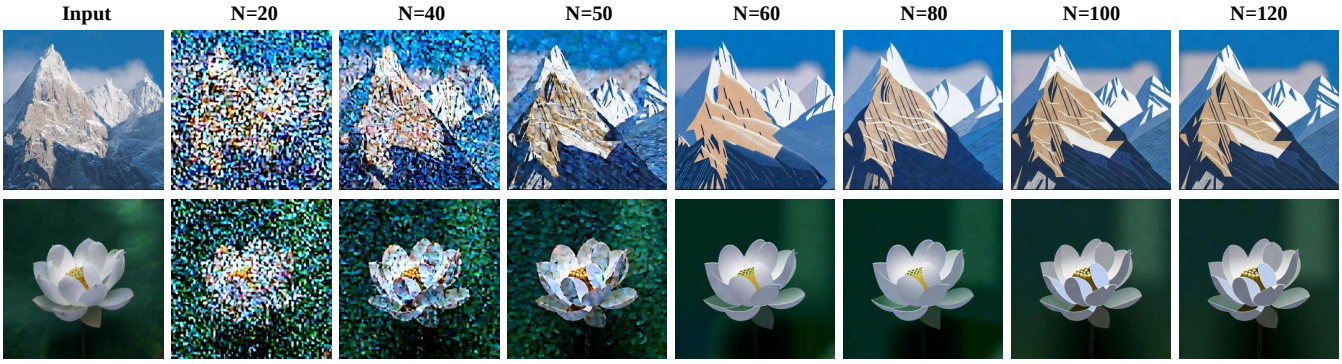

Figure 10: Study on the number of DDIM sampling steps N. N larger than 60 yield a clean cartoon, while greater steps (100 or above) enhance the cartoon effect.

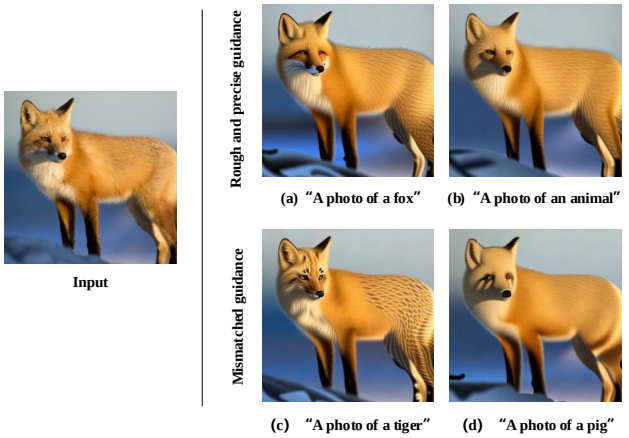

Figure 11: The influence of text guidance. Accurate textual guidance can enhance the conceptual understanding of the diffusion model on the input image, thereby rendering generated images more expressive. conversely, mismatched textual guidance may introduce greater creativity into the generated output.

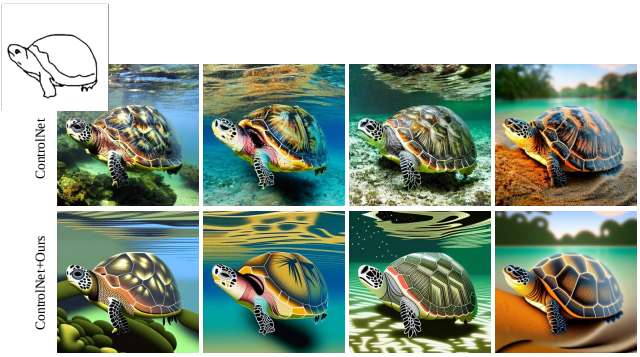

Figure 12: Application on ControlNet.

of DDIM steps larger than 60 yield a clean cartoon, while greater number (100 or above) enhance the cartoon effect.

4.6 Application on ControlNet

As a plug-and-play cartoonize component, the proposed method can be readily applied to the classifier-free guided diffusion model. In this study, we investigated the efficacy of the proposed method in ControlNet [36], a neural network structure designed to regulate pre-trained large diffusion models for accommodating more input conditions. Specifically, we leveraged the Back-D proposed in this work to cartoonize the results of scribble-to-image task in ControlNet and present the findings in Figure 26. The outcomes indicate that the proposed technique is not only easily adaptable to other tasks but also produces a favorable cartoon effect.

5 CONCLUSIONS

In this work, we made a significant discovery that null-text guidance in diffusion model is secretly a cartoon-style creator. Specifically, by iteratively perturbing the null-text guidance with our proposed Rollback disturbance (Back-D) and Image disturbance (Image-D) strategies, we were able to generate cartoons effortlessly and cartoonize specific input images. we systematically investigated and analyzed the impact of various hyper-parameters on the proposed methods, elucidating the appropriate conditions for effective cartoonization. Notably, our approach outperforms existing techniques in terms of generating precise, vivid, and diverse cartoons. Moreover, it is notable that our methods are training-free and easily integrable as a plug-and-play component into any classifier-free guided diffusion model.

REFERENCES

[1] Jihye Back. 2021. Fine-Tuning StyleGAN2 For Cartoon Face Generation. *CoRR* abs/2106.12445 (2021).
[2] Armando Cabrera, Miriam Cha, Prafull Sharma, and Michael Newey. 2022. SAR-to-EO Image Translation with Multi-Conditional Adversarial Networks. *CoRR* abs/2207.13184 (2022).
[3] Jie Chen, Gang Liu, and Xin Chen. 2019. AnimeGAN: A Novel Lightweight GAN for Photo Animation. In *Artificial Intelligence Algorithms and Applications - 11th International Symposium, ISICA 2019, Guangzhou, China, November 16-17, 2019, Revised Selected Papers (Communications in Computer and Information Science, Vol. 1205)*, Kangshun Li, Wei Li, Hui Wang, and Yong Liu (Eds.). Springer, 242–256.
[4] Ming Chen, Hanlu Chu, and Xianglin Wei. 2021. Flocking Control Algorithms Based on the Diffusion Model for Unmanned Aerial Vehicle Systems. *IEEE Trans. Green Commun. Netw.* 5, 3 (2021), 1271–1282.
[5] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. 2018. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In *2018 IEEE Conference on Computer*

Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, 9465–9474.

