V-LASIK: CONSISTENT GLASSES-REMOVAL FROM VIDEOS USING SYNTHETIC DATA

Rotem Shalev-Arkushin
Tel-Aviv University, Lightricks
rotemroo@gmail.com

Aharon Azulay
Lightricks

Tavi Halperin
Lightricks

Eitan Richardson
Lightricks

Amit H. Bermano
Tel-Aviv University

Ohad Fried
Reichman University

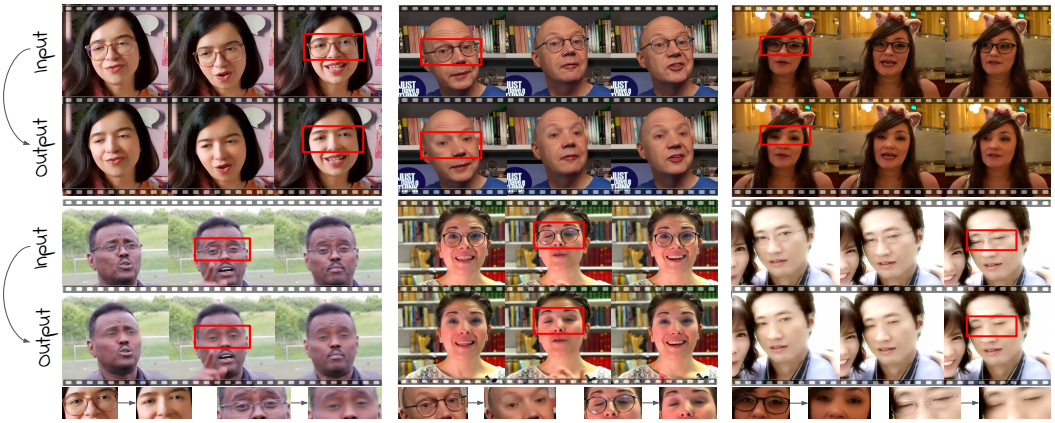

Figure 1: Our method receives an input video of a person wearing glasses, and consistently removes the glasses, while preserving the original content and identity of that person. Our method successfully removes the glasses even when there are reflections (bottom-middle example), heavy makeup (top-right), and eye blinks (bottom-right). Red rectangles are zoomed-in at the bottom row.

ABSTRACT

Recently, diffusion-based generative models showed remarkable image and video editing capabilities. However, local video editing, particularly removal of small attributes like glasses, remains a challenge. Existing methods either alter the videos excessively, generate unrealistic artifacts, or fail to perform the requested edit consistently. In this work, we focus on consistent and identity-preserving glasses-removal in videos, as a case study for video local attribute removal. We demonstrate the generalizability of our method by applying it to facial sticker removal from videos. Due to the lack of paired data, we adopt a weakly supervised approach and generate synthetic imperfect data, using an adjusted pretrained diffusion model. We show that despite data imperfection, by learning from our generated data and leveraging the prior of pretrained models, our model is able to perform the desired edit consistently while preserving the original video content. Furthermore, we suggest a new normalization method, Inside-Out Normalization, which aligns colors in a filled region with colors outside that region. Our approach offers significant improvement over existing methods, showcasing the potential of leveraging synthetic data and strong priors for local video editing tasks.

1 INTRODUCTION

Recent advances in diffusion-based generative models Ho et al. (2020); Sohl-Dickstein et al. (2015) have demonstrated impressive capabilities in image and video editing (Rombach et al., 2022; Zhang

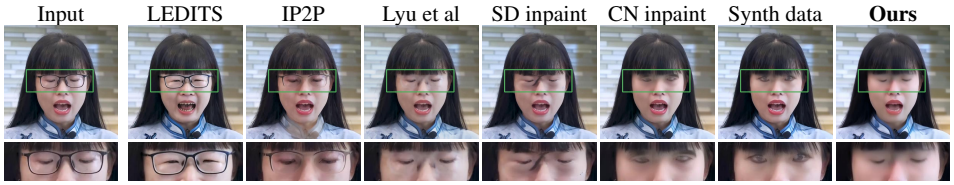

Figure 2: **Glasses-removal from a blinking eye by image editing methods** Image editing methods struggle with blurry images or images of people not looking straight at the camera. Although in our synthetic dataset sample ('Synth data'), eye artifacts appear, our model can still learn from the imperfect data and removes the glasses better than other methods, and better than its training data.

et al., 2023a; Geyer et al., 2023; Yang et al., 2023; Zhang et al., 2023b; Kara et al., 2023; Bar-Tal et al., 2024). While local editing, such as removing attributes without changing the rest of the content, mainly works for images, for videos it remains a challenge. Video frames often contain motion blur and challenging poses that are less common in images. Moreover, videos of people pose additional challenges, as humans are highly sensitive to subtle unrealistic "uncanny" artifacts, and our visual system is more sensitive to motion inconsistencies, which leads to a higher quality bar. Existing image editing and inpainting methods Meng et al. (2021); Zhang et al. (2023a); Tsaban & Passos (2023); Brooks et al. (2023); Couairon et al. (2023); Nitzan et al. (2024) are trained on images that do not contain the aforementioned problems, and therefore do not perform well over video frames; They either change the frames too much, or do not perform the requested edit completely, as illustrated in Fig. 2. Video inpainting methods Zhou et al. (2023); Yu et al. (2023b); Li et al. (2022b); Zhang et al. (2022) typically remove objects by deleting and filling regions of the video, often disregarding the information originally present in those regions. When trying to generate new video content behind an object that never moves, these methods tend to generate a low detailed background, or one with clear unrealistic artifacts. As a representing use-case for local attribute removal, we choose to tackle glasses removal, as it is a particularly challenging example; where some content is never or only partially revealed (e.g. eyes, eyebrows), and is changed across the video (e.g. eyelids change position). Since eyes contribute greatly to identity perception, a consistent and highly realistic generation is required for an identity to be preserved. However, current video inpainting works typically generate smooth results when using small masks, that do not fit the facial case well (see ProPainter Zhou et al. (2023), FGT Zhang et al. (2022) results in Fig. 7). Mask-free video editing methods Geyer et al. (2023); Zhang et al. (2023b); Khachatryan et al. (2023); Kara et al. (2023); Cong et al. (2024) mainly focus on changing the style of the whole video or large parts of it, and not on local changes that should adhere to the original context. Other works Kasten et al. (2021); Bar-Tal et al. (2022) learn an atlas for each new video, leading to long inference times. Furthermore, the goal of these works is to manipulate the appearance of existing objects, not to change the composition of the elements in the scene (Bar-Tal et al., 2022). Hence, their method is not capable of successfully removing attributes such as glasses from faces.

