VideoSPatS: Video SPatiotemporal Splines for Disentangled Occlusion, Appearance and Motion Modeling and Editing

Supplementary Material

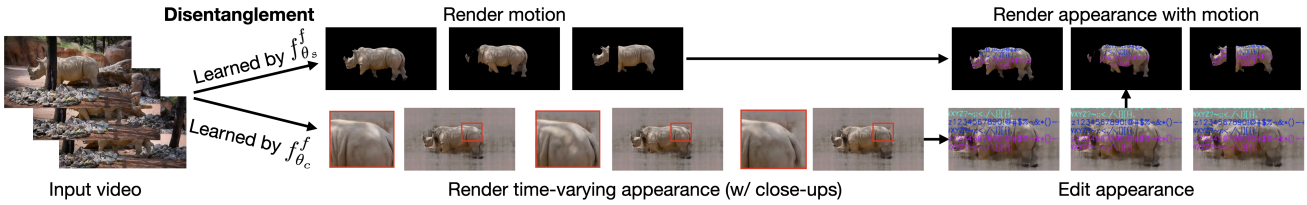

Figure 9. Disentangling motion and appearance.

In this supplemental material, we provide additional implementation details and experimental results. We present further comparisons with previous methods [11, 40], as well as additional results on texture editing, motion editing, and training on longer sequences. We also include additional ablation studies on loss terms, guidance masks, and control points. Finally, we discuss some failure cases and the ethical implications of our method and datasets. Furthermore, we provide several *videos* in the attached `videospats.html`, which are critical for visualizing time-dependent appearance and temporally consistent reconstructions and edits produced by our method. *We strongly encourage readers to open the* `videospats.html` *for the best viewing experience.*

A. Additional details on disentangling motion and appearance.

For the sake of better conceptualization, we provide the additional Fig. 9. As shown in Fig. 9 for the foreground image (rhino), $f^f_{\theta_s}$ learns to model motion (a.k.a. spatial deformation field), while $f^f_{\theta_c}$ learns to model time-dependent appearance (a.k.a. color deformation field). Note that such disentanglement of motion and appearance allows us to perform editing on the rhino that is not distorted by the time-dependent appearance (e.g. shadows). In addition, as we learn a base color and deformation color splines, we can seamlessly blend appearance changes with color edits, as shown in the right-hand side of Fig. 9.

B. Additional implementation details

Our *VideoSPaTS* takes 90 minutes to fit a 512×288, 50 frames video. However, the training time could be reduced with additional engineering efforts, such as replacing MLPs with optimized embedders like those in tiny-cuda-nn [19]. This can potentially provide between 2× and 10× training speed-ups. In addition, our method does not require running the models during inference / editing for every single frame, since a single run suffices to obtain the deformation and color control points, providing further speedups during inference and editing.

We employed periodical positional encoding [35] for our deformation models.

The weights in Eq.(16) of the main paper are empirically set to balance their respective terms with respect to l_{rec}. Specifically:

- $\lambda_{fl} = 100$ is set with a relatively high value as the coordinate error has very small magnitudes in comparison with the color errors in l_{rec}.
- $\lambda_{\mathcal{D}_s} = 0.1$ is set to slightly regularize deformations. See Section D.1 for more details.
- $\lambda_{\mathcal{D}_c} = 0.001$ is set to a relatively low value to regularize color deformation while still allowing it to learn, as such color deformation is enabled after 50% of the training.

C. Additional results

We show additional results in this section. Please refer to the attached `videospats.html` for additional video visualizations.

Scene	CoDeF	Deformable Sprites	Ours
Bear	27.52/0.84	**30.94/0.96**	30.82/0.95
Train	21.53/0.87	27.08/0.94	**27.68/0.92**
Rhino	24.66/0.81	28.60/0.94	**29.35/0.94**
Average	24.57/0.84	28.87/**0.95**	**29.23**/0.94

Table 1. Video editing quantiative results PSNR/SSIM

C.1. Quantitative results

As mere video reconstruction metrics are not indicative of editing performance, we show video editing quantitative results in Table 1 in terms of warping consistency, measured in avgerage PSNR and SSIM between edited and warped

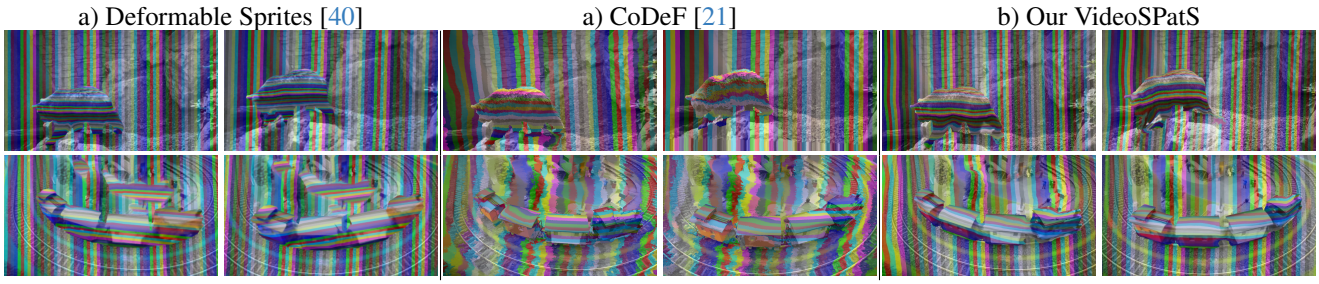

Figure 10. Editing results. Note inconsistencies in (b) as deformation fields incorrectly model time-varying appearance in the bear's fur.

