FreeEyeglass: Training-free and Target-mask-free Eyeglass Transfer for Facial Videos

Anonymous authors
Paper under double-blind review

Abstract

The rise of e-commerce and short-video platforms has fueled demand for realistic video-based virtual try-on. Unlike virtual try-on of clothing, which has been actively studied to date, virtual try-on of eyeglasses is uniquely challenging: they physically interact with facial geometry and strongly affect facial identity, making the faithful preservation of unedited regions especially important. Existing generative editing approaches, such as GAN- and diffusion-based methods, lack reconstruction objectives and often rely on inpainting, which fails to ensure identity consistency. We argue that semantic editing requires not only plausible generation but also faithful reconstruction, making autoencoder-based latent spaces particularly suitable. We introduce a training-free, reference-guided framework for video eyeglass transfer built on Diffusion Autoencoders (DiffAE). By blending semantic features in the encoder and incorporating spatial-temporal self-attention, our method achieves realistic, identity-preserving, and temporally consistent results, and points to the potential of autoencoder-based latent spaces for local video editing. Our implementations and datasets will be released upon acceptance.

1 Introduction

The rapid growth of e-commerce and short-video platforms has created a strong demand for realistic video-based virtual try-on systems. While most existing research has centered on clothing in the video setting (Jiang et al., 2022; Xu et al., 2024; Fang et al., 2024; Wang et al., 2024b; Nguyen et al., 2025), work on other wearable objects has been explored only in the image domain (Miao et al., 2025; Feng et al., 2025). Among them, eyeglasses stand out as a particularly important category. They have long been treated as a standalone research topic in vision and graphics, with prior work on try-on (Zhang et al., 2017; Li et al., 2023), detection (Wu et al., 2002; Bekhet & Alahmer, 2021), removal (Lyu et al., 2022; Zhang & Guo, 2025), and product design (Bai et al., 2021; Plesh et al., 2023). Compared to clothing and other accessories, eyeglasses pose uniquely complex challenges. Geometrically, they must fit precisely on the nose and ears rather than being simply overlaid on the face. They also overlap directly with the eyes and eyebrows, regions known to be critical for facial identity perception (Schyns et al., 2002; Tanaka & Simonyi, 2016). Realistic video eyeglass transfer, therefore, requires not only rendering the glasses plausibly but also preserving identity and maintaining temporal coherence under motion and viewpoint changes. Despite this practical importance, the problem has not been systematically studied to date.

Despite rapid advances in semantic image and video editing, the generative models that are widely adopted, GANs (Karras et al., 2019; 2020b;a) and text-conditional diffusion models (Rombach et al., 2022), are not inherently well-suited for local editing tasks such as object transfer. Trained primarily for generation, these models typically lack reconstruction ability, making them prone to altering unedited regions when adapted to editing. To enforce more localized changes, many object transfer methods (Yang et al., 2023a; Chen et al., 2024b;a; Song et al., 2024; Jiang et al., 2025) train or finetune these models with an inpainting objective, where masked regions in target images are filled with reference content. While this improves locality, they still cannot handle the unique challenges of eyeglass transfer. Target masks inevitably discard identity-related content and spatial information around the eyes, leading to artifacts, poor harmonization, and identity inconsistency. These issues are further amplified in videos, where temporal coherence must also be maintained, and the results often resemble simply copying the reference into the target region.

Figure 1: **FreeEyeglass** is a training-free reference-based video editing method for transferring reference eyeglasses to target facial videos semantically while achieving harmonious local editing and temporal consistency. Given a reference facial image specified with eyeglass position, our method does not require any frame-by-frame accurate masks on the target videos for transferring eyeglasses, which has been required in state-of-the-art inpainting-based editing methods.

These shortcomings indicate that eyeglass transfer requires more than inpainting-based generation. What is needed is an approach that can edit locally while preserving critical identity information. Generative models that lack reconstruction objectives struggle to provide this guarantee. Autoencoders, in contrast, are trained with explicit reconstruction objectives, equipping them with a strong capability to retain unedited content. Diffusion autoencoders (DiffAE) (Preechakul et al., 2022) further show that their latent spaces are compact and semantically structured, enabling natural local edits with minimal distortion. Through extensive experimentation, we observe that DiffAE's semantic latent space not only preserves appearance but also semantically aligns geometry and position. Even with a coarse, face-aligned reference mask without any explicit geometric modeling, DiffAE naturally adapts the reference eyeglasses to the target face, correcting their placement and visible geometry across frames. This observation suggests that the DiffAE's semantic latent space, despite not being trained for any generative editing objectives may in fact be better suited for eyeglass transfer in video, where both identity preservation and local realism are critical.

Building on this insight, we leverage DiffAE as a backbone and introduce a training-free, target-mask-free framework for reference-based eyeglass transfer in video, as shown in Fig. 1. While DiffAE has been explored with text or classifier guidance (Kim et al., 2023), the approach to achieving reference-based editing with this model has not been studied. We address this by designing a feature-blending mechanism in the semantic encoder that combines reference and target features to construct new semantic features for each frame. These features are used in DDIM inversion with noise blending to guide the rendering of eyeglasses on the target face. Furthermore, we extend this framework to video by incorporating spatial-temporal self-attention focused on the editing region to enhance temporal consistency. Our method achieves realistic eyeglass transfer that preserves identity and adapts to pose changes. It demonstrates the potential of compact autoencoder latents for local video editing.

Our contributions are summarized as follows:

- We present the first *training-free* framework that requires *no target-video masks* for **reference-based video eyeglass transfer** to tackle a practically important yet underexplored problem in virtual try-on.
- We show how to adapt diffusion autoencoders for reference-based editing by designing a simple feature-blending strategy in the semantic encoder, combined with spatial-temporal self-attention to ensure natural placement and temporal consistency of eyeglasses.
- We establish a comprehensive benchmark for the video eyeglass transfer task, which will be publicly released (except for the CG face dataset).

2 Related Work

Eyeglasses Eyeglasses are a distinctive accessory that strongly influences facial perception and identity, and have therefore become a standalone research topic in computer vision and graphics. Prior works span several directions, including virtual try-on (Li & Yang, 2011; Yuan et al., 2011; Niswar et al., 2011; Huang et al., 2012; 2013; Tang et al., 2014; Zhang et al., 2017; Li et al., 2023), eyeglass detection (Wu et al., 2002; Bekhet & Alahmer, 2021), removal (Hu et al., 2020; Lee & Lai, 2020; Lyu et al., 2022; Zhang & Guo, 2025; Arkushin et al., 2025), and customized product design (Bai et al., 2021; Plesh et al., 2023). In the context of eyeglass try-on, most methods adopt predefined 3D eyeglass models and composite them onto faces (Li & Yang, 2011; Yuan et al., 2011; Niswar et al., 2011; Huang et al., 2012; 2013; Tang et al., 2014). Recent techniques simulate advanced physical effects, such as refraction (Zhang et al., 2017), or utilize multi-view data for realistic avatar reconstruction (Li et al., 2023). Though effective, they constrain users to predefined shapes and styles. Compared to these prior works, our work addresses the problem of adding and transferring arbitrary reference eyeglasses to facial videos from a single eyeglass reference image, without relying on predefined 3D models or extensive multi-view datasets.

Facial video editing Facial video editing has been explored through latent-space manipulation (Shen et al., 2020; Yao et al., 2021; Patashnik et al., 2021) and temporal consistency techniques such as smoothing and optical flow (Tzaban et al., 2022; Alaluf et al., 2022; Xu et al., 2022), often built on StyleGAN (Karras et al., 2019; 2020b). More recent models, including StyleGANEX (Yang et al., 2023b) and FED-NeRF (Zhang et al., 2024), reduce reliance on cropping and alignment. Kim et al. (2023) introduces a DiffAE-based method that significantly improves reconstruction quality but remains limited to classifier- or text-guided editing. S3Editor (Wang et al., 2024a) offers a model-agnostic self-training framework for semantic disentanglement but similarly lacks fine-grained control. Our work adopts DiffAE (Preechakul et al., 2022) to ensure reconstruction fidelity while enabling reference-guided semantic editing for facial videos.

Inpainting-based image and video editing Inpainting-based editing is a common strategy for object insertion, where masked regions are filled with plausible content guided by a reference. Recent diffusion-backboned models (Song et al., 2023; Yang et al., 2023a; Chen et al., 2024b; Song et al., 2024; Chen et al., 2024a) achieve good semantic coherence, but when applied frame by frame, they struggle with temporal consistency. Training-free variants such as TF-ICON (Lu et al., 2023) reduce the need for fine-tuning but still require accurate masks. Video-oriented extensions, including VideoAnyDoor (Tu et al., 2025) and VACE (Jiang et al., 2025), improve temporal alignment but remain mask-dependent. This mask-based formulation is reasonable for generic object insertion but is ill-suited for eyeglass transfer. As inpainting restricts editing to masked regions, these methods treat the eyeglass area independently of the surrounding face, which often weakens harmonization with facial geometry and identity. Our approach differs by integrating reference features directly into the semantic encoder, allowing the model to adapt eyeglasses to pose and context without relying on explicit target masks, which leads to more natural and temporally consistent results in video.

3 Method: FreeEyeglass

Figure 2 illustrates the overview of our FreeEyeglass pipeline. In this section, we first recap the Diffusion Autoencoder (DiffAE) (Preechakul et al., 2022) and explain how we semantically place the eyeglasses with a pretrained DiffAE.

