Collaborative Score Distillation for Consistent Visual Synthesis

Subin Kim [*1] **Kyungmin Lee** [*1] **Junesuk Choi** [1] **Jongheon Jeong** [1] **Kihyuk Sohn** [2] **Jinwoo Shin** [1]

Abstract

Generative priors of large-scale text-to-image diffusion models enable a wide range of new generation and editing applications on diverse visual modalities. However, when adapting these priors to complex visual modalities, often represented as multiple images (e.g., video), achieving consistency across a set of images is challenging. In this paper, we address this challenge with a novel method, Collaborative Score Distillation (CSD). CSD is based on the Stein Variational Gradient Descent (SVGD). Specifically, we propose to consider multiple samples as "particles" in the SVGD update and combine their score functions to distill generative priors over a set of images synchronously. Thus, CSD facilitates seamless integration of information across 2D images, leading to a consistent visual synthesis across multiple samples. We show the effectiveness of CSD in a variety of tasks, encompassing the visual editing of panorama images, videos, and 3D scenes. Our results underline the competency of CSD as a versatile method for enhancing inter-sample consistency, thereby broadening the applicability of text-to-image diffusion models.[1]

1. Introduction

Text-to-image diffusion models (Saharia et al., 2022; Ramesh et al., 2022; Nichol et al., 2021; Rombach et al., 2022) have been scaled up by using billions of image-text pairs (Schuhmann et al., 2021; 2022) and efficient architectures (Ho et al., 2020; Song et al., 2020b;a; Rombach et al., 2022), showing impressive capability in synthesizing high-quality, realistic, and diverse images with the text given as an input. Furthermore, they have branched into various applications, such as image-to-image translation (Meng et al., 2021; Bar-Tal et al., 2022; Hertz et al., 2022; Kawar et al., 2022; Brooks et al., 2022; Mokady et al., 2022; Voynov et al., 2022), controllable generation (Zhang & Agrawala, 2023), or personalization (Gal et al., 2022; Ruiz et al., 2022). One of the latest applications in this regard is to translate the capability into other complex modalities, viz., beyond 2D images (Ho et al., 2022; Esser et al., 2023) without modifying diffusion models and using modality-specific training data. This paper focuses on adapting the knowledge of pre-trained text-to-image diffusion models to complex visual generative tasks beyond 2D images without necessitating additional training of diffusion models or utilizing modality-specific training data.

We start from an intuition that many complex visual data, e.g., videos and 3D scenes, are represented as a *set of images* constrained by modality-specific consistency. For example, a video is a set of frames requiring temporal consistency, and a 3D scene is built from a set of multi-view frames satisfying view consistency. Unfortunately, image diffusion models do not have a built-in capability to ensure consistency between a set of images for synthesis or editing because their generative sampling process does not take into account the consistency when using the image diffusion model as is. As such, when applying image diffusion models on these complex data without consistency in consideration, it results in a highly incoherent output, as in Figure 2 (Patch-wise Crop), where one can easily identify where images are stitched. Such behaviors are also reported in video editing, thus, recent works (Qi et al., 2023; Khachatryan et al., 2023; Liu et al., 2023; Ceylan et al., 2023) propose to inject video-specific temporal consistency in image diffusion models.

Here, we take attention to an alternative approach, Score Distillation Sampling (SDS) (Poole et al., 2022), which enables the optimization of arbitrary differentiable operators by leveraging the rich generative prior of text-to-image diffusion models. While Poole et al. (2022) has shown the effectiveness of SDS in generating 3D objects from the text by resorting on Neural Radience Fields (Mildenhall et al., 2020) priors which inherently suppose coherent geometry in 3D space by density modeling, it has not been studied for consistent visual synthesis of other modalities.

[*]Equal contribution [1]KAIST [2]Google Research. Correspondence to: Subin Kim <subin-kim@kaist.ac.kr>, Kyungmin Lee <kyungmnlee@kaist.ac.kr>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling

[1]Visualizations are available at the website https://subin-kim-cv.github.io/CSD.

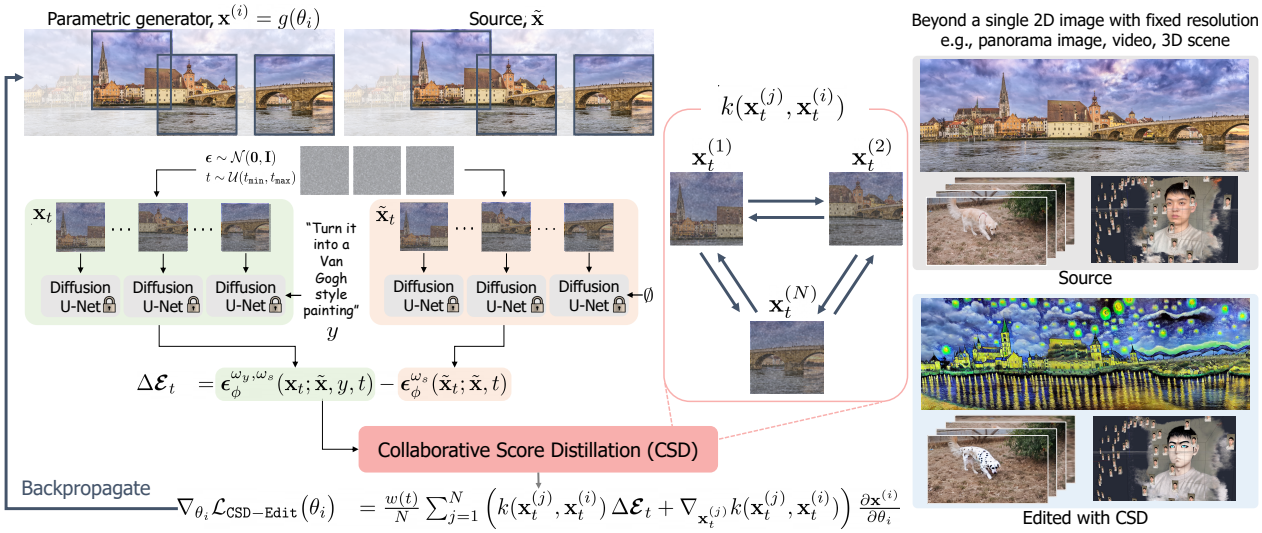

Figure 1: **Method overview**. CSD-Edit enables various visual-to-visual translations with two novel components. First, a new score distillation scheme using Stein variational gradient descent, which considers inter-sample relationships (Section 3.1) to synthesize a set of images while preserving modality-specific consistency constraints. Second, our method edits images with minimal information given from text instruction by subtracting image-conditional noise estimate instead of random noise during score distillation (Section 3.2). By doing so, CSD-Edit is used for text-guided manipulation of various visual domains, e.g., panorama images, videos, and 3D scenes (Section 4).

In this paper, we propose *Collaborative Score Distillation* (CSD), a simple yet effective method that extends the singular of the text-to-image diffusion model for consistent visual synthesis. The crux of our method is two-fold: first, we establish a generalization of SDS by using Stein variational gradient descent (SVGD), where multiple samples share their knowledge distilled from diffusion models to accomplish inter-sample consistency. Second, we present CSD-Edit, an effective method for consistent visual editing by leveraging CSD with Instruct-Pix2Pix (Brooks et al., 2022), a recently proposed instruction-guided image diffusion model (See Figure 1).

We demonstrate the versatility of our method in various applications such as panorama image editing, video editing, and reconstructed 3D scene editing. In editing a panorama image, we show that CSD-Edit obtains spatially consistent image editing by optimizing multiple patches of an image. Also, compared to other methods, our approach achieves a better trade-off between source-target image consistency and instruction fidelity. In video editing experiments, CSD-Edit obtains temporal consistency by taking multiple frames into optimization, resulting in temporal frame-consistent video editing. Furthermore, we apply CSD-Edit to 3D scene editing and generation, by encouraging consistency among multiple views.

Related works. We present a detailed discussion of the related works in Appendix A.

2. Preliminaries

Score distillation sampling. Poole et al. (2022) proposed Score Distillation Sampling (SDS), an alternative sample generation method by distilling the rich knowledge of text-to-image diffusion models. SDS allows optimization of any differentiable image generator, e.g., Neural Radiance Fields (Mildenhall et al., 2020) or the image space itself. Formally, let $\mathbf{x} = g(\theta)$ be an image rendered by a differentiable generator g with parameter θ, then SDS minimizes density distillation loss (Oord et al., 2018) which is KL divergence between the posterior of $\mathbf{x} = g(\theta)$ and the text-conditional density p_ϕ^ω:

$$\mathbb{E}_{t,\epsilon}\left[\alpha_t/\sigma_t\, D_{\text{KL}}\left(q\left(\mathbf{x}_t|\mathbf{x}=g(\theta)\right)\,\|\,p_\phi^\omega(\mathbf{x}_t;y,t)\right)\right]. \quad (1)$$

For an efficient implementation, SDS updates the parameter θ by randomly choosing timesteps $t \sim \mathcal{U}(t_{\text{min}}, t_{\text{max}})$ and forward $\mathbf{x} = g(\theta)$ with noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to compute the gradient as follows:

$$\mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi^\omega(\mathbf{x}_t;y,t) - \boldsymbol{\epsilon}\right)\frac{\partial\mathbf{x}}{\partial\theta}\right]. \quad (2)$$

Remark that the U-Net Jacobian $\partial\epsilon_\phi^\omega(\mathbf{z}_t;y,t)/\partial\mathbf{z}_t$ is omitted as it is computationally expensive to compute, and degrades performance when conditioned on small noise levels. The range of timesteps t_{min} and t_{max} are chosen to sample from not too small or large noise levels, and the guidance scales are chosen to be larger than those used for image generation.