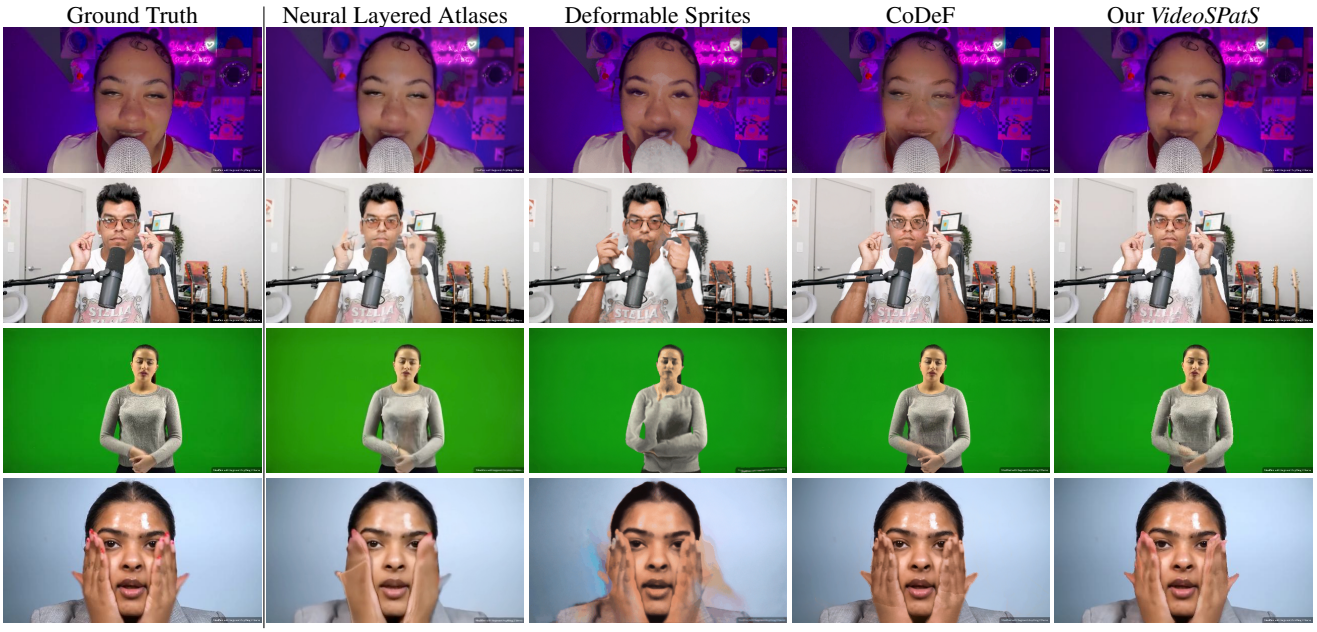

Figure 11. Video reconstruction comparisons with other methods. Our method consistently implicitly reconstructs videos, while the other methods fail on one case (CoDeF) or multiple cases (Layered Atlases, Deformable Sprites).

edited frames. We use RAFT [34] to obtain the original frames' optical flow to warp edited frames at t+n into t. We set n=3 for a significant difference in terms of scene optical flow. Ours outperforms CoDeF[21] by a large margin in terms of PSNR and SSIM, corresponding well to the visuals in Fig. 10. With respect to Deformable Sprites [40], our method outperforms it by **0.4dB** in terms of PSNR, but more importantly, our VideoSpatS can model the time-dependent appearance (e.g. shadows on bear's fur), yielding a more realistic and disentangled reconstruction and editing than the fixed colors in Deformable Sprites [40].

C.2. Canonical spaces and reconstruction

We present additional qualitative results and comparisons with previous methods, including Neural Layered Atlases [11], Deformable Sprites [40] and Codef [21], in terms of video reconstruction and canonical space estimation, as shown in Fig. 11 and Fig. 12, respectively.

Fig. 11 shows that, unlike Neural Layered Atlases and Deformable Sprites, our method consistently yields more detailed reconstructions. Although CoDeF generates very detailed renderings, its canonical spaces are not suitable for editing, as shown in Fig. 12. In contrast, our method generates intuitive canonical spaces that are well-suited for editing.

C.3. Comparisons to diffusion-based methods.

Although flexible for semantic video editing, diffusion-based methods such as [18, 22] are not designed for time-dependent appearance editing or do not support motion editing. Our method, closer to warping-based video modeling, focuses on modeling motion, appearance, and occlusions, so we did not compare to general video editing approaches in the main paper. For completeness, we provide an additional comparison to ReVideo [18] in Fig. 14. Ours keeps original head poses and temporal consistency, and ReVideo changes semantics.