[6] Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 8780–8794.

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV (Lecture Notes in Computer Science, Vol. 13675)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 89–106.

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *CoRR* abs/2208.01618 (2022).

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680.

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *CoRR* abs/2208.01626 (2022).

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *CoRR* abs/2210.02303 (2022).

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[13] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR* abs/2207.12598 (2022).

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 9906)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 694–711.

[15] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-Based Real Image Editing with Diffusion Models. *CoRR* abs/2210.09276 (2022).

[16] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 11451–11461.

[17] Luigi Tommaso Luppino, Michael Kampffmeyer, Filippo Maria Bianchi, Gabriele Moser, Sebastiano Bruno Serpico, Robert Jenssen, and Stian Normann Anfinsen. 2022. Deep Image Translation With an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote. Sens.* 60 (2022), 1–22.

[18] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8162–8171.

[19] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 16784–16804.

[20] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context Encoders: Feature Learning by Inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2536–2544.

[21] Hui Ren, Jia Li, and Nan Gao. 2020. Two-Stage Sketch Colorization With Color Parsing. *IEEE Access* 8 (2020), 44599–44610.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 9351)*, Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (Eds.). Springer, 234–241.

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *CoRR* abs/2208.12242 (2022).

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR* abs/2205.11487 (2022).

[26] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. 2018. A Style-Aware Content Loss for Real-Time HD Style Transfer. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII (Lecture Notes in Computer Science, Vol. 11212)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 715–731.

[27] Jie Shi, Chenfei Wu, Jian Liang, Xiang Liu, and Nan Duan. 2022. DiVAE: Photorealistic Images Synthesis with Denoising Diffusion Decoder. *CoRR* abs/2206.00386 (2022). https://doi.org/10.48550/arXiv.2206.00386

[28] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 2256–2265.

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[30] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 11895–11907.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. *CoRR* abs/2209.14916 (2022).

[33] Xinrui Wang and Jinze Yu. 2020. Learning to Cartoonize Using White-Box Cartoon Representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 8087–8096.

[34] Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. 2022. Learning Fast Samplers for Diffusion Models by Differentiating Through Sample Quality. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

[35] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 82–90.

[36] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *CoRR* abs/2302.05543 (2023).

[37] Zhichao Zhang, Hui Chen, Xiaoqing Yin, and Jinsheng Deng. 2021. Joint Generative Image Deblurring Aided by Edge Attention Prior and Dynamic Kernel Selection. *Wirel. Commun. Mob. Comput.* 2021 (2021), 1391801:1–1391801:14.

[38] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. 2023. MagicFusion: Boosting Text-to-Image Generation Performance by Fusing Diffusion Models. arXiv:2303.13126 [cs.CV]

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

[40] Nan Zhuang and Cheng Yang. 2021. Few-Shot Knowledge Transfer for Fine-Grained Cartoon Face Generation. In *2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 5-9, 2021*. IEEE, 1–6.

APPENDIX

Results of free generation

We present additional examples of free generation in Figures 13 and 14 as a supplement to Figure 6(a) in the main text. Free generation refers to the diffusion model generating cartoon images that are contextually relevant, by leveraging our proposed approach along with a given prompt. Experimental results demonstrate that our method is effective in generating various types of cartoon images. It is noteworthy that the performance of free generation is dependent on the base model's ability to complete the text-to-image task. Our method simply applies a cartoonization process to the generated image based on the base model's output.

Results of Image Cartoonization

To supplement Figure 6(b) in the main text, we present additional cases of image cartoonization using the Rollback disturbance technique in Figures 15, 16, 17 and 18. Furthermore, to complement Figure 6(c) in the main text, Figures 19,20, and 21 were used to introduce additional cases of image cartoonization generated via the Image disturbance technique. Based on extensive testing, we have found that both Rollback disturbance and Image disturbance exhibit pronounced cartoonization effects. Overall, the former achieves a higher degree of cartoonization than the latter (at the expense of more detail is lost). However, we do not assert that the former is superior to the latter as a cartoonization technique. Our view is that Rollback disturbance and Image disturbance are suited to different types of input images, and users are free to choose between them based on their preferences for cartoonization outcomes.

As demonstrated in Figures 22 and 22, which exhibit the cartoonized results of Rollback and Image disturbance, when generating cartoon-style images of animals and sceneries, Image disturbance enhances the expressiveness of the synthesized images by adding semantically meaningful details on top of the input image. Similarly, when it comes to producing cartoonized portraits, Image disturbance achieves high fidelity by capturing more detailed features of the input image.

Diversity Exhibition

We demonstrate the diversity of cartoonization outputs achieved through the usage of Rollback disturbance and Image disturbance in Figures 24 and 25 , respectively. In contrast to conventional cartoonization techniques that enable a one-to-one mapping between input and cartoonized images, our approach enables creative cartoonization with a one-to-many mapping.

Applications on ControlNet

Based on ControlNet[36], we conducted further experiments on the scribble-to-image task and presented the results in Figure 26. The findings robustly demonstrate the efficacy of our proposed method as a plug-and-play cartoonization component that can be readily applied to various generative tasks.

