Supplementary Material for Dilutional Noise Initialization for Diffusion Video Editing

Sunjae Yoon, Gwanhyeong Koo, Ji Woo Hong, and Chang D. Yoo

Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{sunjae.yoon,cd_yoo}@kaist.ac.kr

1 Broader Impacts and Ethic Statements

Visual generative models introduce a spectrum of ethical dilemmas, from the creation of unauthorized counterfeit content and potential privacy breaches to issues surrounding fairness. Given our reliance on the architecture of these models, our work inherits these ethical vulnerabilities. Addressing these concerns is imperative, necessitating the establishment of comprehensive regulations and technical countermeasures. It is incumbent upon researchers, including ourselves, to proactively engineer and implement such safeguards. To promote responsible and ethical utilization, in this paper, we provide detailed specifications about models and data and further explore the adoption of advanced measures including techniques of debiasing [19, 20], explainability [17], and adversarial approaches [6]. These initiatives are part of a concerted effort to ethically navigate the complex terrain of visual generative models, ensuring their development serves the greater good.

2 Limitation and Future Work

Our proposed DNI is designed to alleviate the restrictions of initial latent noise on non-rigid editing, as observed in our experiments. However, there are still many limitations to perform effective non-rigid editing. While existing video diffusion models excel at creating visual shapes, they struggle to generate a continuous and natural sequence of frames. Therefore, it is not feasible to perform movements significantly different from those in the input video, nor is it possible to edit motion in locations that are physically distant from the editing target. Therefore, with the existing video diffusion model, we could only make basic motion changes, such as altering a walking motion to a running motion or modifying arm movements to a limited extent. Our empirical studies found that tuning-based models can handle more dynamic changes, while tuning-free models have greater limitations in non-rigid editing. These limitations are well-known in the video editing community. To address them, motion editing is performed by providing motion guidance [14, 16], conditioning the movement based on other motion videos. We are preparing to embrace this trend and develop effective methods to utilize our motion editing technology based on motion conditioned diffusion model in future work. To be specific, based on the editing system with

motion guidance, we plan to integrate video retrieval technology [1,22] to attentively improve editing effects on the temporal region to be edited. Furthermore, we are currently investigating the possibility of contributions in terms of textual prompt to enhance the editing effects based on the methods of further weighting on words, prompt tuning [18], information-theoretic approach [11,23], and other multimodal methods of video-grounded language systems [9,21].

3 Details of implementation and evaluation

We provide additional details of our experiments, including implementation procedures and evaluation metrics.

3.1 Implementation details

Feature encoding. For video encoding, we employ VQ-VAE [13], which extracts patch-wise features from each frame. For text encoding, we use the CLIP model (ViT-L/14) [10] to generate word token features from the textual prompts.

Usage of source prompt. Given that the DNI framework operates independently of source prompts, the methodology does not rely on the utilization of source prompts to enhance the broader application of DNI. However, as the baselines in our experiments inherently employ source prompts, we outline their respective approaches to utilizing source prompts. The tuning-based models [5,15] use source prompts for tuning with $300 \sim 500$ iterations, while the tuning-free approaches [3,8] can be optionally provided with source prompts as a condition only for noise initialization in DDIM inversion. Both target prompts and source prompts can be possible to select the words for editing guidance. Empirically, weighing more on words about editing is somewhat effective for non-rigid variations in a tuning-based model, and applying dilution directly into video is also effective. In our experiments, all baseline models followed the same methods provided in their public codes, where the baselines [5,15] used source prompts in both tuning and DDIM inversion, while the baselines [3,8] used them only in DDIM inversion. Some flickering effects are attributed to baseline models' temporal modules (e.g., attention, propagation) and DNI follows their properties. The additional results in Sec. 5 are provided along with their used source prompts.

3.2 Evaluation details

Fidelity metric. For measuring the fidelity of the edited video from the input video (see Fig. 1), we applied the same zero mask to the same locations in both the input and output videos. By applying a mask to the edited area, we can assess the consistency and similarity between the input and output videos, particularly in preserving unedited content, using SSIM and LPIPS scores. In our preliminary study, we tested several automatic detectors, such as segment-anything-model [4], to specify areas for editing in both the input and output videos. Despite these efforts, human annotations proved to be the most precise.