We present a more detailed explanation of the preliminaries for our work in Appendix B.1.

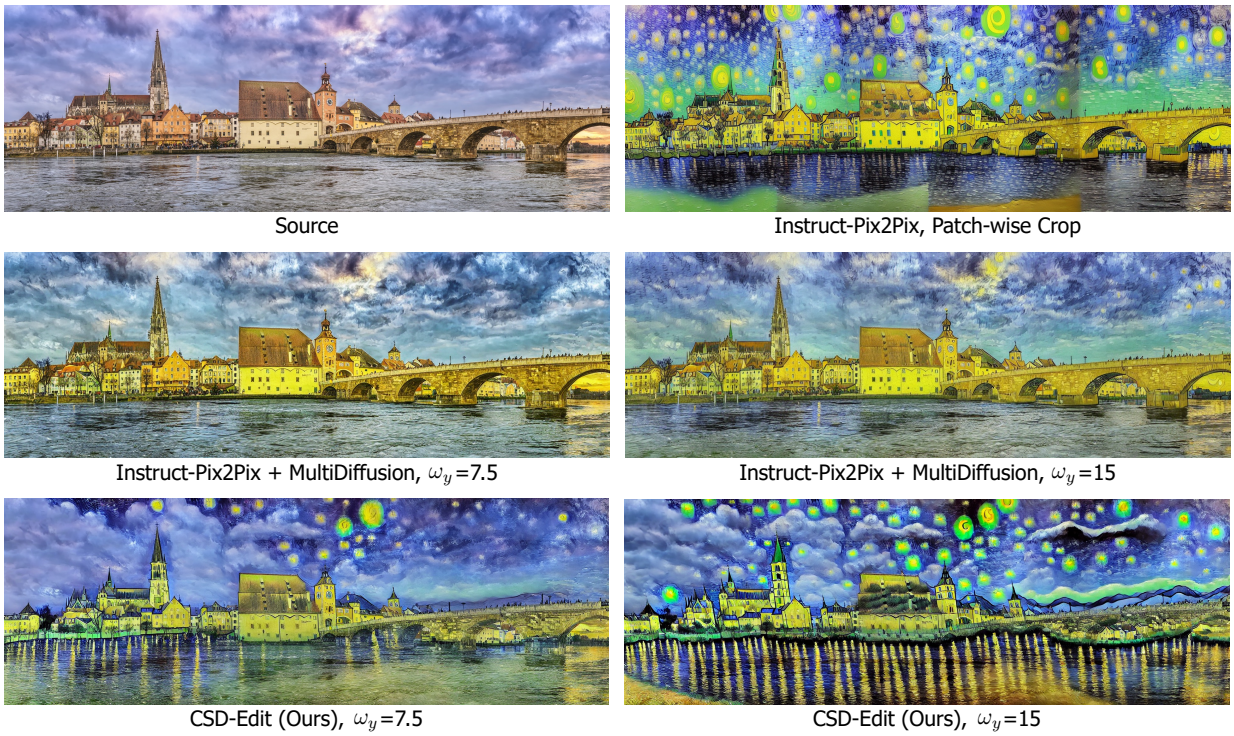

Figure 2: **Panorama image editing**. (Top right) Instruct-Pix2Pix (Brooks et al., 2022) on cropped patches results in inconsistent image editing. (Second row) Instruct-Pix2Pix with MultiDiffusion (Bar-Tal et al., 2023) edits to consistent image, but less fidelity to the instruction, even with high guidance scale ω_y. (Third row) CSD-Edit provides consistent image editing with better instruction-fidelity by setting proper guidance scale.

3. Method

In this section, we introduce *Collaborative Score Distillation* (CSD) for consistent synthesis and editing of multiple samples. We first derive a collaborative score distillation method using Stein variational gradient descent (Section 3.1) and propose an effective image editing method using CSD, i.e., CSD-Edit, that leads to coherent editing of multiple images with instruction (Section 3.2). Lastly, we present various applications of CSD-Edit in editing panorama images, videos, and 3D scenes (Section 4 and Section B.4).

3.1. Collaborative score distillation

Suppose a set of parameters $\{\theta_i\}_{i=1}^N$ that generates images $\mathbf{x}^{(i)} = g(\theta_i)$. Our goal is to update each θ_i by distilling the smoothed densities from the diffusion model by minimizing KL divergence using SVGD demonstrated in Section B.1 so that each θ_i can be updated in sync with updates of other parameters. At each update, CSD samples $t \sim \mathcal{U}(t_{\texttt{min}}, t_{\texttt{max}})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and update each $\theta_i, i \in [N]$ as follows:

$$\nabla_{\theta_i} \mathcal{L}_{\texttt{CSD}}(\theta_i) = \frac{w(t)}{N} \sum_{j=1}^{N} [k(\mathbf{x}_t^{(j)}, \mathbf{x}_t^{(i)})(\boldsymbol{\epsilon}_\phi^\omega(\mathbf{x}_t^{(j)}; y, t) - \boldsymbol{\epsilon}) + \nabla_{\mathbf{x}_t^{(j)}} k(\mathbf{x}_t^{(j)}, \mathbf{x}_t^{(i)})] \frac{\partial \mathbf{x}^{(i)}}{\partial \theta_i}. \tag{3}$$

We note that CSD is a generalization to Score Distillation Sampling (SDS) introduced by Poole et al. (2022) to multiple samples, where SDS is a special case when $N = 1$. For a full derivation and further details, we refer to Appendix B.

As the pairwise kernel values are multiplied by the noise prediction term, each parameter update on θ_i is affected by other parameters, i.e., the scores are mixed with importance weights according to the affinity among samples. The more similar samples tend to exchange more score updates, while different samples tend to interchange the score information less. The gradient of the kernels acts as a repulsive force that prevents the mode collapse of samples. Moreover, we note that Eq. (3) does not make any assumption on the relation between θ_i's or their order besides them being a set of images to be synthesized coherently with each other. As such, CSD is also applicable to arbitrary image generators, as well as text-to-3D synthesis in DreamFusion (Poole et al., 2022), which we compare in Appendix D.2.

3.2. Text-guided editing by collaborative score distillation

In this section, we introduce a text-guided visual editing method using Collaborative Score Distillation (CSD-Edit).

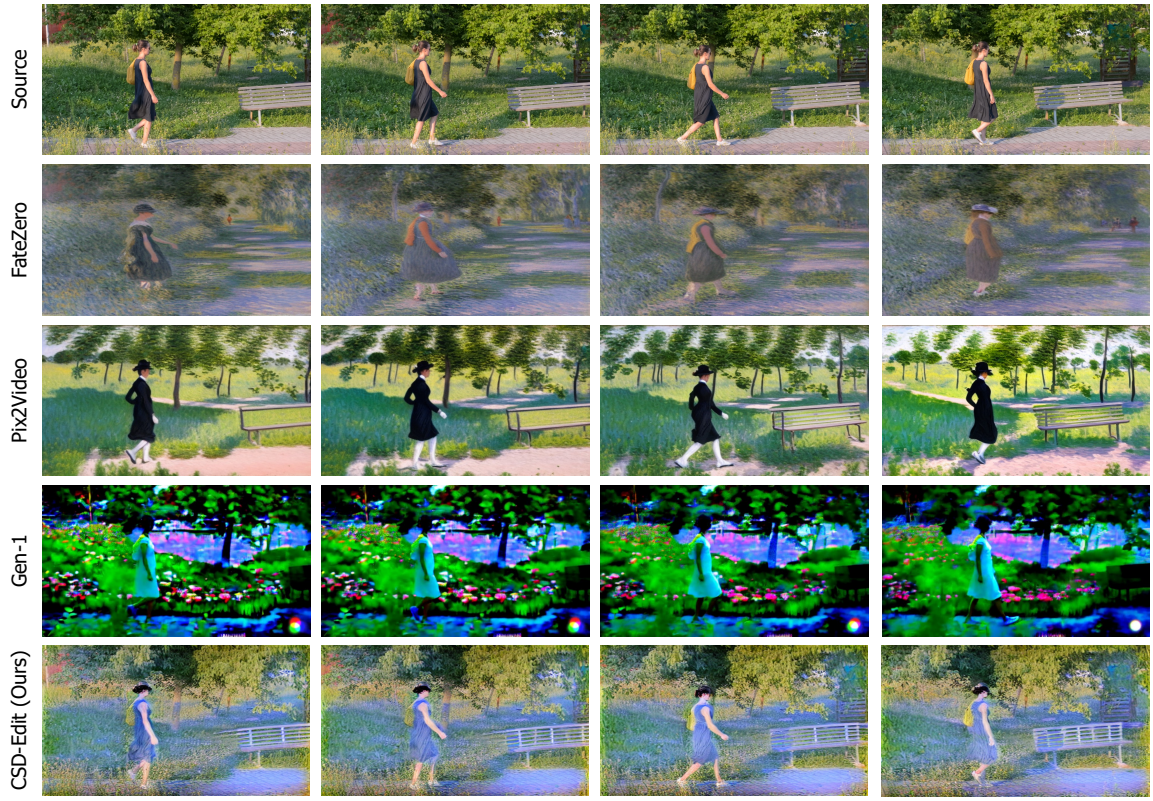

Figure 3: **Video editing**. Qualitative results on the lucia video in DAVIS 2017 (Pont-Tuset et al., 2017). CSD-Edit shows frame-wise consistent editing providing coherent content across video frames e.g., consistent color and background without changes in person. Compared to Gen-1 (Esser et al., 2023), a video editing method trained on a large video dataset, CSD-Edit shows high-quality video editing results reflecting given prompts.

