DiffFERV: Diffusion-based Facial Editing of Real Videos

Xiangyi Chen[1] , **Han Xue**[2✉] , **Li Song**[1✉]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2]School of Computer Science and Technology, Donghua University
chenxiangyi@sjtu.edu.cn, xuehan@dhu.edu.cn, song_li@sjtu.edu.cn

Abstract

Face video editing presents significant challenges, requiring precise preservation of facial identity, temporal consistency, and background details. Existing methods encounter three major challenges: difficulty in achieving accurate facial reconstruction, struggles with challenging real-world videos and reliance on a crop-edit-stitch paradigm that confines editing to localized facial regions. In response, we introduce **DiffFERV**, a novel diffusion-based framework for realistic face video editing that addresses these limitations through three core contributions. (1) A *specialization stage* that extends large Text-to-Image (T2I) models' general prior to faces while retaining their broad generative capabilities. This enables robust performance on non-aligned and challenging face images. (2) *Temporal modeling*, implemented through two distinct attention mechanisms, complements the specialization stage to ensure joint and temporally consistent processing of video frames. (3) Finally, we present a holistic editing pipeline and the concept of *preservation features*, which leverages our model's enhanced priors and temporal mechanisms to achieve faithful edits of entire video frames without the need for cropping, excelling even in real-world scenarios. Extensive experiments demonstrate that DiffFERV achieves state-of-the-art performance in both reconstruction and editing tasks.

1 Introduction

Face video editing aims to modify specific attributes of a face in a video, such as age, gender, or hairstyle, while preserving the original facial identity, motion, and background. It has gained significant attention due to its applications in entertainment, virtual avatars, and content creation.

The advent of GANs [Goodfellow *et al.*, 2020], particularly StyleGAN [Karras *et al.*, 2021], has spurred progress in facial image editing through latent space manipulation [Shen *et al.*, 2020]. Despite their popularity, GAN-based methods face a critical drawback: the inability to accurately reconstruct the original face during GAN inversion [Abdal *et al.*, 2019]. Moreover, when extended to facial videos, GAN-based methods typically rely on per-frame editing followed by smoothing techniques [Yao *et al.*, 2021], and often suffer from limited temporal consistency. On the other hand, Diffusion Models [Ho *et al.*, 2020], which have surpassed GANs in generating high-quality and diverse images, have inspired a range of diffusion-based editing methods. Among them, Diffusion Video Autoencoders (DVA) [Kim *et al.*, 2023] targets Face video editing. It achieves improved reconstruction and editing performance over previous methods.

However, we identify three limitations of DVA and GAN-based methods. First, while GAN methods suffer from poor identity preservation, DVA also fails to maintain intricate facial details despite superior reconstruction ability. Second, existing methods struggle with challenging real videos, typically those with extreme poses or out-of-distribution styles. This is because they are trained on domain-specific datasets that generally lack diversity in real-world variations. Third, previous methods necessitate a crop-edit-stitch pipeline which leads to incapability in handling edits extending beyond the face and introduces risks of stitching artifacts or misalignment between edited face and background. This is because current methods are confined to editing only the facial region due to their reliance on well-aligned, face-centric training data. Fig. 1 showcases these drawbacks.

To address these challenges, we propose **DiffFERV**, a **Diff**usion-based **F**acial **E**diting method for **R**eal **V**ideos. Unlike previous approaches that rely on face-specific training data, DiffFERV leverages the rich generative priors of pre-trained Text-to-Image (T2I) models. We implement a *specialization stage*, where we fine-tune the denoising network on the facial domain while adopting prior preservation techniques. By doing so, we maintain and extend these robust priors to face editing. This stage lays the groundwork to overcome issues of poor generalizability on real-world data and the restrictive cropping paradigm. Furthermore, we complement the specialized network with *temporal modeling*. We leverage contextual frames and optical flow priors to integrate two attention mechanisms that ensure respectively local smoothness and global consistency across edited frames. Finally, we eliminate previous dependence on cropping and external predictors by proposing a holistic editing pipeline. We introduce the concept of *preservation features*: latent inversion features that encode facial details, motion, as well

Figure 1: Comparison on challenging scenarios. Left: man → woman. Right: young. For the left profile (extreme pose) video, baseline deviates from the person's identity and generates blurriness and stitching artifacts in the hair. For the right out-of-distribution video, baseline neglects original facial makeup and produces apparent misalignment between facial and background regions.

as background of the input video. By deliberately reusing them during sampling, our method achieves superior reconstruction and edit capability without requiring facial cropping. Extensive evaluations demonstrate that DiffFERV excels in preserving facial identity, ensuring temporal consistency, especially when handling challenging real-world data. DiffFERV sets a new benchmark for robust, generalizable, and high-quality face video editing. The code is available at https://github.com/MunchkinChen/DiffFERV.

Contributions (1) We successfully adapt pretrained general T2I models to the specialized task of face editing, enabling robust handling of real-world complex scenarios. (2) We equip the specialized image-based model with temporal modeling, ensuring temporally consistent edits. (3) We leverage the rich diffusion latent features and propose a holistic editing pipeline that eliminates the need for face-centric cropping while guaranteeing preservation of motion, background, and facial details. (4) Through extensive qualitative and quantitative experiments, we demonstrate the superiority of DiffFERV over existing GAN- and diffusion-based baselines.