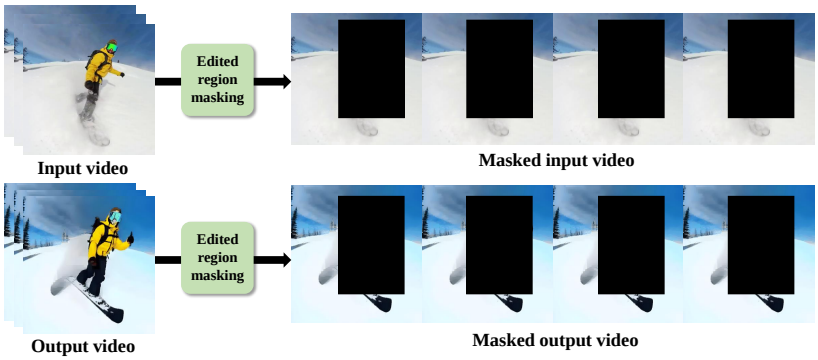

Fig. 1: Illustration of the masking process of input and output videos for measuring fidelity metrics. The same masking is applied to the same location in both the input video and the edited output video.

Human evaluation. We conducted a survey to compare preferences of humans between the results of current editing systems and the results generated incorporated with the DNI framework. Our survey included 36 participants with various backgrounds (e.g., engineering, art) and a mix of native and non-native English speakers.

4 Reverse diffusion process: Closed formulations of KL divergence

The reverse diffusion denoising process aims to approximate $q(x_{t-1}|x_t)$ using parameterized Gaussian transitions $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t))$. Considering the entire sequence of T step-parameterized transitions, these are sequentially formulated as follows:

$$p_\theta(X) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{1}$$

where we take $X = x_{0:T}$ and it starts at normal distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$. To optimize the $p_\theta(X)$, training objective is to maximize log-likelihood $\log(p_\theta(X))$, where we can also apply variational inference by maximizing the variational lower bound $-L_{VLB}$ as given below:

$$-L_{VLB} = \log(p_\theta(X)) - D_{\mathrm{KL}}(q(Z|X)||p_\theta(Z|X)) \leq \log(p_\theta(X)), \tag{2}$$

where D_{KL} is the Kullback-Leibler divergence (KL divergence) and the Z is latent variable by reparametrization trick used in the variational auto-encoder. The q can be any distribution that we can address with ease. We leverage this inequality condition as $-\log(p_\theta(X)) \leq L_{VLB}$. The L_{VLB} can be expanded out

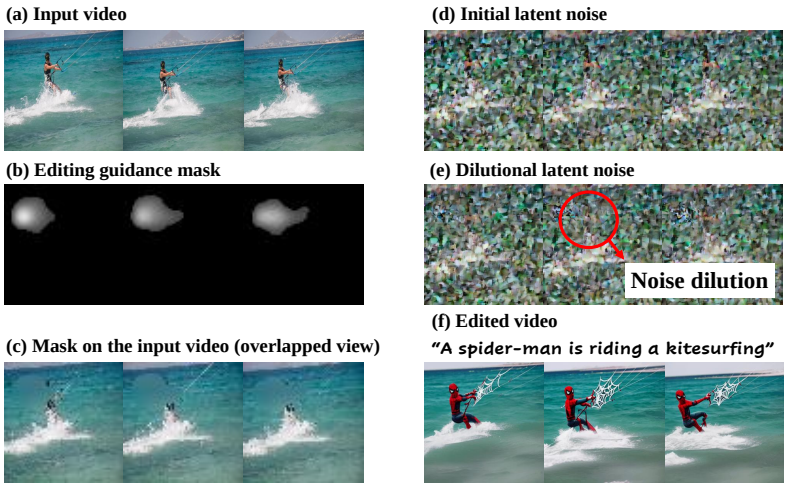

Fig. 2: Detailed results about noise dilution. (a) Input video, (b) Editing guidance mask, (c) Mask on the input video, (d) Initial latent noise, (e) Dilutional latent noise, (f) Edited video with target prompt: "A spider-man is riding a kitesurfing".

