V-LASIK: Consistent Glasses-Removal from Videos Using Synthetic Data

Anonymous Author(s)

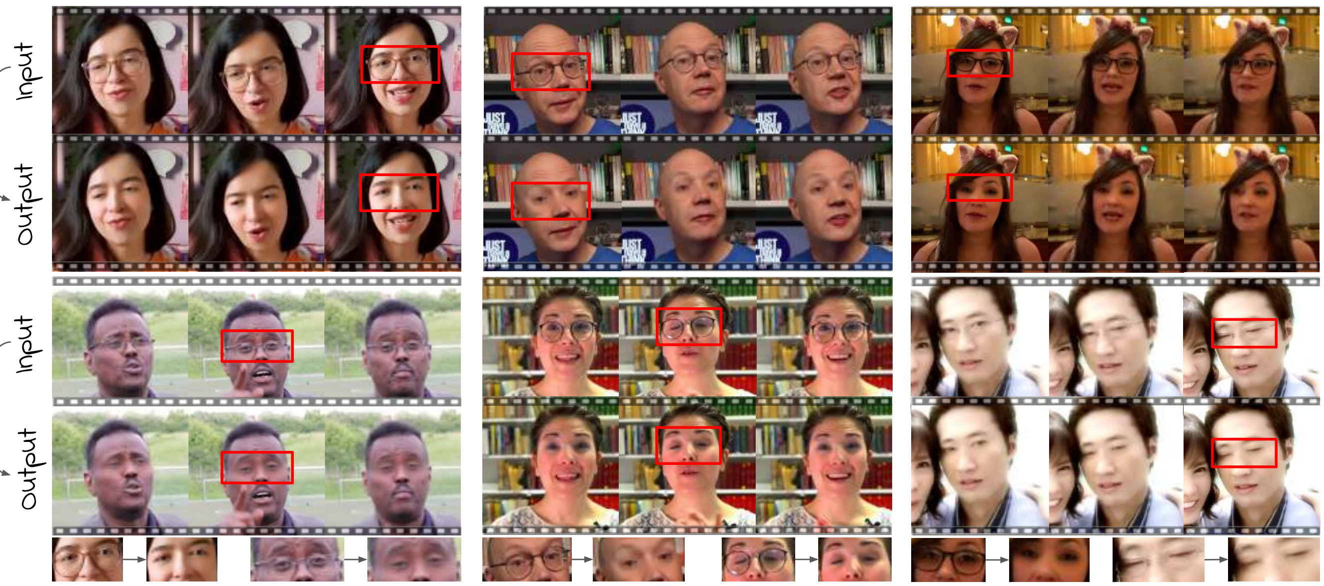

Fig. 1. Our method receives an input video of a person wearing glasses, and consistently removes the glasses, while preserving the original content and identity of that person. Our method successfully removes the glasses even when there are reflections (bottom-middle example), heavy makeup (top-right),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

Diffusion-based generative models have recently shown 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 throughout the video.

In this work, we focus on consistent and identity-preserving removal of glasses in videos, using it as a case study for consistent local attribute removal in 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 diffusion models, our model is able to perform the desired edit consistently while preserving the original video content. Furthermore, we exemplify the generalization ability of our method to other local video editing tasks by applying it successfully to facial sticker-removal. Our approach demonstrates significant improvement over existing methods, showcasing the potential of leveraging synthetic data and strong video 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 [Bar-Tal et al. 2024; Geyer et al. 2023; Kara et al. 2023; Rombach et al. 2022; Yang et al. 2023; Zhang et al. 2023a,b]. While local editing, such as removing and adding attributes without changing the rest of the content, mainly works for images, local video editing 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, so the quality bar is higher. Existing image editing and inpainting methods [Brooks et al. 2023; Couairon et al. 2023; Meng et al. 2021; Nitzan et al. 2024; Tsaban and Passos 2023; Zhang et al. 2023a] are trained on clean images, that do not contain the aforementioned problems, and therefore do not perform well over video frames; They typically either change the original frames too much, or do not perform the requested change completely, as illustrated in Fig. 2.