2 Related works

2.1 Face Image Editing

Advances in Generative Adversarial Networks [Goodfellow *et al.*, 2020] have inspired a plethora of methods for facial image editing. They aim to disentangle and manipulate GAN's rich latent space. Some explore interpretable directions through linear methods such as hyperplane separation [Shen *et al.*, 2020]. Others model non-linear transformations with parameterized networks [Yao *et al.*, 2021]. Some [Patashnik *et al.*, 2021] leverage CLIP [Radford *et al.*, 2021] to optimize latent codes towards open-vocabulary semantic priors. These works require a pre-editing stage of GAN inversion, either optimization- [Abdal *et al.*, 2019] or encoder-based [Tov *et al.*, 2021]. However, these inversion techniques frequently struggle to accurately preserve facial identity, representing a bottleneck for GAN-based methods.

Recent progress in Diffusion Models [Ho *et al.*, 2020] has also driven development of diffusion-based face image editing methods. Most works formulate a face generation process conditioned on guidance features such as semantic embeddings [Preechakul *et al.*, 2022], segmentation masks [Huang *et al.*, 2023], or even aligned StyleGAN latents [Li *et al.*, 2024]. Editing is then addressed by altering the disentangled condition. Per-subject tuning with customization techniques is frequently employed [Lin, 2024] to preserve the original facial identity. Another parallel work, FADING [Chen and Lathuilière, 2023], proposes an additional attribute-aware tuning strategy for pretrained T2I models and then performs diffusion image editing. This specialization-editing approach is similar to ours but is limited to age transformations.

2.2 Face Video Editing

Face video editing (FVE) methods typically extend image baselines with varying scales of temporal consistency applied at different stages. For example, Latent Transformers (Lattrans) [Yao *et al.*, 2021] employ optical-flow-aware cropping, per-frame editing, and Poisson blending for stitching. STIT [Tzaban *et al.*, 2022] hypothesizes an inherently smooth manifold of inversion encoders and enhances global consistency by tuning the generator. TCSVE [Xu *et al.*, 2022b] further optimizes latent codes with explicit temporal guidance. Diffusion Video Autoencoders (DVA) [Kim *et al.*, 2023] is the first to use Diffusion Models for FVE. It extends [Preechakul *et al.*, 2022] to videos by conditioning the diffusion process on facial identity and motion landmarks features. Note that when handling real-world videos, all these methods must first preprocess by cropping and aligning the facial area.

2.3 Diffusion-based Generic Video Editing

Early methods [Wu *et al.*, 2023; Liu *et al.*, 2024] perform one-shot tuning and generate new edits with the overfitted network. Another common paradigm involves first inverting videos into initial noise, then sampling the edited results. Optimization-based inversion techniques [Mokady *et al.*, 2023] are usually employed to ensure accurate reconstruction [Jeong and Ye, 2024]. Other approaches propagate edits from selected anchor frame(s) via feature fusion [Yang *et al.*, 2023] or correspondences matching [Geyer *et al.*,

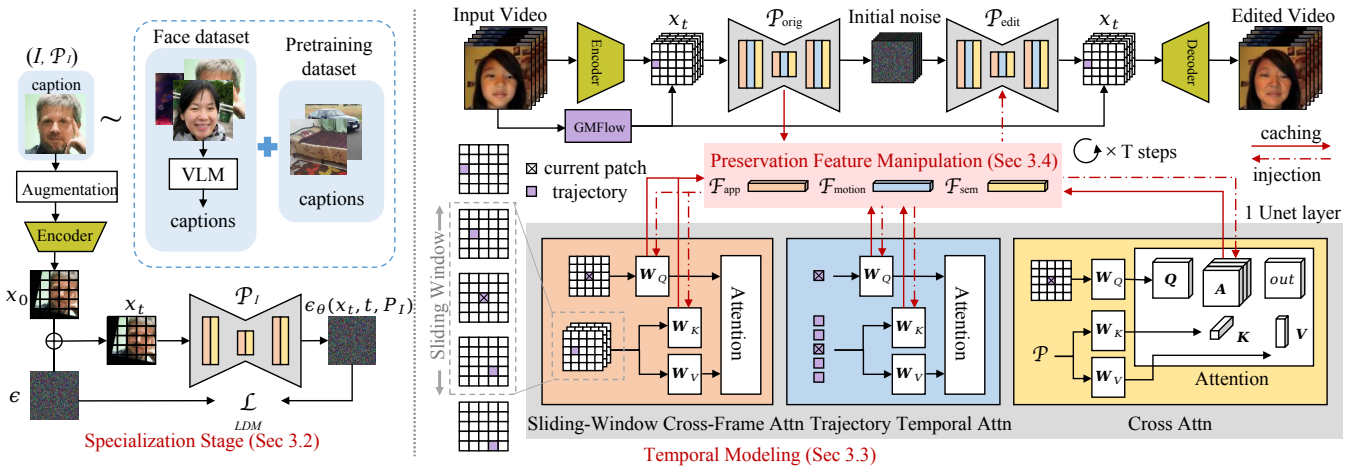

Figure 2: Overview of DiffFERV. Left: *Specialization Stage* (Sec 3.2) where pretrained model's generative priors are extended to face domain. Right: network architecture with *Temporal Modeling* (Sec 3.3) and holistic editing based on *Preservation Feature Manipulation* (Sec 3.4).