as $L_{VLB} = L_T + L_{T-1} + \cdots + L_0$, where they are defined with $1 \leq t \leq T$ as below:

$$\begin{aligned} L_T &= D_{\mathrm{KL}}(q(x_T|x_0)||p_\theta(x_T)), \\ L_t &= D_{\mathrm{KL}}(q(x_t|x_{t+1}, x_0)||p_\theta(x_t|x_{t+1})), \\ L_0 &= -\log(p_\theta(x_0|x_1)). \end{aligned} \tag{3}$$

Therefore the terms about L_t make the closed form of KL divergence under step t with a range of $0 \leq t \leq T$.

5 Qualitative results

In the subsequent sections, we showcase our qualitative results for both non-rigid and rigid editing. The videos used in our experiments are obtained from publicly accessible sources, as cited in [2, 7, 12, 15].

Noise dilution. In Fig. 2, we provide the details about qualitative results of dilutional latent noise and initial latent noise.

Qualitative results and Failure case.

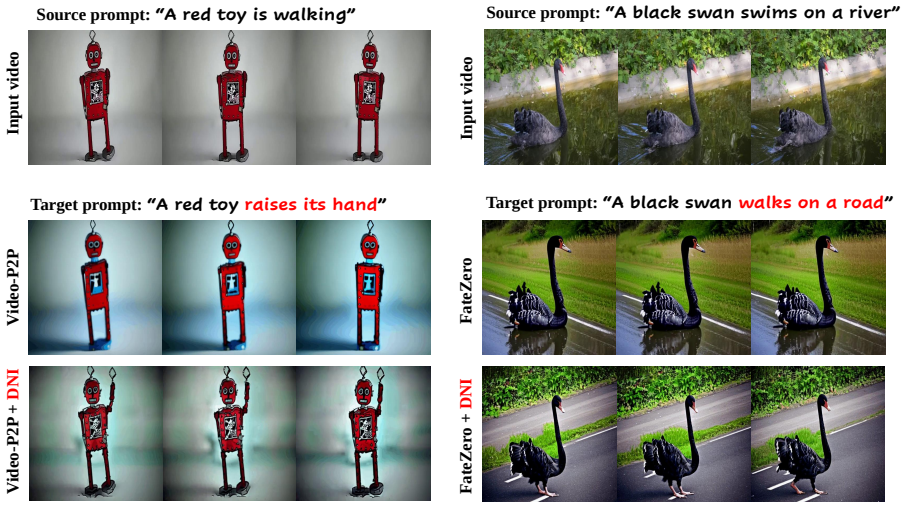

Fig. 3: Results of non-rigid editing about motion change.

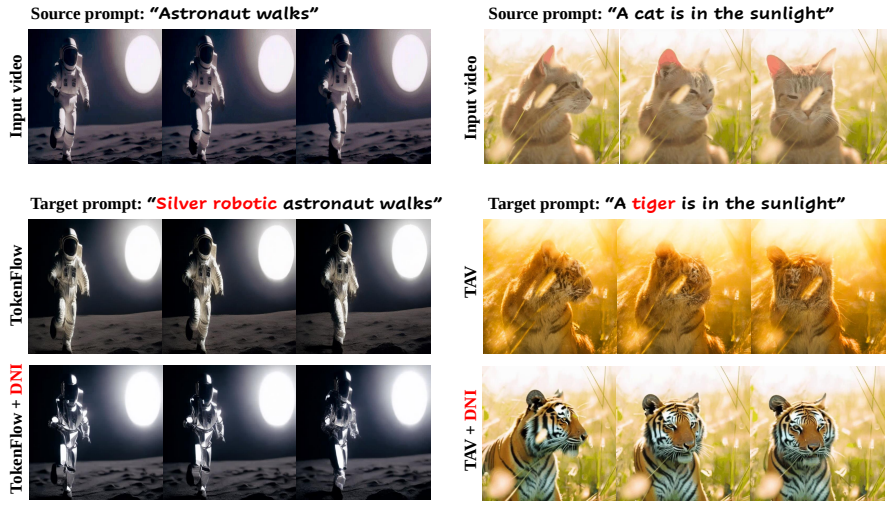

Fig. 4: Results of rigid editing about object overlay.