Video inpainting methods [Li et al. 2022b; Yu et al. 2023a; Zhang et al. 2022; Zhou et al. 2023] typically remove objects by deleting and filling regions of the video, often disregarding the information originally present in those regions. When it comes to generating 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, 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 across frames). 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] and FGT [Zhang et al. 2022] results in Fig. 5).

Mask-free video editing methods [Cong et al. 2024; Geyer et al. 2023; Kara et al. 2023; Khachatryan et al. 2023; Zhang et al. 2023b] mainly focus on changing the style of the whole video, or changing the appearance of large parts of it, and not on local changes that should adhere to the original context. Other works [Bar-Tal et al. 2022; Kasten et al. 2021] learn an atlas for each new video, leading to long inference time. Furthermore, as mentioned in Text2Live [Bar-Tal et al. 2022], the goal of these works is to manipulate the appearance of existing objects, not to change the composition of the elements in the scene. Hence, their method is not capable of successfully removing attributes such as glasses from faces.

In this work, we focus on removing glasses from videos as a case study for local video editing. We introduce a new method that performs this task consistently while preserving the original identity. 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, allowing the model to use 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 eyelid position and other features from the original frame. To overcome these challenges, we finetune a diffusion based image-to-image model over this data. This way we can leverage the prior of the pretrained model, while enhancing its capabilities with our new data. Additionally, the non-masked input allows the model to exploit the fact that it is exposed to useful information within the masked region to achieve the requested edit, while preserving the rest of the content from the original frames. Despite the imperfections in our synthetic data, a model with a strong prior can leverage its prior knowledge to outperform the training data and produce high-quality results. Finally, to get temporally consistent results, we combine our trained model with a motion prior, and 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. We release code and both the stickers and the glasses removal datasets for further research.

2 RELATED WORK

2.1 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; Hertz et al. 2022; Meng et al. 2021; Tsaban and Passos 2023; Tumanyan et al. 2023; Zhang et al. 2023a], including local editing such as removing glasses from images. Moreover, image inpainting methods [Nitzan et al. 2024; Rombach et al. 2022; Yildirim et al. 2023; Zhang et al. 2023a], have also shown impressive results, realistically replacing the content behind a given mask. Additionally, prior work [Lee and Lai 2020; Lyu et al. 2022] tried to tackle the task of removing glasses from images directly. Indeed, 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.

Fig. 2. **Glasses-removal from a blinking eye by image editing methods** Left to right: LEDITS [2023], Instruct pix2pix [2023], Lyu et al. [2022], Stable Diffusion inpaint [2022], ControlNet inpaint [2023a], our synthetic dataset generation result, and our final result. As image editing methods expect high quality images with people looking straight to the camera, they struggle when these constraints are not met. In our dataset result 'Synth data', as a result of the cross-frame attention, eye artifacts appear. However, our model is still able to learn from the imperfect data and remove the glasses better than any out-of-the-box method, and better than the data it was trained on.

2.2 Video editing

Recently, video editing has also developed greatly with video-to-video translation methods [Cong et al. 2024; Geyer et al. 2023; Kara et al. 2023; Khachatryan 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. 5 and in video results in the supplementary material. Another line of works is atlas-based video editing [Bar-Tal et al. 2022; Kasten et al. 2021; Lu et al. 2021; 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. Additionally, while theoretically allowing to remove layers from a video, these methods are limited to relatively simple motion and to static backgrounds.

2.3 Video inpainting

Video inpainting is the task of consistently filling-in new content behind a given mask throughout a video [Gu et al. 2023; Li et al. 2022b; Yu et al. 2023a; Zhang et al. 2022; Zhou et al. 2023]. Such methods

Fig. 3. **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).

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 the face identity, as shown in the ProPainter [Zhou et al. 2023] and FGT [Zhang et al. 2022] results in Fig. 5. While such changes to smooth backgrounds may not significantly affect the perceptual quality of the results, in foreground areas and particularly in faces they are detrimental.