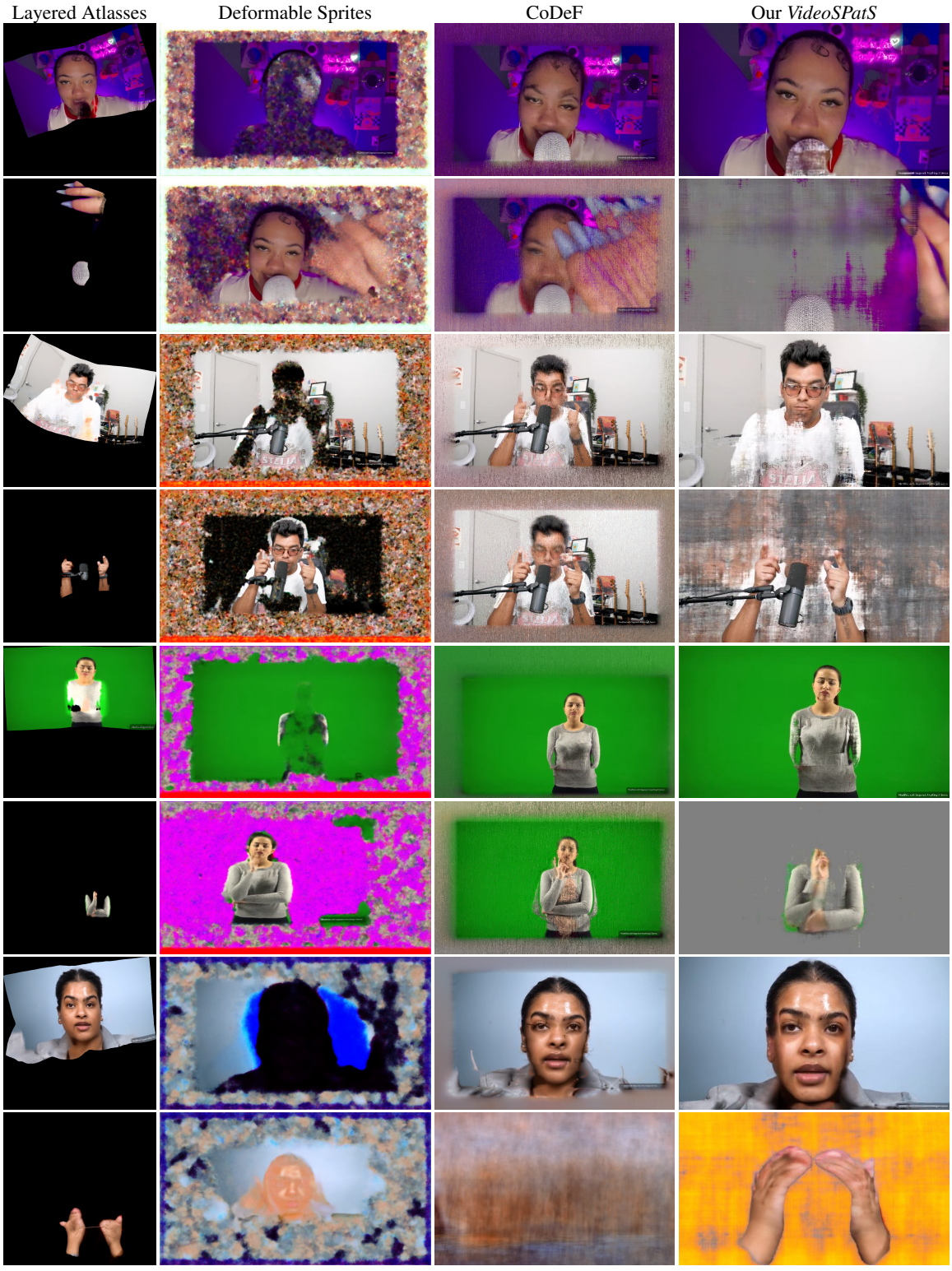

Figure 12. Comparisons of the obtained canonical spaces with other methods. For every two rows, the top row corresponds to the background canonical space, and the bottom row corresponds to the foreground canonical space. Our *VideoSPaTS* consistently yields more editing-intuitive and feasible canonical spaces.