Given source images $\tilde{\mathbf{x}}^{(i)} = g(\tilde{\theta}_i)$ with parameters $\tilde{\theta}_i$, we optimize new target parameters $\{\theta_i\}_{i=1}^N$ with $\mathbf{x}^{(i)} = g(\theta_i)$ such that 1) each $\mathbf{x}^{(i)}$ follows the instruction prompt, 2) preserves the semantics of source images as much as possible, and 3) the obtained images are consistent with each other. To accomplish these, we update each parameter θ_i, initialized with $\tilde{\theta}_i$, using CSD with noise estimate $\boldsymbol{\epsilon}_\phi^{\omega_y,\omega_s}$ of Instruct-Pix2Pix. However, this approach often results in blurred outputs, leading to the loss of details of the source image (see Figure 7). This is because the score distillation term subtracts random noise $\boldsymbol{\epsilon}$, which perturbs the undesirable details of source images.

We handle this issue by adjusting the noise prediction term that enhances the consistency between source and target images. When computing the gradient, subtracting a random noise $\boldsymbol{\epsilon}$ in Eq. (2) is a crucial factor, which helps optimization by reducing the variance of a gradient. Therefore, we amend the optimization by changing the random noise into a better baseline function. Since our goal is to edit an image with only minimal information given text instructions, we set the baseline by the image-conditional noise estimate of the Instruct-Pix2Pix model without giving text instructions on the source image. To be specific, our CSD-Edit is given as follows where we let $k_{ij} = k(\mathbf{x}_t^{(j)}, \mathbf{x}_t^{(i)})$ for brevity.

$$\nabla_{\theta_i}\mathcal{L}_{\texttt{CSD-Edit}}(\theta_i) = \frac{w(t)}{N}\sum_{j=1}^{N}\left[k_{ij}\Delta\boldsymbol{\mathcal{E}}_t^{(i)} + \nabla_{\mathbf{x}_t^{(j)}}k_{ij}\right]\frac{\partial\mathbf{x}^{(i)}}{\partial\theta_i},$$

$$\Delta\boldsymbol{\mathcal{E}}_t^{(i)} = \boldsymbol{\epsilon}_\phi^{\omega_y,\omega_s}(\mathbf{x}_t^{(i)};\tilde{\mathbf{x}},y,t) - \boldsymbol{\epsilon}_\phi^{\omega_s}(\tilde{\mathbf{x}}_t^{(i)};\tilde{\mathbf{x}},t).$$

In Appendix E, we validate our findings on the effect of baseline noise on image editing performance. We notice that CSD-Edit presents an alternative way to utilize Instruct-Pix2Pix in image-editing without any finetuning of diffusion models, by posing as an optimization problem.

4. Experiments

4.1. Text-guided panorama image editing

For the panorama image-to-image translation task, we compare CSD-Edit with variants of Instruct-Pix2Pix. The experimental details are in Appendix C.1. In Figure 4, we plot the CLIP scores of different image editing methods

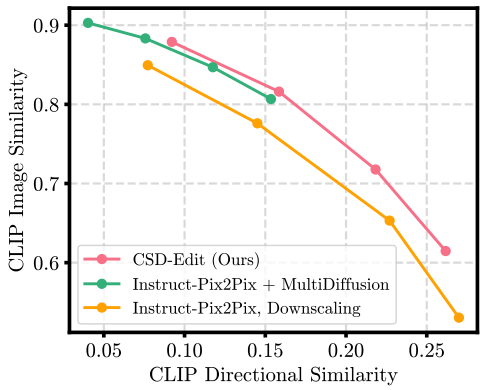

Figure 4: **Panorama image editing**. Comparison of CSD-Edit with baselines at different guidance scales $\omega_y \in \{3.0, 5.0, 7.5, 10.0\}$.

Table 1: **Video editing**. Quantitative comparison of CSD-Edit with baselines on video editing. Bold indicates the best results.

	CLIP Directional Similarity ↑	CLIP Image Consistency ↑	LPIPS ↓
FateZero (Qi et al., 2023)	0.314	0.948	0.267
Pix2Vid (Ceylan et al., 2023)	0.230	0.949	0.283
CSD-Edit (Ours)	**0.320**	**0.957**	**0.236**

Table 2: **3D scene editing**. Quantitative comparison of CSD-Edit with baselines on 3D scene editing. Bold indicates the best results.

	CLIP Directional Similarity ↑	CLIP Image Consistency ↑	LPIPS ↓
IN2N (Brooks et al., 2022)	0.230	0.994	0.048
CSD-Edit (Ours)	**0.239**	**0.995**	**0.043**

with different guidance scales. We notice that CSD-Edit provides the best trade-off between the consistency between source and target images and fidelity to the instruction. Figure 2 provides a qualitative comparison between panorama image editing methods. Remark that Instruct-Pix2Pix + MultiDiffusion is able to generate spatially consistent images, however, the edited images show inferior fidelity to the text instruction even when using a large guidance scale. Additional qualitative results are in Appendix F.

4.2. Text-guided video editing

For the video editing experiments, we primarily compare CSD-Edit with existing zero-shot video editing schemes that employ text-to-image diffusion models. Please refer to Appendix C.2 for a detailed description of the baseline methods and experimental setup. Table 1 summarize a quantitative comparison between CSD-Edit and the baselines. We notice that CSD-Edit consistently outperforms the existing zero-shot video editing schemes in terms of both temporal consistency and fidelity to given text prompts. Moreover, Figure 3 qualitatively demonstrates the superiority of CSD over the baselines on video-stylization and object-aware editing tasks. Impressively, CSD shows comparable editing performance to Gen-1 (Esser et al., 2023) even without training on a large-scale video dataset and any architectural modification to the diffusion model. See Appendix F for additional qualitative results.

4.3. Text-guided 3D scene editing

For the text-guided 3D scene editing experiments, we mainly compare our approach with Instuct-NeRF2NeRF (IN2N) (Haque et al., 2023). Detailed explanations for experiments can be found in Appendix C.3. Table 2 and Figure 12 in Appendix F summarize the comparison between CSD-Edit and IN2N. We notice that CSD-Edit enables a wide-range control of 3D NeRF scenes, such as delicate attribute manipulation (e.g., facial expression alterations) and scene-stylization (e.g., conversion to the animation style). Especially, we notice two advantages of CSD-Edit compared to IN2N. First, CSD-Edit presents high-quality details to the edited 3D scene by providing multi-view consistent training views during NeRF optimization. In Figure 12, one can observe that CSD-Edit captures sharp details of anime character, while IN2N results in blurry face. Second, CSD-Edit is better at preserving the semantics of source 3D scenes, e.g., backgrounds or colors. For instance, we notice that CSD-Edit allows subtle changes in facial expressions without changing the color of the background or adding a beard to the face as shown in Figure 12.

4.4. Text-to-3D generation

We explore the effectiveness of CSD in text-to-3D generation tasks by comparing with DreamFusion (Poole et al., 2022). We refer to Appendix D.2 for experimental setup. In Figure 5, CSD generates higher quality 3D NeRF scene than SDS, due to the coherent 2D diffusion prior distilled during the optimization. We provide more quantitative (Table 3) and qualitative (Figure 13) results in Appendix D.2.

5. Conclusion

In this paper, we propose Collaborative Score Distillation (CSD) for consistent visual synthesis. CSD is built upon Stein variational gradient descent, where multiple samples share their knowledge distilled from text-to-image diffusion models during the update. Furthermore, we propose CSD-Edit that gives us consistent editing of images by distilling essential, yet concise information from instruction-guided diffusion models. We demonstrate the effectiveness of our method in text-guided translation of diverse and complex visual modalities, including high-resolution images, videos, and real 3D scenes, outperforming previous methods both quantitatively and qualitatively.

References

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., and Dekel, T. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pp. 707–723. Springer, 2022.

Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Ceylan, D., Huang, C.-H. P., and Mitra, N. J. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023.

Chen, R., Chen, Y., Jiao, N., and Jia, K. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023.

Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. *arXiv preprint arXiv:2302.11552*, 2023.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017.

Haque, A., Tancik, M., Efros, A., Holynski, A., and Kanazawa, A. Instruct-nerf2nerf: Editing 3d scenes with instructions. 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Hertz, A., Aberman, K., and Cohen-Or, D. Delta denoising score. *arXiv preprint arXiv:2304.07090*, 2023.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th*

European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII, pp. 423–439. Springer, 2022.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Liu, S., Zhang, Y., Li, W., Lin, Z., and Jia, J. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

Melas-Kyriazi, L., Rupprecht, C., Laina, I., and Vedaldi, A. Realfusion: 360 {\deg} reconstruction of any object from a single image. *arXiv preprint arXiv:2302.10663*, 2023.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamorthi, R., and Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. 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 18*, pp. 234–241. Springer, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, K., Han, L., Liu, B., Metaxas, D., and Elgammal, A. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023.

Tang, J. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion.

Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., and Chen, D. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023.

Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., and Tombari, F. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023.

Valevski, D., Kalman, M., Matias, Y., and Leviathan, Y. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022.

Voynov, A., Aberman, K., and Cohen-Or, D. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022.

Wang, C., Jiang, R., Chai, M., He, M., Chen, D., and Liao, J. Nerf-art: Text-driven neural radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022.

Wu, J. Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

Xie, X., Zhou, P., Li, H., Lin, Z., and Yan, S. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022.

