Supplementary Material for Occlusion-robust Stylization for Drawing-based 3D Animation

Sunjae Yoon Gwanhyeong Koo Younghwan Lee Ji Woo Hong Chang D. Yoo
School of Electrical Engineering, KAIST
{sunjae.yoon,kookie,youngh2,jiwoohong93,cd_yoo}@kaist.ac.kr

1. Broader Impacts and Ethic Statements

Alongside recent advances in linguistic generative applications [4, 8, 9, 17], Visual generative applications [2, 5, 7, 18, 20] presents large scope of ethical dilemmas. This includes unauthorized counterfeit, potential privacy and fairness issues. Due to our reliance of the visual generative models [10, 12, 21], our work is also susceptible to these issues. Addressing these challenges is crucial and necessitates the implementation of robust regulations alongside advanced technical safeguards. The responsibility falls on researchers, including ourselves, to proactively design and implement these safeguards. To foster transparency and advocate for ethical use, we will release our source code alongside comprehensive model and data specifications under a license promoting lawful and responsible practices. Additionally, we are studying advanced techniques like learning-based digital forensics [11, 22], digital watermarking [23], debiasing approach [14], regularization [13, 15], causality [16]. These are integral to our strategy for ethically addressing the challenges of visual generative models.

2. Limitation and Future work

We present the technical limitations of our proposed OSF and further address the challenges in drawing-based character 3D animation task. First, the proposed OSF enhances the edges used as pior input for stylization, but it is limited by the inability of edges to fully capture the intricate details of drawing properties. To be specific, human-hand drawings often include a large range of thicknesses in their lines; however, the provided edge priors fail to distinguish these nuances, and our method also does not address this limitation either. Edges serve as an effective prior for generating contours and textures in drawing stylization. However, developing edge priors with enhanced sensitivity to line thickness could further elevate these effects. This also represents a promising direction for our future work.

Another persistent challenge lies in the task of this drawing animation's inputs. The model is tasked with generating 3D objects in various poses based on a single character drawing. While it excels at preserving the visual information presented in the original drawing (*e.g.* the front view), it struggles significantly with generating unseen perspectives, such as the side and back views in Figure 1, due to a lack of training data, resulting in noticeably lower quality outputs than ones from front view. Although using multiple images could resolve this issue, it is not always feasible. Consequently, achieving consistent 3D synthesis remains a fundamental challenge (if it were perfect, stylization would be unnecessary). Thus our future work aims to restore sharpness in unseen views using a diffusion-based high-frequency recovery methods [19], while enhancing flexibility through automatic 3D pose estimation [1] instead of relying on predefined poses. From an efficiency perspective, current diffusion-based image-to-3D methods incur significant overhead, which motivates our interest in more efficient diffusion methods [3, 6].

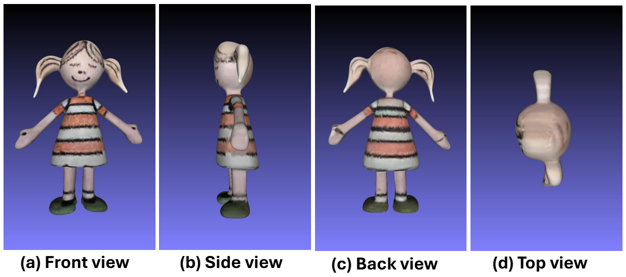
(a) Front view (b) Side view (c) Back view (d) Top view

Figure 1. 3D reconstructions from drawing image using image-to-3D diffusion (Wonder3D) and reconstruction model (Nerf). The reconstruction qualities are irregular according to different views due to the insufficient visual information about target drawing (only a single drawing image is given).

3. Further qualitative results

Explanation.
Figure 2 shows input edges used in OSF corresponding to the samples in Figure 8.
Figure 3 is results on same motion with various drawings.
Figure 4 is results on drawing with various motions.