2024]. Some works also use auxiliary structural guidance from the original video, such as depth maps, edges [Yang *et al.*, 2024], or optical flow [Cong *et al.*, 2024; Liang *et al.*, 2024]. These generic methods focus more on global style changes or object swaps and do not address details such as facial identity and background, thus underperforming face experts in FVE.

3 Methodology

In this section, we provide a comprehensive description of DiffFERV's methodology. Section 3.1 introduces the necessary preliminaries. Section 3.2 describes the specialization stage, which extends a general generative model to the facial domain. Section 3.3 elaborates on the techniques employed to endow the specialized model with temporal modeling capabilities. Finally, Section 3.4 presents our holistic editing framework, which utilizes preservation features to achieve precise reconstructions and faithful modifications. Fig. 2 provides an overview of the proposed pipeline.

3.1 Preliminaries

Diffusion Probabilistic Models [Ho *et al.*, 2020] learn to approximate a data distribution by reversing a Markovian noise corruption process. It is composed of a forward and a reverse process. The forward process is a Gaussian noise perturbation to data point $\mathbf{x}_0$ over T timesteps:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (1)$$

with $\bar{\alpha}_t$ the noise schedule. The reverse process learns to denoise step by step through a parameterized model $\epsilon_\theta(\mathbf{x}_t, t)$:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}) \quad (2)$$

where $\mu_\theta(\mathbf{x}_t, t)$ is mean function and σ_t variance term.

Latent Diffusion Models (LDM) [Rombach *et al.*, 2022a] are a type of Diffusion Models that operate in the latent space of an image auto-encoder [Kingma, 2013] $\mathcal{D}(\mathcal{E}(\cdot))$ to achieve lower computation complexity. Our work is based on the publicly available Stable Diffusion. In particular, it adopts a U-Net architecture for the denoising network $\epsilon_\theta(\mathbf{x}_t, t, \psi(\mathcal{P}))$, where the generation is conditioned on text prompt $\mathcal{P}$ encoded by text encoder $\psi(\cdot)$.

Attention mechanism [Vaswani, 2017] is a key component in ϵ_θ. It computes the relationship between query, key, and value representations.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{AV} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key, and value matrices and $\mathbf{A}$ is the attention map. In Stable Diffusion, each U-Net layer contains a self-attention and a cross-attention block. Self-attention captures dependencies within the image features: $\mathbf{Q} = \mathbf{x}\mathbf{W}_Q, \mathbf{K} = \mathbf{x}\mathbf{W}_K, \mathbf{V} = \mathbf{x}\mathbf{W}_V$ where $\mathbf{x}$ is the latent image feature and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ learned projectors. Cross-attention, on the other hand, computes the key and value from $\psi(\mathcal{P})$ to integrate text conditions into image features.

3.2 Specializing a General T2I Model for Face Editing

Pretrained T2I models [Rombach *et al.*, 2022a] are trained on massive text-image datasets [Schuhmann *et al.*, 2022].They inherently possess the visual diversity and semantic richness needed to manage complex, real-world scenarios. We aim to harness this rich generative prior to enhance face editing, which is currently limited to well-aligned faces and struggles with complex cases. To this end, we introduce a specialization stage to extend a general T2I model to handle faces more proficiently. This stage is illustrated on the left side of Fig. 2. It involves fine-tuning the T2I model's denoising network $\epsilon_\theta(\mathbf{x}_t, t, \psi(\mathcal{P}))$ with a high-quality face image dataset. For each image I, we use a Vision Language Model (VLM) to generate a descriptive text prompt $\mathcal{P}_I$. The model is fine-tuned on the curated image-text pairs using the Latent Diffusion Loss [Rombach *et al.*, 2022a].

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{x}_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\text{Uniform}(1,T)}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \psi(\mathcal{P}_I))\|_2\right] \tag{4}$$

To retain the broad generative priors established during pretraining—critical for handling diverse real-world scenarios—we carefully design the specialization process to avoid catastrophic forgetting. Specifically, we incorporate a portion of the pretraining data into the fine-tuning data, maintaining a balance between learning face-specific features and retaining the model's original versatility. To mitigate overfitting on cropped, centered faces, we apply zooming and rotation augmentations to the training face data. As a result, the specialization stage establishes the foundation to address the previous limitations of poor generalizability to real-world data and the restriction to processing only cropped and aligned faces.

3.3 Temporal Modeling

After the specialization stage, our specialized T2I model can handle proficiently diverse facial images. However, directly applying it to individual video frames with image editing techniques introduces temporal inconsistencies.