Zhang, L. and Agrawala, M. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Appendix

Website: https://icmlw2023csd.github.io/

A. Related works

Following remarkable success of text-to-image diffusion models (Rombach et al., 2022; Ho et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Balaji et al., 2022), numerous works have attempted to exploit rich knowledge of text-to-image diffusion models for various visual editing tasks including images (Meng et al., 2021; Couairon et al., 2022; Kawar et al., 2022; Valevski et al., 2022; Brooks et al., 2022; Hertz et al., 2022; Mokady et al., 2022), videos (Wu et al., 2022; Ceylan et al., 2023), 3D scenes (Haque et al., 2023), etc. However, extending existing image editing approaches to more complex visual modalities often faces a new challenge; consistency between edits, e.g., spatial consistency in high-resolution images, temporal consistency in videos, and multi-view consistency in 3D scenes. While prior works primarily focus on designing task-specific methods (Liu et al., 2023; Qi et al., 2023; Ceylan et al., 2023) or model fine-tuning for complex modalities (Wu et al., 2022), we present a modality-agnostic novel method for editing, effectively capturing consistency between samples.

The most related to our work is DreamFusion (Poole et al., 2022), which introduced Score Distillation Sampling (SDS) for the creation of 3D assets, leveraging the power of text-to-image diffusion models. Despite the flexible merit of SDS to enable the optimization of arbitrary differentiable operators, most works mainly focus on applying SDS to enhance the synthesis quality of 3D scenes by introducing 3D specific frameworks (Lin et al., 2022; Tsalicoglou et al., 2023; Melas-Kyriazi et al., 2023; Chen et al., 2023; Tang et al., 2023). Although there exists some work to apply SDS for visual domains other than 3D scenes, they have limited their scope to image editing (Hertz et al., 2023), or image generation (Song et al., 2022). Here, we clarify that our main focus is not to improve the performance of SDS for a specific task, but rather to shift the focus to generalizing it from a new perspective in a principled way. To the best of our knowledge, we are the first to center our work on the generalization of SDS and introduce a novel method that simply but effectively adapts text-to-image diffusion models to diverse high-dimensional visual syntheses beyond a single 2D image with a fixed resolution. Importantly, our approach does not require modifications to the diffusion model or the use of modality-specific training data.

B. Technical details

In this section, we introduce preliminaries and provide detailed explanations on the proposed methods, CSD and CSD-Edit.

B.1. Preliminaries

Diffusion models. Generative modeling with diffusion models consists of a forward process q that gradually adds Gaussian noise to the input $\mathbf{x}_0 \sim p_{\texttt{data}}(\mathbf{x})$, and a reverse process p which gradually denoises from the Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Formally, the forward process $q(\mathbf{x}_t|\mathbf{x}_0)$ at timestep t is given by $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$, where σ_t and $\alpha_t^2 = 1 - \sigma_t^2$ are pre-defined constants designed for effective modeling (Song et al., 2020b; Kingma et al., 2021; Karras et al., 2022). Given enough timesteps, reverse process p also becomes a Gaussian and the transitions are given by posterior q with optimal MSE denoiser (Sohl-Dickstein et al., 2015), i.e., $p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mathbf{x}_t - \hat{\mathbf{x}}_\phi(\mathbf{x}_t; t), \sigma_t^2 \mathbf{I})$, where $\hat{\mathbf{x}}_\phi(\mathbf{x}_t; t)$ is a learned optimal MSE denoiser. Ho et al. (2020) proposed to train an U-Net (Ronneberger et al., 2015) autoencoder $\boldsymbol{\epsilon}_\phi(\mathbf{x}_t; t)$ by minimizing following objective:

$$\mathcal{L}_{\texttt{Diff}}(\phi; \mathbf{x}) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\big[w(t)\|\boldsymbol{\epsilon}_\phi(\mathbf{x}_t; t) - \boldsymbol{\epsilon}\|_2^2\big], \quad \mathbf{x}_t = \alpha_t \mathbf{x}_0 + \alpha_t \boldsymbol{\epsilon} \tag{4}$$

where $w(t)$ is a weighting function for each timestep t. Text-to-image diffusion models (Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Nichol et al., 2021) are trained by Eq. (4) with $\boldsymbol{\epsilon}_\phi(\mathbf{x}_t; y, t)$ that estimates the noise conditioned on the text prompt y. At inference, those methods rely on Classifier-free Guidance (CFG) (Ho & Salimans, 2022), which allows higher quality sample generation by introducing additional parameter $\omega_y \geq 1$ as follows:

$$\boldsymbol{\epsilon}_\phi^\omega(\mathbf{x}_t; y, t) = \boldsymbol{\epsilon}_\phi(\mathbf{x}_t; t) + \omega_y\big(\boldsymbol{\epsilon}_\phi(\mathbf{x}_t; y, t) - \boldsymbol{\epsilon}_\phi(\mathbf{x}_t; t)\big) \tag{5}$$

By setting the appropriate guidance scale $\omega_y > 0$, one can improve fidelity to the text prompt at the cost of diversity. Throughout the paper, we refer $p_\phi^{\omega_y}(\mathbf{x}_t; y, t)$ a conditional distribution of a text y.

Instruction-based image editing by Instruct-Pix2Pix. Recently, many works have demonstrated the capability of diffusion models in editing or stylizing images (Meng et al., 2021; Kawar et al., 2022; Bar-Tal et al., 2022; Hertz et al., 2022; Brooks

et al., 2022). Among them, Brooks et al. (2022) proposed Instruct-Pix2Pix, where they finetuned Stable Diffusion (Rombach et al., 2022) models with the source image, text instruction, edited image (edited by Prompt-to-Prompt (Hertz et al., 2022)) triplet to enable instruction-based editing of an image. Given source image $\tilde{\mathbf{x}}$ and instruction y, the noise estimate at time t is given as

$$\begin{aligned}\boldsymbol{\epsilon}_\phi^{\omega_s,\omega_y}(\mathbf{x}_t;\tilde{\mathbf{x}},y,t) = \boldsymbol{\epsilon}_\phi(\mathbf{x}_t;t) &+ \omega_s\big(\boldsymbol{\epsilon}_\phi(\mathbf{x}_t;\tilde{\mathbf{x}},t) - \boldsymbol{\epsilon}_\phi(\mathbf{x}_t;t)\big) \\ &+ \omega_y\big(\boldsymbol{\epsilon}_\phi(\mathbf{x}_t;\tilde{\mathbf{x}},y,t) - \boldsymbol{\epsilon}_\phi(\mathbf{x}_t;\tilde{\mathbf{x}},t)\big),\end{aligned} \tag{6}$$

where ω_y is CFG parameter for text as in Eq. (5) and ω_s is an additional CFG parameter that controls the fidelity to the source image $\tilde{\mathbf{x}}$.

Stein variational gradient descent. The original motivation of Stein variational gradient descent (SVGD) (Liu & Wang, 2016) is to solve a variational inference problem, where the goal is to approximate a target distribution from a simpler distribution by minimizing KL divergence. Formally, suppose p is a target distribution with a known score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ that we aim to approximate, and $q(\mathbf{x})$ is a known source distribution. Liu & Wang (2016) showed that the steepest descent of KL divergence between q and p is given as follows:

$$\mathbb{E}_{q(\mathbf{x})}\big[\mathbf{f}(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \mathrm{Tr}(\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}))\big], \tag{7}$$

where $\mathbf{f} : \mathbb{R}^D \to \mathbb{R}^D$ is any smooth vector function that satisfies $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{f}(\mathbf{x}) = 0$. Remark that Eq. (7) becomes zero if we replace $q(\mathbf{x})$ with $p(\mathbf{x})$ in the expectation term, which is known as Stein's identity (Gorham & Mackey, 2017). Here, the choice of the critic $\mathbf{f}$ is crucial in its convergence and computational tractability. To that end, Liu & Wang (2016) proposed to constrain $\mathbf{f}$ in the Reproducing Kernel Hilbert Space (RKHS) which yields a closed-form solution. Specifically, given a positive definite kernel $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$, Stein variational gradient descent provides the greedy directions as follows:

$$\mathbf{x} \leftarrow \mathbf{x} - \eta\Delta\mathbf{x}, \quad \Delta\mathbf{x} = \mathbb{E}_{q(\mathbf{x}')}\big[k(\mathbf{x},\mathbf{x}')\nabla_{\mathbf{x}'} \log p(\mathbf{x}') + \nabla_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\big], \tag{8}$$

with small step size $\eta > 0$. The SVGD update in Eq. (8) consists of two terms that play different roles: the first term moves the particles towards the high-density region of target density $p(\mathbf{x})$, where the direction is smoothed by kernels of other particles. The second term acts as a repulsive force that prevents the mode collapse of particles. One can choose different kernel functions, while we resort to a standard Radial Basis Kernel (RBF) kernel $k(\mathbf{x},\mathbf{x}') = \exp(-\frac{1}{h}\|\mathbf{x}-\mathbf{x}'\|_2^2)$ with bandwidth $h > 0$.

B.2. CSD derivation.