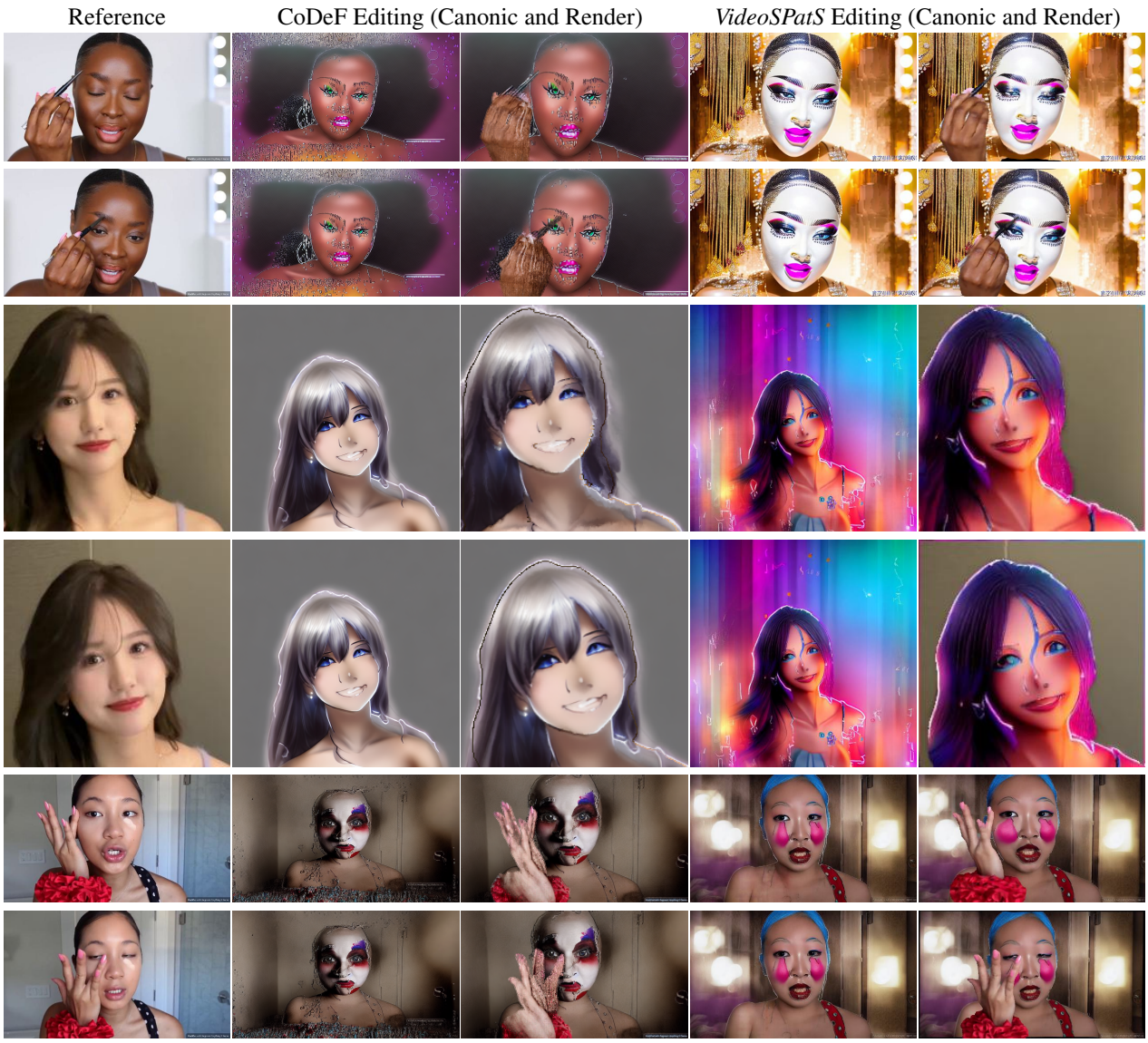

Figure 13. Additional editing results. The consistency of our canonical spaces allows for better deep editing than that of CoDeF.

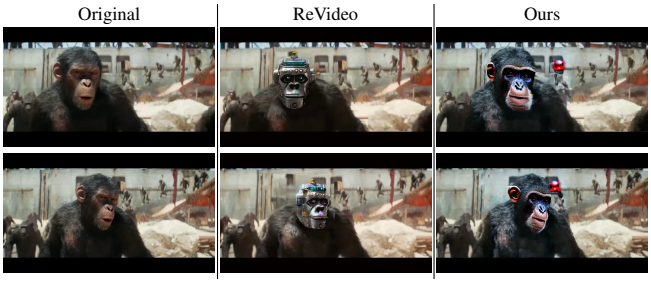

Figure 14. Comparison to ReVideo [18].

C.4. Texture editing

We provide additional editing results in Fig 13. We use ControlNet [42] to apply editing on the canonical space. Note that the inconsistencies of canonical spaces in CoDeF prevent ControlNet from generating a high quality edit, as shown in the first and last rows of Fig 13. In contrast, our method generates more edit-friendly canonical spaces that are translated into higher-quality, temporally-consistent images.

C.5. Motion editing

By modifying the precomputed control points, we can smoothly perform motion editing. For instance, we can

Figure 15. Additional results on motion editing by control points. See our videos for a better visualization.

select every m control point of each foreground pixels and apply a vertical offset. Thanks to the spline nature of our deformation fields, we can smoothly transfer this new motion into the rendered video. Thanks to our spline deformation fields, instead of rendering frames where the foreground is instantly "teleporting" to the offset location, our motion-edited frames are smoothly rendered without discontinuities. Additional motion edits, such as amplification and diminishing of motion, are shown in the attached videos as well as in Fig. 15.

C.6. Experiments on long sequences

While most of the experiments mentioned above were conducted with videos of 50 frames, our method also performs well on longer sequences. Fig. 16 presents additional results on sequences of 10 seconds. Our method is capable of capturing the long-range correspondences in longer videos.

D. Additional ablation studies

D.1. Spatial regularization loss

We show the effects of the Spatial Splines Deformation Regularization loss, $l_{\mathcal{D}_s}$, in Fig. 17. Although the contribution of the regularization loss is minimal to the canonical space and final reconstruction, it still helps maintain a better aspect ratio between the canonical space and the observed space. This is because it encourages similar deformations between neighboring pixel locations, preventing the "squeeze" of the canonical space, as observed in the "without $l_{\mathcal{D}_s}$" column of Fig. 17.

D.2. Color regularization loss

Fig. 18 depicts the effects of the Color Deformation Regularization loss, l_{D_c}, showing that not regularizing P_c can lead to potential entanglement between motion and appearance in the canonical space, as shown in the bent finger on the rightmost image.