3.1 Preliminary: Diffusion Autoencoder (DiffAE)

Our method is built on DiffAE proposed in Preechakul et al. (2022). In DiffAE, given an input image $\mathbf{x}$, the semantic encoder $\text{Enc}(\mathbf{x})$ maps it to a semantically meaningful latent $\mathbf{z}_{\text{sem}}$. A stochastic latent $\mathbf{z}_{\text{sto}}$, which captures the remaining stochastic details to achieve a near-perfect reconstruction, is then computed through the deterministic DDIM inversion (Song et al., 2021) using a conditional diffusion model $\text{DDIM}(\mathbf{z}_{\text{sto}}, \mathbf{z}_{\text{sem}})$. Taking $\mathbf{z}_{\text{sto}}$ and $\mathbf{z}_{\text{sem}}$ as input, the conditional diffusion model $\text{DDIM}(\mathbf{z}_{\text{sto}}, \mathbf{z}_{\text{sem}})$ reconstructs an image $\hat{\mathbf{x}}$ through the generative DDIM process. A typical DiffAE model uses a U-Net architecture (Ronneberger et al., 2015) similar to the diffusion model proposed in Dhariwal & Nichol (2021), which consists of multiple ResBlocks (He et al., 2016) and self-attention blocks (Vaswani et al., 2017). The semantic encoder shares the same architecture as the U-Net encoder and conditions the diffusion U-Net by adaptive group normalization (Dhariwal & Nichol, 2021; Huang & Belongie, 2017).

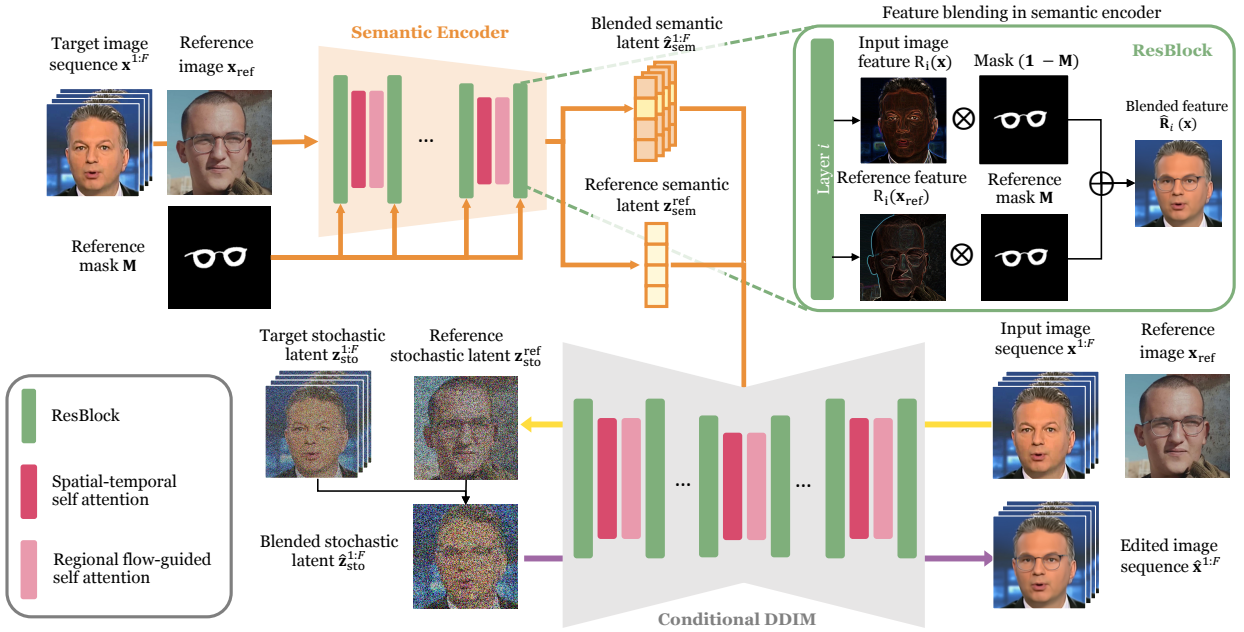

Figure 2: **Overview of our FreeEyeglass pipeline.** Given a target video (*i.e.*, target image sequence) and a reference image of desired eyeglasses, we blend the features of the reference eyeglasses and the target image sequence to obtain a blended semantic latent sequence. We then compute the stochastic latent sequences with the input images and semantic latent sequences through conditional DDIM inversion and construct a blended stochastic latent sequence. Using the blended stochastic latent and semantic latent sequences, we can obtain the final edited image sequence with our desired eyeglasses semantically and naturally placed through condition DDIM sampling.

3.2 Feature Blending for Eyeglass Transfer

The semantic encoder $\mathrm{Enc}(\cdot)$ is designed to encode the semantic information of the input image. In our case, the input images are the aligned faces from video frames. The semantic encoder maps these input images to a latent space where high-level semantic features are captured. Our approach involves blending the feature maps of the target frames with those of the reference eyeglasses at each ResBlock of the semantic encoder. By doing so, we aim to incorporate the semantics of the reference eyeglasses into the semantic latent of the target frame. The blending of feature maps is performed using a binary mask at each ResBlock. We represent each ResBlock i as a function $\mathbf{R}_i(\mathbf{x})$. Let a target frame be $\mathbf{x}$, a reference image be $\mathbf{x}_{\text{ref}}$, and let $\mathbf{M} \in \{0,1\}^{h_i \times w_i}$ be the binary mask indicating the eyeglass region. Here, h_i and w_i are the spatial dimensions of the feature map $\mathbf{R}_i(\cdot)$, and the mask is bilinearly downsampled to match this resolution. We compute the blended feature map $\hat{\mathbf{R}}_i(\mathbf{x})$ as follows:

$$\hat{\mathbf{R}}_i(\mathbf{x}) = \mathbf{M} \odot \mathbf{R}_i(\mathbf{x}_{\text{ref}}) + (\mathbf{1} - \mathbf{M}) \odot \mathbf{R}_i(\mathbf{x}), \tag{1}$$

where $\odot$ denotes element-wise multiplication. The mask $\mathbf{M}$ ensures that the features from the reference eyeglasses are only blended into the specified regions of the target frame.

After the feature blending process, we obtain a new semantic latent $\hat{\mathbf{z}}_{\text{sem}}$ that merges the semantic information from the target frame $\mathbf{x}$ and the reference image $\mathbf{x}$ref. While integrating the reference eyeglasses into the semantic latent allows the model to place eyeglasses on the target image semantically, this operation can introduce local noise and boundary artifacts due to feature mismatches between the target and reference features. To mitigate these artifacts, we also blend the stochastic latent codes, which capture residual information from the input image $\mathbf{x}$ that is not represented in the semantic latent $\hat{\mathbf{z}}_{\text{sem}}$. By mixing the stochastic latent of the reference image latent $\mathbf{z}_{\text{sto}}^{\text{ref}}$ with that of the target image $\mathbf{z}_{\text{sto}}$, we smooth the transition between reference and target features and suppress boundary inconsistencies. We construct a stochastic-latent blending mask $\tilde{\mathbf{M}}$ by combining the original reference mask $\mathbf{M}$ and its Gaussian-blurred version $\mathrm{Blur}(\mathbf{M})$:

$$\tilde{\mathbf{M}} = \beta \mathbf{M} + \gamma \, \mathrm{Blur}(\mathbf{M}), \tag{2}$$

where $\beta \geq 0$ and $\gamma \geq 0$ are scalar parameters that control the intensity and smoothness of the blending between the stochastic latent codes. We clamp the $\tilde{\mathbf{M}}$ values to ensure they lie within the $[0, 1]$ range.We smooth only the border of the blending mask, keeping the interior binary so that high-frequency details are preserved. Using the constructed mask $\tilde{\mathbf{M}}$, we perform alpha blending of the stochastic latent codes of the target and reference images:

$$\hat{\mathbf{x}}_T = (1 - \tilde{\mathbf{M}}) \odot \mathbf{z}_{\text{sto}} + \tilde{\mathbf{M}} \odot \mathbf{z}_{\text{sto}}^{\text{ref}}, \tag{3}$$

where $\odot$ denotes element-wise multiplication. Using the blended stochastic latent $\hat{\mathbf{z}}_{\text{sto}}$ as the starting noise of the DDIM sampling, we ensure that the generated image not only semantically includes the eyeglasses but also preserves their specific shape and color, improving the preservation of fine details in the edited images.

3.3 Self-attention for Temporal Consistency

Building on our success in achieving realistic and semantically consistent transfer of eyeglasses in images, we extend our method to facial videos. Video editing introduces the critical challenge of maintaining temporal consistency, as humans are highly sensitive to discrepancies across frames. Naively applying our image-based method frame by frame results in the eyeglasses appearing slightly different in each frame, leading to a jarring, flickering effect. As illustrated in Fig. 2, we extend the DiffAE-based architecture for video editing by incorporating 3D convolutions and two self-attention layers: a spatial-temporal self-attention layer and a regional flow-guided self-attention layer to enhance temporal consistency.

Extending DiffAE for video editing The original U-Net architecture in DiffAE consists of a series of 2D convolutional residual blocks and spatial self-attention blocks. To adapt this architecture for video editing, we inflate the 2D U-Net into a pseudo-3D U-Net to accommodate the temporal dimension, similar to previous works (Cong et al., 2024; Wu et al., 2023). Specifically, we replace each 2D convolutional layer with a pseudo-3D convolutional layer. For a 3×3 kernel, we adjust it to a $1 \times 3 \times 3$ kernel. We also expand the spatial self-attention blocks to spatial-temporal self-attention blocks, often used in video diffusion models, by using features from the entire video as queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$. However, applying full spatial-temporal self-attention can lead to undesired modifications in regions outside the eyeglasses area as the model attends to irrelevant embeddings. Thus, we introduce a regional self-attention mechanism that uses optical flows to mitigate this problem.

Regional self-attention with optical flows We aim to insert eyeglasses without altering other facial details. To this end, we employ a regional self-attention strategy. While similar local attention has been explored in prior works (Zhang et al., 2022; Cong et al., 2024), we adapt it to our framework by restricting attention to the eyeglass region. In the target video, we define a rough, predefined region of interest (ROI) of the eyeglasses area on the aligned face, computed as the bounding box of the aligned reference mask. We denote the set of all pixels in the ROI bounding box as $\mathcal{C}_{\text{roi}}$, which are appropriately downsampled to fit the spatial dimension of the self-attention block.