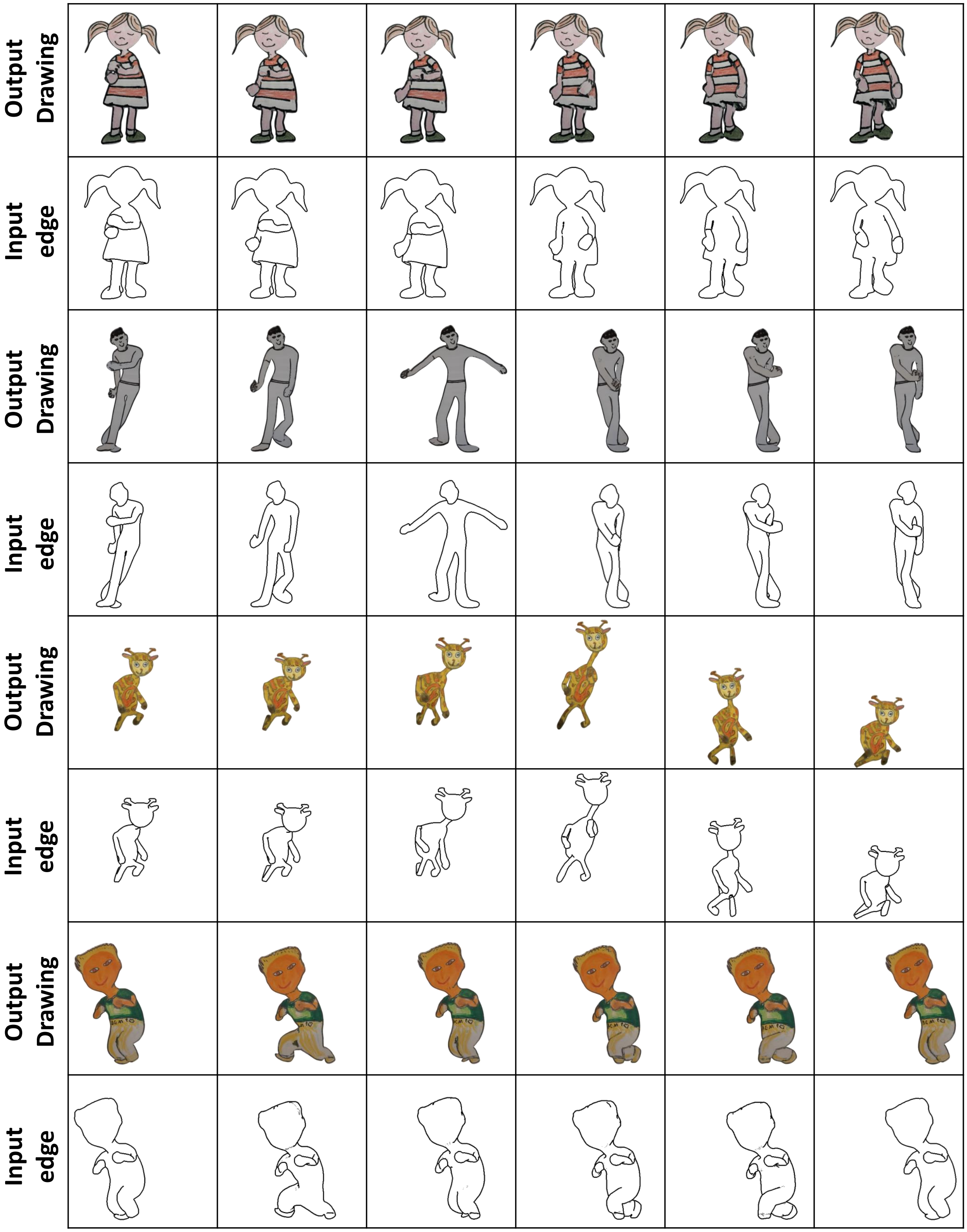

Figure 2. Qualitative results about edges used in OSF of Figure 8 in the main paper.