To address this, we extend the model into a spatial-temporal architecture for consistent cross-frame editing. We first add a pseudo-temporal channel to the original 2D convolutions. We then incorporate two spatial-temporal attention schemes to ensure respectively local temporal continuity and global temporal consistency. The two mechanisms are illustrated in the orange and blue blocks in Fig. 2.

Sliding-Window-based Cross-Frame Attention

Given a video of m frames, where the latent feature for frame i is denoted by $\mathbf{v}_i$, naive per-frame editing scheme computes self-attention with each frame attending to itself. Formally,

$$\mathbf{Q} = \mathbf{v}_i\mathbf{W}_Q, \mathbf{K} = \mathbf{v}_i\mathbf{W}_K, \mathbf{V} = \mathbf{v}_i\mathbf{W}_V \tag{5}$$

To extend this spatial-only self-attention to the temporal domain, some previous work [Jeong and Ye, 2024] adopt dense spatio-temporal attention where each frame attends to all other frames. Differently, we propose a more efficient Sliding-Window-based Cross-Frame Attention (SWCFA). In our approach, each frame attends only to a fixed number of its neighboring frames within a defined window length:

$$\begin{aligned}
\mathbf{Q}_{\text{SWCFA}} &= \mathbf{v}_i\mathbf{W}_Q \\
\mathbf{K}_{\text{SWCFA}} &= \left[\mathbf{v}_{\max(i-\frac{w}{2},1)}\cdots\mathbf{v}_{\min(i+\frac{w}{2},m)}\right]\mathbf{W}_K \\
\mathbf{V}_{\text{SWCFA}} &= \left[\mathbf{v}_{\max(i-\frac{w}{2},1)}\cdots\mathbf{v}_{\min(i+\frac{w}{2},m)}\right]\mathbf{W}_V
\end{aligned} \tag{6}$$

where $[\cdot]$ denotes concatenation and w the window size. SWCFA achieves efficient bidirectional temporal modeling and ensures local continuity in each temporal adjacency.

Trajectory-based Temporal Attention

While SWCFA ensures smooth transitions between adjacent frames, it falls short in capturing long-term temporal dependencies. To enhance global consistency, previous approaches include an additional fixed anchor frame in the cross-frame attention [Wu *et al.*, 2023]. This yields suboptimal results when there are discrepancies between the anchor frame and other frames. Others [Guo *et al.*, 2024] introduce new temporal layers that perform 1D attention along the temporal axis. Formally, for patch p on the i-th frame, and its feature $\mathbf{v}_{i,p}$:

$$\begin{aligned}
\mathbf{Q}_{\text{temp}} &= \mathbf{v}_{i,p}\mathbf{W}_Q \\
\mathbf{K}_{\text{temp}} &= [v_{1,p}\ldots v_{m,p}]\,\mathbf{W}_K \\
\mathbf{V}_{\text{temp}} &= [v_{1,p}\ldots v_{m,p}]\,\mathbf{W}_V
\end{aligned} \tag{7}$$

Despite being effective, this strategy necessitates extensive additional training of the temporal layers on video data.

Differently, we draw inspiration from recent works [Yang *et al.*, 2024; Cong *et al.*, 2024] that utilize optical flow priors to enforce temporal consistency and introduce Trajectory-based Temporal Attention (TTA), a method to enhance global consistency without additional training.

We first predict the optical flow of the input video to derive a set of temporal displacement trajectories. We follow the post-processing proposed by [Cong *et al.*, 2024] to ensure that each frame patch is uniquely assigned to a single trajectory. Temporal attention is then computed along these trajectories. In other words, each patch attends to all patches on the same temporal trajectory. For a given trajectory $\{p_1\ldots p_i\ldots p_m\}$ where p_i denotes the patch index on the i-th frame,

$$\begin{aligned}
\mathbf{Q}_{\text{TTA}} &= \mathbf{v}_{i,p_i}\mathbf{W}_Q \\
\mathbf{K}_{\text{TTA}} &= [v_{1,p_1}\ldots v_{m,p_m}]\,\mathbf{W}_K \\
\mathbf{V}_{\text{TTA}} &= [v_{1,p_1}\ldots v_{m,p_m}]\,\mathbf{W}_V
\end{aligned} \tag{8}$$

TTA leverages the natural motion prior of the input video to aggregate content efficiently along the entire temporal axis, thereby enhancing global temporal consistency.

3.4 Holistic Editing via Preservation Feature Manipulation

Given a real face video and a desired editing direction, FVE aims to achieve accurate, consistent edits while preserving the facial identity, motion, and background information. Existing methods rely on a crop-edit-stitch approach, depending on external face recognizors and landmark detectors to preserve these relevant information. In contrast, we propose leveraging *preservation features* that are inherently encoded in the intermediate features during diffusion inversion. These features retain the input video's facial details, motion, as well as background information, thus eliminating the need for cropping or external predictors.

Inversion with Preservation Feature Caching

A common paradigm of diffusion-based editing is to first invert an image or video with DDIM inversion [Song *et al.*, 2021] and then begin editing from the inverted noise. However, DDIM inversion leads to inaccurate reconstructions of real videos due to accumulated errors amplified by classifier-free guidance [Mokady *et al.*, 2023]. Such inaccuracies are particularly problematic for face editing tasks, where fine-grained preservation of original visual details is crucial.

