ReNeg: Learning Negative Embedding with Reward Guidance

Supplementary Material

1. Implementation Details of the Pilot Study

Let $f_\theta : \mathcal{N} \times \mathcal{C} \rightarrow \mathcal{X}$ be a pretrained diffusion model parameterized with θ, mapping a D-dimensional Gaussian distribution $\mathcal{N}$ to a D-dimensional image distribution, conditioned on the embedding space $\mathcal{C}$, where $D = h \times w \times c$.

Definition 1: Parameter efficiency. *Let us sample N random condition embeddings from $\mathcal{C}$ and map them to images using f_θ, assembling a result tensor $X \in \mathbb{R}^{N \times D}$. When N is large enough, X can be considered a nice sampling of the learned distribution. We define the parameter efficiency $E(\theta)$ w.r.t. parameter θ as the average rate of change X under a small perturbation. Formally,*

$$E(\theta) := \frac{1}{NDd_\theta} \left\| \frac{\partial X}{\partial \theta} \right\|_F . \tag{1}$$

Note that $\mathcal{J} = \frac{\partial X}{\partial \theta} \in \mathbb{R}^{d_\theta \times ND}$ is commonly referred to as the *Jacobian matrix*, with each element reflecting the rate of change between pair-wise X_i and θ_j. We compute the Frobenius norm of $\mathcal{J}$ and average it with the total number of matrix elements NDd_θ. In our implementation, we leverage the automatic differentiation capabilities of PyTorch, *i.e.*, `torch.autograd.grad()` to compute the derivatives. Empirically, we set $N = 5000$, as further increases in N did not lead to notable changes in the results. For better clarity, let us first consider the scenario when $N = 1$. In this case, we calculate each column of $\mathcal{J}$ iteratively. Each gradient computation yields a set of gradients with respect to all d_θ. However, since the value of D is typically quite large, *e.g.*, $D = 64 \times 64 \times 16 = 65536$ in our case, it is computationally infeasible to compute all columns of the Jacobian matrix directly. To address this, we perform a random sampling strategy, selecting 16 columns at random. For $N > 1$, we iterate over N samples and concatenate the resulting Jacobian matrices along the row dimension. Empirically, we observe that random sampling along the D dimension provides a good approximation of the full Jacobian matrix.

We argue that an efficient set of parameters should have a large $E(\theta)$ so that a small amount of gradient descent steps can change the learned distribution remarkably. We observe for a pretrained model, both the full model parameter θ_0 and LoRA parameters θ_l have low parameter efficiency. This is likely because they start from a well-pretrained checkpoint, where the model has already converged to a local minimum. In contrast, tuning the negative embedding n exhibits remarkably higher parameter efficiency, *i.e.*, $E(n)$ is 2 to 5 order larger in magnitudes than $E(\theta_0)$ and $E(\theta_l)$.

2. Benchmark for ControlNet and T2V models

Benchmark for Evaluating Pose-Conditioned ControlNet. We collect a total of 20 different poses from the HuggingFace repositories 'sayakpaul/poses-controlnet-dataset' [1] and 'raulc0399/open_pose_controlnet' [2]. Then, we assign five text prompts to each pose to create the text-pose condition pairs. The assigned text prompts are listed as follows:

- A girl in the playground.
- A doctor in the hospital.
- A chef in the kitchen.
- An artist in the studio.
- A runner in the park.

The final evaluation benchmark consists of 100 text-pose pairs. The qualitative and quantitative performance of the proposed ReNeg on this dataset is presented in Fig. 1 and Tab. 5 of the main paper.

Benchmark for Evaluating T2V models. Following the evaluation protocol of VBench [5], we curated a subset of text prompts to evaluate the performance of the proposed ReNeg on ZeroScope [2] and VideoCrafter2 (VC2) [3]. The evaluation dataset consists of various categories and scenes, providing a comprehensive benchmark for testing model generalization and robustness.