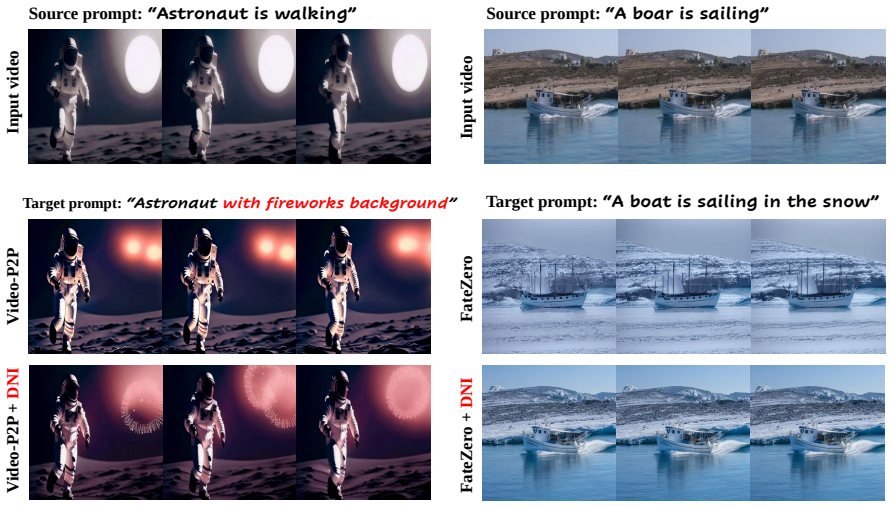

Fig. 5: Results of rigid editing about style transfer.

References

1. Barrios, W., Soldan, M., Ceballos-Arroyo, A.M., Heilbron, F.C., Ghanem, B.: Localizing moments in long video via multimodal guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13667–13678 (2023)
2. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
3. Geyer, et al, M.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
5. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
7. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
8. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
9. Qu, M., Chen, X., Liu, W., Li, A., Zhao, Y.: Chatvtg: Video temporal grounding via chat with video dialogue large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1847–1856 (2024)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
11. Seo, S., Lee, J.Y., Han, B.: Information-theoretic bias reduction via causal view of spurious correlation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2180–2188 (2022)
12. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
13. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
14. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for realistic human dance generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9326–9336 (2024)
15. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)
16. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)
17. Yoon, H.S., Tee, J.T.J., Yoon, E., Yoon, S., Kim, G., Li, Y., Yoo, C.D.: Esd: Expected squared difference as a tuning-free trainable calibration measure. arXiv preprint arXiv:2303.02472 (2023)
18. Yoon, H.S., Yoon, E., Tee, J.T.J., Hasegawa-Johnson, M., Li, Y., Yoo, C.D.: C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. arXiv preprint arXiv:2403.14119 (2024)
19. Yoon, S., Hong, J.W., Eom, S., Yoon, H.S., Yoon, E., Kim, D., Kim, J., Kim, C., Yoo, C.D.: Counterfactual two-stage debiasing for video corpus moment retrieval. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
20. Yoon, S., Hong, J.W., Yoon, E., Kim, D., Kim, J., Yoon, H.S., Yoo, C.D.: Selective query-guided debiasing for video corpus moment retrieval. In: European Conference on Computer Vision. pp. 185–200. Springer (2022)
21. Yoon, S., Kim, D., Yoon, E., Yoon, H.S., Kim, J., Yoo, C.D.: Hear: Hearing enhanced audio response for video-grounded dialogue. arXiv preprint arXiv:2312.09736 (2023)
22. Yoon, S., Koo, G., Kim, D., Yoo, C.D.: Scanet: Scene complexity aware network for weakly-supervised video moment retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13576–13586 (2023)
23. Yoon, S., Yoon, E., Yoon, H.S., Kim, J., Yoo, C.D.: Information-theoretic text hallucination reduction for video-grounded dialogue. arXiv preprint arXiv:2212.05765 (2022)