The multi-step denoising process of Diffusion Models generates intermediate features across timesteps, which can be viewed as a high-dimensional *latent space*. (Note that this *latent space* refers not to the VAE latent space where LDM operates, but to the union of intermediate network features

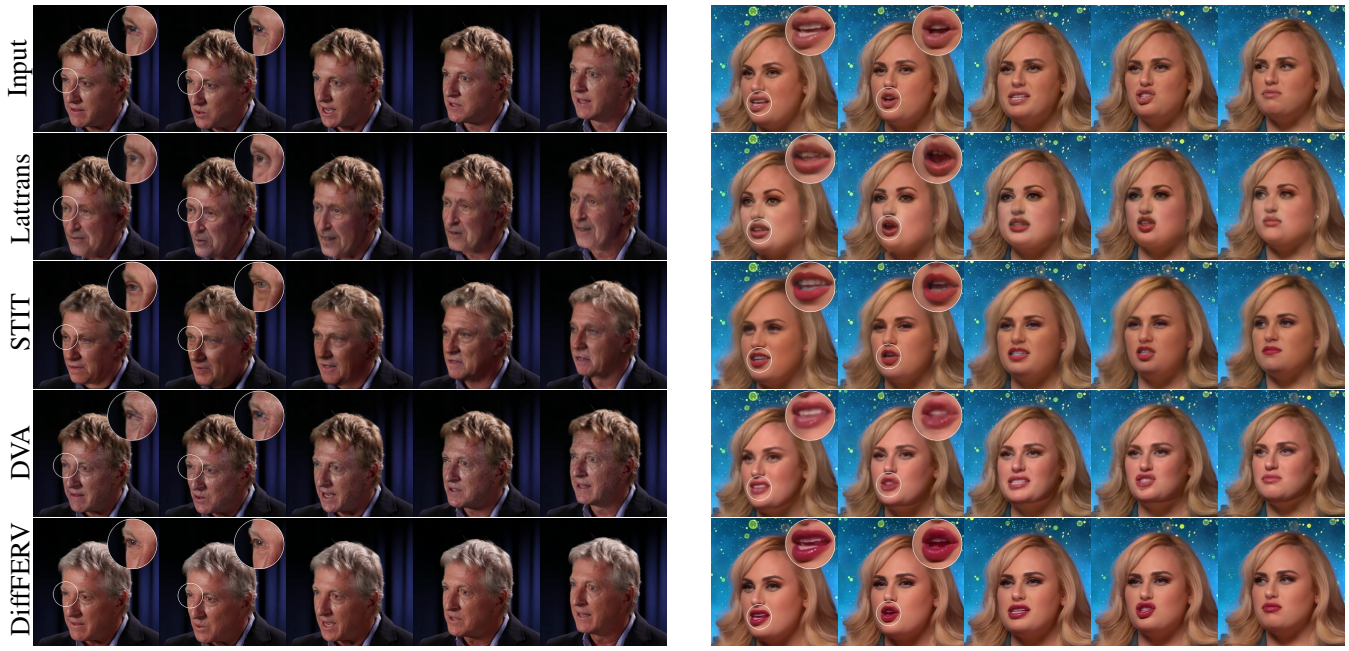

Figure 3: Comparison of global edit results (+age) and local edit results (+lipstick). Lattrans and STIT produce inconsistent identities, while DVA struggles with detail preservation and introduces artifacts. DiffFERV preserves facial details accurately and achieves realistic edits.

across timesteps.) We hypothesize that this latent space encodes both fine-grained facial information and background details. By identifying and leveraging these features, we can (1) accurately reconstruct the original content and (2) perform high-fidelity edits without cropping (since the background is also encoded in these features). We refer to these features as *preservation features*.

Our hypothesis aligns with findings in prior studies on image generation [Tumanyan *et al.*, 2023], which demonstrate that intermediate attention maps during the generation process contain detailed spatial information about the generated content. On top of these insights, we further claim—and empirically validate—that (1) intermediate attention features during *inversion* encode fine-grained spatial details of the original real video, and (2) when inverting with our model with temporal modeling proposed in Section 3.3, these features additionally maintain inter-frame correspondence, preserving consistency and motion in the input video.

Based on these two findings, we perform T-step DDIM inversion on the input video using our temporally enhanced specialized model and a text description $\mathcal{P}_{\text{orig}}$ of the input video. During the inversion process, we identify three distinct types of preservation features and cache them at each diffusion step t.