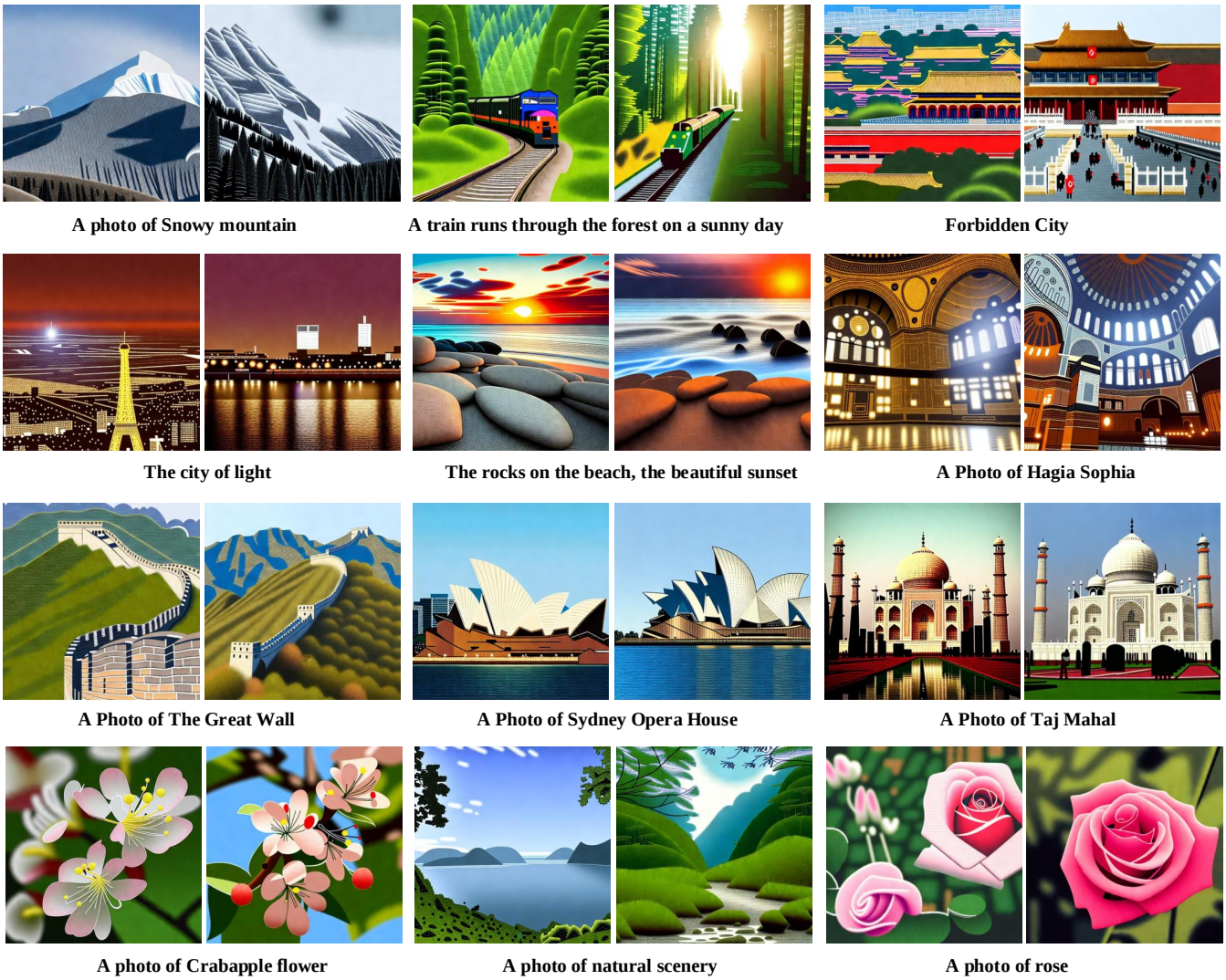

Figure 13: Free generation of cartoons using Rollback disturbance. The image is generated based on textual prompts below.

A koala is climbing a tree | A photo of elephant | A photo of giraffes on the grassland

A zebra is drinking by the river | A photo of tortoise | A rabbit is eating carrot

The squirrel picked pine nuts | An eagle caught a fish in the sea | An owl stood on the branch

A photo of Anne Hathaway | A photo of David Beckham | A photo of Dwayne Johnson

A photo of Robert Downey Jr. | A photo of Taylor Lautner | A photo of Taylor Swift

A photo of chrysanthemum | Bridge_flowing_water | A photo of Country road

Figure 14: Free generation of cartoons using Rollback disturbance.