In this work, we introduce a new method that consistently removes glasses from videos while preserving the content of the original video. To the best of our knowledge, no paired data exists for this task; Therefore, we cannot finetune a model directly to solve it. To tackle this problem, we generate imperfect synthetic data, and use it to train a model in a weakly supervised manner. We show that by using an adjusted pretrained image inpainting model, we can generate imperfect data pairs, and learn from them. Our method is closest to video inpainting in the sense that it removes something from the video and replaces it with different content. However, similarly to recent video editing works, the input to our model is a mask-free image, exposing it to facial details that are occluded for inpainting models. As glasses often contain colored lens and reflections that change across frames, which affect the way the eye behind them looks, they are a challenging attribute to remove from faces. Hence, we incorporate cross-frame attention layers in our data generation model, which allows aggregating information from different frames. Still, as presented in Fig. 2, the generated data ('Synth data') is imperfect and contains many artifacts, such as not always preserving the eyelids positions from the original frame. To overcome these challenges, we finetune a pretrained diffusion based image-to-image model over this data. This way, we can leverage its prior, while enhancing its capabilities with our new data. Despite imperfections in our synthetic data, the model can leverage its prior knowledge to outperform the training data and produce high-quality results. For temporally consistent results, we combine our trained model with a motion prior. Finally, for smooth and content-faithful edits, we suggest a novel normalization method, Inside-Out Normalization (ION)

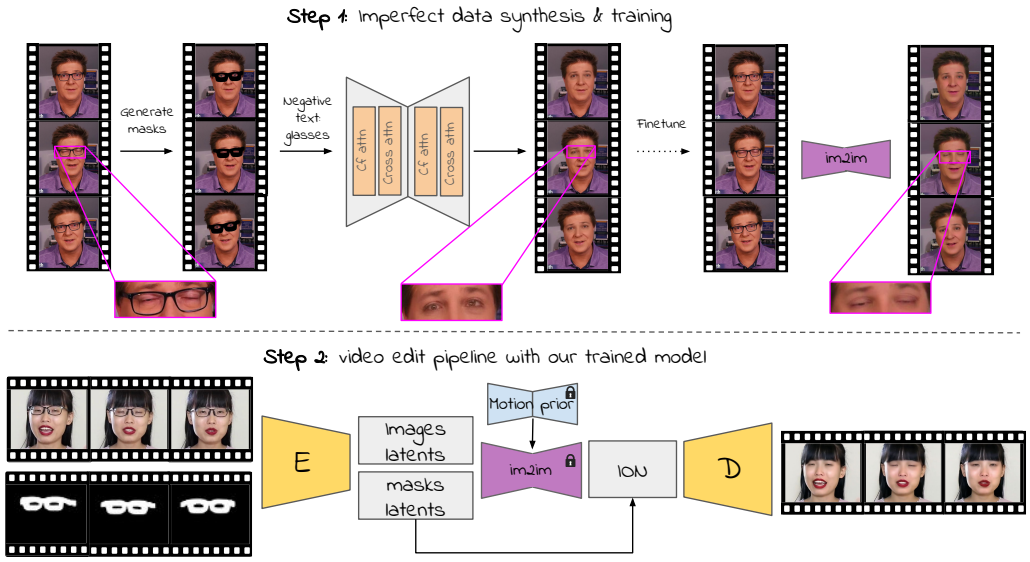

Figure 3: **Method overview: Step 1:** we create a synthetic paired dataset by generating glasses masks for each video frame and inpainting them with cross-frame attention. The generated data in this step is imperfect; e.g. in the input middle frame, the person blinks, however its generated pair has open eyes. Nevertheless, this data is good enough for finetuning an image-to-image (im2im) diffusion model and achieving satisfactory results, due to the strong prior of the model. **Step 2:** We incorporate our trained model with a motion prior to generate temporally consistent unseen videos without glasses. Before decoding, we apply a new normalization method — Inside-Out Normalization (ION), to better align the statistics within the masked area and outside the mask.

that allows matching the colors of a masked region with the area outside the mask. We achieve state-of-the-art results on video glasses-removal, surpassing all current video editing and inpainting methods. To demonstrate that our method can be generalized to other local video editing tasks, we successfully apply it to the task of removing stickers from faces.

2 METHOD

Given an input video of a person wearing glasses, our goal is to remove the glasses while preserving all other information. As illustrated in Fig. 3, our approach consists of three stages: data generation, training, and editing. First, as no paired data is available for this task, we generate a synthetic paired dataset using videos of people wearing glasses. Next, to get realistic video frames without glasses, where all other parts of the frame remain similar to the original frame, and the identity of the edited person is preserved, we finetune an image-to-image (im2im) diffusion model over our synthetic dataset. Finally, we incorporate our trained model with a pretrained motion module into a video editing pipeline, to obtain temporally consistent videos without glasses.