Appearance Features We define appearance features $\mathcal{F}_{\text{app}}$ as the key and query embeddings $\mathbf{Q}_{\text{SWCFA}}, \mathbf{K}_{\text{SWCFA}}$ in SWCFA (Equation 6), which capture fine-grained spatial details of both facial (therefore identity) and background information. These features are also temporally aware due to SWCFA's cross-frame modeling nature.

$$\mathcal{F}_{\text{app}} = \left[\mathbf{Q}^{\text{inv}}_{\text{SWCFA},t}, \mathbf{K}^{\text{inv}}_{\text{SWCFA},t}\right]_{t=1\ldots T}$$

Motion Features We define motion features $\mathcal{F}_{\text{motion}}$ as the key $\mathbf{Q}_{\text{TTA}}$ and query $\mathbf{K}_{\text{TTA}}$ in TTA (Equation 8), which preserve the temporal feature correspondence along trajectories across the entire video, thereby retaining motion.

$$\mathcal{F}_{\text{motion}} = \left[\mathbf{Q}^{\text{inv}}_{\text{TTA},t}, \mathbf{K}^{\text{inv}}_{\text{TTA},t}\right]_{t=1\ldots T}$$

Semantic Features As demonstrated by [Hertz *et al.*, 2023], cross-attention maps $\mathbf{A}_{\text{cross}}$ encode spatial correspondence between image patches and text semantics. Retaining them during inversion helps preserve the original semantic layout. Therefore, we utilize them as our semantic features,

$$\mathcal{F}_{\text{sem}} = \left[\mathbf{A}^{\text{inv}}_{\text{cross},t}\right]_{t=1\ldots T}$$

$\mathcal{F}_{\text{app}}, \mathcal{F}_{\text{motion}}, \mathcal{F}_{\text{sem}}$ then serve as a foundation for reconstructing the original video during edit sampling.

Sampling with Preservation Feature Injection

We start the editing sampling with the initial noise obtained from DDIM inversion. We use a new text prompt $\mathcal{P}_{\text{edit}}$ that specifies the desired changes. The cached $\mathcal{F}_{\text{app}}, \mathcal{F}_{\text{motion}}, \mathcal{F}_{\text{sem}}$ are incorporated during sampling to recover the input video's identity, motion, background as well as semantic layout.

At each sampling step t, cached $\mathcal{F}_{\text{app}}[t] = (\mathbf{Q}^{\text{inv}}_{\text{SWCFA},t}, \mathbf{K}^{\text{inv}}_{\text{SWCFA},t})$ and $\mathcal{F}_{\text{motion}}[t] = \left(\mathbf{Q}^{\text{inv}}_{\text{TTA},t}, \mathbf{K}^{\text{inv}}_{\text{TTA},t}\right)$ are injected into the sampling process by overriding their counterparts $\left(\mathbf{Q}^{\text{edit}}_{\text{SWCFA},t}, \mathbf{K}^{\text{edit}}_{\text{SWCFA},t}\right)$ and $\left(\mathbf{Q}^{\text{edit}}_{\text{TTA},t}, \mathbf{K}^{\text{edit}}_{\text{TTA},t}\right)$ in the SWCFA and TTA attentions. For $\mathcal{F}_{\text{sem}}[t] = \mathbf{A}^{\text{inv}}_{\text{cross},t}$, we adopt the strategy from [Hertz *et al.*, 2023]: for text tokens shared between $\mathcal{P}_{\text{orig}}$ and $\mathcal{P}_{\text{edit}}$, the cross-attention maps in the editing path $\mathbf{A}^{\text{edit}}_{\text{cross}}$ are replaced with the cached maps $\mathbf{A}^{\text{inv}}_{\text{cross}}$ to retain the original semantic layout. Otherwise, cross-attention

Figure 4: Comparison of reconstruction results. Note that DiffFERV is the only method to successfully reconstruct the person's necklace.

Model	MSE ↓	SSIM ↑	LPIPS ↓
psp	0.070	0.701	0.140
e4e	0.086	0.662	0.182
PTI	0.055	0.758	0.138
DVA	**0.010**	0.983	0.017
DiffFERV	**0.010**	**0.985**	**0.008**

Table 1: Comparison of reconstruction metrics

maps for novel words in the editing prompt are preserved in the editing path. The preservation feature caching and injection are depicted on the right side of Fig. 2.

We notice that the features captured at different timesteps during inversion exhibit varying levels of granularity: larger t focus on low-level details like textures, while smaller t emphasize higher-level structural components. To balance fidelity and editing effectiveness, we introduce a timestep threshold τ_{app}: $\mathcal{F}_{\text{app}}[t]$ is injected only when $t > (1 - \tau_{\text{app}})T$. We use a higher τ_{app} for texture-level edits (e.g., mild age changes, hair color adjustments) and lower τ_{app} for shape-altering edits (e.g., gender changes, hairstyle modifications).

Note that our method addresses edits effectively without over-aligning to the original video. This is because the editing guidance specified in $\mathcal{P}_{\text{edit}}$ is injected via the cross-attention values $\mathbf{V}^{\text{edit}}_{\text{cross}}$, which are untouched during preservation feature manipulation. Additionally, the value vectors in the spatio-temporal attentions $\mathbf{V}^{\text{edit}}_{\text{SWCFA}}, \mathbf{V}^{\text{edit}}_{\text{TTA}}$ also remain unaltered. This essentially allows the original structure and motion to guide the aggregation of new semantic edits.

4 Experiments

4.1 Implementation Details

For specialization, we initialize with the pretrained weights of Stable Diffusion 1.5 [1]. We utilize the FFHQ dataset [Karras *et al.*, 2019] as our training dataset. We employ Pixtral [2] for automatic captioning. Within our dataset, we integrate 10% of image-text pairs sampled from the LAION-2B-en [Rombach *et al.*, 2022b] dataset . We opt for Adam [Kingma, 2014] optimizer with a batch size of 8 and a learning rate of $2.5e-6$.