Consider a set of parameters $\{\theta_i\}_{i=1}^N$ which generates images $\mathbf{x}^{(i)} = g(\theta_i)$. For each timestep $t \sim \mathcal{U}(t_{\texttt{min}}, t_{\texttt{max}})$, we aim at minimizing the following KL divergence

$$D_{\mathrm{KL}}\big(q(\mathbf{x}_t^{(i)}|\mathbf{x}^{(i)} = g(\theta_i))\|p_\phi(\mathbf{x}_t;y,t)\big)$$

for each $i = 1, 2, \ldots, N$ via SVGD using Eq. (8). To this end, we approximate the score function, (i.e., the gradient of log-density) by the noise predictor from the diffusion model as follows:

$$\nabla_{\mathbf{x}_t^{(i)}} \log p_\phi(\mathbf{x}_t^{(i)};y,t) \approx -\frac{\boldsymbol{\epsilon}_\phi(\mathbf{x}_t^{(i)};y,t)}{\sigma_t}.$$

Then, the gradient of the score function with respect to parameter θ_i is given by

$$\nabla_{\theta_i} \log p_\phi(\mathbf{x}_t^{(i)};y,t) = \nabla_{\mathbf{x}_t^{(i)}} \log p_\phi(\mathbf{x}_t^{(i)};y,t)\frac{\partial \mathbf{x}_t^{(i)}}{\partial \theta_i} \approx -\frac{\alpha_t}{\sigma_t}\boldsymbol{\epsilon}_\phi(\mathbf{x}_t^{(i)};y,t)\frac{\partial \mathbf{x}^{(i)}}{\partial \theta}, \tag{9}$$

for each $i = 1, \ldots N$. Finally, to derive CSD, we plug Eq. (9) to Eq. (8) to attain Eq. (3). Also, we subtract the noise $\boldsymbol{\epsilon}$, which helps to reduce the variance of the gradient for better optimization. Following DreamFusion (Poole et al., 2022), we do not compute the Jacobian of U-Net. At a high level, CSD takes the gradient update on each $\mathbf{x}^{(i)}$ using SVGD and updates θ_i by simple chain rule without computing the Jacobian. This formulation makes CSD as a straightforward generalization to SDS for multiple samples and leads to an effective gradient for optimizing consistency among the batch of samples.

B.3. CSD-Edit derivation.

As mentioned above, we subtract the random noise to reduce the variance of CSD gradient estimation. This is in a similar manner to the variance reduction in policy gradient (Schulman et al., 2015), where having a proper baseline function results in faster and more stable optimization. Using this analogy, our intuition is built upon that setting a better baseline function can ameliorate the optimization of CSD. Thus, in image-editing via CSD-Edit, we propose to use image-conditional noise estimate as a baseline function. This allows CSD-Edit to optimize the latent driven by only the influence of instruction prompts. We notice that similar observations were proposed in Delta Denoising Score (DDS) (Hertz et al., 2023), where they introduced an image-to-image translation method that is based on SDS, and the difference of the noise estimate from target prompt and that from source prompt are used. Our CSD can be combined with DDS by changing the noise difference term as follows:

$$\Delta\boldsymbol{\mathcal{E}}_t = \boldsymbol{\epsilon}_\phi(\mathbf{x}_t; y_{\texttt{tgt}}, t) - \boldsymbol{\epsilon}_\phi(\tilde{\mathbf{x}}_t; y_{\texttt{src}}, t),$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are target and source images, $y_{\texttt{tgt}}$ and $y_{\texttt{src}}$ are target and source prompts. However, we found that CSD-Edit with InstructPix2Pix is more amenable in editing real images as it does not require a source prompt. Finally, we remark that CSD-Edit can be applied to various text-to-image diffusion models such as ControlNet (Zhang & Agrawala, 2023), which we leave for future work.

B.4. CSD-Edit for various complex visual domains

Panorama image editing. Diffusion models are usually trained on a fixed resolution (e.g., 512×512 for Stable Diffusion (Rombach et al., 2022)), thus when editing a panorama image (i.e., an image with a large aspect ratio), the editing quality significantly degrades.= Otherwise, one can crop an image into smaller patches and apply image editing on each patch. However this results in spatially inconsistent images (see Figure 2, Patch-wise Crop, and Figure 8 in Appendix E). To that end, we propose to apply CSD-Edit on patches to obtain spatially consistent editing of an image, while preserving the semantics of the source image. Following (Bar-Tal et al., 2023), we sample patches of size 512×512 that overlap using small stride and apply CSD-Edit on the latent space of Stable Diffusion (Rombach et al., 2022). Since we allow overlapping, some pixels might be updated more frequently. Thus, we normalize the gradient of each pixel by counting the appearance.

Video editing. Editing a video with an instruction should satisfy the following: 1) temporal consistency between frames such that the degree of changes compared to the source video should be consistent across frames, 2) ensuring that desired edits in each edited frame are in line with the given prompts while preserving the original structure of source video, and 3) maintaining the sample quality in each frame after editing. To meet these requirements, we randomly sample a batch of frames and update them with CSD-Edit to achieve temporal consistency between frames.

3D scene editing. We consider editing a 3D scene reconstructed by a Neural Radiance Field (NeRF) (Mildenhall et al., 2020), which represents volumetric 3D scenes using 2D images. To edit reconstructed 3D NeRF scenes, it is straightforward to update the training views with edited views and finetune the NeRF with edited views. Here, the multi-view consistency between edited views should be considered since inconsistencies between edits across multiple viewpoints lead to blurry and undesirable artifacts, hindering the optimization of NeRF. To mitigate this, Haque et al. (2023) proposed Instruct-NeRF2NeRF, which performs editing on a subset of training views and updates them sequentially at training iteration with intervals. However, image-wise editing results in inconsistencies between views, thus they rely on the ability of NeRF in achieving multi-view consistency. Contrary to Instruct-NeRF2NeRF, we update the dataset with multiple consistent views through CSD-Edit, which serves as better training resources for NeRF, leading to less artifacts and better preservation of source 3D scene.

C. Implementation details

Setup. For the experiments with CSD-Edit, we use the publicly available pre-trained model of Instruct-Pix2Pix (Brooks et al., 2022)[2] by default. We perform CSD-Edit optimization on the latent space of Stable Diffusion (Rombach et al., 2022) autoencoder. We use SGD optimizer with step learning rate decay, without adding weight decay. We set $t_{\texttt{min}} = 0.2$ and $t_{\texttt{max}} = 0.5$, where original SDS optimization for DreamFusion used $t_{\texttt{min}} = 0.2$ and $t_{\texttt{max}} = 0.98$ because we do not generally require a large scale of noise in editing. We use the guidance scale $\omega_y \in [3.0, 15.0]$ and image guidance scale $\omega_s \in [1.5, 5.0]$.

[2] https://github.com/timothybrooks/instruct-pix2pix

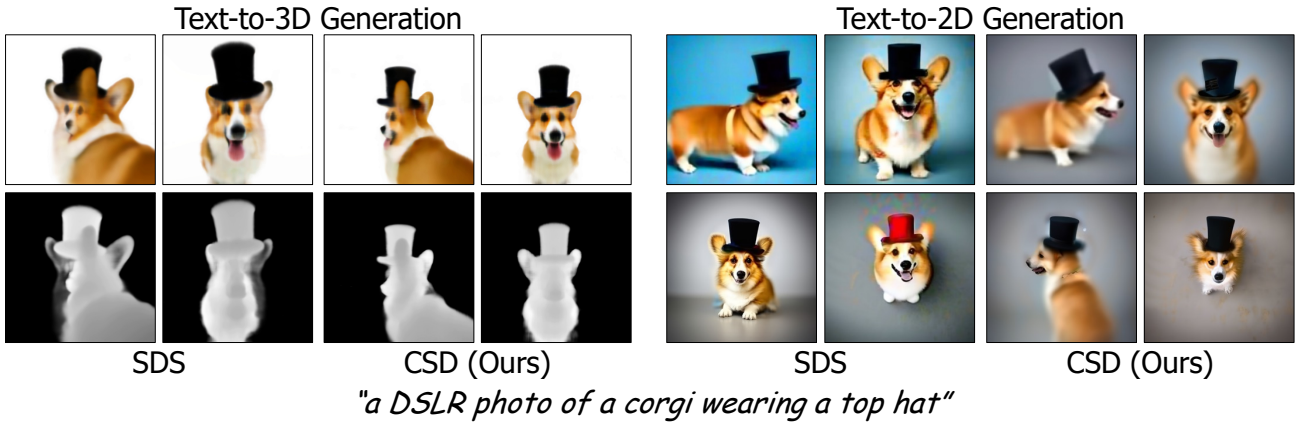

Figure 5: **Text-to-3D generation**. (Left) CSD helps capturing coherent geometry in synthesizing 3D object. (Right) CSD generates coherent images conditioned on view-dependent prompts.

We find that our approach is less sensitive to the choice of image guidance scale, yet a smaller image guidance scale exhibits more sensitivity to editing. All experiments are conducted on AMD EPYC 7V13 64-Core Processor and a single NVIDIA A100 80GB. Throughout the experiments, we use OpenCLIP (Ilharco et al., 2021) `ViT-bigG-14` model for evaluation.

C.1. Panorama image editing

To edit a panorama image, we first encode into the Stable Diffusion latent space (i.e., downscale by 8), then use a stride size of 16 to obtain multiple patches. Then we select a B batch of patches to perform CSD-Edit. Note that we perform CSD-Edit and then normalize by the number of appearances as mentioned in Section B.4. Note that our approach performs well even without using small batch size, e.g., for an image of resolution 1920×512, there are 12 patches and we use $B = 4$.

For experiments, we collect 32 panorama images and conduct 5 artistic stylizations: "turn into Van Gogh style painting", "turn into Pablo Picasso style painting", "turn into Andy Warhol style painting", "turn into oriental style painting", and "turn into Salvador Dali style painting". We use learning rate of 2.0 and image guidance scale of 1.5, and vary the guidance scale from 3.0 to 10.0.

C.2. Video editing

We edit video sequences in DAVIS 2017 (Pont-Tuset et al., 2017) by sampling 24 frames at the resolution of 1920×1080 from each sequence. Then, we resize all frames into 512×512 resolution and encode all frames each using Stable Diffusion (Rombach et al., 2022) autoencoder. We use learning rate $[0.25, 2]$ and optimize them for $[200, 500]$ iterations.