2.4 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 [Brooks et al. 2023; Li et al. 2022a; Lyu et al. 2022; Peebles et al. 2022; Ravuri and Vinyals 2019] use 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 the 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. [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.

3 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. 4, our approach to removing glasses from videos 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, we fine-tune an image-to-image diffusion model over our synthetic dataset, 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. Finally, we incorporate our trained model with a pretrained motion module, into a video editing pipeline, to obtain temporally consistent videos without glasses.

3.1 Paired Data Generation

As illustrated in Fig. 2, current image editing and inpainting methods do not perform well on the task of removing glasses 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. 2023b] 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]:

$$\mathrm{softmax}\left(\frac{Q \cdot [K_0, ..., K_k]^T}{\sqrt{d}}\right) \cdot [V_1, ..., V_k] \quad (1)$$

where each query frame (Q) attends to k different values (V_i), corresponding to k different reference frames in our case. 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. 3. 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. 3. 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.

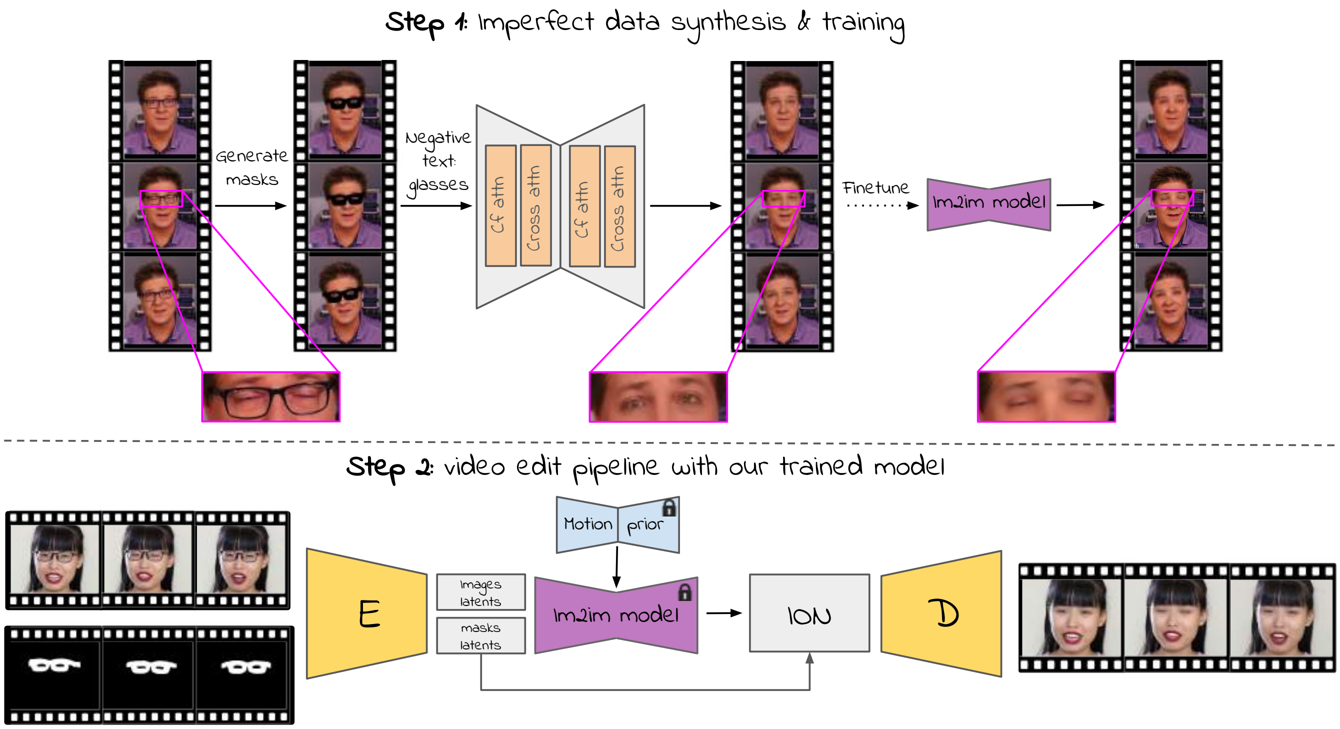

Fig. 4. **Method overview: Step 1:** we create an imperfect synthetic paired dataset by generating glasses masks for each video frame and inpainting it. We inpaint each frame using an adjusted version ControlNet inpaint [Zhang et al. 2023a]. We replace the self-attention layers with cross-frame attention (cf attn) and use blending between the generated latent images and the noised masked original latent images at each diffusion step. The generated data in the first step is imperfect; e.g. in the middle frame, the person blinks, however its generated pair has open eyes. Nevertheless, the data is good enough for finetuning an image-to-image diffusion model and achieving satisfactory results, due to the strong prior of the model. **Step 2:** Given our trained model for the task of removing glasses from images, we incorporate it with a motion prior module to generate temporally consistent videos without glasses from previously unseen videos. To obtain the original frame colors, at each diffusion step we blend the generated frames with the noised original masked latent images, and before decoding, we apply an Inside-Out Normalization (ION), to better align the statistics within the masked area and the area outside of the mask.

3.2 Model fine tuning

After the data generation stage, we obtain pairs of frames with and without glasses. However, our generated frames are not temporally consistent, and also contain per-frame artifacts in many cases, as featured in Figs. 2 and 4. For example, the eyelid positions, i.e. closed or open, sometimes change after the inpainting process. Despite that, the generated data is good enough for finetuning a pretrained image-to-image diffusion model for the task of removing glasses 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.