For temporal modeling, we configure window length to $w = 3$ for SWCFA and leverage GMFlow [Xu *et al.*, 2022a] for optical flow prediction in TTA. During editing, we use DDIM [Song *et al.*, 2021] sampling and inversion with $T = 50$ timesteps. A negative prompt [Ban *et al.*, 2025] scheme is adopted, where the original prompt serves as the negative prompt to enhance editing effectiveness, with guidance scale set to 5. We use $\tau_{\text{app}} = 0.9$ for texture-level edits and $\tau_{\text{app}} = 0.7$ for shape-altering edits.

4.2 Evaluation Protocol

We evaluate DiffFERV on CelebV-HQ [Zhu *et al.*, 2022]. We include both reconstruction and editing tasks. We devise two protocols to evaluate editing performance. (1) Global editing: age manipulation (+−age), gender transformation (+−man), and emotion changes (+−smiling) (2) Local editing: hairstyle (+blond, +bang), makeup (+lipstick), and accessories (+glasses).

Metrics We employ MSE, SSIM [Wang *et al.*, 2004], and LPIPS [Zhang *et al.*, 2018] to evaluate reconstruction accuracy across multiple scales. For editing tasks, evaluation spans four dimensions: (1) **faithfulness**, assessed using Non-target Attribute Preservation Rate (NAPR) and Identity Preservation (IDP) scores [Yao *et al.*, 2021] (2) **effectiveness**, measured by Target Attribute Change Rate (TACR) [Yao *et al.*, 2021] (3) **temporal consistency**, using temporally-local (TL-ID) and temporally-global (TG-ID) identity preservation metrics [Tzaban *et al.*, 2022], and (4) **editing quality**, gauged by CLIP-Score [Wang *et al.*, 2023] for realism.

4.3 Comparisons with State-of-the-Art Methods

For reconstruction, we compare against GAN-based inversion methods psp [Richardson *et al.*, 2021], e4e [Tov *et al.*, 2021], PTI [Dong *et al.*, 2023] and diffusion-based DVA [Kim *et al.*, 2023]. In editing tasks, we benchmark DiffFERV against three state-of-the-art face video editing methods: Latent Transformers (Lattrans) [Yao *et al.*, 2021], STIT [Tzaban *et al.*, 2022], and Diffusion Video Autoencoders (DVA). We adhere to each method's respective official implementation, including the crop-edit-stitch process.

Qualitative Results

In Fig. 4, we present reconstruction results. All three GAN methods fail to preserve facial identity and lose background information. DVA and DiffFERV perform better in reconstruction, with DiffFERV excelling in detail preservation: it is the only to accurately recover the necklace detail.

Fig. 3 provides a visual comparison of global edits (+age) and local edits (+lipstick). Lattrans and STIT produce outputs that are visually different from the original identity, and create inconsistencies in identities across frames in the aging case. DVA achieves better identity consistency but still fails to maintain eye detail in the aging case and incorrectly reproduces the mouth shape in the lipstick case. Additionally, DVA introduces cropping artifacts, including unnatural jawline seams in the lipstick example. In contrast, our method preserves facial details with precision and achieves realistic edits. Notably, for the aging case, DiffFERV is the only method that addresses consistent changes even beyond the facial region, such as adding white hair, highlighting the benefits of our holistic editing approach.

Quantitative Results

Table 1 shows that DiffFERV achieves the highest scores across all reconstruction metrics. This validates our quali-

[1] https://huggingface.co/ruwnayml/stable-diffusion-v1-5

[2] https://huggingface.co/mistralai/Pixtral-12B-2409

Model	Faithfulness				Effectiveness	
	Global		Local		Global	Local
	IDP↑	NAPR↑	IDP↑	NAPR↑	TACR↑	TACR↑
Lattrans	0.515	**0.908**	0.602	0.868	0.829	**0.559**
STIT	0.512	0.887	0.536	0.893	0.845	0.457
DVA	0.559	0.839	0.641	0.851	0.834	0.305
DiffFERV	**0.563**	0.865	**0.677**	**0.915**	**0.870**	0.492

Table 2: Comparison of faithfulness and effectiveness metrics

Model	Temporal Consistency				Quality	
	Global		Local		Global	Local
	TL-ID↑	TG-ID↑	TL-ID↑	TG-ID↑	CLIP↑	CLIP↑
Lattrans	0.690	0.664	0.696	0.659	0.693	0.698
STIT	0.674	0.635	0.627	0.601	0.643	0.673
DVA	0.666	0.618	0.658	0.621	0.611	0.676
DiffFERV	**0.753**	**0.706**	**0.735**	**0.706**	**0.707**	**0.719**

Table 3: Comparison of temporal consistency and quality metrics

tative observation and proves that our proposed preservation feature manipulation improves reconstruction capability.