We adapt FLATTEN's flow-guided temporal attention to trajectories that pass through our predefined ROI. Instead of attending to the full frame, it attends only to tokens along these ROI-related trajectories, guided by the estimated optical flow. For a video sequence of length S, we can obtain a trajectory $\mathcal{T}$ for a pixel in coordinates (x_0, y_0) in the first frame by deriving its coordinates in all subsequent frames as

$$\mathcal{T} = \{(x_0, y_0), ..., (x_s, y_s), ..., (x_S, y_S)\}. \tag{4}$$

We only consider trajectories for which some or all coordinates satisfy $(x_s, y_s) \in \mathcal{C}_{\text{roi}}$, and denote them as $\{\mathcal{T}\}_{\text{roi}} \subset \{\mathcal{T}\}$.

We compute the self-attention of a selected motion trajectory $\mathcal{T} \in \{\mathcal{T}\}_{\text{roi}}$ in a similar manner to FLATTEN (Cong et al., 2024). Let the query $\mathbf{Q}$ of a pixel (x_s, y_s) of frame s be the embeddings $\mathbf{h}(x_s, y_s)$. The corresponding embeddings of the remaining coordinates in the same trajectory, $\mathcal{T}^- = \mathcal{T} - (x_s, y_s)$ are concatenated as

$$\mathbf{H}(\mathcal{T}^-) = [..., \mathbf{h}(x_{s-1}, y_{s-1}), \mathbf{h}(x_{s+1}, y_{s+1}), ...]. \tag{5}$$

Regional optical-flow-guided self-attention is thus computed using the dimensionality of the embeddings d as

$$\mathbf{Q} = \mathbf{h}(x_s, y_s), \quad \mathbf{K} = \mathbf{V} = \mathbf{H}(\mathcal{T}^-), \quad \text{RA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{6}$$

By constraining the attention to the eyeglasses region and guiding it with optical flow, we ensure that each embedding in our ROI attends only to its temporally corresponding embeddings across frames. This approach improves the temporal consistency of the eyeglasses while preventing unwanted artifacts in unrelated areas.

3.4 Overall Video Editing Pipeline

We adopt a video editing pipeline proposed in Yao et al. (2021), which comprises three main stages: (1) face alignment and cropping, (2) latent feature encoding, manipulation, and decoding, and (3) unalignment and merging of edited frames into the original video. Given an input video sequence $\mathbf{v} \in \mathbb{R}^{1\times C\times F\times H\times W}$, where C is the number of channels, F is the number of frames, and H, W denote the height and width, respectively, we first align and square-crop the face regions in each frame and obtain a cropped sequence $\mathbf{x}_{1:F}$ Similarly, we process a reference eyeglasses image to obtain its aligned and cropped face region $\mathbf{x}_{\text{ref}}$. This alignment ensures that the facial features are spatially consistent across frames and with the reference. To enhance temporal consistency in long video sequences that cannot be input at once, we process the video in batches and sample two neighboring frames, one immediately before and one immediately after each batch. These neighboring frames provide additional temporal context, allowing for smoother transitions between batches. The cropped face sequences $\mathbf{x}_{1:F}$ and the cropped reference image $\mathbf{x}_{\text{ref}}$ are then concatenated as $\mathbf{x}_{1:F,\text{ref}}$ and passed through our semantic encoder $\text{Enc}(\mathbf{x}_{1:F,\text{ref}})$. We apply our feature blending strategy and yield the following semantic latent codes:

$$\hat{\mathbf{z}}_{\text{sem}}^{1:F} = \text{Enc}(\mathbf{x}_{1:F}), \quad \mathbf{z}_{\text{sem}}^{\text{ref}} = \text{Enc}(\mathbf{x}_{\text{ref}}). \tag{7}$$

Next, we obtain the stochastic latent representations $\mathbf{z}_{\text{sto}}^{1:F}$ and $\mathbf{z}_{\text{sto}}^{\text{ref}}$ via conditional DDIM inversion using $\hat{\mathbf{z}}_{\text{sem}}^{1:F}$ and $\mathbf{z}_{\text{sem}}^{\text{ref}}$. We construct the blended stochastic latent codes $\hat{\mathbf{z}}_{\text{sto}}^{1:F}$ using $\mathbf{z}_{\text{sto}}^{1:F}$ and $\mathbf{z}_{\text{sto}}^{\text{ref}}$ by Equation (3). The edited frames are then generated by passing the blended stochastic latent codes $\hat{\mathbf{z}}_{\text{sto}}^{1:F}$ and semantic latent codes $\hat{\mathbf{z}}_{\text{sem}}^{1:F}$ through our conditional DDIM model. To ensure temporal coherence, we incorporate regional optical-flow-guided self-attention in all self-attention blocks, applying it exclusively to the target video sequence. Finally, we unalign the edited face regions and paste them back into the original video.

4 Experiments

We evaluate our method with various baselines to confirm its effectiveness. We describe our implementation details and report further experiments in the appendix.

4.1 Settings

Datasets Since there is no existing dataset tailored for evaluating semantic eyeglass transfer, we construct a benchmark from CelebV-HQ (Zhu et al., 2022). For target facial videos, we randomly select videos without visible glasses that are at least 120 frames long, *i.e.*, 4 seconds in 30 fps, and use the first 120 consecutive frames of these videos as target facial videos. To obtain reference eyeglasses, we select videos with clearly visible glasses and apply the following quality filters: an average VSFA score (Li et al., 2019) above 0.8, mean luminance > 90, head pitch and yaw within $\pm 15^\circ$, and unique identity per sample. After filtering, we randomly selected 100 unique pairs of glasses, comprising 75 eyeglasses and 25 sunglasses, to serve as reference eyeglasses. Finally, we prepare eyeglass and sunglass masks using Grounded-SAM (Ren et al., 2024). Specifically for eyeglass masks, we refine them by subtracting the segmented eye region from the segmented eyeglasses region. We use the aligned mask's bounding box as the ROI for our regional flow-guided self-attention. In total, we construct 100 pairs of target facial videos and reference glasses for our main experiment. We also render a small-scale CG face dataset of four target identities, each with one pair of eyeglasses and one pair of sunglasses. We render eight ground-truth videos in total for the target identities.

Table 1: Baseline requirements. Most existing approaches depend on training and/or explicit masks. In contrast, **our method is training-free and target-mask-free**, requiring only a reference eyeglass and the target video.

Method	Target Mask	Training	Inputs	Backbone
TF-ICON (Lu et al., 2023)	Yes	No	Image + Ref + Mask	Diffusion
Paint-by-Example (Yang et al., 2023a)	Yes	Yes	Image + Ref + Mask	Diffusion
ObjectStitch (Song et al., 2023)	Yes	Yes	Image + Ref + Mask	Diffusion
AnyDoor (Chen et al., 2024b)	Yes	Yes	Image + Ref + Mask	Diffusion
MimicBrush (Song et al., 2024)	Yes	Yes	Image + Ref + Mask	Diffusion
OmniTry (Feng et al., 2025)	No	Yes	Image + Ref	Diffusion
DVAE (Kim et al., 2023)	No	Yes	Video + Classifier/Text	DiffAE
VideoEditGAN (Tzaban et al., 2022)	No	Yes	Video + Classifier	GAN
FLATTEN (Cong et al., 2024)	No	Yes	Video + Text	Diffusion
RAVE (Kara et al., 2024)	No	Yes	Video + Text	Diffusion
VidToMe (Li et al., 2024)	No	Yes	Video + Text	Diffusion
FRESCO (Yang et al., 2024)	No	Yes	Video + Text	Diffusion
RF-Solver-Edit (Wang et al., 2025)	No	Yes	Video + Text	Diffusion
VACE (Jiang et al., 2025)	Yes	Yes	Video + Ref + Text + Mask	Diffusion
FreeEyeglass (Ours)	**No**	**No**	Video + Ref	DiffAE

Ref: Reference Image

Table 2: **Quantitative results** on our evaluation benchmark. The top rows show the methods for image editing, while the bottom rows are for video editing. Eye preservation is evaluated only on eyeglasses. Values closer to 1.0 indicate better in TL-ID and TG-ID. **Bold** and underline indicate the best and second-best results. While our method achieves state-of-the-art results in many cases, it yields a remarkably better *Unified* score S_{edit} that balances identity preservation and editing quality.

Methods	Editing Fidelity				Temporal Consistency				Eye Preservation		S_{edit} ↑
	FID_{CLIP} ↓	FVD ↓	CLIP-I ↑	DINO-I ↑	CLIP-F ↑	E_{warp} ↓	TL-ID −	TG-ID −	$\text{LPIPS}_{\text{eye}}$ ↓	SSIM_{eye} ↑	
Object Stitch (Song et al., 2023)	11.626	294.12	0.882	0.624	0.957	0.0182	0.957	0.887	0.183	0.814	48.336
TF-ICON (Lu et al., 2023)	26.394	2382.6	0.802	0.441	0.868	0.0462	0.222	0.216	0.214	0.804	17.457
Paint-by-Example (Yang et al., 2023a)	12.915	295.03	0.877	0.650	0.957	0.0182	0.956	0.884	0.191	0.810	48.090
Anydoor (Chen et al., 2024b)	18.469	1007.7	0.850	0.517	0.949	0.0279	0.779	0.748	0.242	0.781	30.452
MimicBrush (Chen et al., 2024a)	13.399	362.31	**0.909**	**0.739**	0.959	0.0188	0.960	0.899	0.211	0.788	48.334
OmniTry (Feng et al., 2025)	11.776	415.29	0.857	0.588	0.952	0.0173	0.892	0.684	0.183	0.813	49.670
VideoEditGAN (Xu et al., 2022)	10.443	382.93	0.846	0.547	0.958	0.0167	0.964	0.930	0.149	0.840	50.770
DVAE Classifier (Kim et al., 2023)	13.624	1190.9	0.819	0.433	0.951	0.0169	0.788	0.882	**0.124**	**0.869**	48.396
FLATTEN (Cong et al., 2024)	52.022	1522.7	0.780	0.300	0.952	0.0154	0.860	0.718	0.175	0.839	50.809
RAVE (Kara et al., 2024)	43.152	1035.0	0.833	0.526	**0.965**	0.0240	0.916	0.850	0.215	0.797	34.085
VidToME (Li et al., 2024)	36.361	960.97	0.828	0.499	0.959	0.0201	0.931	0.802	0.200	0.805	41.111
FRESCO (Yang et al., 2024)	53.046	973.35	0.787	0.486	0.952	0.0262	0.859	0.821	0.161	0.827	30.086
RF-Solver-Edit (Wang et al., 2025)	21.507	485.35	0.833	0.492	0.944	0.0205	0.946	0.894	0.164	0.838	40.627
VACE (Jiang et al., 2025)	9.947	223.57	0.858	0.634	0.961	0.0167	**0.974**	**0.942**	0.163	0.836	53.136
FreeEyeglass (Ours)	**9.839**	**206.37**	0.865	0.542	0.962	**0.0152**	0.969	0.885	0.124	0.868	**56.976**

Baselines We compare against a wide range of state-of-the-art image and video editing methods. Table 1 summarizes the requirements of all baselines in terms of training, masks, and input modalities. For reference-based image editing, we include TF-ICON, Paint-by-Example, ObjectStitch, AnyDoor, and MimicBrush, which insert a reference object into a background image through inpainting. We also evaluate OmniTry, a concurrent method targeting wearable object try-on. As these are image editing methods, we apply them frame by frame to videos.