3. Inference Details

For T2I generation, the handcrafted negative prompts include: *distorted, ugly, blurry, low resolution, low quality, bad, deformed, disgusting, overexposed, simple background, plain background, grainy, underexposed, too dark, too bright, too low contrast, too high contrast, broken, macabre, artifacts, oversaturated.* **For T2V generation**, the handcrafted negative prompts include: *blurry, pixelated, noisy, glitches, watermarks, compression artifacts, overexposed, underexposed, color banding, unnatural colors, inconsistent lighting, incorrect white balance, low resolution, lack of detail, soft focus, bad hands, extra limbs, distorted facial features, incorrect proportions, missing body parts, misrendered objects, incorrect object sizes, deformed shapes, unrecognizable items, cluttered scenes, poor framing, unbalanced composition, elements cut off, inconsistent perspective, flattened depth, incorrect scaling, temporal flickering, stuttering motion, inconsistent frame rates, jerky transitions, inconsistent backgrounds, changing light-*

[1] https://huggingface.co/datasets/sayakpaul/poses-controlnet-dataset

[2] https://huggingface.co/datasets/raulc0399/open_pose_controlnet

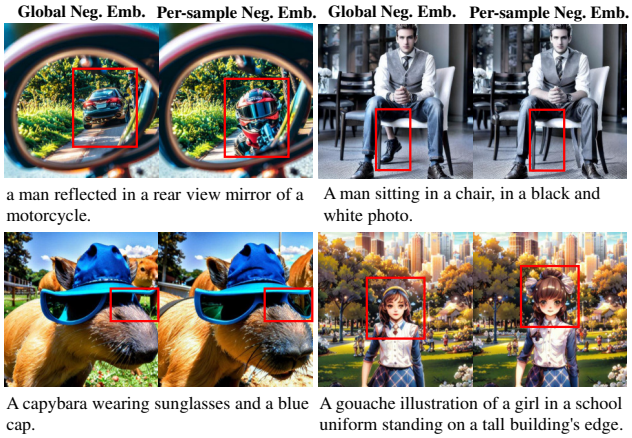

Figure 1. Comparison of results using global negative embedding and per-sample negative embedding. Red boxes highlight the improvement in image details achieved by the per-sample negative embedding. Best viewed zoomed in.

ing between frames, disjointed scenes, ugly, unattractive visuals, clashing styles, inconsistent art style, unharmonious color schemes, abrupt cuts, poor transitions, repeated frames, illogical sequences, lack of continuity, incorrect setting, inappropriate elements for the theme, illegible text, misaligned text, unintended logos. Compared to that, the proposed ReNeg, which optimizes negative embeddings, demonstrates superior performance. The optimized negative embedding exhibits strong generalization capabilities when applied to SD1.4, SD2.1, ControlNet [9], ZeroScope [2], and VideoCrafter2 [3]. During inference, we adopt the DDIM sampler [7] with 30 steps for SD1.4 and SD1.5, setting classifier-free guidance (CFG) weight to 7.5. For ControlNet, we use the UniPC Scheduler [10] with 20 steps. For T2V models, we employ the DDIM sampler for ZeroScope with 40 inference steps and a CFG weight of 9.0. The video resolution is 576×320. VideoCrafter2 is configured with 50 inference steps and a CFG weight of 12.

4. Transferability between SD1.4 and SD1.5

We conduct an additional experiment to further validate the generalization capability of the proposed ReNeg across different SD versions. Specifically, we optimize the negative embedding on SD1.4 and SD1.5 and evaluate the performance on other SD versions. Both quantitative and qualitative results are provided in Tab. 1 and Fig. 2. The proposed ReNeg significantly enhances performance when the obtained negative embedding is transferred across different foundation models. This highlights the flexibility of our method, which can be seamlessly applied to various models sharing the same text encoder as SD1.5. Compared to handcrafted negative prompts, our method achieves performance gains exceeding 5% over the original SD baseline.