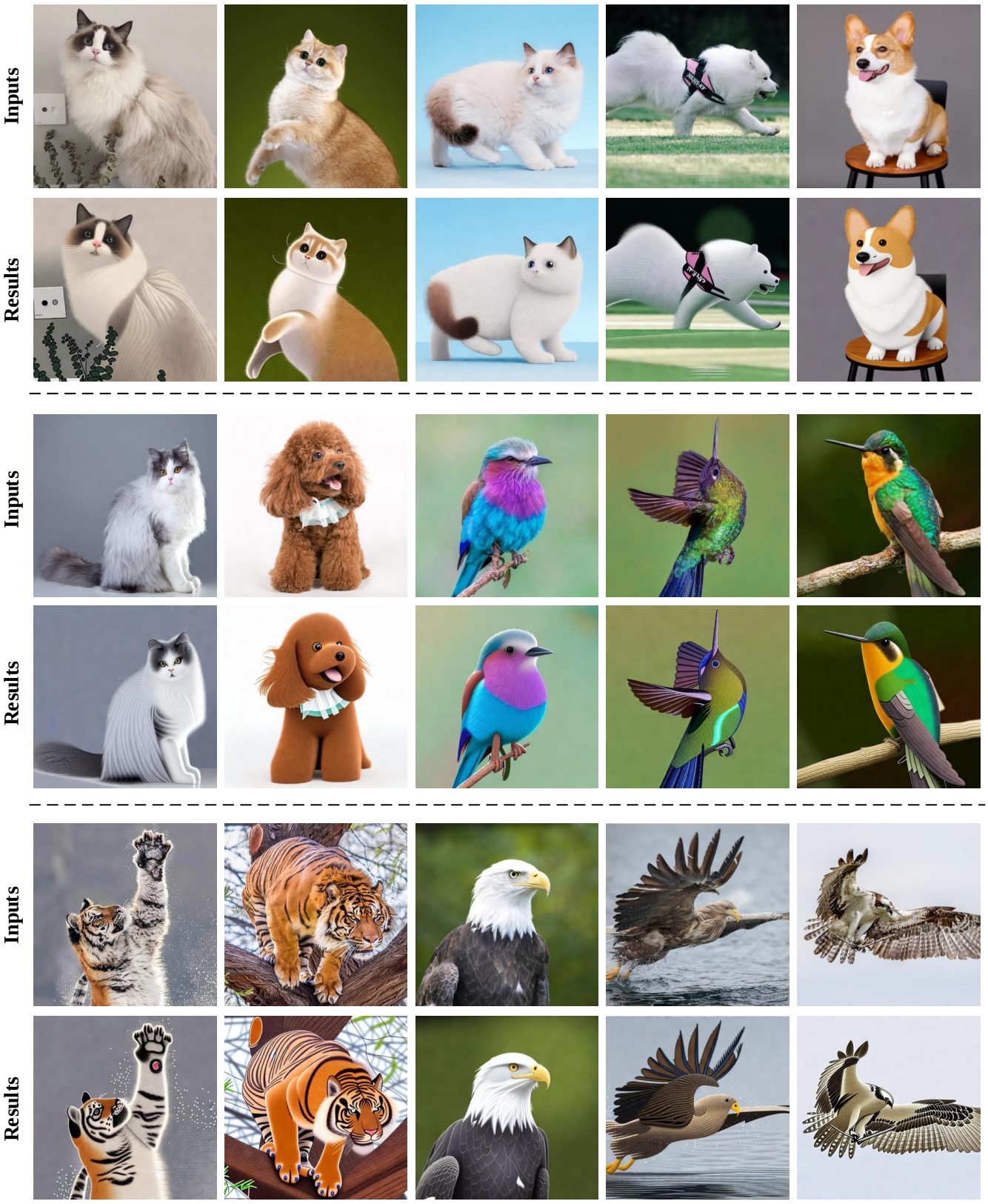

Figure 15: The results of Image cartoonization using Rollback disturbance.

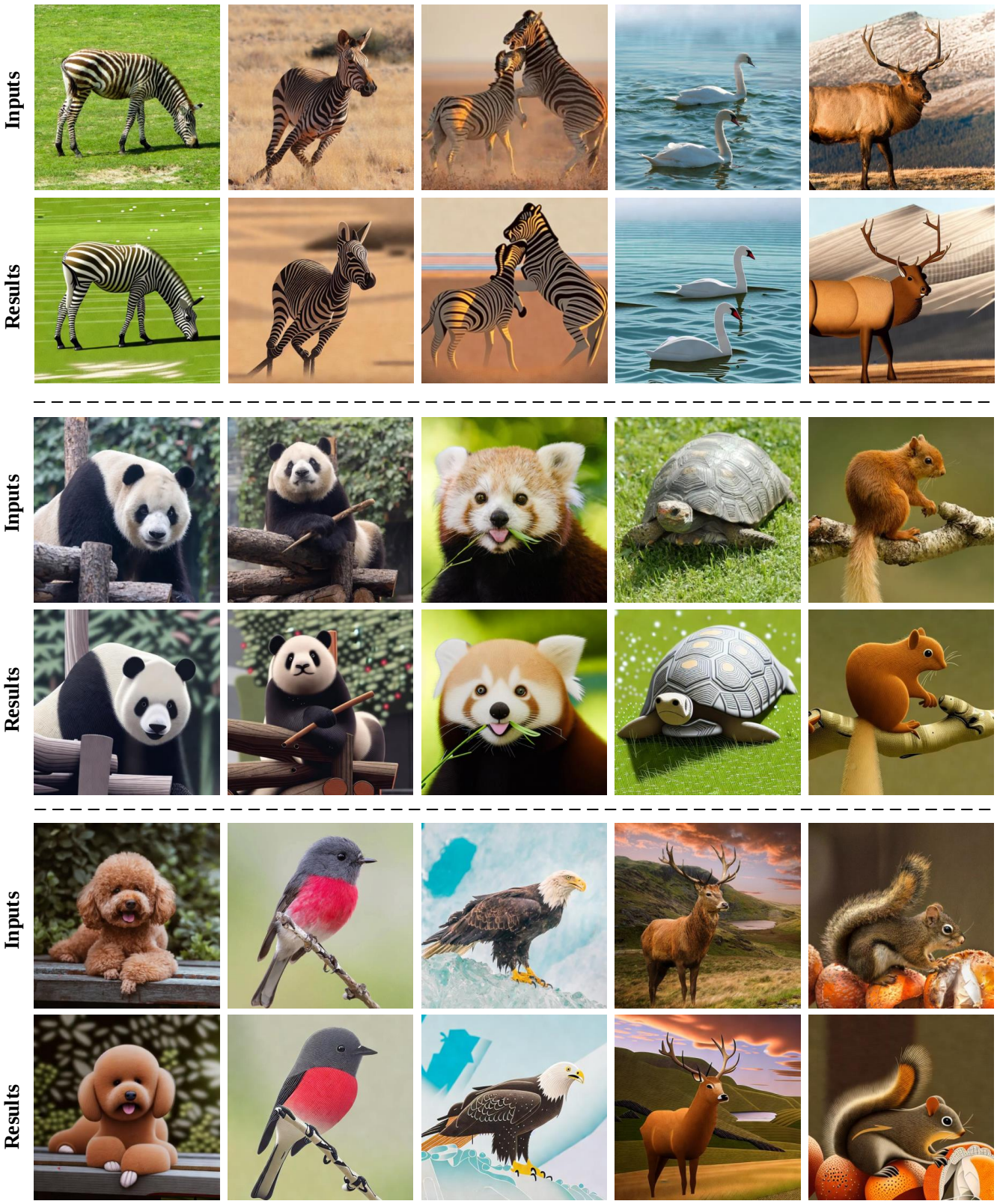

Figure 16: The results of Image cartoonization using Rollback disturbance.

Figure 17: The results of Image cartoonization using Rollback disturbance.