C.3. 3D scene editing

Following Instruct-NeRF2NeRF (Haque et al., 2023), we first pre-train NeRF using the *nerfacto* model from NeRFStudio (Tancik et al., 2023), training it for 30,000 steps. Next, we re-initialize the optimizer and finetune the pre-trained NeRF model with edited train views. In contrast to Instruct-NeRF2NeRF, which edits one train view with Instruct-Pix2Pix after every 10 steps of optimization of NeRF, we edit a batch of train views (batch size of 16) with CSD-Edit after every 2000 steps of update. The batch is randomly selected among the train views without replacement.

D. Additional experiments

D.1. Compositional editing

Recent works have shown the ability of text-to-image diffusion models in compositional *generation* of images handling multiple prompts (Du et al., 2023; Liu et al., 2022). Here, we show that CSD-Edit can extend this ability to compositional *editing*, even at panorama-scale images which require a particular ability to maintain far-range consistency. Specifically, we demonstrate that one can edit a panorama image to follow different prompts on different regions while keeping the overall context uncorrupted.

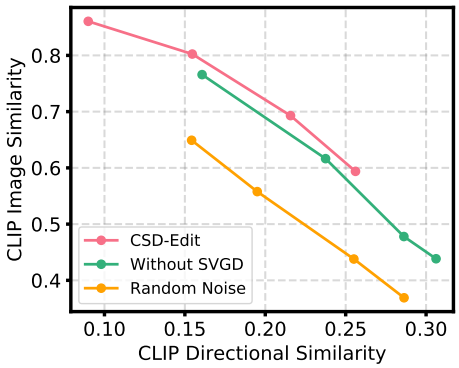

Figure 6: **Ablation study**. Ablation study on the components of CSD-Edit at different guidance scales $\omega_y \in \{3.0, 5.0, 7.5, 10.0\}$.

Table 3: **Text-to-3D**. Quantitative comparison between CSD and SDS under on text-to-3D generation via DreamFusion (Poole et al., 2022)

	CLIP Similarity Color ↑	CLIP Similarity Geo ↑	FID ↓
SDS (Poole et al., 2022)	0.437	0.322	259.4
CSD (Ours)	**0.447**	**0.345**	**247.1**

Given multiple textual prompts $\{y_k\}_{k=1}^K$, the compositional noise estimate is given by

$$\epsilon_\phi(\mathbf{x}_t; \{y_k\}_{k=1}^K, t) = \sum_{k=1}^K \alpha_k \epsilon_\phi^\omega(\mathbf{x}_t; y_k, t),$$

where α_k are hyperparameters that regularize the effect of each prompt. When applying compositional generation to the panorama image editing, the challenge lies in obtaining image that is smooth and natural within the region where the different prompts are applied. To that end, for each patch of an image, we set α_k to be the area of the overlapping region between the patch and region where prompt y_k is applied. Also, we normalize to assure $\sum_k \alpha_k = 1$. In Figure 10, we illustrate some examples on compositional editing of a panorama image. For instance, given an image, one can change into different weathers, different seasons, or different painting styles without leaving artifacts that hinder the spatial consistency of an image.

D.2. Text-to-3D generation with CSD

We explore the effectiveness of CSD in text-to-3D generation tasks following DreamFusion (Poole et al., 2022). We follow the most of experimental setups from those conducted by Poole et al. (2022); train a coordinate MLP-based NeRF architecture from scratch using text-to-image diffusion models. Our experiments in this section are based on Stable-DreamFusion (Tang, 2022), a public re-implementation of DreamFusion, given that currently the official implementation of DreamFusion is not available on public. Also, since the pixel-space diffusion model that DreamFusion used (Poole et al., 2022) is not publicly available, we used an open-source implementation of pixel-space text-to-image diffusion model.[3]

Setup. We use vanilla MLP based NeRF architecture (Mildenhall et al., 2020) with 5 ResNet (He et al., 2016) blocks. Other regularizers such as shading, camera and light sampling are set as default in (Tang, 2022). We use view-dependent prompting given the sampled azimuth angle and interpolate by the text embeddings. We use Adan (Xie et al., 2022) optimizer with learning rate warmup over 2000 steps from 10^{-9} to 2×10^{-3} followed by cosine decay down to 10^{-6}. We use batch size of 4 and optimize for 10000 steps in total, where most of the case sufficiently converged at 7000 to 8000 steps. For the base text-to-image diffusion model, we adopt `DeepFloyd-IF-XL-v1.0` since we found it way better than the default choice of Stable Diffusion in a qualitative manner. While the original DreamFusion (Poole et al., 2022) used guidance scale of 100 for their experiments, we find that guidance scale of 20 works well for `DeepFloyd`. We selected 30 prompts used in DreamFusion gallery[4] and compare their generation results via DreamFusion from the standard SDS and those from our proposed CSD. We use one A100 (80GB) GPU for each experiment, and it takes ∼5 hours to conduct one experiment.

For CSD implementation, we use LPIPS (Zhang et al., 2018) as a distance of RBF kernel, as we empirically found it to work well over the ℓ_2-norm. The LPIPS is computed between two rendered views of size 64×64. For the kernel bandwidth, we use $h = \frac{\texttt{med}^2}{\log B}$, where `med` is a median of the pairwise LPIPS distance between the views, B is the batch size. For evaluation,

[3] https://github.com/deep-floyd/IF
[4] https://dreamfusion3d.github.io/gallery.html

we render the scene at the elevation at 30 degrees and capture at every 30 degrees of azimuth angle. Then we compute the CLIP image-text similarity between the rendered views and input prompts. We measure similarities for both textured views (RGB) and textureless depth views (Depth). We also report Frechet Inception Distance (FID) between the RGB images and ImageNet validation dataset to evaluate the quality and diversity of rendered images compared to natural images.

Results. In Table 3, we report the evaluation results of CSD on text-to-3D generation comparison to DreamFusion. Note that CSD presents better CLIP image-text similarities in both RGB and Depth views. Additionally, CSD achieves a lower FID score, indicating a better quality of generated samples. Despite using the same random seed for generating both CSD and DreamFusion, resulting in similar shapes and colors, CSD still manages to capture finer details in its generations. This can be attributed to the ability of CSD to remove blurry artifacts in the synthesized 3D NeRF scene, which are often a consequence of inconsistent view distillation. This is demonstrated in the right of Figure 5, where we confirm that CSD produces more coherent images when conditioned on view-dependent prompts that were used in DreamFusion.

In Figure 13, we additionally provide a qualitative comparison between the baseline DreamFusion (SDS) and our approach. We empirically observe three benefits of using CSD over SDS. First, CSD provides better quality compared to SDS. SDS often suffers from the Janus problem, where multiple faces appear in a 3D object. We found that CSD often resolves the Janus problem by showing consistent information during training. See the first row of Figure 13. Second, CSD can give us better fine-detailed quality. The inconsistent score distillation often gives us blurry artifacts or undesirable features left in the 3D object. CSD can handle this problem and results in a higher-quality generation, e.g., Figure 13 second row and Figure 5. Lastly, CSD can be used for improving diversity. One problem of DreamFusion, as acclaimed by the authors, is that it lacks sample diversity. Thus, it often relies on changing random seeds, but it largely alters the output. On the other hand, we show that CSD can obtain an alternative sample with only small details changed, e.g., Figure 13 third row. Even when SDS is successful, CSD can be used in generating diverse samples. Also in Figure 5, we verify that the CSD generates more coherent images when conditioned on view-dependent prompts which were used in DreamFusion.

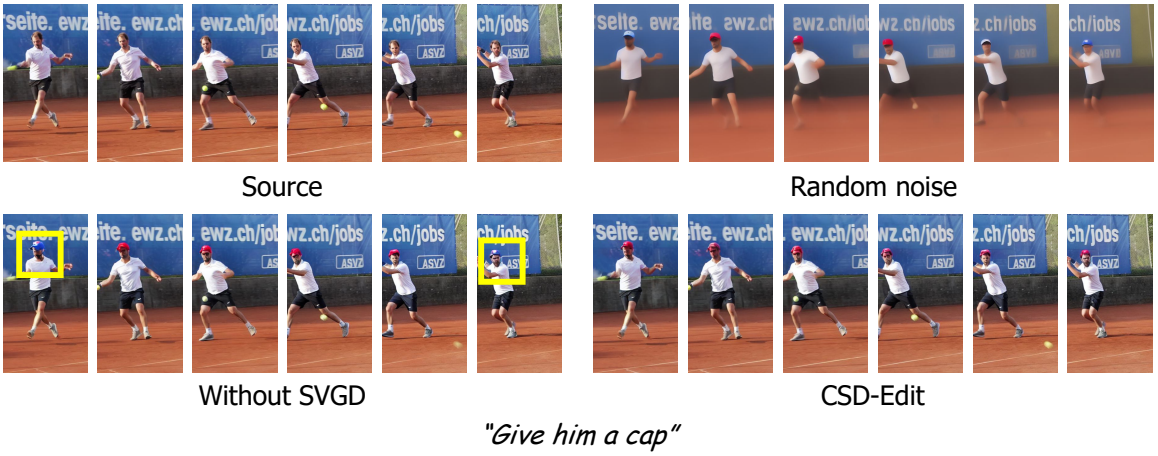

Figure 7: **Ablation study**. Given a source video (top left), CSD-Edit without SVGD results in inconsistent frames (bottom left), and subtracting random noise in CSD-Edit results in loss of details (top right). CSD-Edit obtains consistency between frames without loss of semantics (bottom right).