Paired Data Generation: As illustrated in Fig. 2, current image editing and inpainting methods do not remove glasses well from video frames. Therefore, we wish to train a model for this task. However, to the best of our knowledge, no relevant paired dataset currently exists. Hence, we create a synthetic dataset of paired video frames, with and without glasses, using videos from the CelebV-Text dataset Yu et al. (2023a) where people wear glasses. For each video frame, we generate a glasses mask using a face parser Zheng et al. (2022) that identifies the glasses in it. The eyes are a key component in the identity of a person, and the current position of the eyelids is crucial to generating a result that is consistent with the original video, that might contain eye-closure and blinking. Therefore, we want to give the model information about the eyes and eyelids, to keep their original appearance. To do that, we make eye "holes" in the glasses masks, using the identified eye landmarks. Then, we inpaint the glasses area by applying an adjusted inpainting diffusion model over each video frame and its generated mask. We make two adjustments to the inpainting model: First, to achieve smoother and more realistic transitions between the frame and the inpainted part,

we blend the latent feature vectors with a noised encoded version of the masked original frame at each diffusion step, as suggested in Blended Latent Diffusion (Avrahami et al., 2023). Second, to increase the edit consistency across frames from the same video, we replace the self-attention layers of the model with a cross-frame attention (Khachatryan et al., 2023). For each video frame, we perform a cross-frame attention with k reference frames from the same video. We use multiple reference frames as sometimes information such as the eyebrows or the eye color is hidden behind the glasses or their reflections. Using cross-frame attention with multiple frames allows for generalization from all k frames. For this purpose, we adopt the cross-frame attention mechanism suggested by TokenFlow Geyer et al. (2023):

$$\text{softmax}(\frac{Q \cdot [K_0, ..., K_k]^T}{\sqrt{d}}) \cdot [V_1, ..., V_k] \tag{1}$$

where each query frame (Q) attends to k different values (V_i), corresponding to k. As a result, even when face parts are occluded, the information about the occluded parts can be retrieved from the reference frames if they are revealed there. An example of occluding reflections is presented on the right side of Fig. 5. In this example, there are reflections over the glasses, hiding the true eye colors behind them. Without using cross-frame attention, reflections and artifacts are present on the output result. However, when using cross-frame attention, the true eye color is preserved and no reflections are shown in the result. Moreover, using cross-frame attention also helps when the masks are not exact and do not cover all of the glasses in all video frames, as in the left example of Fig. 5. In this case, glasses remnants appear when no cross-frame attention is used, due to an incomplete mask. However, they are removed when we use cross-frame attention, where the masks of the reference frames cover the uncovered part.

Model fine tuning: We generated pairs of frames with and without glasses. However, they are not temporally consistent, and contain per-frame artifacts in many cases, as featured in Figs. 2 and 3. Despite that, they are good enough for finetuning a pretrained im2im model for glasses-removal from faces. These models have strong prior for generating realistic looking images that are similar to the original ones and preserve their small and delicate details. Therefore, after finetuning, the model learns the task of removing glasses from our data, while preserving fine details such as eye color and eyelid positions.

Video editing pipeline: To consistently remove glasses from unseen videos, we integrate our trained model with a pretrained motion module (Guo et al., 2024). When applied directly to video frames, the results of this model often exhibit different colors from the original input. To overcome this issue, we can use the masks from the data generation stage, and retrieve the original values in areas that are outside of the mask. For a smooth result, at each diffusion step we perform a gradual blending between the noised masked input and the generated result, such that the area within the mask changes completely, and the areas that surround the mask change less, gradually decreasing the amount of change as the pixels are further from the mask. Additionally, inspired by AdaIN Huang & Belongie (2017), we apply a new normalization method, we dub Inside-Out Normalization (ION), aimed at aligning the statistics of the masked area with those of the area outside the mask. This way, the original video colors are preserved, and the transition between the masked area and the non-masked area is smoother. Formally, we calculate the mean and standard deviation of the masked area and the non-masked area, μ_m, σ_m and $\mu_{\bar{m}}, \sigma_{\bar{m}}$ respectively, and normalize the values of the latent features in the area inside the mask by calculating:

$$ION(x) = \sigma_{\bar{m}} \frac{x - \mu_m}{\sigma_m} + \mu_{\bar{m}} \tag{2}$$

3 EXPERIMENTS

We evaluate our results both quantitatively and qualitatively, testing three aspects: 1. **Fidelity** to the required edit, i.e. glasses-removal. 2. **Identity and content preservation** of the original video, i.e. removing glasses while leaving the rest of the video intact. 3. **Realism** in terms of temporal consistency and per-frame. As we perform local video editing, we compare our results with the results of the SOTA video editing methods: TokenFlow Geyer et al. (2023), RAVE Kara et al. (2023), and Text2Video-Zero Khachatryan et al. (2023). TokenFlow and Text2Video-Zero incorporate SDEdit Meng et al. (2021) and instructPix2Pix Brooks et al. (2023), respectively, to allow for local attribute editing, and RAVE uses CN "depth-zoe" Zhang et al. (2023a) for that purpose. We also compare to

the SOTA video inpainting works ProPainter Zhou et al. (2023) and FGT Zhang et al. (2022) as our work is similar to inpainting works in the sense that it tries to replace a part of the video. Finally, we compare to our video editing pipeline with CN inpaint Zhang et al. (2023a) instead of our trained model, to emphasize that existing image inpainting models do not perform well on this task, even when combined in our video editing pipeline. Our experiments, including a comprehensive user study, demonstrate that our results are favored over all other methods across all tested aspects.

3.1 Experimental details

We generate our data pairs as described in Sec. 2 from the dataset CelebV-Text (Yu et al., 2023a). We train our model over 1296 of those videos, and test over 144 unseen videos. More implementation details including used models can be found in Appendix C.