D.3. Levels of guidance mask

In the main paper, we show that our method can refine the guidance mask. Fig. 20 provides additional results on different levels of degradation of the guidance mask. In this supplemental study, our motivation is to show the robustness of the proposed method when the guidance mask is imperfect. As shown in Fig. 20 our proposed model can capture the foreground motion even with a rough mask. Although our method cannot recover the mask when it is too heavily degraded (last row in Fig. 20), it still succeeds with smaller degradation levels, supporting our design choices in Section 3.3.

D.4. Number of control points

Fig. 21 provides additional ablation studies on the number of control points. While the best fit can be obtained with the number of control points equal to the number of frames, our method can also reasonably reconstruct the scene with fewer control points.

D.5. Number of iterations

While performance optimization was not the research focus of this work, we acknowledge the processing time can be accelerated using faster neural representations (e.g. hash encodings [20]), optimized learning libraries (e.g. PyTorch Lightning [6]), and quantization (half-precision). We provide additional ablation studies on the effects of training iterations in Fig. 19. As observed, reasonable results can be obtained with 30K iterations (<30min), with only a 1dB drop w.r.t. the fully trained model (∼90min).

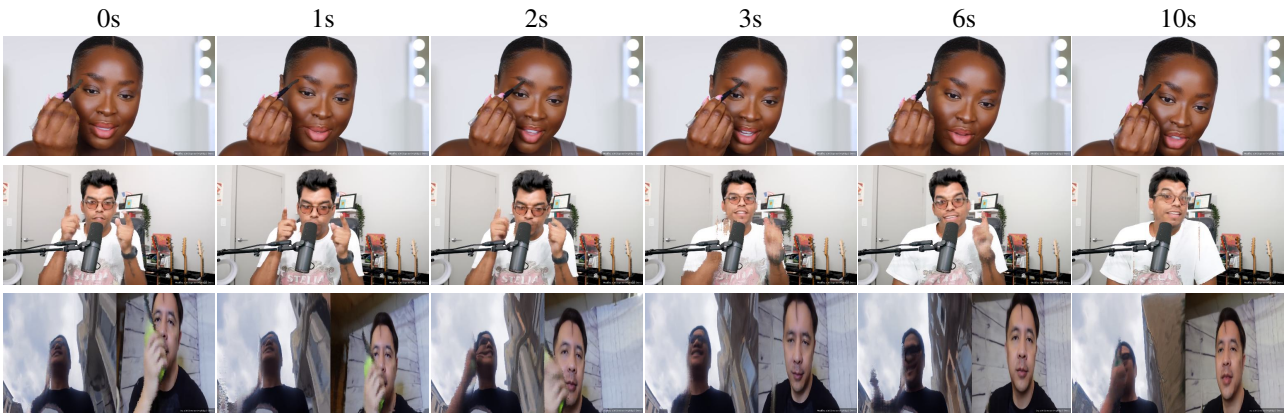

Figure 16. Additional results on long sequences. Our method can capture long-range relationships in long video sequences (10s).

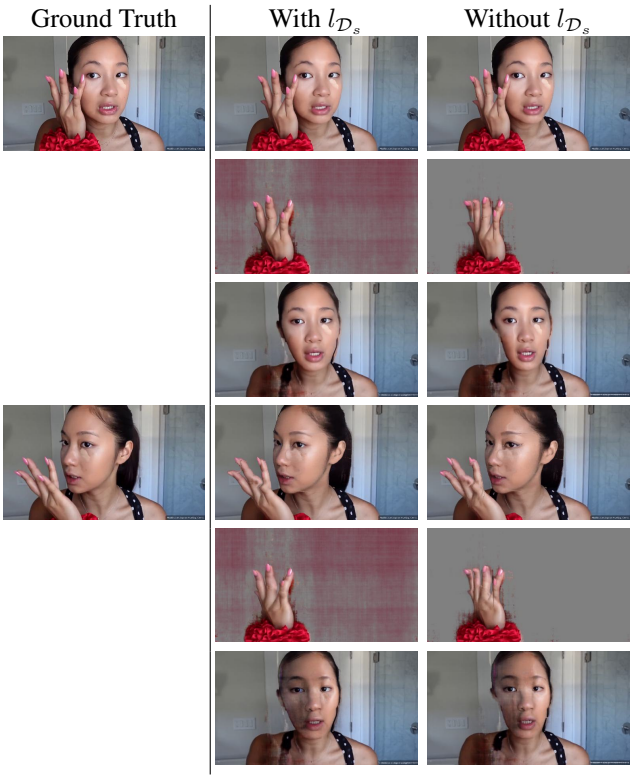

Figure 17. Effects of $l_{\mathcal{D}_s}$. *From top to bottom:* Composited images, foreground occluder canonical spaces, and background face canonical spaces. Our model without $l_{\mathcal{D}_s}$ yields a slightly squeezed canonical space, with respect to the observed frames and our model with $l_{\mathcal{D}_s}$.

E. Failure cases

Fig. 22 illustrates examples of failure cases. In the top two rows, our method fails to reconstruct a feasible canonical space for the background face. This is because the relative size of the facial region with respect the amount and complexity of the motion is very small. A work-around for this issue would be to crop the images around the face region and run our method again. In the bottom two rows, the amount of motion is too large for our model to capture. In these cases, the brush goes from one side to the other and also rotates showing different faces of it, inducing two brushes on our estimated canonical space. A potential solution would consist on modeling the brush with different layers when it is on one side or the other.