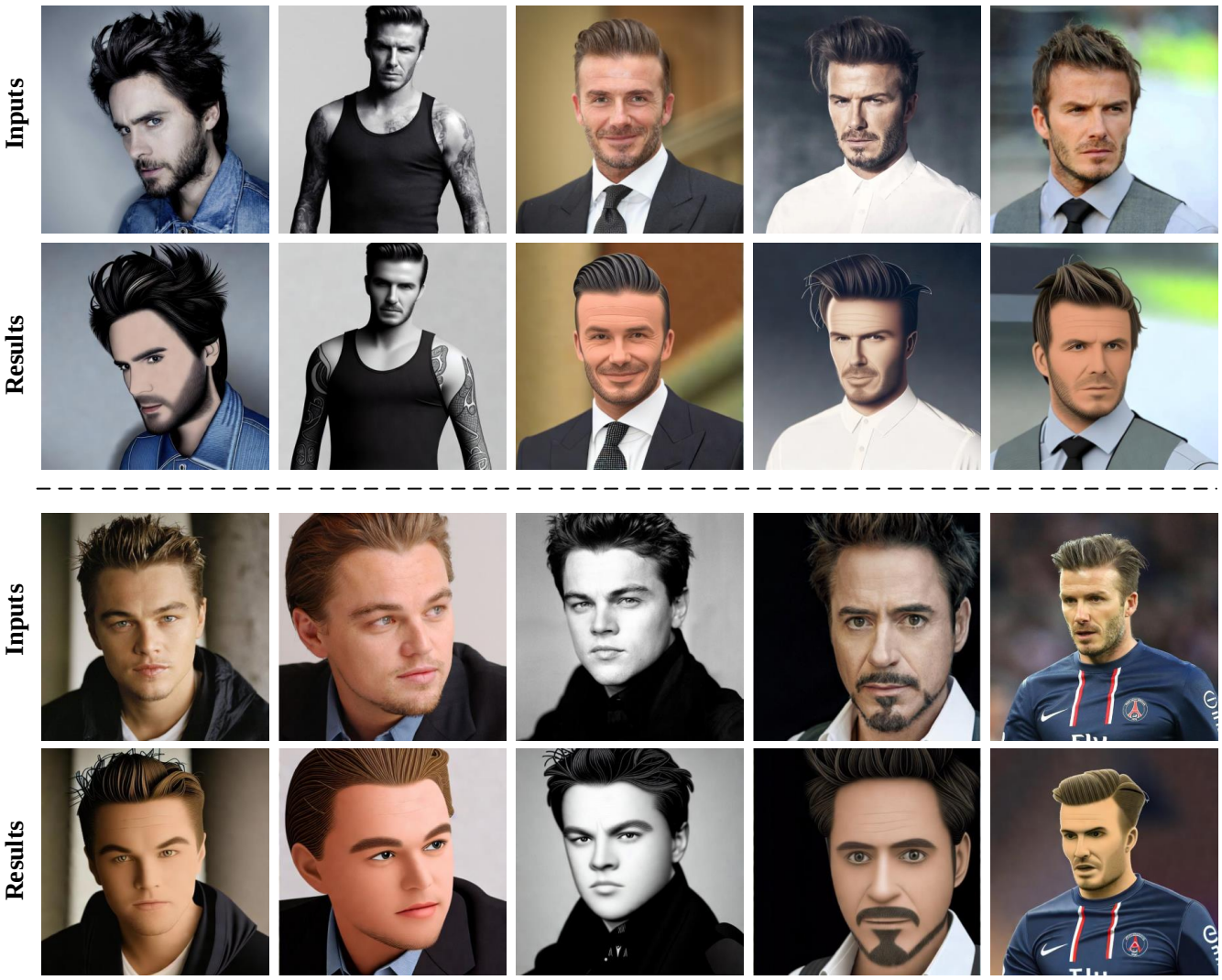

Figure 18: The results of Image cartoonization using Rollback disturbance.

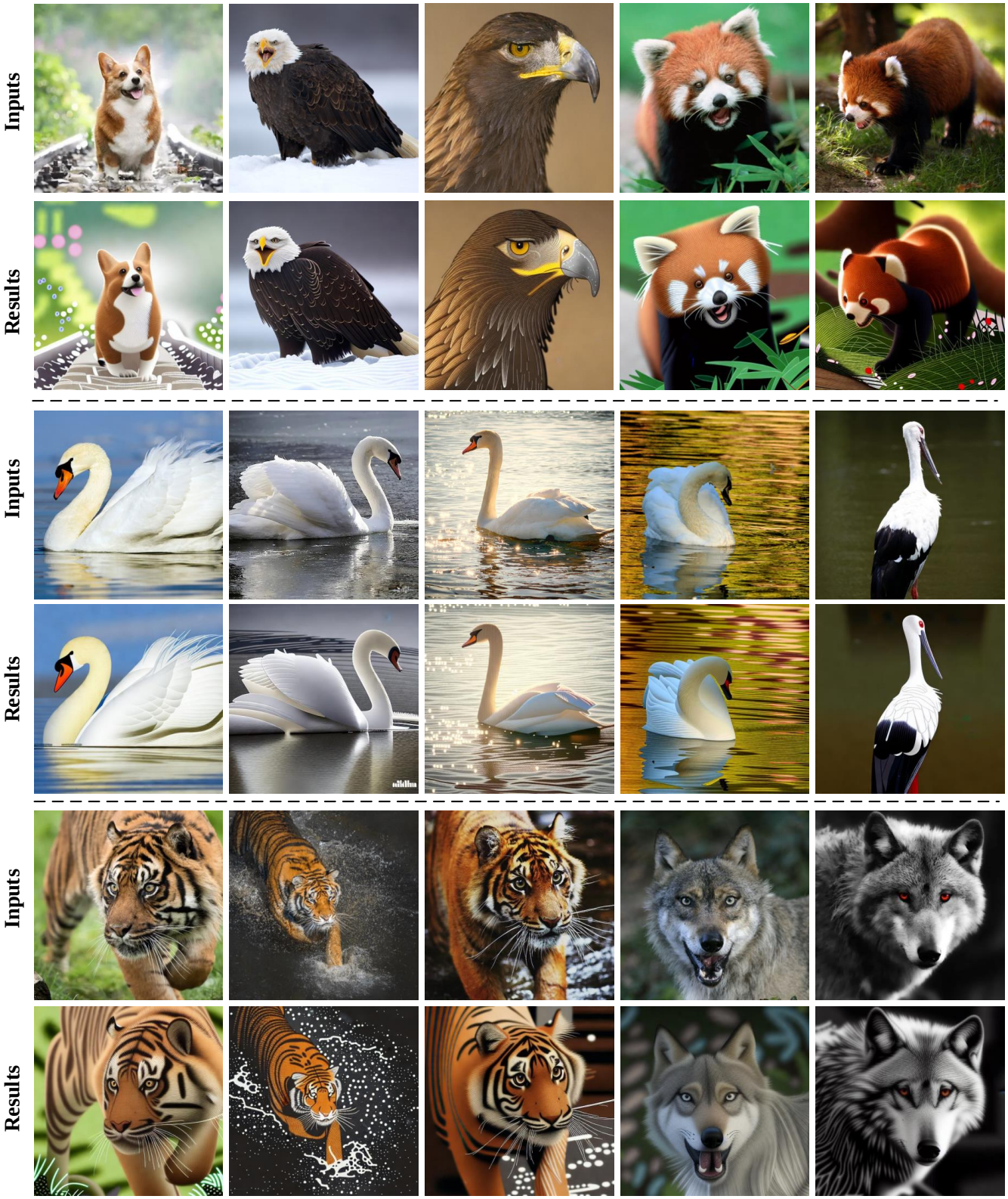

Figure 19: The results of Image cartoonization using Image disturbance.