3.2 Qualitative evaluation

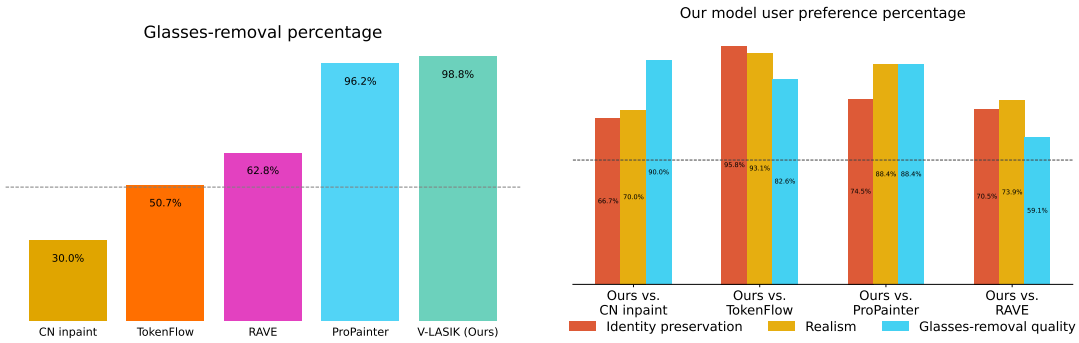

Figure 4: **User study results:** First, we ask users which models remove the glasses from the input video (left). Then, when both models remove the glasses, we ask which one better preserves the identity of the person, which one is more realistic, and which result contains less glasses-remnants (right). Compared to all examined models, users preferred our results in all measured aspects.

Visual results of our method are presented in Figs. 7 and 8. As shown, existing video editing and inpainting methods struggle with removing glasses from a person. Moreover, even when they do remove the glasses, they tend to generate artifacts and unrealistic results, leave glasses-remnants, or do not preserve the identity of the person or its original eyelid positions. In contrast, our method successfully removes the glasses while preserving the identity and content of the original video. To evaluate our results and compare to other methods qualitatively we performed a thorough user study with 57 users, answering a total of 627 comparisons. In the study we compared our results with the results of TokenFlow Geyer et al. (2023), RAVE Kara et al. (2023), ProPainter Zhou et al. (2023), and our editing pipeline with CN inpaint Zhang et al. (2023a). We did not compare to Text2Video-Zero, because it only removes the glasses from the input video in 4% of the test videos, as exemplified in Fig. 7. The user study tested three aspects: glasses-removal, resemblance to the identity in the original video, and realism of the result. For each video of a person wearing glasses in the survey, we presented the users with two other videos at a time: our result, and a competing method result, and asked in which of the videos the glasses are removed. If both removed the glasses for that video, we additionally asked which video contains less remnants of glasses, which one looks more realistic, and which one better preserves the identity of the person in the original video. We ask for glasses-remnants because we noted that even when other models remove the glasses from the video, they often leave parts of the glasses or their reflections, (e.g. Fig. 7, RAVE results). The results of our user study are presented in Fig. 4. The left table shows the percentage of videos for which the users thought each model removed the glasses from the input video. Users thought our model removed the glasses in 98.9% of the cases, more than any other model. The right table shows the percentage of users that preferred our results over the compared models. Users preferred our results over all other models, in all the measured aspects: identity preservation, realism, and glasses-removal quality.

Table 1: **Quantitative results:** We compare our results to different video inpainting and editing methods, where CN inpaint* is CN inpaint embedded in our video editing pipeline instead of our model. We present results for two versions of our model — with and without masks, as elaborated in Sec. 3.3. ΔG stands for the average difference between glasses pixels in the original video frames and the edited one. ID stands for ID score, $ID \cdot \Delta G$ stands for the tradeoff between them, and E_{warp} stands for optical flow warp error.

Method	$\Delta G_{(\times.01)} \uparrow$	$ID_{(\times.1)} \uparrow$	$ID \cdot \Delta G_{(\times.01)} \uparrow$	$E_{warp(\times 10^{-4})} \downarrow$
ProPainter Zhou et al. (2023)	**4.0**	7.5	**3.0**	3.9
FGT Zhang et al. (2022)	**4.0**	7.1	2.8	4.0
CN inpaint* Zhang et al. (2023a)	3.0	8.0	2.3	4.1
TokenFlowGeyer et al. (2023)	2.0	6.9	1.4	5.7
RAVE Kara et al. (2023)	2.0	7.4	1.5	6.8
T2V-Zero Khachatryan et al. (2023)	0.3	**9.0**	0.3	8.1
V-LASIK (ours)	**4.0**	7.6	**3.0**	**3.6**
V-LASIK masked (ours)	**4.0**	7.6	**3.0**	4.2

3.3 Quantitative evaluation

We compare the results of different video editing and inpainting methods to the results of our model in two versions — with and without masks. Both options exist for our model and there is a trade-off between them — the masked version better preserves the original video colors, while the non-masked version better removes the glasses and thus is more temporally consistent (more details in Appendix C.3). These differences are small, hence as shown in Tab. 1, they only slightly affect the results. We use the following measurements to asses the criteria of fidelity, identity preservation, and temporal consistency: **Fidelity:** We find the pixels that contain glasses in an image using a face parser Zheng et al. (2022) that detects glasses, and apply it to each video frame. Then, we calculate the average difference between the number of pixels with glasses in the original frame and the edited one, normalized by the total number of pixels per frame, and report it as ΔG in Tab. 1. As the ΔG scores in Tab. 1 show, our method removes the glasses better than all other video editing methods, and is on-par with the inpainting methods Zhou et al. (2023); Zhang et al. (2022). However, as exemplified in Fig. 7 and by the results of our user study (Fig. 4), although they remove the glasses from most videos, as our method does, they often generate unrealistic results with artifacts around the eyes, which our model does not generate. **Identity preservation:** we use an ID score (ID in Tab. 1), which is defined by the average cosine similarity between the face embeddings of the video frames, generated by the face recognition model Arcface Deng et al. (2019). We note that neither of these metrics is complete on its own, as an unchanged video would get a perfect ID score, and a random video without glasses would get a very high ΔG score. For example, as Text2Video-Zero does not remove the glasses from most videos, it achieves a very high ID score. For this reason, we follow prior work Kara et al. (2023); Cong et al. (2024) and also look at $ID \cdot \Delta G$, which quantifies the trade-off between removing the glasses from the video, and remaining faithful to the identity of the person in the original video. When looking at $ID \cdot \Delta G$, our model achieves the best results, together with ProPainter Zhou et al. (2023). However, as mentioned, the results of our user-study in Fig. 4 show that our model outperforms ProPainter and all other methods both in terms of glasses-removal, and in terms of realism and identity preservation. **Temporal consistency:** following previous work Geyer et al. (2023); Zhou et al. (2023); Lei et al. (2020); Lai et al. (2018) we report warp error (E_{warp} in Tab. 1). We calculate optical flow using RAFT Teed & Deng (2020) between each consecutive pair of frames, warp the former towards the latter, and calculate a masked MSE loss (masking occlusions) between the warped frame and the second frame. As shown in Tab. 1, our non-masked model produces the most temporally consistent results.