Table 1. Transferability results of negative embeddings between SD1.4 and SD1.5 on the HPSv2 benchmark. The evaluation metric used is HPSv2.1. *Neg. Emb. (SD1.5)* indicates that SD1.4 performs inference using the negative embedding optimized for SD1.5. The best scores for each model are highlighted in **bold**.

Model	Animation	Concept Art	Painting	Photo	Average
SD1.4	25.86	24.78	24.62	25.48	25.19
w/ Handcrafted Prompt	27.27	26.39	26.45	26.90	26.75
w/ ReNeg (SD1.5)	**30.50**	**30.91**	**31.28**	**28.77**	**30.37**
SD1.5	25.92	24.66	24.65	25.62	25.21
w/ Handcrafted Prompt	27.29	26.33	26.39	27.01	26.76
w/ ReNeg (SD1.4)	**30.79**	**31.03**	**31.65**	**28.97**	**30.61**

Table 2. Results of combining our negative embedding with the recaptioned positive prompt on the HPSv2 benchmark. The evaluation metric is HPSv2.1. The best scores for each model are highlighted in **bold**.

Model	Animation	Concept Art	Painting	Photo	Average
SD1.5	25.92	24.66	24.65	25.62	25.21
+ N^*	27.29	26.33	26.39	27.01	26.76
+ ReNeg	31.37	31.67	32.00	**29.27**	31.08
+ ReNeg + Promptist	**31.97**	**32.37**	**32.77**	28.89	**31.50**

Table 3. Human evaluation of ReNeg compared to base models. Our approach obtains better human preference over all compared methods.

	Human Preference	Prompt Alignment	Aesthetic
Ours vs. SD1.5	79.80%	65.26%	83.95%
Ours vs. Handcrafted Prompt	72.58%	68.75%	79.17%

Moreover, the visualization results in Fig. 2 reveal that the generated images exhibit superior aesthetic quality and are visually more appealing, further validating the universality and robust generalization capability of our approach.

5. Additional Results

5.1. Human Evaluations

To demonstrate the effectiveness of the learned negative embeddings, we conduct human evaluations comparing the *win rates* of our method and the baselines across three dimensions. The results in Tab. 3 show that users prefer the images synthesized by our method, as evidenced by win rates exceeding 50%.

5.2. Comparison between Global and Per-sample Negative Embedding

As illustrated in Fig.1, we present additional comparisons between global and per-sample negative embeddings. The results reveal that per-sample negative embedding effectively refines the generated outputs based on the given positive prompt, resulting in outputs that align more closely with the textual descriptions. Moreover, the refined outputs exhibit significantly improved visual quality.