For video editing, we consider both facial and general text-guided approaches. Facial video editing baselines include Diffusion Video Autoencoder (DVAE) and VideoEditGAN. Text-guided video editing baselines include FLATTEN, RAVE, VidToMe, FRESCO, and RF-Solver-Edit with HunyuanVideo (Kong et al., 2024) as backbone. We further include VACE 1.3B, a concurrent multi-modal video editing framework that requires text prompts, target masks, and reference images. We could not run the 14B variant of VACE due to memory limits on A100 80GB GPUs. For text-based baselines, we use GPT-4o (API version 2024-05-13) (OpenAI, 2024) to generate fine-grained descriptions of reference eyeglasses. Implementation details and prompt preparation are provided in the appendix. We use the implementations and pretrained models provided by the authors for all baselines except FLATTEN to generate edited videos. Due to GPU memory constraints, FLATTEN cannot process a video of 120 frames simultaneously, so we split each video into three chunks.

Evaluation metrics Since our benchmark lacks ground truth, we assess results using three categories of metrics: *editing fidelity*, *temporal consistency*, and *eye preservation*.

Editing fidelity: We use the Fréchet Inception Distance (FID) computed on CLIP features (Kynkäänniemi et al., 2023) to assess overall perceptual quality and Fréchet Video Distance (FVD) (Skorokhodov et al., 2022) to evaluate video realism. We use all frames in the videos for FVD calculation. We compute CLIP-I and DINO-I scores following Wei et al. (2024), using average cosine similarity between each edited frame and the reference eyeglasses image to assess semantic alignment with the reference. CLIP-I uses CLIP ViT-B/32 (Radford et al., 2021), while DINO-I uses DINOv2 ViT-S/16 (Oquab et al., 2024); in both cases, we align frames and crop the eyeglasses region.

Temporal consistency: We compute the average cosine similarity between consecutive frame embeddings using CLIP (CLIP-F) to capture feature-level smoothness, and Warp Error (Lai et al., 2018) that measures the pixel-wise difference between adjacent frames warped using the optical flow from the original video. We also compute TL-ID and TG-ID (Tzaban et al., 2022), which quantify identity preservation across adjacent frames or across all frames, respectively.

Eye preservation: We crop the eyeglasses region for all frames and report $\text{LPIPS}_{\text{eye}}$ and SSIM_{eye} for the eyeglasses regions of source frames and edited frames, similar to Feng et al. (2025). We conduct the evaluation exclusively on 75 eyeglass pairs, as sunglasses typically obscure the eye region.

Figure 3: **Visual results** on transferring different eyeglasses compared with our baselines. Our method successfully inserts and swaps eyeglasses from the reference eyeglasses image into the target video, compared to existing baselines. Please refer to the supplementary video for more results.

We also report a *unified evaluation* score S_{edit}, proposed by FLATTEN, defined as the ratio CLIP-I/E_{warp}, providing a single metric that balances semantic fidelity and temporal consistency.
For the CG face dataset, we compute PSNR, MS-SSIM, LPIPS (Zhang et al., 2018), and MSE between ground-truth frames and edited frames generated by ours and the baselines.

4.2 Main Results

We present our quantitative results in Table 2 and our visual results compared with our baselines in Fig. 3. We achieve the best unified score S_{edit} among all baselines, which reflects our clear advantage in balancing editing fidelity and temporal consistency. We also achieve superior FID_{CLIP}, FVD, and eye preservation scores $\text{LPIPS}_{\text{eye}}$ and SSIM_{eye}, which shows our strength in preserving original video content and overall realism. Qualitative results support these findings. Our method transfers eyeglasses while preserving facial

identity and maintaining temporal coherence. Inpainting-based editing baselines reproduce eyeglass details but fail to maintain facial identity. Even the latest works, such as OmniTry and VACE, fail to maintain the eye regions (*e.g.* Eyeglass #1 in Fig. 3). Video editing methods guided by text or classifiers are temporally stable but cannot capture the reference eyeglasses. We adopt DVAE with classifier guidance as our main DVAE baseline in Table 2 following its original video editing setup. We further analyze DVAE with text guidance and the original image-based DiffAE with classifier guidance in Sec. 4.6 to understand how the DiffAE backbone behaves under different guidance types. We also include a supplementary video for full comparison.

Our method scores slightly lower on CLIP-I and DINO-I, as these metrics favor methods that closely match the reference, yielding high scores when directly copying eyeglasses. Inpainting-based baselines, including Object Stitch, Paint-by-example, Anydoor, and MimicBrush, use CLIP or DINO features during training, which provides them an advantage. In contrast, our method has no prior exposure to these embeddings. Finally, our framework requires neither per-frame target-mask annotations nor model training. Despite being training-free and target-mask-free, it delivers results that are competitive with, or superior to, those of training-based state-of-the-art methods.

4.3 Evaluation on CG-rendered Ground Truth

We present our quantitative results on CG-rendered scenes in Table 3. Our method outperforms all baselines on PSNR, SSIM, and MSE, and achieves competitive performance on LPIPS, demonstrating superior fidelity across both pixel-level and perceptual metrics. Text-based video editing often over-modifies the appearance and style of the video when applied to local regions, *i.e.*, eyeglass transfer, thus leading to unsatisfactory results. This emphasizes the importance of reference-based editing for achieving precise semantic control.

Table 3: **Quantitative results** on the CG face dataset. Our method outperforms on most metrics. **Bold** and underline indicate the best and second-best results.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	MSE ↓
Object Stitch	26.495	0.956	0.0401	481.30
TF-ICON	10.596	0.200	0.503	17617
Paint-by-Example	26.482	0.957	0.0392	<u>460.58</u>
Anydoor	24.457	0.948	0.0422	774.08
MimicBrush	26.527	<u>0.964</u>	**0.0301**	467.81
OmniTry	24.794	0.943	0.0404	679.84
DVAE Classifier	25.626	0.942	0.118	572.16
VideoEditGAN	24.008	0.912	0.0715	843.26
FLATTEN	16.857	0.780	0.212	4245.6
RAVE	15.942	0.799	0.175	6440.2
VidToME	17.802	0.813	0.193	3411.6
FRESCO	18.537	0.798	0.148	2779.3
RF-Solver-Edit	18.548	0.595	0.431	2936.0
VACE	<u>26.688</u>	0.958	<u>0.0375</u>	463.53
FreeEyeglass (Ours)	**28.425**	**0.967**	0.0387	**307.21**

SSIM: MS-SSIM

4.4 Robustness and Generalization Analysis

Robustness to lighting variations Our method remains stable under strong illumination mismatches between the reference eyeglasses and the target video frames. As in Fig. 4, it preserves correct eyeglass geometry and placement when transferring from a front-lit indoor reference to targets captured under bright outdoor sunlight, strong backlighting, and dim indoor side-lit. Across all cases, the transferred glasses integrate cleanly into each scene without artifacts.

Figure 4: Transferring eyeglasses to target frames under **different illuminations** (*e.g.*, outdoor sunlight, backlit, and side-lit).

Robustness to pose change We analyze FreeEyeglass's performance across varying head yaw angles. As in Fig. 5, our method remains stable up to approximately ±40–50° of pose change, beyond which the transfer becomes less reliable due to strong geometric mismatch between the reference and target.

Occlusion handling We evaluate FreeEyeglass under several types of partial occlusions. As shown in Fig. 6, the method produces reasonable results in cases involving hair bangs, dynamic hand motion, and rigid objects such as headphones or microphones. Our method preserves the target identity and eyeglass appearance, while maintaining plausible geometry and placement. We attribute this robustness to the strong reconstruction priors of DiffAE, which enable the model to infer a consistent facial structure despite partial visibility.