3.4 Ablations

In this section we demonstrate the importance of different parts in our method. First, the importance of cross-frame attention is presented in Fig. 5. The left example presents the case where the glasses mask is imperfect. In that case, removing the glasses without cross-frame attention may sometimes leave glasses-remnants. When using cross-frame attention, the glasses are fully removed. The right example presents the case where reflections hide the color of the eyes. In that case, when not using

Figure 5: **Cross-frame attention importance in data generation.** Cross-frame attention helps removing glasses remnants, even when the mask is not perfect (left example) and reducing glasses reflections (right example).

Figure 6: **Importance of ION.** our goal is to remove the glasses while preserving the original colors. When not normalizing, the colors may be very different from the original ones. When using AdaIN, the colors are closer to the originals, but are still different. When using our suggested ION, the original colors are preserved.

cross-frame attention, artifacts resulting from the reflections appear in the result. However, when using cross-frame attention, the true eye color is preserved and no reflections appear in the result.

Next, the importance of using our suggest ION is presented in Fig. 6. When not using any form of normalization, the colors are changed completely from the original video colors after removing the glasses. When using AdaIN, the results are more similar to the original ones but are still not exactly the same, and the transition between the masked area and the non-masked area is visible and not smooth. When using our proposed ION, the original video colors are preserved, and the transition between the masked area and the non-masked area is smoother.

4 Discussion

We explored the potential of local video editing through learning from imperfect synthetic data without paired data. Our results surpass existing methods, consistently and realistically removing glasses from videos while preserving the identity of an individual. We focus on the challenging case of removing glasses from videos, however we show our method can also be applied to other local video editing tasks such as removing stickers from faces (Appendix B). We hypothesize it would work similarly for any other local attribute, and we encourage future work to pursue this direction. While the goal of this work is to allow editing of owned or licensed videos only, we acknowledge it may be misused to alter videos without consent, and contribute to the spread of misinformation. We condemn such usage, and actively work on systems that detect synthetic and edited media.

5 Acknowledgements

We thank Yotam Nitzan for fruitful discussions. This project was funded in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT), and by ISF grant numbers 1574/21, 1337/22.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pp. 707–723. Springer, 2022.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. In *International Conference on Learning Representations (ICLR)*, 2024.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 75–82. IEEE, 2018.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2023.

Bohai Gu, Yongsheng Yu, Heng Fan, and Libo Zhang. Flow-guided diffusion for video inpainting. *arXiv preprint arXiv:2311.15368*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. *arXiv preprint arXiv:2312.04524*, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

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.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 170–185, 2018.

Yu-Hui Lee and Shang-Hong Lai. Byeglassesgan: Identity preserving eyeglasses removal for face images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 243–258. Springer, 2020.

Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020.

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21330–21340, 2022a.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022b.

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*, pp. 4507–4515, 2021.

Junfeng Lyu, Zhibo Wang, and Feng Xu. Portrait eyeglasses and shadow removal by leveraging 3d synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3429–3439, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Thao Nguyen, Anh Tran, and Minh Hoai. Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yotam Nitzan, Zongze Wu, Richard Zhang, Eli Shechtman, Daniel Cohen-Or, Taesung Park, and Michaël Gharbi. Lazy diffusion transformer for interactive image editing. *arXiv preprint arXiv:2404.12382*, 2024.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2085–2094, 2021.

William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13481, 2022.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2287–2296, 2021.

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*, pp. 10684–10695, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Mohammed Suhail, Erika Lu, Zhengqi Li, Noah Snavely, Leonid Sigal, and Forrester Cole. Omnimatte3d: Associating objects and their effects in unconstrained monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 630–639, 2023.

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*, pp. 402–419. Springer, 2020.

Linoy Tsaban and Apolinário Passos. Ledits: Real image editing with ddpm inversion and semantic guidance. *arXiv preprint arXiv:2307.00522*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

Yangyang Xu, Shengfeng He, Kwan-Yee K Wong, and Ping Luo. Rigid: Recurrent gan inversion and editing of real face videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13691–13701, 2023.

Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *European Conference on Computer Vision*, pp. 357–374. Springer, 2022.

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023.

Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13789–13798, 2021.

Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14805–14814, 2023a.

Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023b.

Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, pp. 74–90. Springer, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023a.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations*, 2023b.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18697–18709, 2022.

Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023.

A VISUAL RESULTS

Figs. 7 and 8 show visual comparisons between our model and other video editing and inpainting methods. Other methods either do not remove the glasses from the input, or generate deformations and artifacts around the eyes. In contrast, our method removes the glasses successfully. Video results and more comparisons are presented in the attached video.