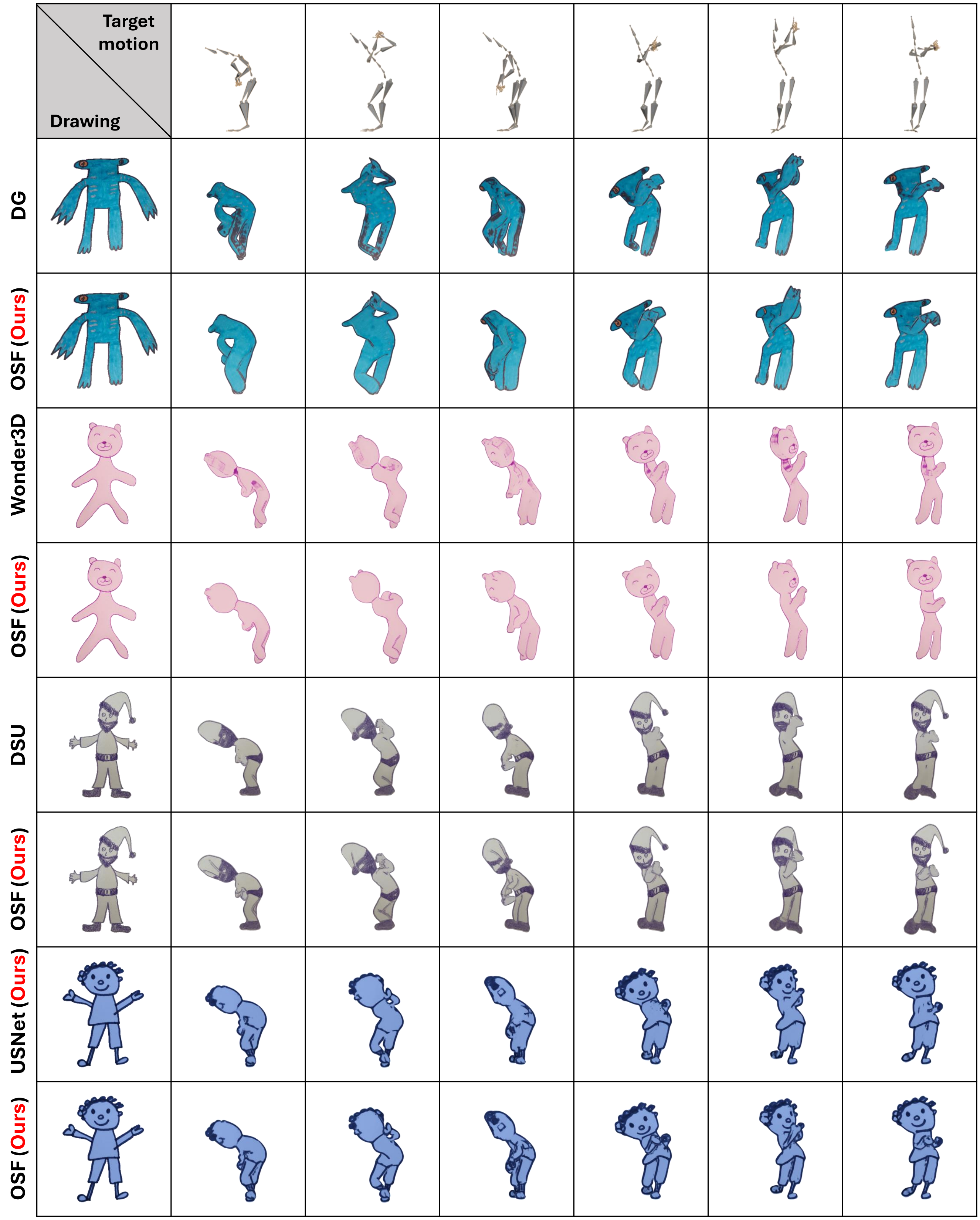

Figure 3. Qualitative results about drawing animation on various character drawings with the same motion. DG: DreamGaussian, DSU: DrawingSpinUp. OSF uses stylization baseline of USNet.