Generalization to other facial attributes Although FreeEyeglass is designed for eyeglass transfer, we push the limits of the framework by applying the same mask-only blending strategy to other local facial

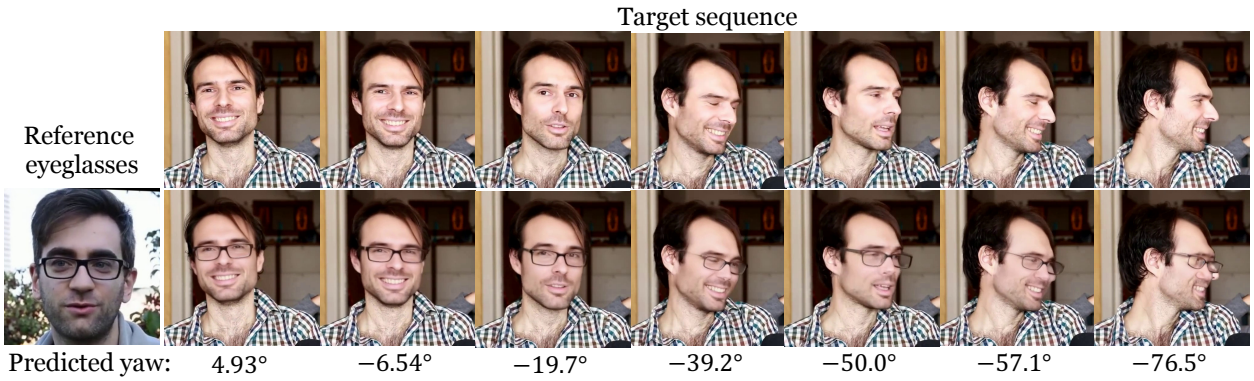

Figure 5: **Eyeglass transfer under head pose variation.** Edited results across increasing yaw angles. Our method remains stable up to roughly ±40–50°, beyond which misalignment and artifacts occur.

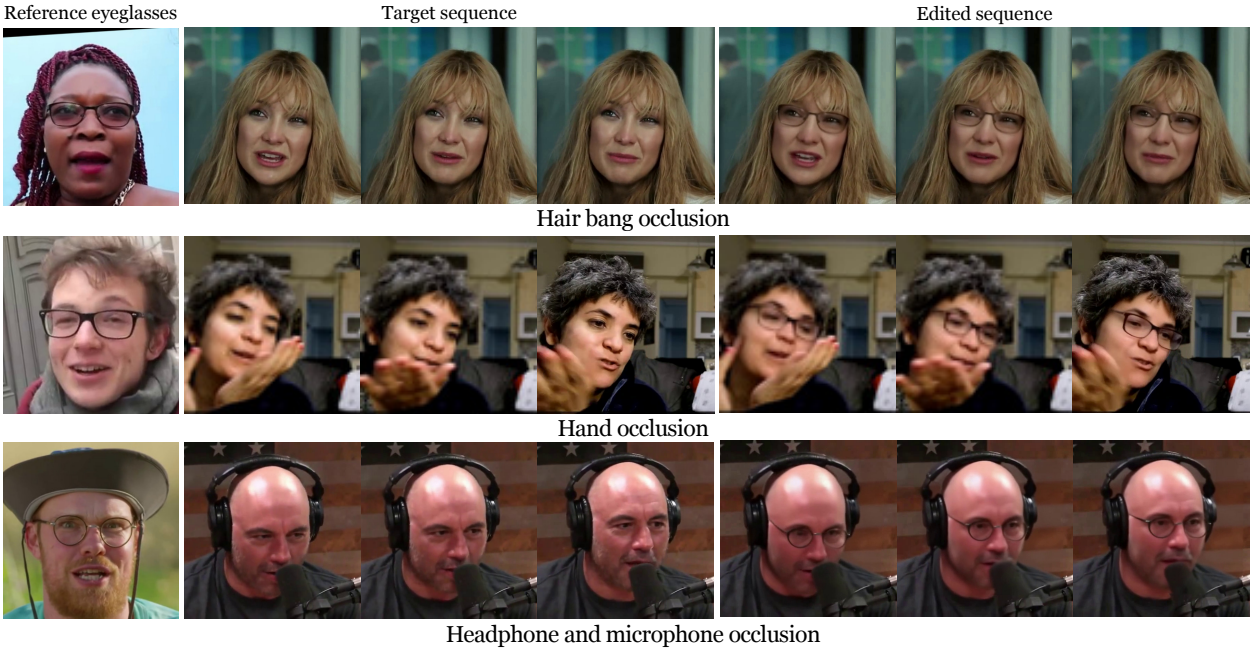

Figure 6: **Occlusion handling results.** FreeEyeglass handles cases when the target face is partially occluded. We present three representative occlusions, including hair bangs, hand motion, and object occlusion over the face, where the transferred eyeglasses remain well-aligned and visually coherent.

attributes, including eyebrows, noses, and moustaches (Fig. 7). Without any model modification, the method can inject these attributes into the target sequence with reasonable geometric alignment. These examples suggest potential applicability to other face-centric attributes. We also explore sequential multi-attribute editing and present the results in Sec. C.4.

4.5 Ablation Study

We conduct an ablation study across various settings to demonstrate the effectiveness of our method. We ablate our method by switching off different proposed strategies: (1) w/o feature blending, (2) w/o blended stochastic latent $\hat{\mathbf{z}}_{\text{sto}}$, and different self-attention variants, including per-frame, spatial–temporal, flow-guided, and their regional counterparts. Figure 8 shows that feature blending is essential for inserting the reference eyeglasses; the "w/o feat. blend" variant yields a lower FVD simply because it produces almost no edits and therefore remains closer to the target distribution. Using feature blending without stochastic blending yields the most faithful results for the eyeglasses (Table 4), but introduces noticeable noise in Fig. 8. Adding a fixed ROI boundary to per-frame SA or STSA reduces FVD by limiting global interference, but it also lowers CLIP-I because the static ROI often misaligns with the eyeglasses, causing partial corruption of their shapes. In

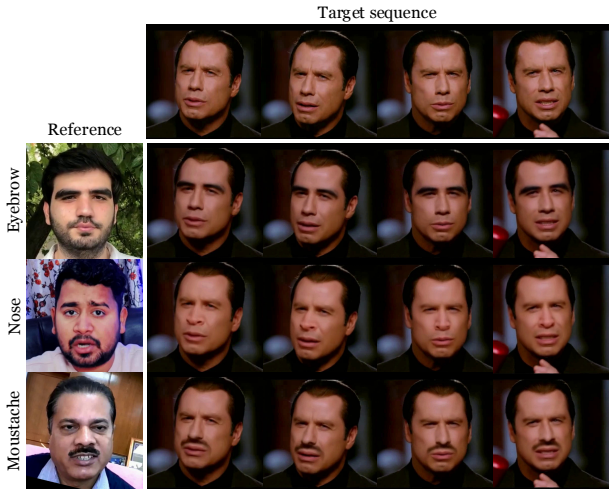

Figure 7: **Transfer of other facial attributes.** FreeEyeglass can transfer other face-centric attributes.

Table 4: **Ablation study** on different components of our method. **Bold** and underline indicate the best and second-best results.

Settings	FVD↓	CLIP-I↑	E_{warp} ↓	TL-ID−
Our full model	206.37	0.865	0.0152	0.969
w/o feat. blend.	**167.65**	0.819	**0.0149**	**0.970**
w/o blended $\hat{\mathbf{z}}_{\text{sto}}$	280.73	**0.868**	0.0162	0.946
Per-frame SA	226.37	0.863	0.0151	0.968
STSA	226.51	0.864	0.0151	0.968
FlowSA	223.61	0.864	0.0151	0.969
Regional Per-frame SA	215.03	0.849	0.0152	0.966
Regional STSA	193.68	0.851	0.0152	0.958

Feat. blend.: feature blending strategy
SA: self-attention
STSA: spatial-temporal self-attention
FlowSA: flow-guided self-attention

Table 5: **Performance of DiffAE-based methods** on the eyeglass transfer task. DVAE text fails to apply edits despite strong reconstruction-oriented scores.

Methods	Editing Fidelity				Temporal Consistency				Eye Preservation		S_{edit} ↑
	FID_{CLIP} ↓	FVD ↓	CLIP-I ↑	DINO-I ↑	CLIP-F ↑	E_{warp} ↓	TL-ID −	TG-ID −	$\text{LPIPS}_{\text{eye}}$ ↓	SSIM_{eye} ↑	
DiffAE Classifier (Preechakul et al., 2022)	8.397	295.59	0.854	0.532	0.960	0.0167	0.962	0.866	0.166	0.850	51.072
DVAE Text (Kim et al., 2023)	**6.009**	**152.12**	0.805	0.417	0.955	0.0160	0.960	**0.916**	**0.0809**	**0.900**	51.629
DVAE Classifier (Kim et al., 2023)	13.624	1190.9	0.819	0.433	0.951	0.0169	0.788	0.882	0.124	0.869	48.396
FreeEyeglass (Ours)	9.839	206.37	**0.865**	**0.542**	**0.962**	**0.0152**	**0.969**	0.885	0.124	0.868	**56.976**

Table 6: **Comparison** with the inflated version of MimicBrush (Chen et al., 2024a). Values closer to 1.0 indicate better in TL-ID and TG-ID.

Methods	Editing Fidelity				Temporal Consistency				Eye Preservation		S_{edit} ↑
	FID_{CLIP} ↓	FVD ↓	CLIP-I ↑	DINO-I ↑	CLIP-F ↑	E_{warp} ↓	TL-ID −	TG-ID −	$\text{LPIPS}_{\text{eye}}$ ↓	SSIM_{eye} ↑	
MimicBrush (Chen et al., 2024a)	13.399	362.31	0.909	0.739	0.959	0.0188	0.960	**0.899**	0.211	0.788	48.334
Inflated MimicBrush (Chen et al., 2024a)	13.193	648.59	**0.910**	**0.744**	0.955	0.0226	0.867	0.843	0.211	0.788	40.229
FreeEyeglass (Ours)	**9.839**	**206.37**	0.865	0.542	**0.962**	**0.0152**	**0.969**	0.887	**0.124**	**0.868**	**56.976**

contrast, combining the same ROI boundary with flow-guided attention maintains both semantic correctness and temporal alignment, as it tracks motion trajectories passing through the ROI and keeps the attended features aligned with the moving eyeglasses across frames. We also compare different locations to apply feature blending within DiffAE (see Sec. C.5). Results confirm that applying blending in the semantic encoder yields the best identity-preserving edits, whereas applying it elsewhere causes artifacts or copy-paste effects.