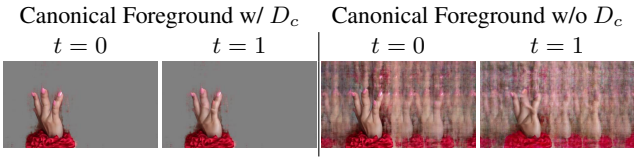

Figure 18. Effects of l_{D_c}. $t = 0$: start, $t = 1$: end of the video.

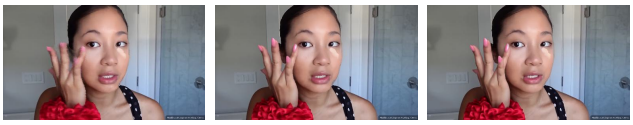

Figure 19. Effects of Iteration # on reconstruction PSNR.

F. Ethical implications

The use of ControlNet in conjunction with our proposed method to modify the appearance of video content may raise ethical concerns around authenticity and potential misuse, such as creating misleading information. To address these concerns, we advocate for the responsible and transparent use of this technology, ensuring that any modifications are clearly indicated and used ethically.

Our collected dataset from publicly available YouTube videos contains exclusively *Creative Commons* licensed videos, with the corresponding URLs provided in the

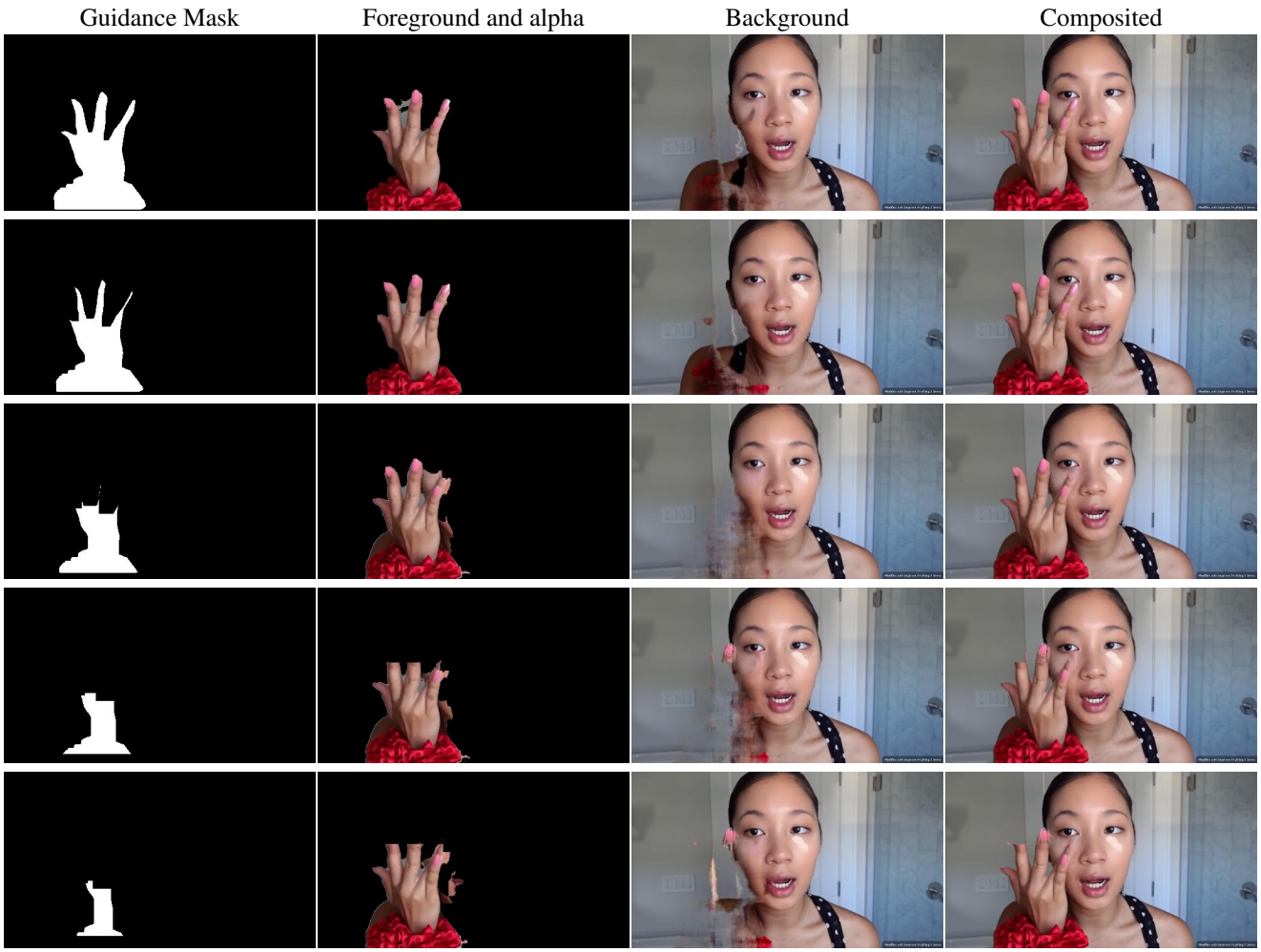

Figure 20. Additional ablation studies on guidance mask. Original guidance mask from SAM2[28] is eroded by 5, 11, 21, 31, and 41 pixels. Even under extreme erosion, our method can still reasonably separate the occluder foreground and the face background.