3.3 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, this model's results often exhibit slightly different colors from the original input. To overcome this issue, we can use the masks from the data generation stage to retrieve the original values on 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 and Belongie 2017], we apply a new normalization function, we dub Inside-Out Normalization (ION), where we aim to align the statistics of the masked area with those of the non-masked area. 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 we normalize the values of the latent features in the area inside the mask by calculating:

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

ION allows a smooth transition between the areas inside and outside of the mask by moving the statistics of the latent masked area toward those of the non-masked area.

4 IMPLEMENTATION DETAILS

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 $1e-5$. 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].

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 Fig. 9, 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.

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.

General. We use Stable Diffusion 1.5 [?] as our backbone in all the method steps, as at the time of development there was no compatible motion module for SDXL.

5 EXPERIMENTS

We evaluate our results both qualitatively and quantitatively, testing three aspects:
1. **Fidelity** to the required edit, i.e glasses-removal.
2. **Identity and content preservation** of the original video. We want to remove the glasses while leaving the rest of the video intact.
3. **Realism** of the result, by means of temporal consistency and realism 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 our model 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 some part of the video. Additionally, we compare our results to the results of 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 enough on this task, even when combined in our video editing pipeline. Our quantitative and qualitative evaluations, including a comprehensive user study, demonstrate that our results are favored over all other methods across all tested aspects.

5.1 Experimental details

We generate our data pairs as described in Sec. 3.1 from the dataset CelebV-Text [Yu et al. 2023b]. We train our model over 1296 of those videos, and test over 144 unseen videos.

5.2 Qualitative evaluation

Visual results of our method are presented in Figs. 5 and 6 and in the supplementary material. As illustrated in the figures, existing video editing and inpainting methods struggle with performing the required local edit, i.e., removing the glasses of the 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.

We perform a thorough user study to compare our results with the results of TokenFlow, RAVE, ProPainter, and our editing pipeline with CN inpaint. In the user study, we do not compare our results to those of Text2Video-Zero, as it only removes the glasses from the input video in 4% of the test videos, as featured in Fig. 5.