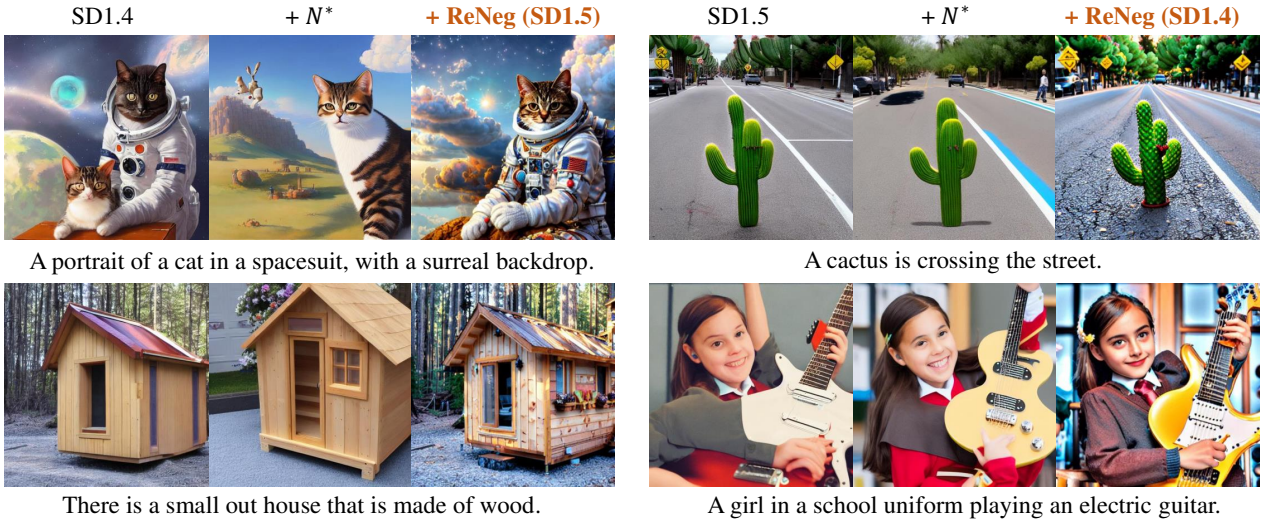

Figure 2. Qualitative transferability results of negative embeddings between SD1.4 and SD1.5. '$+N^*$' and '+ReNeg' indicate improved results with handcrafted negative prompts and our negative embedding of the corresponding SD model, respectively.