4.6 Analysis of DiffAE-based Editing Methods

We evaluate the DiffAE-based editing mechanisms, original DiffAE with classifier guidance, and DVAE with text guidance in this section. DiffAE with classifier guidance can insert eyeglasses and produce plausible results. However, because the classifier provides only a coarse binary signal (*e.g.*, "wearing glasses"), the generated eyeglasses do not match the reference, as reflected in lower CLIP-I and DINO-I scores. We also observe a weaker temporal consistency. These behaviours reflect the limits of DiffAE's original editing mechanism, as it is not designed for reference-guided or temporally aligned control. DVAE with text guidance almost reconstructs the original frames, yielding deceptively strong scores on metrics that compare against the input (*e.g.*, FID_{CLIP} and FVD), as summarized in Table 5. However, eyeglasses are barely added under the default guidance level (0.5), and stronger guidance forces the faces to collapse (see Fig. 9). In contrast, DVAE with classifier guidance can insert eyeglasses, but the results are less faithful due to the coarse nature of classifier signals. These findings suggest that while DiffAE models possess the capacity for editing, the success of the edit depends strongly on the type and strength of the guidance. Effective and reliable reference-guided control is therefore non-trivial, even with a DiffAE backbone, and motivates our design of FreeEyeglass.

4.7 Comparison with Inflated Reference-based Baseline

We further investigate inflating MimicBrush's U-Net (Chen et al., 2024a), the most reference-faithful image editing baseline, from 2D to 3D, following FLATTEN. Our goal is to determine whether simply extending

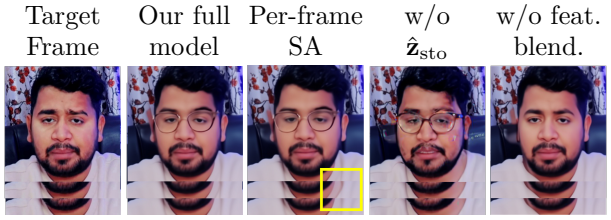

Figure 8: **Visual results** of ablation study. The yellow box denotes the region of temporal inconsistency. Zoom in for a clear comparison.

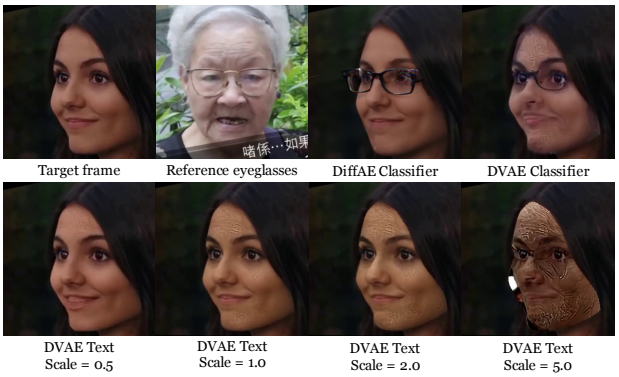

Figure 9: **Visual results** of DiffAE classifier guidance, DVAE classifier, and text guidance.

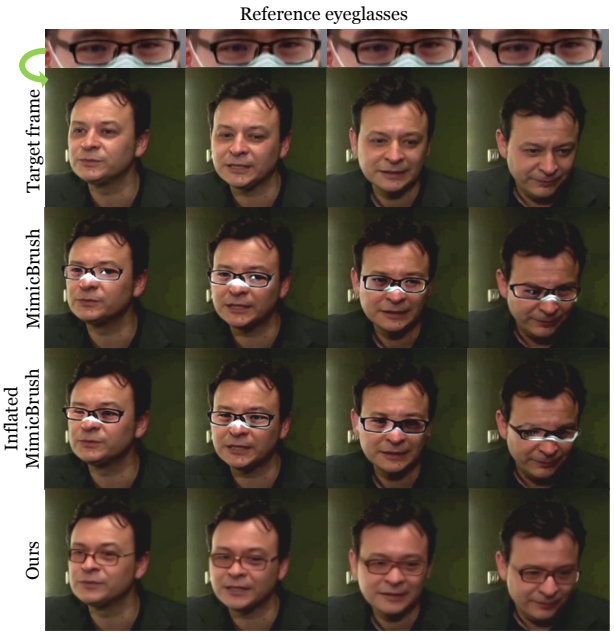

Figure 10: **Visual results** on temporal consistency with MimicBrush and inflated MimicBrush (Chen et al., 2024a).

reference-based image editing methods into the temporal domain straightforwardly improves their performance in video tasks. However, this naive extension worsened temporal consistency and video quality as shown in Table 6 and Fig. 10. These results indicate that temporal adaptation of existing image-based methods is non-trivial and does not necessarily yield better outcomes, likely due to temporal attention mismatches.

5 Conclusion

This work targets realistic video eyeglass transfer, where preserving facial identity is critical. To move beyond inpainting-style generation, we propose a training-free, reference-guided framework on reconstruction-oriented DiffAE. Our method blends semantic features into the encoder and uses regional flow-guided attention to propagate eyeglass appearance while maintaining temporal coherence. Experiments show that our method successfully transfers the reference eyeglasses into the target video, creating harmonized results that accurately reflect the original eyeglasses. Meanwhile, our method faces challenges when swapping eyeglasses that differ significantly in style or completely removing eyeglasses from an image (see Sec. G). It is because fine details of the eyeglasses that the semantic encoder fails to capture are embedded into the stochastic latent space, making them difficult to manipulate. Improving the semantic encoder to better capture eyeglass semantics may address this issue.

Broader impact This work proposes a method for reference-guided eyeglass transfer in facial videos and has potential societal implications related to privacy, misuse, and fairness. Our experiments use publicly available datasets licensed for use and a private CG dataset created with informed consent, solely for research purposes. No personally identifying information is collected in our user study, and the authors record no new human data. As the method operates on facial videos, it could be misused to alter a person's appearance without consent or to create deceptive content. Although the method does not aim to modify facial identity or expressions, realistic accessory edits may still mislead viewers if used maliciously. Any real-world deployment should therefore involve appropriate consent, transparency, and safeguards, such as labeling edited content with watermarks or metadata. The method relies on a DiffAE pretrained on the FFHQ dataset. While we introduce no new training data, the learned representation may reflect existing dataset biases, potentially affecting performance across demographic groups. We acknowledge this limitation and view fairness evaluation as an important direction for future work.

References

Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with StyleGAN3. In *Proceedings of European Conference on Computer Vision Workshops (ECCVW)*, pp. 204–220, 2022.

Rotem Shalev Arkushin, Aharon Azulay, Tavi Halperin, Eitan Richardson, Amit Haim Bermano, and Ohad Fried. V-LASIK: Consistent glasses-removal from videos using synthetic data. In *Proceedings of International Conference on Learning Representations Workshops (ICLRW)*, 2025.

Xiaobo Bai, Omar Huerta, Ertu Unver, James Allen, and Jane E Clayton. A parametric product design framework for the development of mass customized head/face (eyewear) products. *Applied Sciences*, 11 (12):5382, 2021.

Saddam Bekhet and Hussein Alahmer. A robust deep learning approach for glasses detection in non-standard facial images. *IET Biometrics*, 10(1):74–86, 2021.

Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 84010–84032, 2024a.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602, 2024b.

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 *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8780–8794, 2021.

Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024.

Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. Omnitry: Virtual try-on anything without masks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. Unsupervised eyeglasses removal in the wild. *IEEE Transactions on Cybernetics*, 51(9):4373–4385, 2020.

Szu-Hao Huang, Yu-I Yang, and Chih-Hsing Chu. Human-centric design personalization of 3D glasses frame in markerless augmented reality. *Advanced Engineering Informatics*, 26(1):35–45, 2012.

Wan-Yu Huang, Chaur-Heh Hsieh, and Jeng-Sheng Yeh. Vision-based virtual eyeglasses fitting system. In *Proceedings of IEEE International Symposium on Consumer Electronics (ISCE)*, pp. 45–46, 2013.

Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1510–1519, 2017.

Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10799–10808, 2022.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6507–6516, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020b.

Gyeongman Kim, Hajin Shim, Hyunsu Kim, Yunjey Choi, Junho Kim, and Eunho Yang. Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6091–6100, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *Proceedings of International Conference on Learning Representations (ICLR)*, 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 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 *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 243–258, 2020.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 2351–2359, 2019.

Juan Li and Jie Yang. Eyeglasses try-on based on improved poisson equations. In *Proceedings of International Conference on Multimedia Technology (ICMT)*, pp. 3058–3061, 2011.

Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. MEGANE: Morphable eyeglass and avatar network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12769–12779, 2023.

Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. VidToMe: Video token merging for zero-shot video editing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7486–7495, 2024.

Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-based training-free cross-domain image composition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2294–2305, 2023.

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

Yingmao Miao, Zhanpeng Huang, Rui Han, Zibin Wang, Chenhao Lin, and Chao Shen. Shining yourself: High-fidelity ornaments virtual try-on with diffusion model. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 359–368, 2025.

Hung Nguyen, Quang Qui-Vinh Nguyen, Khoi Nguyen, and Rang Nguyen. Swifttry: Fast and consistent video virtual try-on with diffusion models. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 6200–6208, 2025.

Arthur Niswar, Ishtiaq Rasool Khan, and Farzam Farbiz. Virtual try-on of eyeglasses using 3D model of the head. In *Proceedings of International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI)*, pp. 435–438, 2011.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, pp. 1–31, 2024.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2085–2094, 2021.

Richard Plesh, Peter Peer, and Vitomir Struc. GlassesGAN: Eyewear personalization using synthetic appearance discovery and targeted subspace modeling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16847–16857, 2023.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10619–10629, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Philippe G Schyns, Lizann Bonnar, and Frédéric Gosselin. Show me the features! understanding recognition from the use of visual information. *Psychological Science*, 13(5):402–409, 2002.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(4):2004–2018, 2020.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626–3636, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18310–18319, 2023.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian L. Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel G. Aliaga. IMPRINT: Generative object compositing by learning identity-preserving representation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8048–8058, 2024.

James W Tanaka and Diana Simonyi. The "parts and wholes" of face recognition: A review of the literature. *Quarterly Journal of Experimental Psychology*, 69(10):1876–1889, 2016.