E. Ablation study

We present an ablation study on (a) the effect of SVGD and (b) subtracting random noise in CSD-Edit in panorama image editing experiments. Following the experimental setup in Section 4.1, we apply 5 different artistic stylizations to 16 images with CSD-Edit, without SVGD, and without subtracting the image-conditional noise estimate. We measure the CLIP image similarity and CLIP directional similarity for the evaluation. The results of the ablation study are plotted in Figure 6.

Note that CSD-Edit without SVGD radically alters the image due to the lack of consistency regularization. Furthermore, if the baseline noise is set to be random that is injected into both the source and target images, each frame becomes blurry, leading to the loss of original structures. This results in a significant drop in both CLIP image similarity and CLIP directional similarity, as details from the source image are lost. These effects are detailed in Figure 6.

To further verify the role of communication between samples using SVGD, we conduct an additional video editing experiment. As shown in Figure 7, CSD-Edit consistently edits a source video by adding a red cap on a man's head when given the instruction "give him a cap." In the absence of SVGD, the edits between frames become inconsistent; for example, both blue and red caps appear inconsistently in the edited frames. We additionally present the qualitative results of our ablation study in Figure 8.

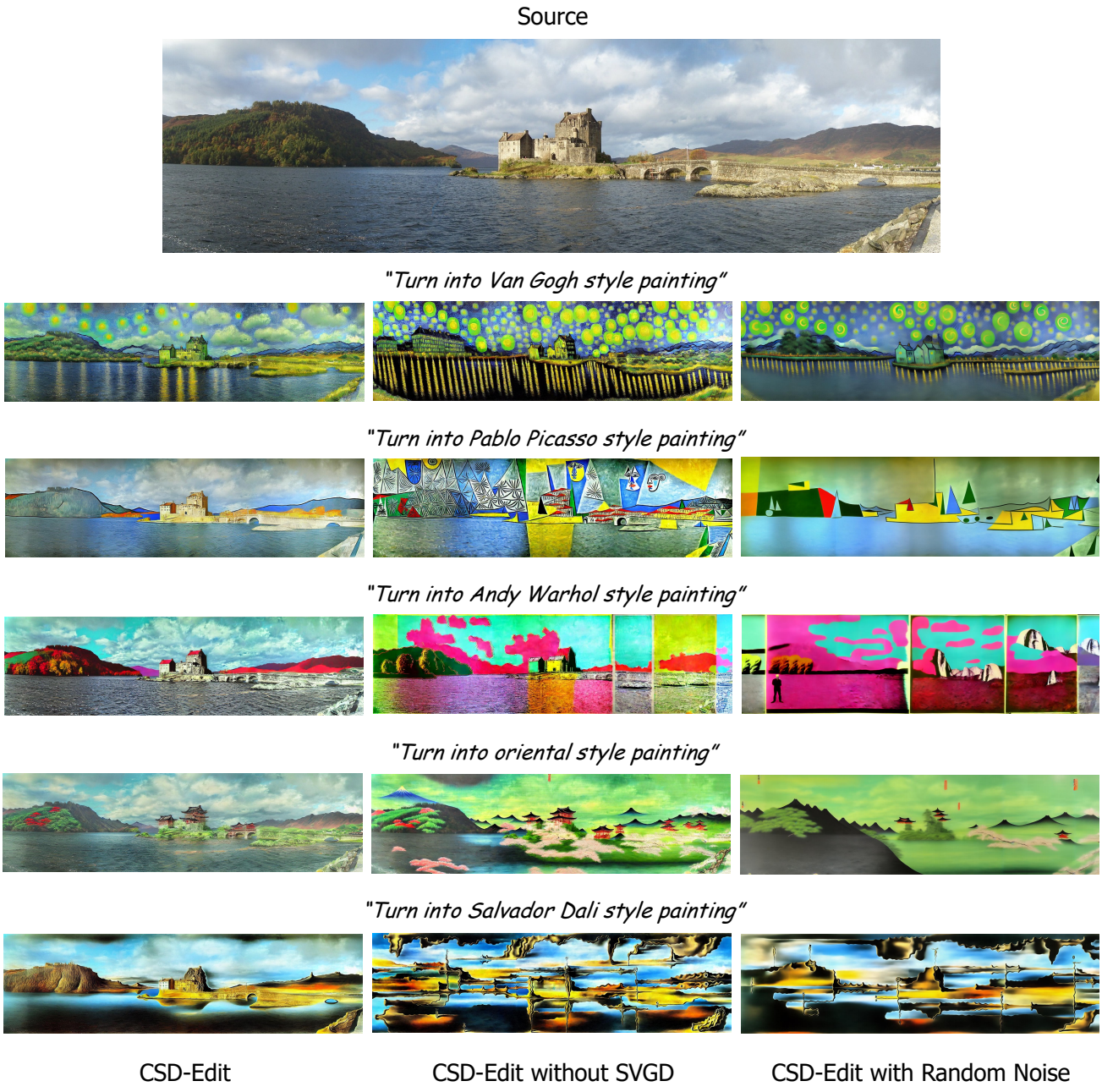

Figure 8: **Ablation study: SVGD and random noise**. As illustrated, edits across different patches are not consistent without SVGD. Also, when using random noise as baseline noise, it loses the content and the detail of the source image.

F. More qualitative results

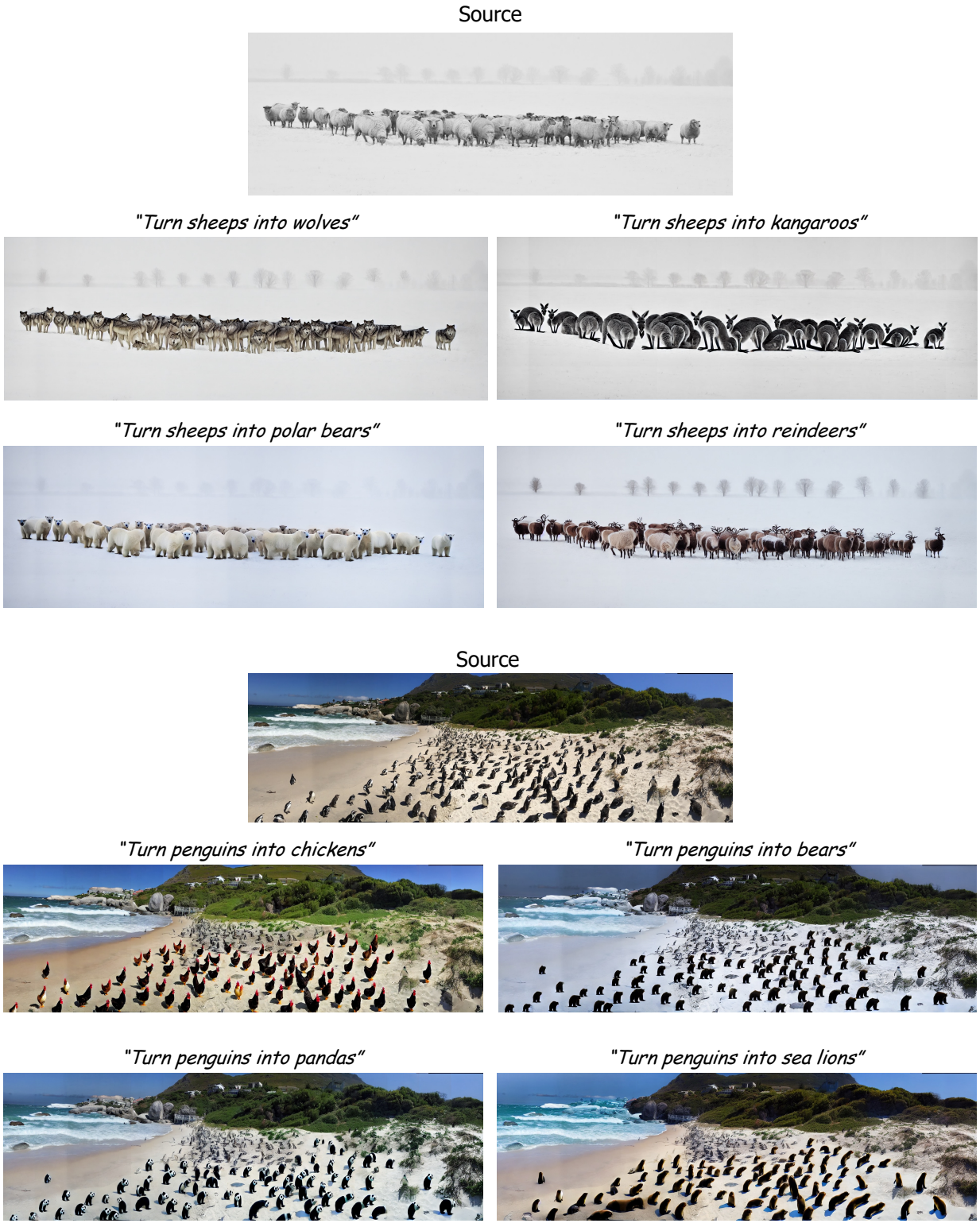

Figure 9: **Object editing**. CSD-Edit can edit many objects in a wide panorama image consistently in accordance with the given instruction while preserving the overall structure of source images.

Figure 10: **Compositional image editing**. CSD-Edit demonstrates the ability to edit consistently and coherently across patches in panorama images. This provides the unique capability to manipulate each patch according to different instructions while maintaining the overall structure of the source image. Remarkably, CSD-Edit ensures a smooth transition between patches, even when different instructions are applied.