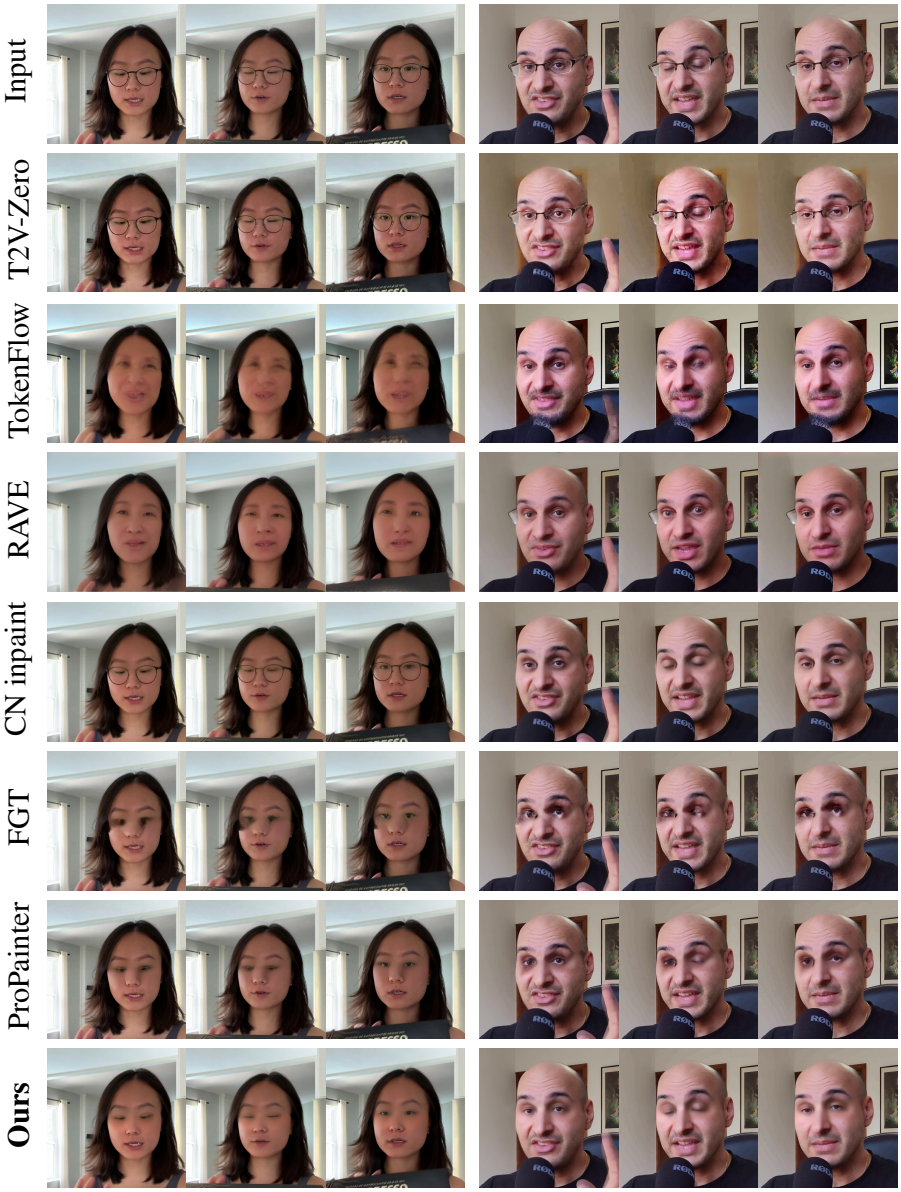

Figure 7: **Visual comparisons:** We compare our results to different video editing and inpainting methods. Other methods often struggle with glasses-removal, and even when they do remove the glasses, they tend to leave glasses remnants (e.g. RAVE right example), generate artifacts (e.g. FGT, ProPainter examples, TokenFlow left example), do not preserve the identity of the person (e.g. RAVE Left example), or their eyelids position (e.g. RAVE both examples).