Difei Tang, Juyong Zhang, Ketan Tang, Lingfeng Xu, and Lu Fang. Making 3D eyeglasses try-on practical. In *Proceedings of IEEE International Conference on Multimedia & Expo Workshop (ICMEW)*, 2014.

Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 402–419, 2020.

Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. In *Proceedings of ACM SIGGRAPH Conference Papers*, pp. 151:1–151:11, 2025.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in time: GAN-based facial editing of real videos. In *Proceedings of ACM SIGGRAPH Asia Conference Papers*, pp. 29:1–29:9, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

Guangzhi Wang, Tianyi Chen, Kamran Ghasedi, HsiangTao Wu, Tianyu Ding, Chris Nuesmeyer, Ilya Zharkov, Mohan Kankanhalli, and Luming Liang. S3Editor: A sparse semantic-disentangled self-training framework for face video editing. *arXiv preprint arXiv:2404.08111*, 2024a.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. In *Proceedings of International Conference on Machine Learning (ICML)*, 2025.

Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. Gpd-vvto: Preserving garment details in video virtual try-on. In *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 7133–7142, 2024b.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. DreamVideo: Composing your dream videos with customized subject and motion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6537–6549, 2024.

Haiyuan Wu, Genki Yoshikawa, Tadayoshi Shioyama, T Lao, and T Kawade. Glasses frame detection with 3D hough transform. In *Proceedings of IEEE International Conference on Pattern Recognition (ICPR)*, pp. 346–349, 2002.

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 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7623–7633, 2023.

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

Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. In *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 3199–3208, 2024.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023a.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. StyleGANex: StyleGAN-based manipulation beyond cropped aligned faces. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21000–21010, 2023b.

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. FRESCO: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8703–8712, 2024.

Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13789–13798, 2021.

Miaolong Yuan, Ishtiaq Rasool Khan, Farzam Farbiz, Arthur Niswar, and Zhiyong Huang. A mixed reality system for virtual glasses try-on. In *Proceedings of International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI)*, pp. 363–366, 2011.

Haitao Zhang and Jingtao Guo. Erat: Eyeglasses removal with attention. *Pattern Recognition*, 158:110970, 2025.

Hao Zhang, Yu-Wing Tai, and Chi-Keung Tang. FED-NeRF: Achieve high 3D consistency and temporal coherence for face video editing on dynamic NeRF. *arXiv preprint arXiv:2401.02616*, 2024.

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

Qian Zhang, Yu Guo, Pierre-Yves Laffont, Tobias Martin, and Markus Gross. A virtual try-on system for prescription eyeglasses. *IEEE Computer Graphics and Applications*, 37(4):84–93, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 650–667, 2022.

Appendix

This appendix provides additional discussion and experimental results. We first present a discussion of related work on text-based editing in Sec. A. We describe further implementation details in Sec. B. We then report extended experiments, including a temporal consistency analysis (Sec. C.1), robustness to mask inaccuracies, viewpoint differences, and sequential attribute editing in Sec. C.2, Sec. C.3, and Sec. C.4, followed by an ablation on the location of feature blending (Sec. C.5). A user study appears in Sec. D, followed by computational analysis in Sec. E and a discussion of limitations in Sec. G. Section F describes how we prepare text prompts for the text-based video-editing baselines. More visual results and comparisons with our baselines are available in the supplementary video. We *strongly* encourage readers to refer to it as well.

A Related Work on General Text-based Video Editing

Recent text-based video editing methods (Cong et al., 2024; Kara et al., 2024; Li et al., 2024; Yang et al., 2024; Wang et al., 2025) enable high-level video manipulation via language prompts. However, these approaches struggle with fine-grained semantic control, particularly in tasks like eyeglass placement, due to the inherent ambiguity in textual descriptions. Our work adopts a reference-based approach for precise and consistent semantic editing across video frames.

B Implementation Details

We implement our method on the DiffAE FFHQ256 model (Preechakul et al., 2022). As preprocessing, we align all facial videos and reference images similar to the FFHQ dataset (Karras et al., 2019), except we pad black pixels to the missing background and crop them into the resolution of 256^2. We use RAFT (Teed & Deng, 2020) to estimate the optical flow of each video. For our main experiment, we generate videos of 120 frames at 30 fps using a batch size of 120. In generating a blended mask for stochastic latent, we use a Gaussian blur with kernel size $(25, 25)$ and sigma value $(25, 25)$. We set $\beta = 1, \gamma = 1$ for eyeglasses and $\beta = 0, \gamma = 2$ for sunglasses.

Our system includes several tunable parameters, such as the temporal attention configuration and the number of denoising steps, that can be adjusted to further suppress minor artifacts. In practice, users can choose these settings to match their needs, which trades computational cost for additional visual refinement.

C Further experiments

C.1 Temporal Consistency Analysis

We compare the temporal consistency of edited videos with OmniTry and VideoEditGAN in Fig. A. The edited eyeglasses of OmniTry change between frames because it does not aim to maintain temporal consistency. While VideoEditGAN generates temporally consistent eyeglasses, it does not match the reference closely due to the language ambiguity involved in text-based control. Additionally, facial expressions are also altered. In contrast, our approach successfully replicates the reference eyeglasses and maintains temporal consistency across frames. This is also reflected in our quantitative results, where we achieve the state-of-the-art scores in temporal consistency. Please refer to the supplementary video for a full comparison.

Figure A: **Visual results** on temporal consistency with OmniTry (Feng et al., 2025) and VideoEditGAN (Xu et al., 2022).

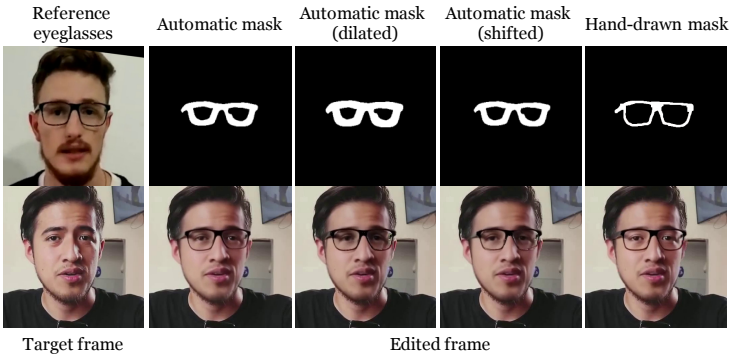

Figure B: We evaluate **the sensitivity of our method to imperfect reference masks**. From left to right: the reference eyeglasses, the automatic mask, a dilated version of the mask (expanded by 2 px), a shifted version (translated 5 px to the right), and a rough hand-drawn mask. The automatic mask is generated by our mask preparation pipeline using Grounded-SAM on the aligned reference frame.

Figure C: **Reference-target viewpoint mismatch.** Examples where the reference eyeglasses are captured from a noticeably different viewpoint than the target frames (top: left-facing reference vs. right-facing target; bottom: pitch-angle mismatch).

C.2 Robustness to Mask Inaccuracies

Since our method relies on a 2D reference mask to localize the eyeglasses region, we assess its sensitivity to imperfect masks. We use the same automatic mask generated by our reference-mask segmentation pipeline and evaluate three representative perturbations: (1) dilation (+2 px), (2) spatial misalignment (5-px right shift), and (3) rough hand-drawn masks. As shown in Fig. B, the edited results remain stable across all cases. The transferred eyeglasses maintain correct geometry and placement without introducing noticeable artifacts. This demonstrates that our mask-only blending strategy tolerates moderate boundary errors and does not require pixel-accurate masks.

C.3 Robustness to Reference-Target Viewpoint Mismatch

We test our method when the reference eyeglasses are captured from a noticeably different viewpoint than the target frames (*e.g.*, left-facing reference vs. right-facing target). Despite the large pose gap, our method can still place the eyeglasses with correct orientation and produce visually coherent edits (Fig. C).

C.4 Sequential Editing

To evaluate whether our method can compose multiple attribute edits, we apply it sequentially to different facial regions. As shown in Fig. D, we first transfer a reference nose and then apply eyeglass transfer to the edited frames. The two edits remain stable and well aligned, indicating that the method supports clean multi-attribute editing through simple sequential application.

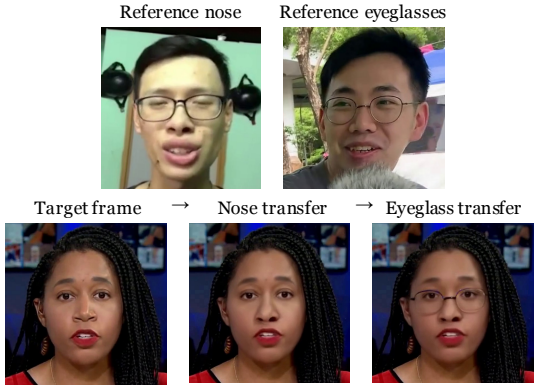

Figure D: **Sequential attribute editing.** We apply our method twice: first applying a nose transfer, followed by an eyeglass transfer to the edited result.

C.5 Ablation Studies

We conduct another ablation study on the location where our feature blending strategy is applied. We apply our proposed feature-blending strategy across different components of the DiffAE: Conditional DDIM only, stochastic encoder only, conditional DDIM + semantic encoder, and semantic encoder only (Ours). Figure E shows visual results, where *Sto. Enc.* refers to the stochastic encoder. *DDIM* refers

Table A: Criteria used in our user study.

Criterion	Question
Identity Preservation	Rank the videos in order from the one that looks most like the same person as the ID image.
Eyeglass Preservation	Rank the videos in order from the one that looks most like the same eyeglasses compared to the reference eyeglasses image.
Eyeglass Semantic Fidelity	Rank the videos in order from the one where the eyeglasses appear most naturally worn (physically natural).
Temporal Consistency	Rank the videos in order from the one with the most temporal consistency (least flickering).
Overall Video Fidelity	Rank the videos in order from the one that looks most like a real video captured by a camera.

to conditional DDIM. *Sem. Enc.* refers to the semantic encoder. Applying feature blending in conditional DDIM results in unwanted artifacts and a copy-and-paste effect, *i.e.*, there are no arms for the transferred eyeglasses. Applying only to the stochastic encoder does not show successful editing results. Using feature blending in the semantic encoder achieves semantic edits.