The user study tests three aspects: glasses-removal, resemblance to the identity in the original video, and realism of the result. For each video in the survey, we ask the users in which videos are the glasses removed. If both models removed the glasses for that video, we additionally ask 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 remnants of glasses, because we noted that even when other models remove the

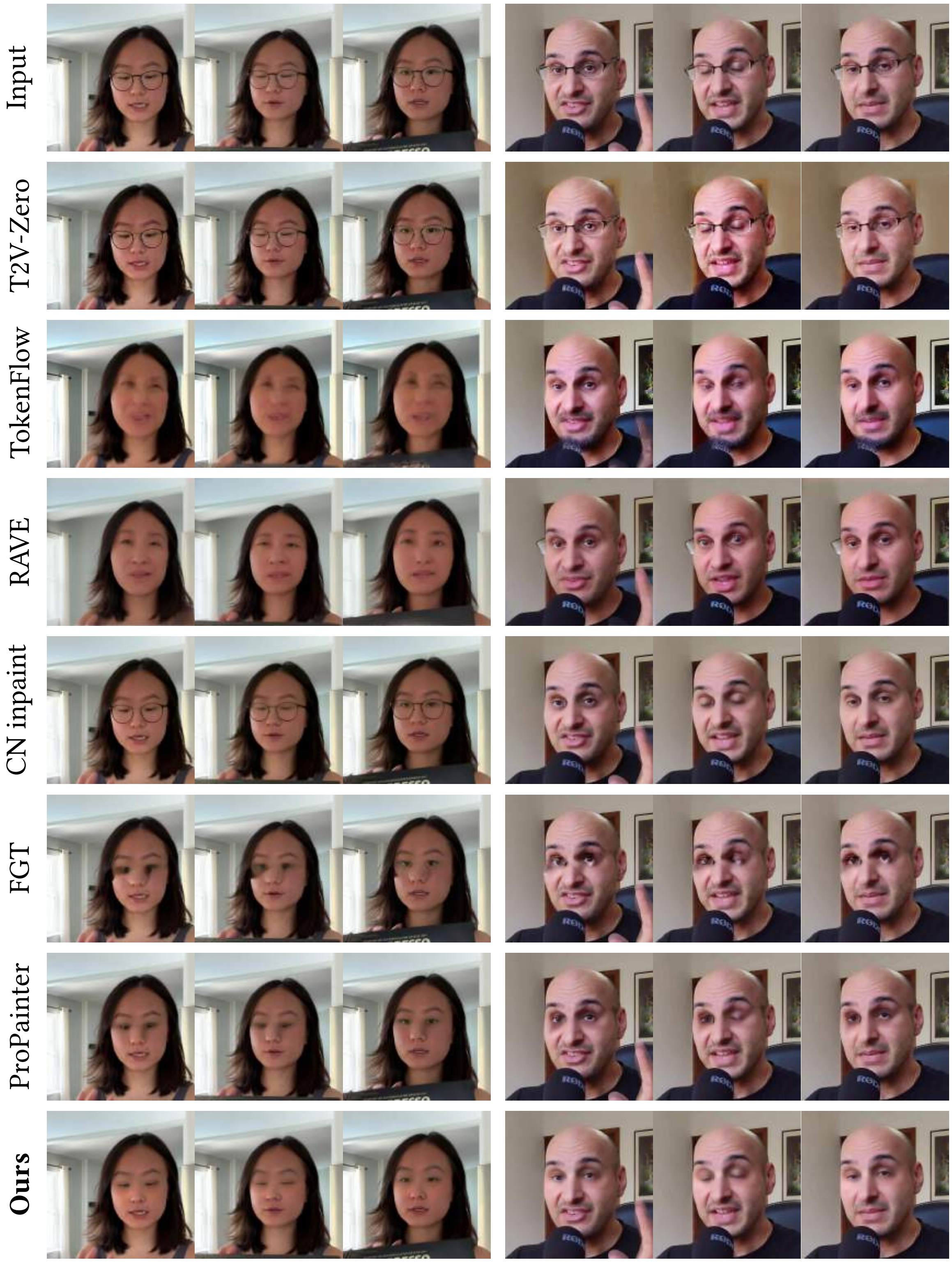

Fig. 5. **Visual comparisons:** We compare our results to different video editing and inpainting methods. Other methods often struggle with glasses-removal, and even when they do remove the glasses, they tend to leave glasses remnants (e.g. RAVE right example), generate artifacts (e.g. FGT, ProPainter 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).