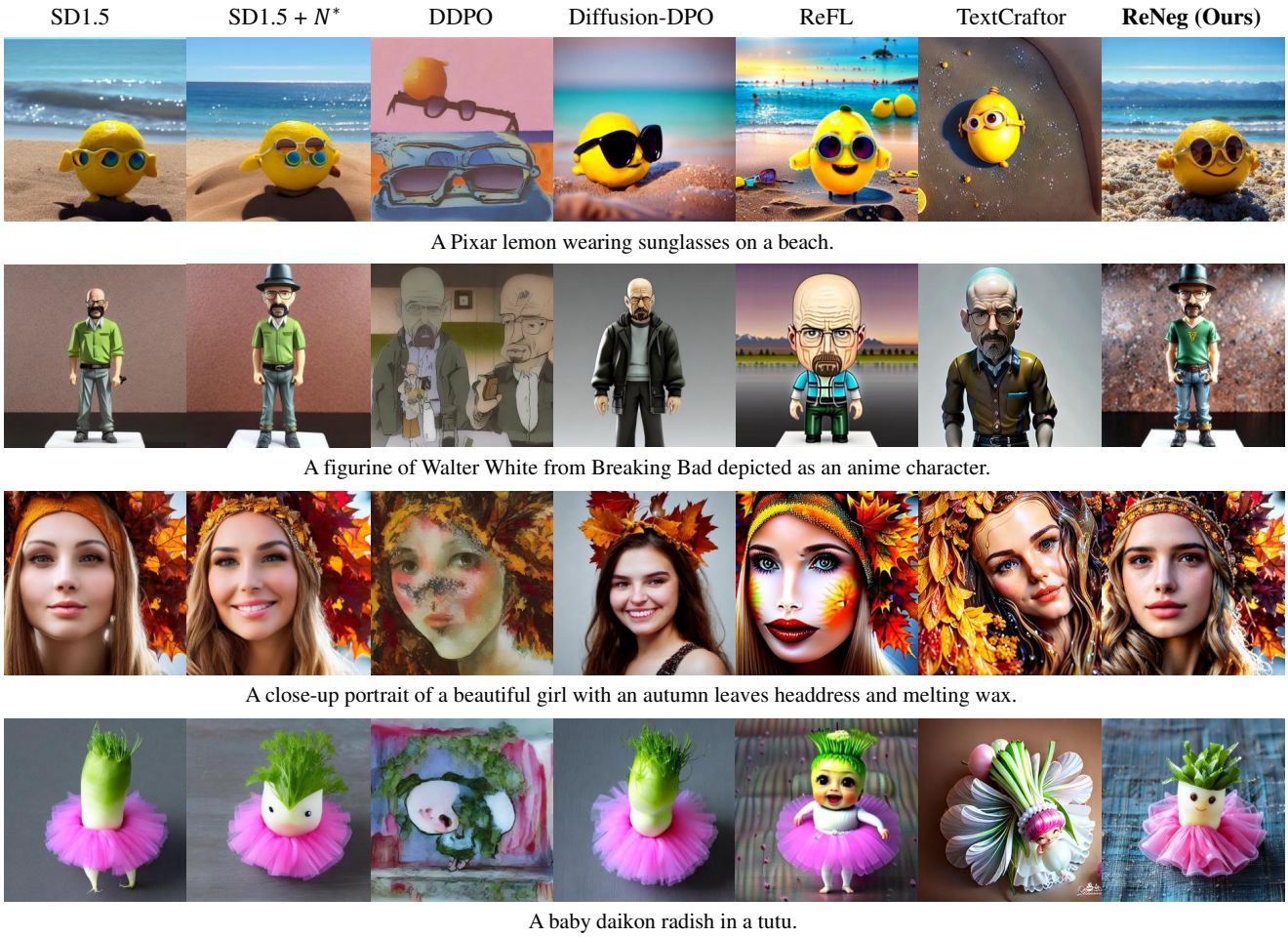

Figure 3. Comparison with methods requiring tuning all model parameters of SD. Prompts are sourced from the HPSv2 and Parti-Prompts benchmarks. All images are generated at a resolution of 512×512, using the same initial noise and seed to ensure fairness.

Figure 4. Qualitative results of transferring our ReNeg to ControlNet under different conditions.