D User Study

We conduct a user study to complement automatic metrics, which cannot fully capture human preferences. We ask 20 participants to rank our edited videos and the baseline results using three representative baselines: MimicBrush, VideoEditGAN, and RAVE. We randomly pick 10 videos from our evaluation benchmark and ask our participants to rank according to five criteria (1) Identity Preservation, (2) Eyeglass Preservation, (3) Eyeglass Semantic Fidelity, *i.e.*, is the eyeglasses transfer semantically correct, (4) Temporal Consistency, and (5) Overall Video Fidelity. We provide the details of the criteria used in our user study in Table A. Table B reports the results of our user study. Our method consistently achieves the best or second-best rankings in all criteria. MimicBrush performs best in eyeglass preservation but falls short in other criteria, such as eyeglass semantic fidelity and overall video fidelity. VideoEditGAN achieves satisfying results in eyeglass semantic fidelity but ranks the lowest in eyeglass preservation. These findings highlight the strengths of our approach: while baseline methods exhibit significant weaknesses in specific criteria, our method achieves robust performance across all metrics, making it the most well-balanced solution for eyeglass transfer in facial videos.

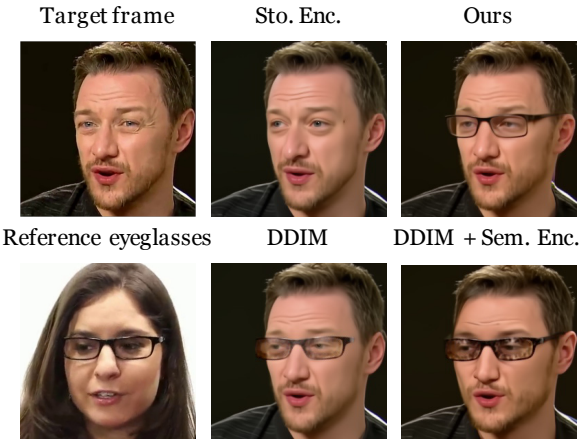

Figure E: **Ablations** on applying the feature blending in different components of DiffAE.

E Computational Analysis

We include a detailed computational time analysis for a 120-frame video on an NVIDIA A100 GPU in Table C. Our results show that the inference time of our method is comparable to that of the baseline methods, with no significant additional overhead, as it imposes no extra computational cost on the diffusion generative process.

Table B: **User study** results. We show the number of first-place rankings, mode, median, and average rank for each criterion. *Sem.* stands for semantic. **Bold** represents the best, and underline represents the second-best results.

Proportion of First-Place Rankings ↑				
Method	MimicBrush Chen et al. (2024a)	VideoEditGAN Xu et al. (2022)	RAVE Kara et al. (2024)	Ours
ID Preservation	0.205	0.300	0.145	**0.350**
Eyeglasses Preservation	**0.370**	0.170	0.215	0.245
Eyeglasses Sem. Fidelity	0.155	**0.355**	0.210	0.280
Temporal Consistency	0.185	0.320	0.140	**0.355**
Overall Video Fidelity	0.145	0.340	0.135	**0.380**

Mode Rank ↓				
Method	MimicBrush Chen et al. (2024a)	VideoEditGAN Xu et al. (2022)	RAVE Kara et al. (2024)	Ours
ID Preservation	3	**1**	4	**1**
Eyeglasses Preservation	**1**	4	4	3
Eyeglasses Sem. Fidelity	4	**1**	3	**1**
Temporal Consistency	3	2	4	**1**
Overall Video Fidelity	4	**1**	4	**1**

Median Rank ↓				
Method	MimicBrush Chen et al. (2024a)	VideoEditGAN Xu et al. (2022)	RAVE Kara et al. (2024)	Ours
ID Preservation	3	**2**	3	**2**
Eyeglasses Preservation	3	3	**2**	**2**
Eyeglasses Sem. Fidelity	3	**2**	3	**2**
Temporal Consistency	3	**2**	3	**2**
Overall Video Fidelity	3	**2**	3	**2**

Average Rank ↓				
Method	MimicBrush Chen et al. (2024a)	VideoEditGAN Xu et al. (2022)	RAVE Kara et al. (2024)	Ours
ID Preservation	2.585	2.310	2.890	**2.215**
Eyeglasses Preservation	**2.170**	2.780	2.640	2.410
Eyeglasses Sem. Fidelity	2.930	**2.220**	2.490	2.360
Temporal Consistency	2.655	2.205	2.970	**2.170**
Overall Video Fidelity	2.920	2.150	2.820	**2.110**

Table C: **Computational time analysis**. We compare the inference time taken to edit a 120-frame video on an NVIDIA A100 80GB GPU. Time is measured in seconds.

Method	Time (s)	Method	Time (s)
ObjectStitch (Song et al., 2023)	657	DVAE Classifier (Kim et al., 2023)	721
TF-ICON (Lu et al., 2023)	7505	DVAE Text (Kim et al., 2023)	7798
Paint-by-Example (Yang et al., 2023a)	368	VideoEditGAN (Xu et al., 2022)	3846
Anydoor (Chen et al., 2024b)	826	FLATTEN (Cong et al., 2024)	2603
MimicBrush (Chen et al., 2024a)	583	RAVE (Kara et al., 2024)	1157
OmniTry (Feng et al., 2025)	7444	VidToME (Li et al., 2024)	633
VACE (Jiang et al., 2025)	412	RF-Solver-Edit (Wang et al., 2025)	899
FreeEyeglass (Ours)	392	FRESCO (Yang et al., 2024)	1063

F Text Prompt Preparation

We use GPT 4o (OpenAI, 2024), API version 2024-05-13, to generate fine-grained descriptions of reference eyeglasses for text-based editing. We generate a fine-grained description for each reference eyeglasses and the target video, respectively. This section provides the details of generating text descriptions. We generate a fine-grained description for each reference eyeglasses and the target video, respectively. We specify our instructions with system prompts as listed in Table D. For the descriptions of target videos, we use the first frame of each video during generation. We prepare the input prompts by concatenating the descriptions of the target videos with those of the reference eyeglasses. We also impose character limits: a maximum of 100 characters for target video descriptions and 150 for reference eyeglass descriptions. To encourage photorealistic outputs, we prefix the concatenated prompts with the term *Realistic* (except for the Diffusion Video Autoencoder (Kim et al., 2023)). These prompts are then used as input for all text-based editing baselines. For FLATTEN (Cong et al., 2024), following their default configurations, we also use a negative prompt, "A person not wearing eyeglasses, deformed." Despite augmenting the prompt with the word *Realistic* and employing a negative prompt, we observe that FLATTEN frequently produced videos with a portrait drawing style rather than a photorealistic style. We suspect this tendency is due to FLATTEN's use of the Stable Diffusion 2.1 model, whereas other methods, RAVE (Kara et al., 2024) and VidToMe (Li et al., 2024), utilize the widely adopted Stable Diffusion 1.5 model.

Table D: System prompts and user prompts used for generating fine-grained descriptions.

System Prompt for **Eyeglasses** Descriptions
Task description: You will receive a facial image that is wearing a pair of eyeglasses. You must give a detailed description of the eyeglasses. Instructions: Step 1. Look at the image and locate where is the eyeglasses. Step 2. State the color of the frame and also the color of the lens. Determine the shape, type, and material of the frame. If the frame is half-rim, state whether the top or bottom is rimless. If the frame has a full rim, state whether the frame is thin or thick. If there are any distinctive features of the eyeglasses, mention it also, if no just skip it. Step 3. Output into a text description that only describes the eyeglasses you see. Tips: The eyeglasses may be an eyeglasses or a sunglasses. The shape of the eyeglasses frame includes square, boston, wellington, round, oval, aviator, browline, fox, crown panto, and geometric. The fox frame is equivalent to the cateye frame. The type of eyeglasses frame describes how the frame covers the lens and can be categorized into the full rim, half rim, rimless, and under-rim. Output: Do not output any description that is not about the eyeglasses in the image. The final output should only focus on the eyeglasses.
User Prompt for **Eyeglasses** Descriptions
Here is the image.
System Prompt for **Target Video** Descriptions
Task description: You will receive a facial image. You must give a detailed description of the image.
User Prompt for **Target Video** Descriptions
Here is the image.

G Limitations and Future Work

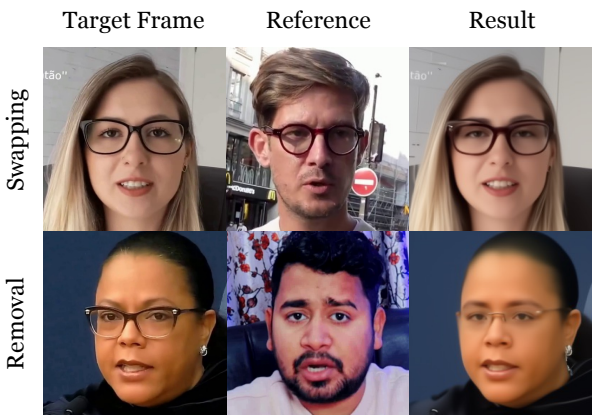

Figure F: Failure cases of our method.

We illustrate common failure modes of our method in Fig. F. One challenge for the swapping scenario arises when the target frame contains thick or visually dominant original glasses. In such cases, our method can transfer the color of the reference eyeglasses but may leave behind the original shape (see top row). For eyeglass removal, we use an aligned face without eyeglasses as the reference. While most of the frame is removed, subtle parts such as the bridge and arms sometimes remain (see bottom row). Moreover, our method transfers the lens appearance from the reference without estimating target-scene lighting. Physically consistent reflections would require environment mapping or illumination modeling, which lie outside the scope of a training-free reconstruction framework. Existing baselines also do not synthesize illumination-consistent reflections (see Fig. 3 in the main paper). We regard reflection-aware generation as an orthogonal direction for future work.