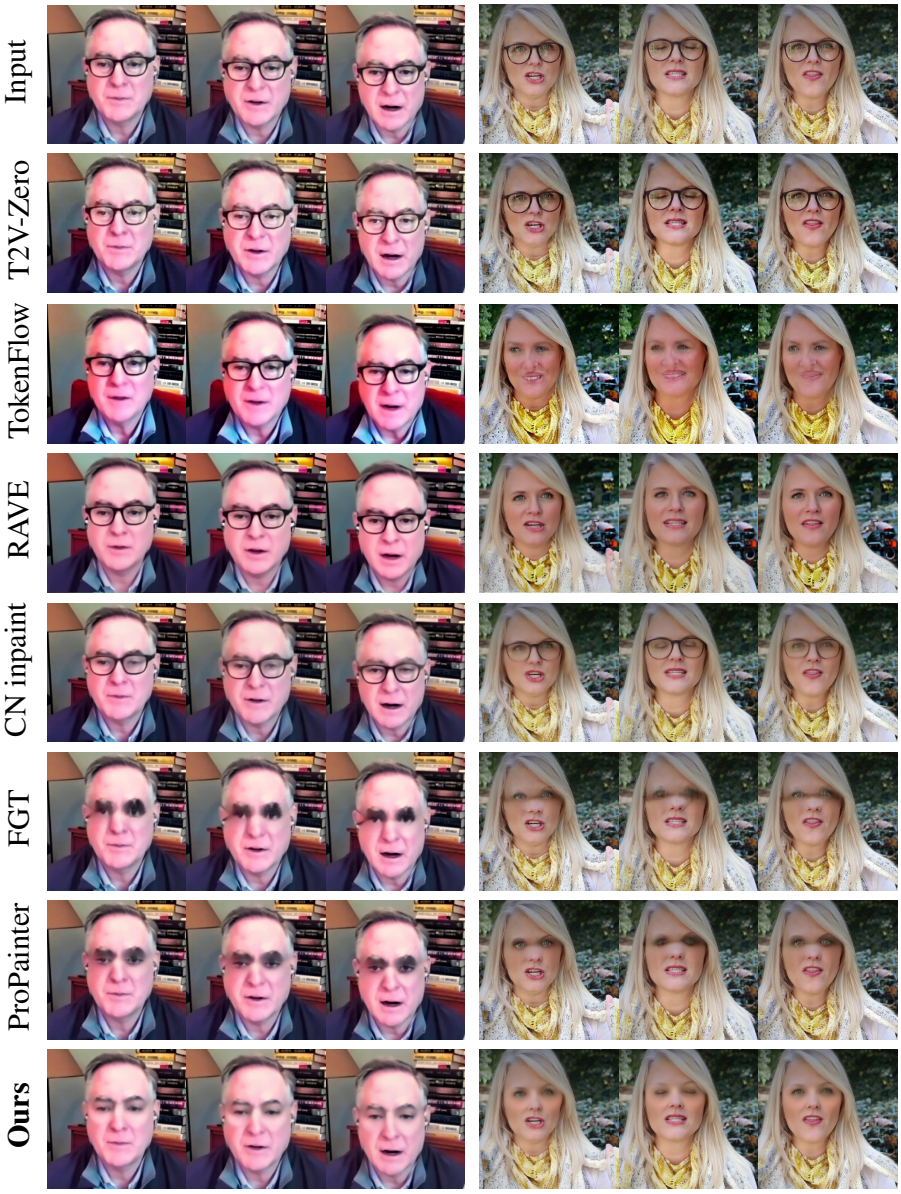

Figure 8: **Visual comparison:** We compare our results to different video editing and inpainting methods. Other methods often struggle with glasses-removal, and even when they do remove the glasses, they tend to either change the identity completely (e.g. TokenFlow right example), generate artifacts such as black areas around the eyes (e.g. FGT, ProPainter both examples), or do not preserve the eyelids position (e.g. TokenFlow, RAVE right example)

.

Figure 9: **Stickers results.** These examples show the generalization ability of our method. We apply it to removing facial stickers from videos, and show our model successfully removes different stickers from different locations over the face.

B Generalization Experiment

In this work, we focus on glasses as a case study. We chose glasses as they are a particularly complex task, as explained in Sec. 1. However, our method can be generalized to other types of local video editing. To show our method works for a different use-case, we generate a synthetic dataset of videos by applying stickers from the Stickers dataset Nguyen et al. (2021) on faces from the CelebV-Text Yu et al. (2023a) dataset, imitating real facial stickers, tattoos or synthetic features added by social-media apps, and show that our method is able to remove them. For this task, we detect the face in each frame, and estimate its 3D shape using a trained neural network that predicts the coefficients of the Basel 3D face model Gerig et al. (2018). Then, we project and render the sticker texture on the original frame. For each generated clip, we sample one sticker and place it at a random position in the uv coordinate space, which is mapped to the surface of the face. We generate 1274 clips this way, each 40-frames long. After the dataset is created, we train an image-to-image model over the stickers dataset and run our video editing pipeline using this model over the test videos. Some results are presented in Fig. 9. As shown, our model is able to seamlessly remove stickers from videos.

C Implementation Details

C.1 Data Generation & Training Specifics

For the glasses masks generation, we use Facer Zheng et al. (2022) with `retinaface/mobilenet` as a face detector, `farl/celebm/448` as a face parser, and `farl/ibug300w/448` as a face aligner. We first find the glasses mask using the parser, and then make eye holes in it based on the eye landmarks found by the face aligner. To generate the holes, for each eye, we connect the eye landmarks into one connected component, dilate it with a (10,10) kernel, so that we keep enough of the eyes information, and remove the component from the mask. Additionally, after resizing the mask to match the size of the latent vectors, we dilate the mask with a kernel of (3,3), and then blur it with a (3,3) kernel as well, to make sure we include all glasses pixels, and not too much from the rest of the image. For the inpainting process, we use CN inpaint Zhang et al. (2023a) as our model, with latent blending with a blending ratio of 0.9. We use 2 reference frames for the cross-frame attention: the first and middle frames in each video, since usually the person moves and changes position throughout the video so more information is gathered this way. If more memory is available, more reference frames will probably give better results. Moreover, for consistency we also use the same noise (encoding of the first reference frame) for all the video. We use CN Tile Zhang et al. (2023a) as the model we finetune over our dataset, with batch size 8, learning rate $1\mathrm{e}{-5}$. To avoid learning the artifacts of our imperfect data, and avoid forgetting the prior knowledge of CN Tile, we stop the training at an early stage, as suggested by DVP Lei et al. (2020).

C.2 Editing Pipeline Specifics

We use the pretrained motion module of AnimateDiff Guo et al. (2024) with context length 16, context overlap 4 as our motion prior module. As shown in the right example of the limitations figure of the main paper, the motion module tends to smooth the frames to get a more temporally

consistent result, hence sometimes the results using this module get blurry. To avoid blurriness, we use only some of the motion layers and not all of them. Specifically, we remove the first 5 output motion layers. This way, we get a more realistic and less blurry video. For blending we use gradual values between 0 and 0.7 as mask values, so that pixels outside of the mask can also change a little bit for a smoother result.

C.3 Original Colors vs. Glasses-Removal Trade-off

As an option in our video editing pipeline, we blend original pixels back into areas outside the masked region. If we do not blend them back, the colors in the edited video may not match those of the original video exactly. However, as the masks are not perfect, glasses remnants are sometimes left within the non-masked region, causing glasses remnants to appear in the result when using masks. To avoid it, we use dilation and different blending values outside of the mask. If the original colors are not as important and glasses-removal is of higher priority, one can use higher dilation for the mask, lower mask blending values, or even not use the masks at all.

C.4 General

We use Stable Diffusion 1.5 Rombach et al. (2022) as our backbone in all the method steps.