In Table 2, we present faithfulness and effectiveness metrics. DiffFERV achieves the overall best performance in faithfulness, with a notably higher identity preservation score compared to baselines. For NAPR and TACR, we observe a trade-off, as no single method dominates both metrics. While Lattrans achieves the highest global NAPR and local TACR and DiffFERV leads in global TACR and local NAPR, Lattrans exhibits significantly lower IDP, underscoring DiffFERV's superior overall performance. Table 3 highlights DiffFERV's large-margin improvement in temporal consistency and editing quality metrics. This proves the effectiveness of our specialization stage and temporal modeling, in ensuring high-quality and coherent edits.

4.4 Ablation Studies

Specialization Stage Table 4 and Fig. 5 present a comparison of results using the original SD1.5 weights versus our specialized model. We observe that specialization leads to substantial improvements in editing effectiveness and fidelity for global edits, although it results in a lower faithfulness metric for local edits. We posit that this discrepancy arises because the unspecialized model struggles to generate the necessary local changes effectively and produces outputs that closely resemble the original. Fig. 5 validates our hypothesis, proving that the specialized model excels in generating face-related concepts that were previously unmanageable.

Figure 5: Ablation of specialization stage (spec.)

Model	Global		Local		Global	Local
	IDP↑	NAPR↑	IDP↑	NAPR↑	TACR↑	TACR↑
w/o Spec.	0.545	0.864	**0.707**	**0.923**	0.845	0.442
DiffFERV	**0.563**	**0.865**	0.677	0.915	**0.870**	**0.492**

Table 4: Ablation of specialization stage (spec.)

Figure 6: Ablation of temporal modeling

Temporal Modeling We evaluate the contributions of SWCFA and TTA to temporal consistency and identity fidelity. Table 5 and Fig. 6 show that SWCFA is critical for maintaining local temporal continuity, while TTA further enhances global consistency. Interestingly, this temporal context also improves faithfullness and editing success rate, as observed in Fig. 6. We hypothesize that this occurs because the cross-frame awareness and optical flow prior from the original video aids in managing challenging frames by supplying richer temporal contextual information.

Time Threshold for Appearance Feature Caching Fig. 7 displays the results of aging edits at varying τ_{app} thresholds. As τ_{app} increases, edits align more closely with the original face but exhibit weaker transformations. Users can freely adjust this parameter according to their need for a trade-off between editing effectiveness and faithfulness.

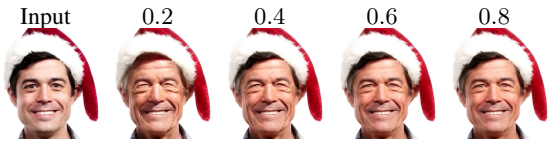

Figure 7: Comparison of editing results at different τ_{app}

SWCFA	TTA	Temporal Consistency				Faithfulness	
		Global		Local		Global	Local
		TL-ID↑	TG-ID↑	TL-ID↑	TG-ID↑	IDP↑	IDP↑
×	×	0.643	0.620	0.619	0.599	0.499	0.628
✓	×	0.721	0.680	0.693	0.648	0.520	0.639
✓	✓	**0.753**	**0.706**	**0.735**	**0.706**	**0.563**	**0.677**

Table 5: Ablation of temporal modeling

Acknowledgements

This work was partly supported by the Fundamental Research Funds for the Central Universities, the MoE-China Mobile Research Fund Project (MCM20180702) and National Key R&D Project of China (2019YFB1802701), Shanghai Key Laboratory of Digital Media Processing and Transmission under Grant 22DZ2229005, 111 project BP0719010.

References

[Abdal *et al.*, 2019] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[Ban *et al.*, 2025] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision*, pages 190–206. Springer, 2025.

[Chen and Lathuilière, 2023] Xiangyi Chen and Stéphane Lathuilière. Face aging via diffusion-based editing. *British Machine Vision Conference*, 2023.

[Cong *et al.*, 2024] Yuren Cong, Mengmeng Xu, christian simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*, 2024.

[Dong *et al.*, 2023] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023.

[Geyer *et al.*, 2024] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

[Hertz *et al.*, 2023] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020.

[Huang *et al.*, 2023] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2023.

[Jeong and Ye, 2024] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[Karras *et al.*, 2021] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, December 2021.

[Kim *et al.*, 2023] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kingma, 2014] DP Kingma. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*, 2014.

[Li *et al.*, 2024] Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Liang *et al.*, 2024] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Lin, 2024] Haonan Lin. Dreamsalon: A staged diffusion framework for preserving identity-context in editable face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Liu *et al.*, 2024] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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

[Preechakul *et al.*, 2022] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[Radford *et al.*, 2021] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

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

[Rombach *et al.*, 2022b] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[Shen *et al.*, 2020] 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*, 44(4):2004–2018, 2020.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[Tov *et al.*, 2021] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

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

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

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transa ctions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2023] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

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

[Xu *et al.*, 2022a] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

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

[Yang *et al.*, 2023] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[Yang *et al.*, 2024] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

[Zhang *et al.*, 2018] 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 the IEEE conference on computer vision and pattern recognition*, 2018.

[Zhu *et al.*, 2022] 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 *European conference on computer vision*, pages 650–667. Springer, 2022.