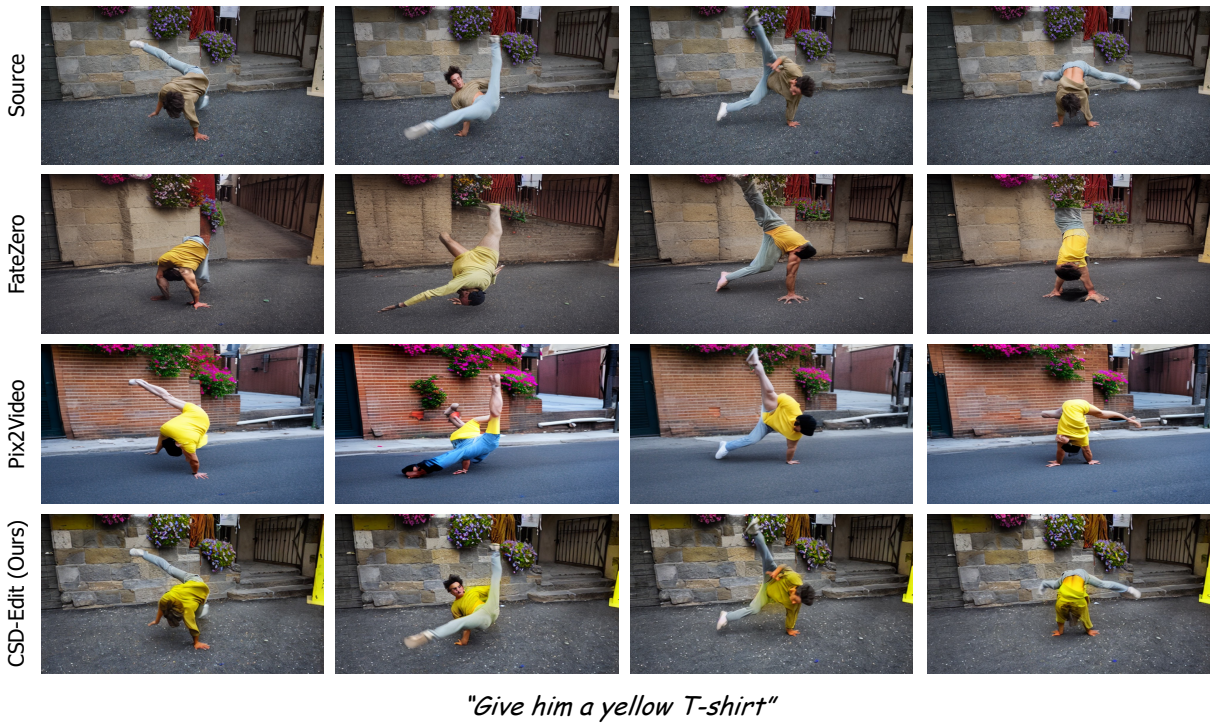

Figure 11: **Video editing**. CSD-Edit demonstrates various editing from an object (e.g., tiger) to attributes (e.g., color) while providing consistent edits across frames and maintaining the overall structure of a source video.

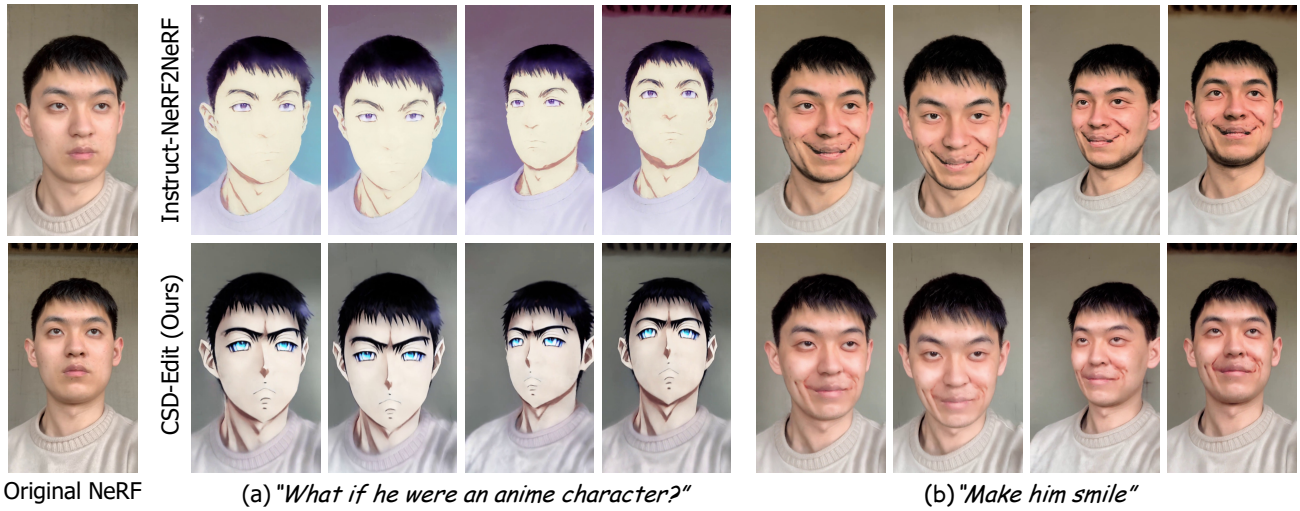

Figure 12: **3D NeRF scene editing**. Visualizing novel-views of edited Fangzhou NeRF scene (Wang et al., 2022). CSD-Edit leads to high-quality editing of 3D scenes and better preserves semantics of source scenes, e.g., obtains sharp facial details (left) and makes him smile without giving beard (right).

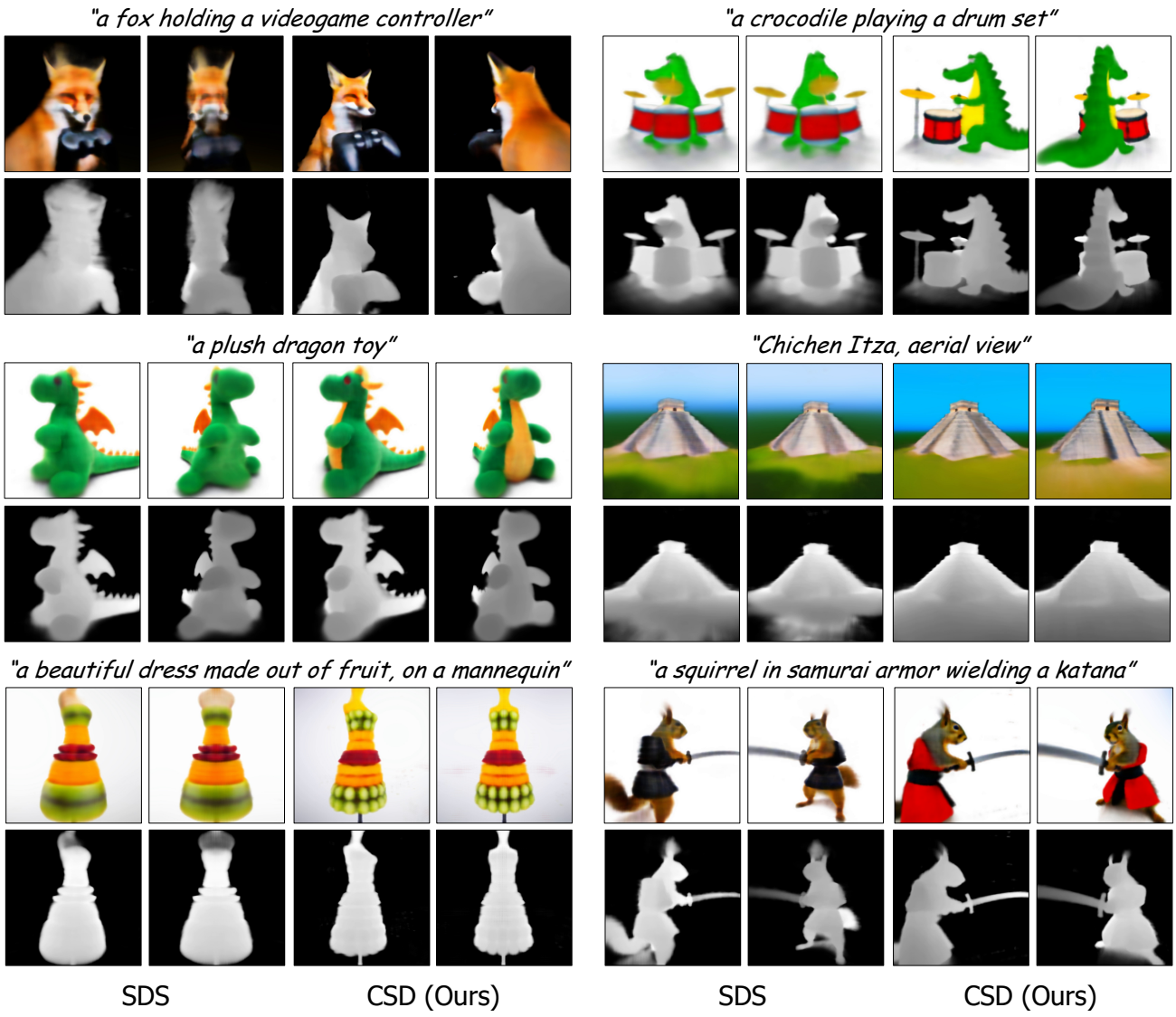

Figure 13: **Text-to-3D generation examples**. (First row) CSD helps to capture coherent geometry compared to using SDS. (Second row) CSD allows learning finer details than SDS. (Third row) CSD can provide diverse and high-quality samples without changing random seeds.