D Limitations

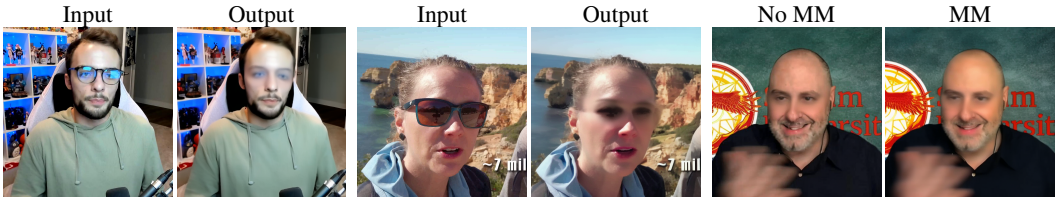

Figure 10: **Limitations.** Left: strong reflections. Middle: Dark sunglasses. In these cases, as the eyes are not exposed, the model cannot guess the right eye color. Moreover, our data was not clean and contained sunglasses videos which were treated the same way as eye glasses in the data generation process. Hence, the model generates dark areas instead of eyes for dark sunglasses videos. Right: Eye blurriness that comes from the motion module.

Although our method removes glasses reflections in many cases, it struggles with strong reflections, as in the left example in Fig. 10, where the true eyes are never visible. Additionally, our method struggles with dark sunglasses, as in the middle example in Fig. 10, where identifying eye color is difficult. Moreover, as current glasses detectors do not differ between eye and sun glasses, our dataset was not clean and contained sunglasses videos as well. As we used the same kind of masks for all our data, if the input is a video with sunglasses, our model often generates dark areas where the eyes should be, as is shown the middle example in Fig. 10. This could be solved by cleaning the data and only training over eye-glasses videos. Finally, the motion module tends to smooth the frames to get a more temporally consistent result. As a result, our outputs also tend to be a bit blurry, as shown in the right example in Fig. 10. To reduce the blurriness to a minimum, we use only some of the motion layers. See supplemental material for details.

E Related Work

E.1 Image editing

Since the advent of diffusion models for image generation and editing Ho et al. (2020), real image editing has rapidly advanced and showed remarkable results Brooks et al. (2023); Couairon et al. (2023); Tsaban & Passos (2023); Hertz et al. (2022); Tumanyan et al. (2023); Zhang et al. (2023a); Meng et al. (2021), including local editing such as glasses-removal. Image inpainting methods Nitzan et al. (2024); Rombach et al. (2022); Zhang et al. (2023a); Yildirim et al. (2023), also showed

impressive results, realistically replacing content behind a given mask. Moreover, prior work Lyu et al. (2022); Lee & Lai (2020) tackled the task of image glasses-removal directly. These works perform well over images. However, as shown in Fig. 2 they fail when applied to video frames of people, that do not always look directly at the camera, and constantly move, causing motion blur and other artifacts. Moreover, when applied frame-by-frame, the generated content differs between frames, resulting in temporal inconsistencies. Prior to diffusion models, Generative Adversarial Networks (GANs) have been widely used for facial image editing (Karras et al., 2019; 2020; Yao et al., 2021; Patashnik et al., 2021; Richardson et al., 2021). However, In addition to temporal consistency issues, they operate on aligned crops of faces, thus when applied to videos, faces must be stitched back to the original frames which is challenging and often yields unsatisfactory results.

E.2 Video editing

Recently, video editing has developed greatly with video-to-video translation methods Geyer et al. (2023); Khachatryan et al. (2023); Cong et al. (2024); Kara et al. (2023); Qi et al. (2023); Wu et al. (2023); Zhang et al. (2023b) that focus on transforming the entire frame into a different style, while trying to preserve temporal consistency in the generated videos. Some of them, such as RAVE Kara et al. (2023), or TokenFlow Geyer et al. (2023) with SDEdit Meng et al. (2021) also perform local editing. However, in the case of glasses-removal, where a person moves across the video and the requested edit is small and delicate, many artifacts such as face deformations and inconsistencies are generated. Examples are presented in Fig. 7 and in video results in the supplementary material. Earlier work used GANs to perform local editing of faces (Tzaban et al., 2022; Xu et al., 2022; 2023). Another line of works is atlas-based video editing (Kasten et al., 2021; Lu et al., 2021; Bar-Tal et al., 2022; Suhail et al., 2023). These methods learn an atlas for each video, to apply changes to either the background or foreground of the video. The initial atlas reconstruction requires excessive computational resources and long running times. Moreover, as these methods were designed to change the appearance of existing objects, they are not meant for adding or removing attributes that were not part of the original video, such as glasses, and do not perform these changes well.

E.3 Video inpainting

Video inpainting is the task of consistently filling-in new content behind a given mask throughout a video (Yu et al., 2023b; Zhang et al., 2022; Zhou et al., 2023; Gu et al., 2023; Li et al., 2022b). Such methods work well when the object moves throughout the video, and the model is able to fill-in the background from other video frames, where the background is visible. However, they struggle with filling-in completely occluded areas, especially when they are part of the foreground, such as faces. Particularly, areas behind glasses are only partially revealed by head motion, and sometimes glasses cause optical deformations that hide the background. Slight changes to face features in the "background" of glasses may yield unrealistic results, or alter face identity, as shown in the ProPainter Zhou et al. (2023) and FGT Zhang et al. (2022) results in Fig. 7. While such changes to smooth backgrounds may not affect the quality of the results significantly, in foreground areas, particularly in faces, they are detrimental.

E.4 Learning from synthesized data

Deep learning models require large amount of data to learn from, however paired data is not always available for the task at hand. Hence, several works Li et al. (2022a); Peebles et al. (2022); Ravuri & Vinyals (2019); Brooks et al. (2023); Lyu et al. (2022) used generative models for data generation to train models. These works usually rely on the high quality of the generated data, and learn solely from it. In this work, we accept that generated data is usually imperfect, and is insufficient to achieve the required result on its own, thus we take advantage of strong priors of pretrained models and use our generated data for fine-tuning. This way, the model learns to generate results that are superior to the data it was trained on, while performing the relevant task. Lyu et al. (2022) specifically explores synthetic data for the task of glasses-removal from images. However, their data acquisition process requires face scanning and 3D data, which is hard to acquire, unlike our method which does not require any special equipment or effort.