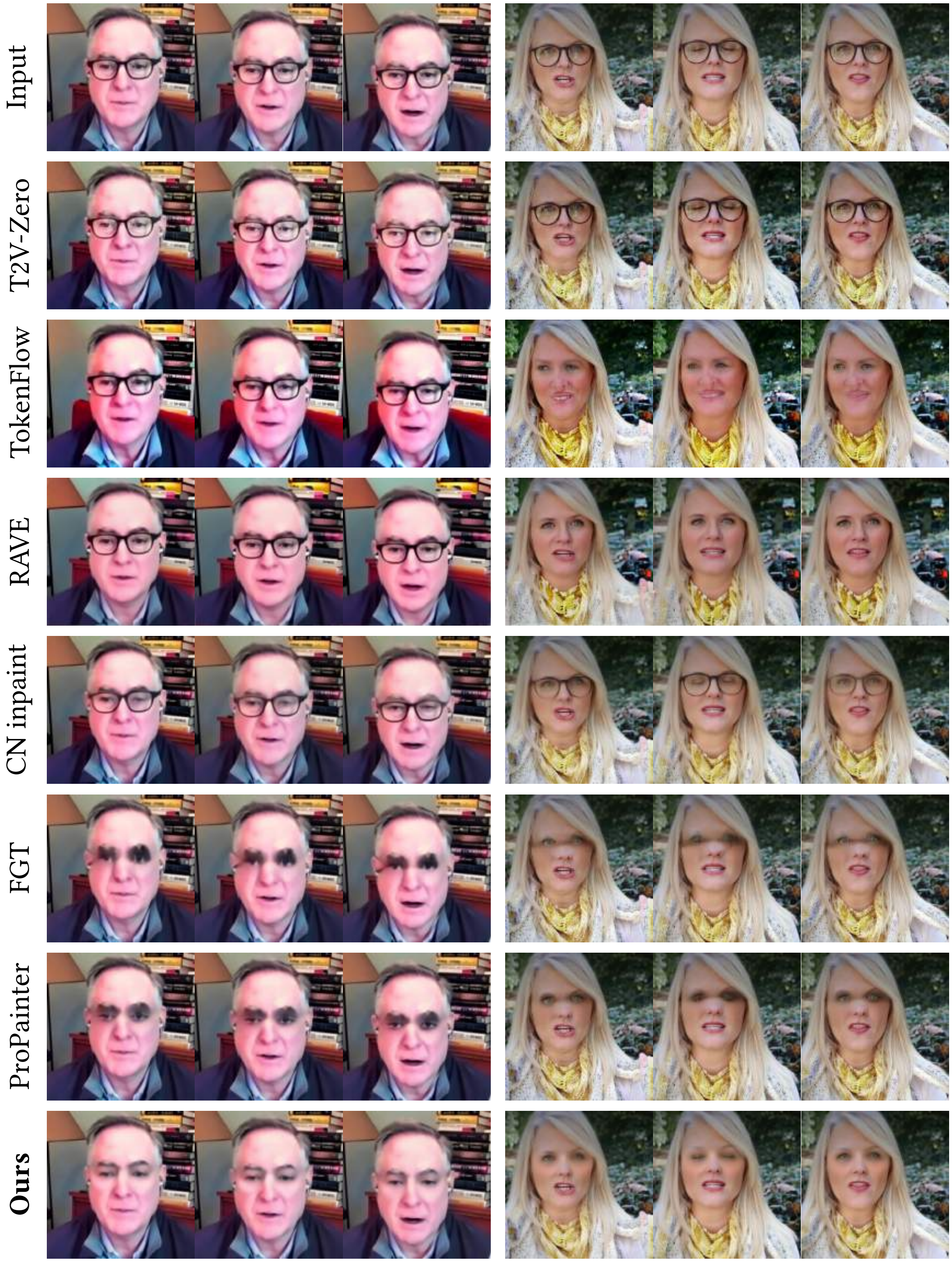

Fig. 6. **Visual comparison:** We compare our results to different video editing and inpainting methods. Other methods often struggle with glasses-removal, and even when they do remove the glasses, they tend to either change the identity completely (e.g. TokenFlow right example), generate artifacts such 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)

.