`urls.json` file. Authors of these videos are free to contact us upon publication (due to the anonymous nature of submission) to have their videos removed from this dataset or paper results.

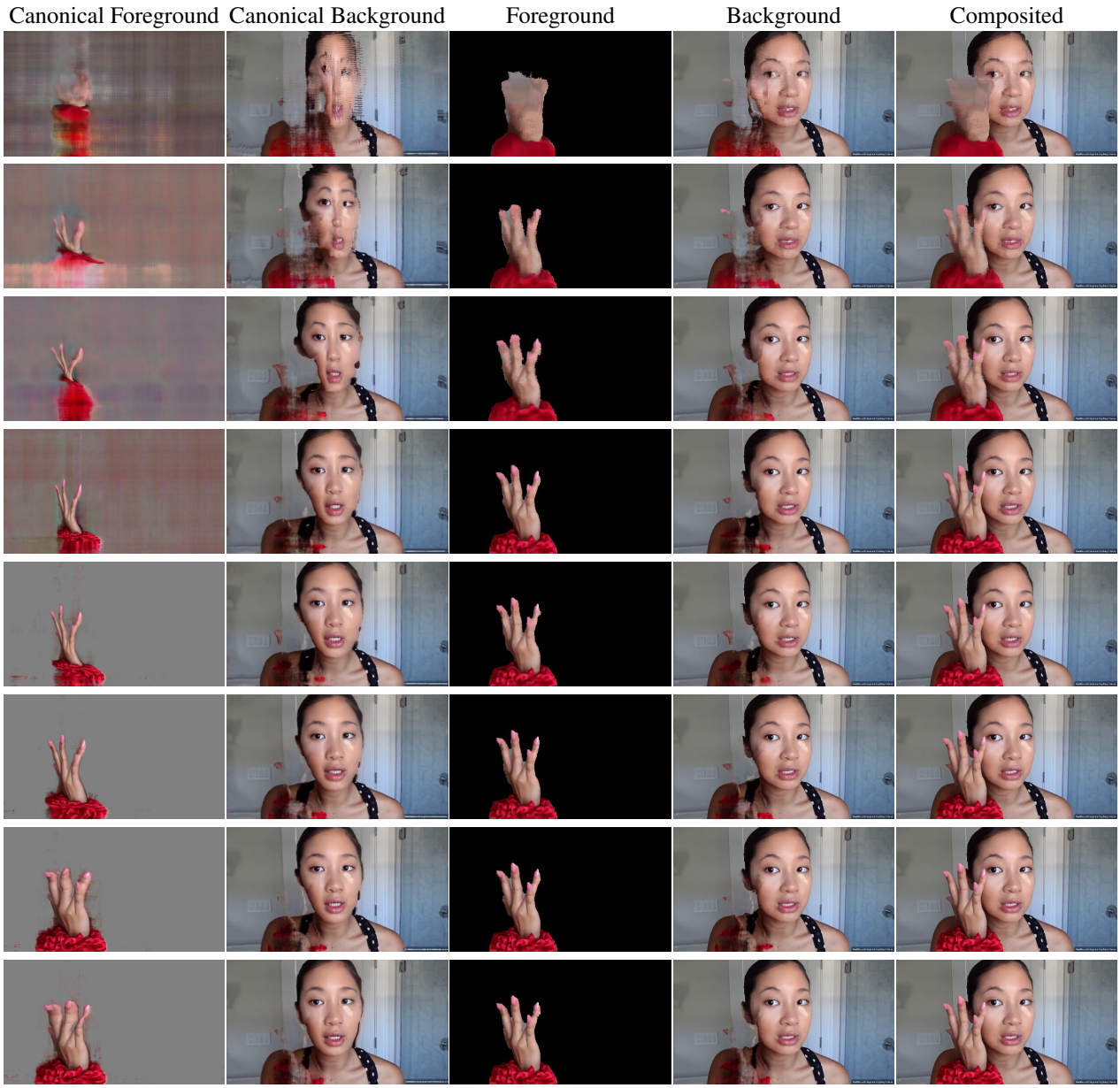

Figure 21. Additional ablation studies on number of control points. *From top to bottom:* 2 (24.195dB), 4 (26.132dB), 8 (28.433dB), 16 (32.209dB), 20 (32.721dB), 30 (34.280dB), 41 (37.010dB), and 82 (36.606dB) control points for a video of 41 frames.