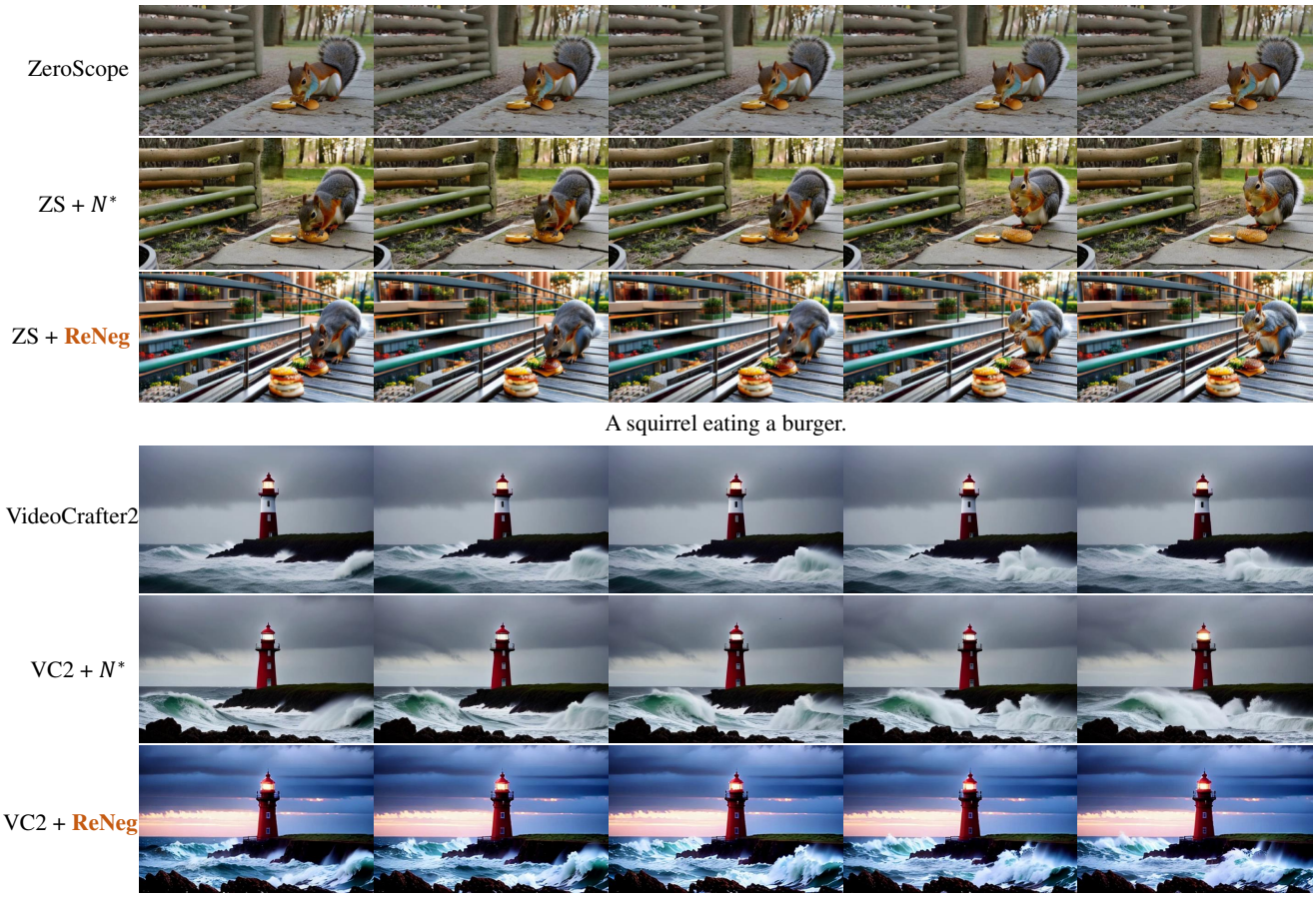

Figure 5. Qualitative results of transferring our ReNeg to various T2V models. "ZS" and "VC2" represent ZeroScope and VideoCrafter2.