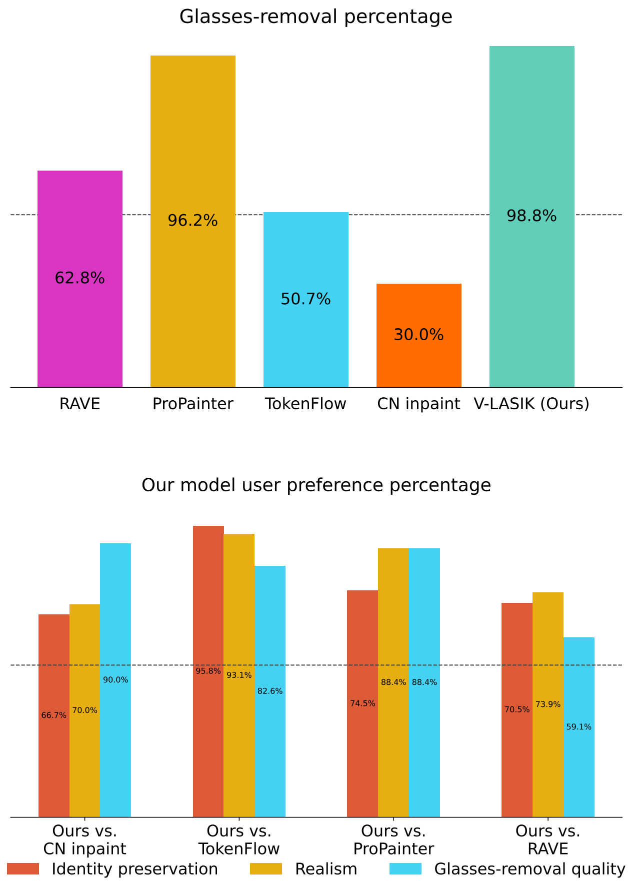

Fig. 7. **User study results:** First, we ask users which models remove the glasses from the input video (top). Then, when both models remove the glasses, we ask which one better preserves the identity of the person, which one is more realistic, and which result has less remnants of glasses (bottom). Compared to all model examined, the users preferred our results through all measured aspects.

glasses from the video, they often leave parts of the glasses or their reflections, e.g. as presented in the results of RAVE in Fig. 5. The user study contained side-by-side video pairs of our results vs. the results of different models for the same input (11 per model, 44 pairs in total), and it was answered by 57 users. The results of our user study are presented in Fig. 7. The top table shows the percentage of videos for which the users thought each model removed the glasses from the input video. The users thought our model removed the glasses in 98.9% of the cases, more than any other model. The bottom table shows the percentage of users that preferred our results over the tested models. Fig. 7 shows that the users preferred our results over all other models, in all the measured aspects: identity preservation, realism, and quality of glasses-removal.

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. 5.3. We test the fidelity of the results by checking ΔG, the average difference between glasses pixels in the original video frames and the edited one. We test the identity preservation of the edited video using ID score, and the tradeoff between them using $ID \cdot \Delta G$. Moreover, we test the temporal consistency of the generated videos using the optical flow warp error E_{warp}.