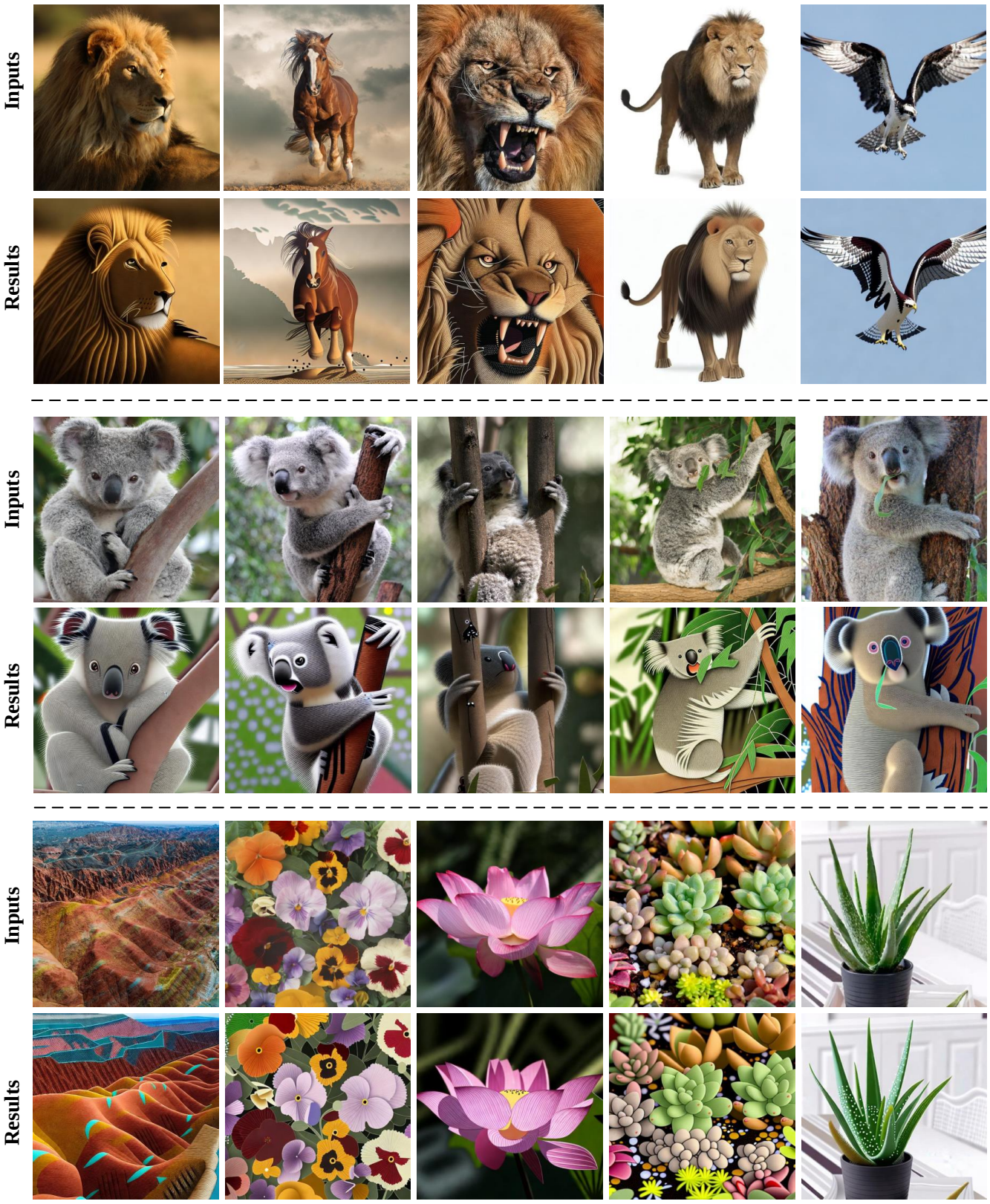

Figure 20: The results of Image cartoonization using Image disturbance.

Figure 21: The results of Image cartoonization using Image disturbance.