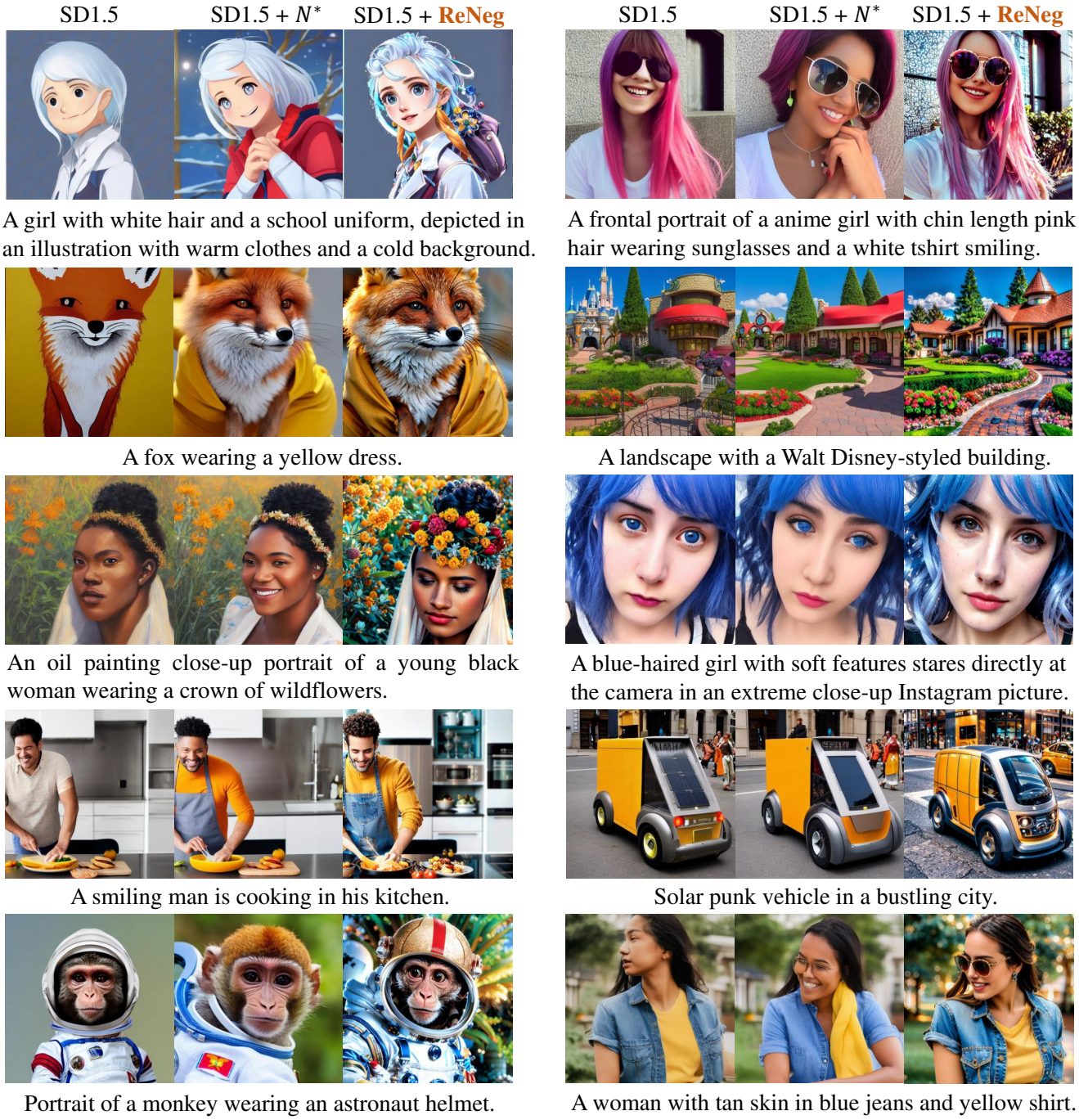

Figure 6. More examples generated by ReNeg. Given the same text prompt and fixed seed, the images above correspond to outputs from the original SD1.5, SD1.5 with a handcrafted negative prompt, and SD1.5 with our proposed negative embedding, respectively.

5.3. Comparison with Methods Requiring Finetuning of SD

As shown in Fig. 3, the visual quality of images generated by Diffusion-DPO [8] appears comparable to SD1.5. However, the performance of DDPO-based models [1] declines dramatically, which is primarily due to their finetuning on specific subdomains, such as animals, limiting their ability to generalize to other prompts. While TextCraftor [6] generally yields high-quality images, it suffers from a notable drawback: a lack of realism, with an overall generation style leaning toward oil painting. In contrast, our method generates more appealing imagery, featuring vivid colors and

well-balanced compositions that align closely with human aesthetic preferences.

5.4. Combine with Recaption-based Methods

As shown in Tab. 2 our method can achieve additional performance gains by combining with positive prompt refinement. Provided that the recaptioned positive prompts are reasonable, incorporating our negative embedding significantly enhances visual appeal. However, in some cases, the improved prompts generated by Promptist [4] may lack rationality, limiting their effectiveness.

5.5. Transfer to ControlNet and T2V Models

To showcase the generalization capability of our negative embedding, we present additional generation results of ControlNet under various conditions and the outputs from different T2V models. As illustrated in Fig. 4 and Fig. 5, our method can be seamlessly transferred to ControlNet and other T2V models, demonstrating its adaptability and effectiveness across diverse scenarios. For more intuitive visualizations, please refer to the accompanying video file, *video.mp4*.

5.6. More Results of ReNeg

We provide additional visual results, as shown in Fig. 6. It is evident that the use of our negative embedding significantly enhances the aesthetic quality of the generated images.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 5

[2] cerspense. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 1, 2

[3] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 1, 2

[4] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *NeurIPS*, 2024. 6

[5] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 1

[6] Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Textcraftor: Your text encoder can be image quality controller. In *CVPR*, 2024. 5

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[8] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. 5

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[10] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 36, 2024. 2