Method	$\Delta G_{(\times.01)} \uparrow$	$ID_{(\times.1)} \uparrow$	$ID \cdot \Delta G_{(\times.01)} \uparrow$	$E_{warp(\times 10^{-4})} \downarrow$
ProPainter [2023]	**4.0**	7.5	**3.0**	3.9
FGT [2022]	**4.0**	7.1	2.8	4.0
CN inpaint* [2023a]	3.0	8.0	2.3	4.1
TokenFlow[2023]	2.0	6.9	1.4	5.7
RAVE [2023]	2.0	7.4	1.5	6.8
T2V-Zero [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

5.3 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. As mentioned in Sec. 4 both options exist for our model and there is a tradeoff 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. These differences are small, hence as shown in Tab. 1, they only slightly affect the quantitative results.

Fidelity: to test the fidelity of our results, we measure the average difference between the number of pixels with glasses in the original videos vs. the edited ones. To find the pixels that contain glasses in an image, we use 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 ProPainter [Zhou et al. 2023] and FGT [Zhang et al. 2022]. However, as illustrated in Figs. 5 and 6 and by the results of our user study in Fig. 7, 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: to test 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 [Khachatryan et al. 2023] does not remove the glasses from most videos, and does not change the videos by much, it achieves a very high ID score. For this reason, we follow prior work [Kara et al. 2023; ?] 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. 7 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: to evaluate temporal consistency, we follow previous work [Geyer et al. 2023; Lai et al. 2018; Lei et al. 2020; Zhou et al. 2023] and calculate the warp error (E_{warp} in Tab. 1). We use RAFT [Teed and Deng 2020] to calculate optical flow 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, the non-masked version of our model produces the most temporally consistent results, compared to all other methods.

5.4 Generalization

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.

Fig. 8. **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.

5.4.1 Stickers. 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. 2023b] 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 the same 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. 8. As shown, our model is able to seamlessly remove the stickers from the videos.

6 LIMITATIONS

Although our method removes glasses reflections in many cases, it struggles with strong reflections, as in the left example in Fig. 9, where the true eyes are never visible. Additionally, our method struggles with dark sunglasses, as in the middle example in Fig. 9, 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. 9. 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. 9. As discussed in Sec. 4, by using only some of the motion layers, we reduce the blurriness to a minimum.

Fig. 9. **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.

7 SOCIETAL IMPACT

While the goal of this work is to allow editing of owned or licensed videos only, we acknowledge the fact it may be misused to altering videos without consent, and contribute to the spread of misinformation. We condemn such usage, and we are actively working on systems that detect synthetic and edited media.

8 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 individual's identity. 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. We hypothesize it would work similarly for any other local attribute, and we encourage future work to pursue this direction.

REFERENCES

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

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

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

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

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

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2023. DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*.

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

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

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *The Twelfth International Conference on Learning Representations*.

Bohai Gu, Yongsheng Yu, Heng Fan, and Libo Zhang. 2023. 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. 2024. 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. 2022. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

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

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

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. 2023. RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models. *arXiv preprint arXiv:2312.04524* (2023).

Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–12.

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

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

Yu-Hui Lee and Shang-Hong Lai. 2020. 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*. Springer, 243–258.

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

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

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

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

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

Thao Nguyen, Anh Tran, and Minh Hoai. 2021. 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)*.

Yotam Nitzan, Zongze Wu, Richard Zhang, Eli Shechtman, Daniel Cohen-Or, Taesung Park, and Michaël Gharbi. 2024. Lazy Diffusion Transformer for Interactive Image Editing. *arXiv preprint arXiv:2404.12382* (2024).

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

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

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

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

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

Mohammed Suhail, Erika Lu, Zhengqi Li, Noah Snavely, Leonid Sigal, and Forrester Cole. 2023. Omnimatte3D: Associating Objects and Their Effects in Unconstrained Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 630–639.

Zachary Teed and Jia Deng. 2020. 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*. Springer, 402–419.

Linoy Tsaban and Apolinário Passos. 2023. 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. 2023. 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*. 1921–1930.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. 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*. 7623–7633.

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.

Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. 2023. 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. 2023b. CelebV-Text: A Large-Scale Facial Text-Video Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14805–14814.

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

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

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding Conditional Control to Text-to-Image Diffusion Models.

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

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

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