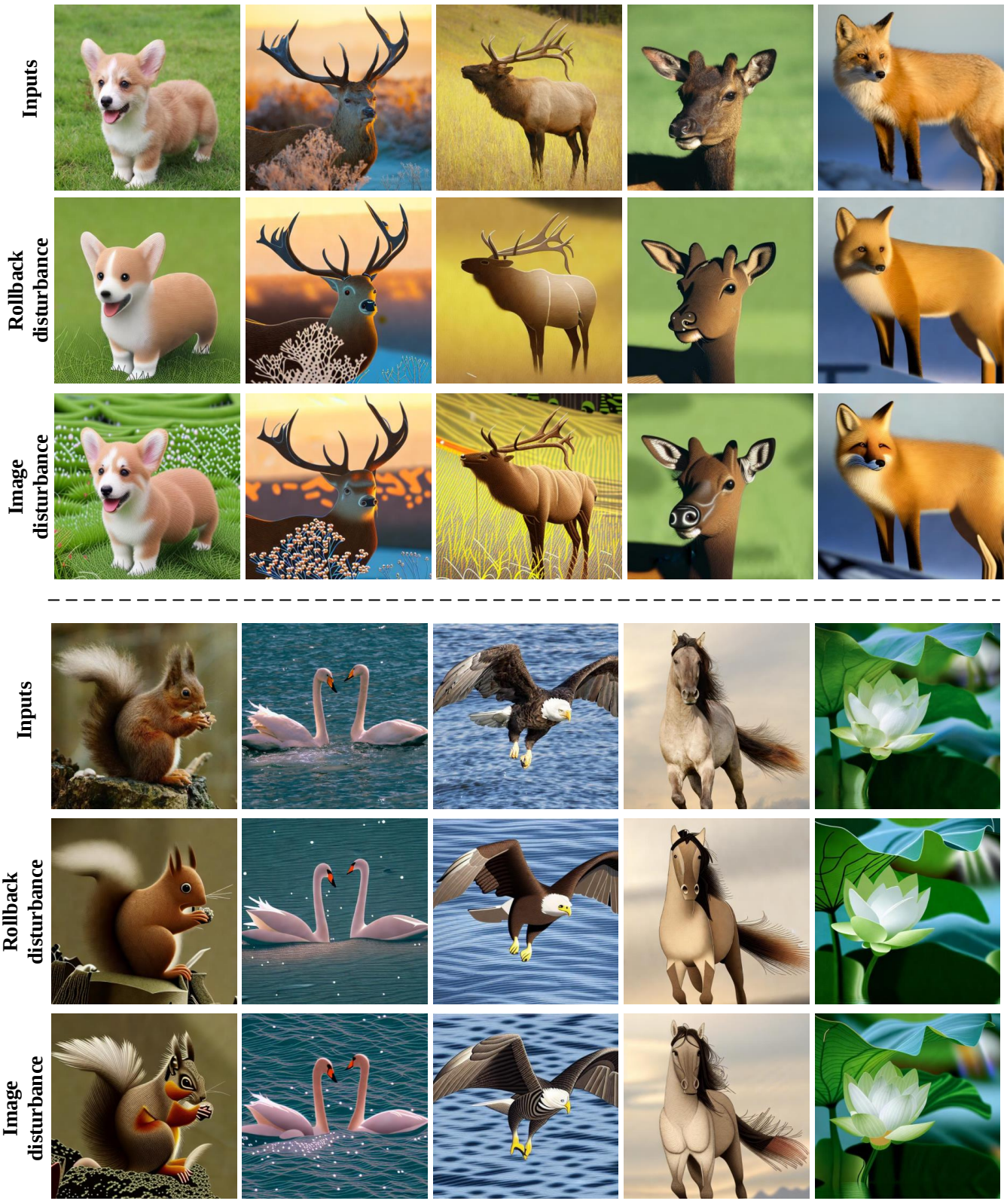

Figure 22: Comparison with Rollback disturbance and Image disturbance.

Figure 23: Comparison with Rollback disturbance and Image disturbance.

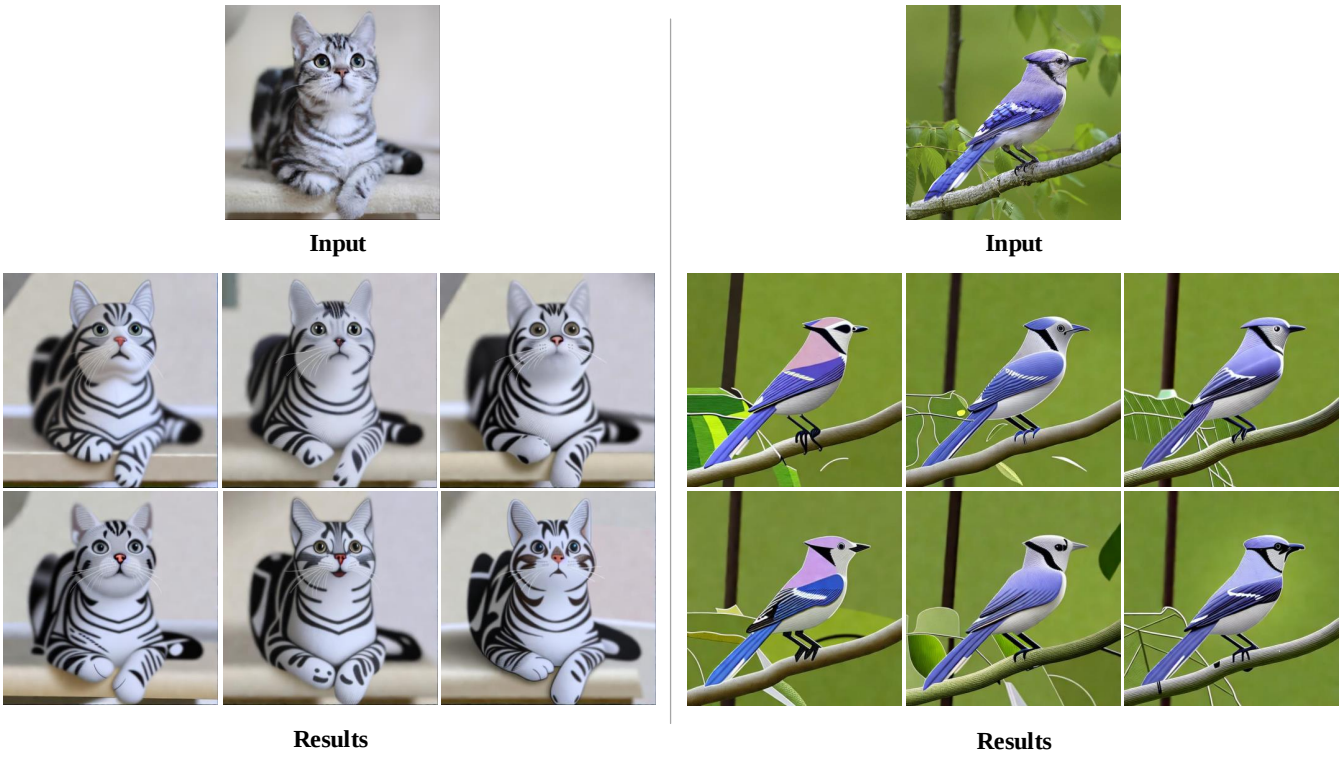

Figure 24: Diversity results demonstration based on Rollback disturbance.