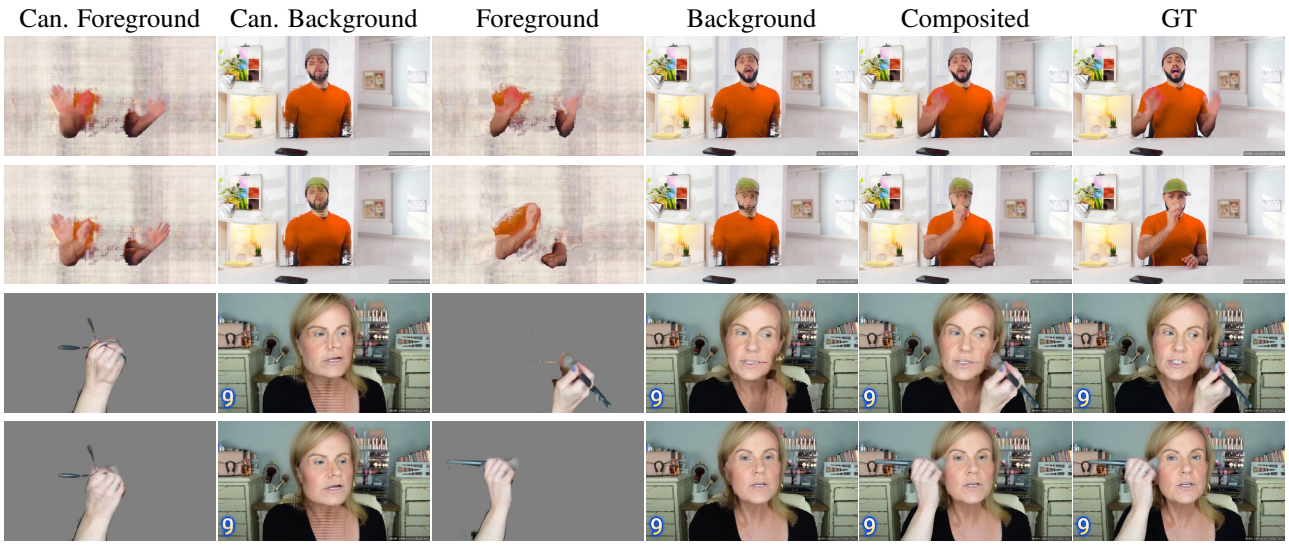

Figure 22. Failure cases. *Top two rows:* The dynamic region in background image (face region) is too small. *Bottom two rows:* Too large foreground motion and self-occlusion (opposite sides of brush) cause a double brush effect in foreground canonical space.

References

[1] Richard H Bartels, John C Beatty, and Brian A Barsky. *An introduction to splines for use in computer graphics and geometric modeling*. Morgan Kaufmann, 1995. 3

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 1

[3] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*, pages 25–36. Springer, 2004. 3

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, 2023. 3

[5] Ilya Chugunov, David Shustin, Ruyu Yan, Chenyang Lei, and Felix Heide. Neural spline fields for burst image fusion and layer separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25763–25773, 2024. 2, 3

[6] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. 13

[7] Gerald Farin. *Curves and surfaces for CAGD: a practical guide*. Elsevier, 2001. 3

[8] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. 3

[9] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 3

[10] Nebojsa Jojic and Brendan J Frey. Learning flexible sprites in video layers. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I. IEEE, 2001. 2

[11] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 1, 2, 3, 4, 5, 9, 10

[12] Junhua Liao, Haihan Duan, Xin Li, Haoran Xu, Yanbing Yang, Wei Cai, Yanru Chen, and Liangyin Chen. Occlusion detection for automatic video editing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2255–2263, 2020. 2

[13] Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, and Ayush Saraf. Omnimatterf: Robust omnimatte with 3d background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23471–23480, 2023. 2

[14] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *ACM Transactions on Graphics*, 39(6), 2020. 1, 2

[15] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4507–4515, 2021. 1, 2

[16] Julien NP Martel, David B Lindell, Connor Z Lin, Eric R Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *arXiv preprint arXiv:2105.02788*, 2021. 3

[17] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3, 6

[18] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2024. 2, 10, 12

[19] Thomas Müller. tiny-cuda-nn, 2021. 9

[20] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 1, 4, 13

[21] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8089–8099, 2024. 1, 2, 3, 4, 5, 6, 7, 10

[22] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 10

[23] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 2

[24] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 3

[25] M Pawan Kumar, Philip HS Torr, and Andrew Zisserman. Learning layered motion segmentations of video. *International Journal of Computer Vision*, 76:301–319, 2008. 2

[26] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 7

[27] Alex Rav-Acha, Pushmeet Kohli, Carsten Rother, and Andrew Fitzgibbon. Unwrap mosaics: A new representation for video editing. In *ACM SIGGRAPH 2008 papers*, pages 1–11, 2008. 2

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6, 15

[29] 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 (CVPR)*, pages 10684–10695, 2022. 6

[30] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 3

[31] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 1, 3

[32] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 1, 3

[33] Genmo Team. Mochi 1: A new sota in open-source video generation. https://github.com/genmoai/models, 2024. 1

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4, 7, 10

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 9

[36] John YA Wang and Edward H Adelson. Representing moving images with layers. *IEEE transactions on image processing*, 3(5):625–638, 1994. 2

[37] Yiran Xu, Zhixin Shu, Cameron Smith, Seoung Wug Oh, and Jia-Bin Huang. In-n-out: Faithful 3d gan inversion with volumetric decomposition for face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7225–7235, 2024. 2

[38] Wenqi Yang, Zhenfang Chen, Chaofeng Chen, Guanying Chen, and Kwan-Yee K Wong. Deep face video inpainting via uv mapping. *IEEE Transactions on Image Processing*, 32: 1145–1157, 2023. 2

[39] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[40] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2666, 2022. 1, 2, 7, 9, 10

[41] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2437–2447, 2023. 2

[42] 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. 12

[43] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10477–10486, 2023. 2

[44] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004. 2