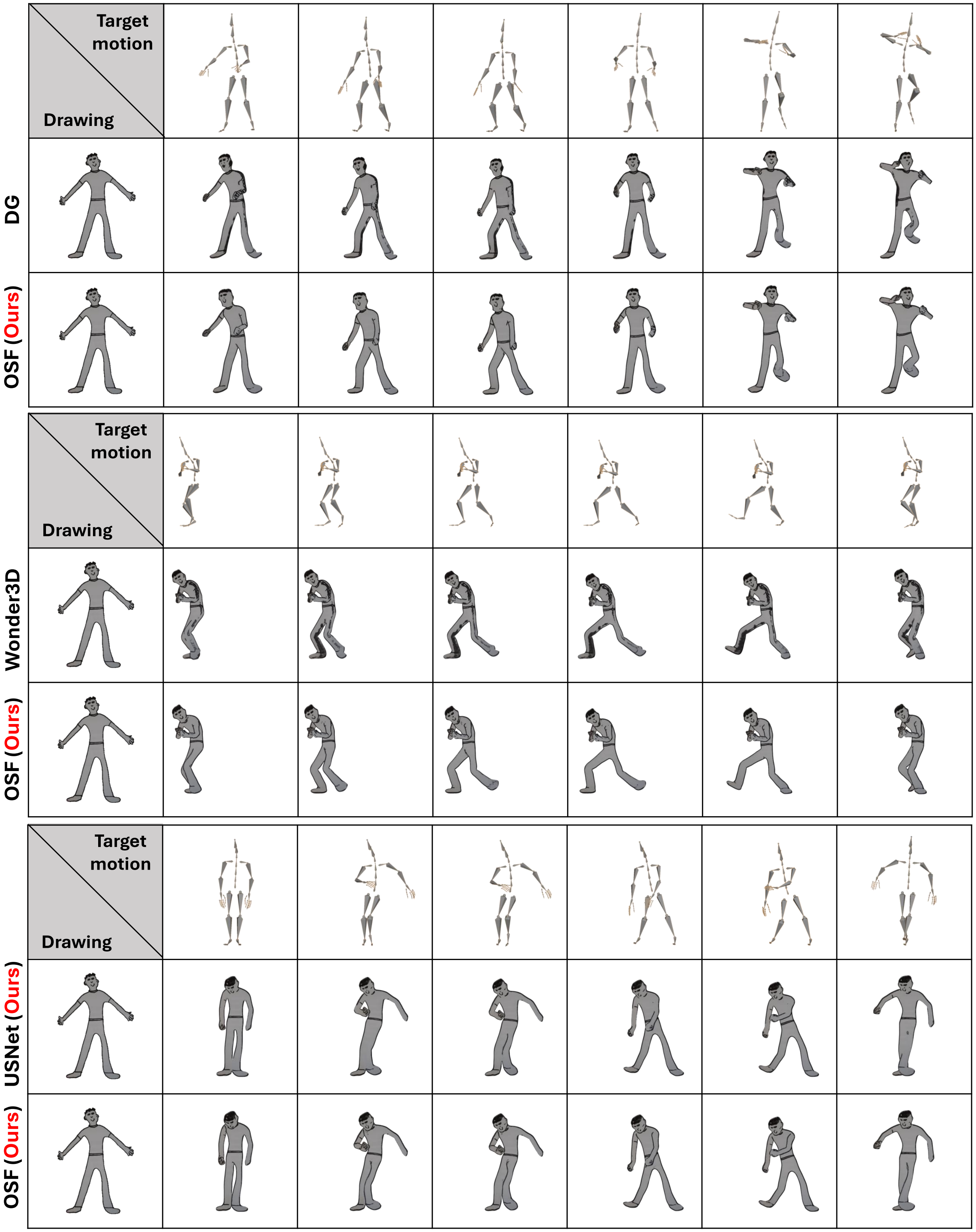

Figure 4. Qualitative results about drawing animation on various motions with the same character drawing. DG: DreamGaussian, DSU: DrawingSpinUp. OSF uses stylization baseline of USNet.

References

[1] Ji Woo Hong, Sunjae Yoon, Junyeong Kim, and Chang D Yoo. Joint path alignment framework for 3d human pose and shape estimation from video. *IEEE Access*, 11:43267–43275, 2023. 1

[2] Ji Woo Hong, Tri Ton, Trung X Pham, Gwanhyeong Koo, Sunjae Yoon, and Chang D Yoo. Ita-mdt: Image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28284–28294, 2025. 1

[3] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022. 1

[4] Junyeong Kim, Sunjae Yoon, Dahyun Kim, and Chang D Yoo. Structured co-reference graph attention for video-grounded dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1789–1797, 2021. 1

[5] Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D Yoo. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In *European Conference on Computer Vision*, pages 363–379. Springer, 2024. 1

[6] Gwanhyeong Koo, Sunjae Yoon, and Chang D Yoo. Wavelet-guided acceleration of text inversion in diffusion-based image editing. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4380–4384. IEEE, 2024. 1

[7] Gwanhyeong Koo, Sunjae Yoon, Younghwan Lee, Ji Woo Hong, and Chang D Yoo. Flowdrag: 3d-aware drag-based image editing with mesh-guided deformation vector flow fields. *arXiv preprint arXiv:2507.08285*, 2025. 1

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[11] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[13] Hee Suk Yoon, Eunseop Yoon, John Harvill, Sunjae Yoon, Mark Hasegawa-Johnson, and Chang D Yoo. Smsmix: Sense-maintained sentence mixup for word sense disambiguation. *arXiv preprint arXiv:2212.07072*, 2022. 1

[14] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D Yoo. Selective query-guided debiasing for video corpus moment retrieval. In *European Conference on Computer Vision*, pages 185–200. Springer, 2022. 1

[15] Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang D Yoo. Information-theoretic text hallucination reduction for video-grounded dialogue. *arXiv preprint arXiv:2212.05765*, 2022. 1

[16] Sunjae Yoon, Ji Woo Hong, Soohwan Eom, Hee Suk Yoon, Eunseop Yoon, Daehyeok Kim, Junyeong Kim, Chanwoo Kim, and Chang D Yoo. Counterfactual two-stage debiasing for video corpus moment retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[17] Sunjae Yoon, Dahyun Kim, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chnag D Yoo. Hear: Hearing enhanced audio response for video-grounded dialogue. *arXiv preprint arXiv:2312.09736*, 2023. 1

[18] Sunjae Yoon, Gwanhyeong Koo, Ji Woo Hong, and Chang D Yoo. Dni: Dilutional noise initialization for diffusion video editing. In *European Conference on Computer Vision*, pages 180–195. Springer, 2024. 1

[19] Sunjae Yoon, Gwanhyeong Koo, Geonwoo Kim, and Chang D Yoo. Frag: Frequency adapting group for diffusion video editing. *arXiv preprint arXiv:2406.06044*, 2024. 1

[20] Sunjae Yoon, Gwanhyeong Koo, Younghwan Lee, and Chang Yoo. Tpc: Test-time procrustes calibration for diffusion-based human image animation. *Advances in Neural Information Processing Systems*, 37:118654–118677, 2024. 1

[21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1

[22] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018. 1

[23] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018. 1