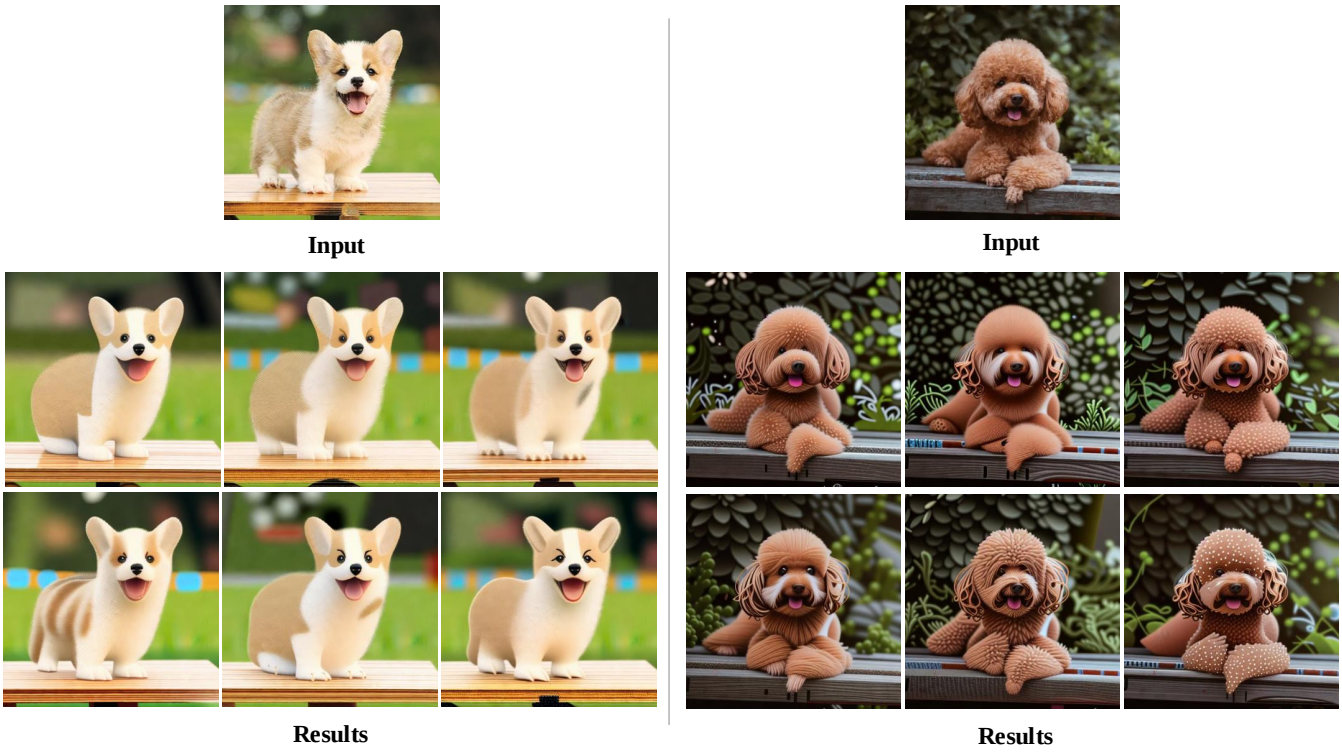

Figure 25: Diversity results demonstration based on Image disturbance.

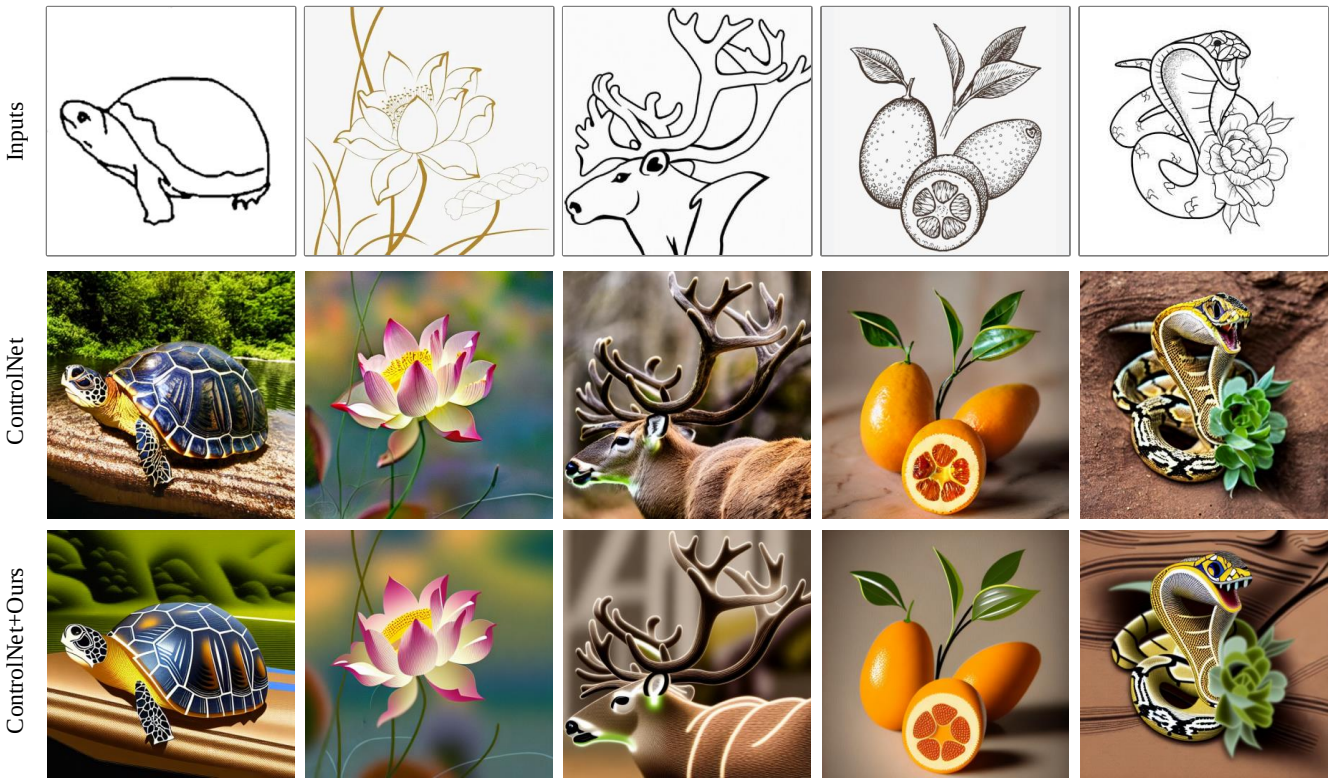

Figure 26: Applications on ControlNet