IP-FaceDiff: Identity-Preserving Facial Video Editing with Diffusion

Tharun Anand[1], Aryan Garg[2], Kaushik Mitra[1]
[1]Indian Institute Of Technology Madras, [2]University of Wisconsin Madison
{ed20b068, kmitra}@smail.iitm.ac.in, agarg54@wisc.edu

Figure 1. **Qualitative Facial Attribute Editing.** Our method enables precise, localized facial edits whilst maintaining low computational latency. Through explicit identity preservation optimization, we ensure that global facial identity features are retained. Furthermore, our approach ensures temporal consistency and robust generalization across a wide range of facial video editing tasks.

Abstract

Facial video editing has become increasingly important for content creators, enabling the manipulation of facial expressions and attributes. However, existing models encounter challenges such as poor editing quality, high computational costs and difficulties in preserving facial identity across diverse edits. Additionally, these models are often constrained to editing predefined facial attributes, limiting their flexibility to diverse editing prompts. To address these challenges, we propose a novel facial video editing framework that leverages the rich latent space of pre-trained text-to-image (T2I) diffusion models and fine-tune them specifically for facial video editing tasks. Our approach introduces a targeted fine-tuning scheme that enables high quality, localized, text-driven edits while ensuring identity preservation across video frames. Additionally, by using pre-trained T2I models during inference, our approach significantly reduces editing time by 80%, while maintaining temporal consistency throughout the video sequence. We evaluate the effectiveness of our approach through extensive testing across a wide range of challenging scenarios, including varying head poses, complex action sequences, and diverse facial expressions. Our method consistently outperforms existing techniques, demonstrating superior performance across a broad set of metrics and benchmarks.

1. Introduction

Digital content generation and editing have experienced remarkable strides since the advent of diffusion models [18, 19, 30, 43]. These generative models have been progressively trained to enhance input image clarity but often lack precise control over the output. However, recent advancements in conditional diffusion models [10, 32] have addressed this limitation, opening up exciting avenues for multi-modal or text-controlled generation [7, 11, 37, 38, 40]. These innovative approaches have demonstrated effective-

ness and improved quality across various tasks, including image generation [10, 15, 21], diverse image editing scenarios [4, 12, 17], and even video editing applications [6, 33, 36].

Facial video editing differs from conventional video editing, where changes are often applied to broader scenes. Our approach focuses on the unique challenges of facial editing, such as executing precise localized modifications, maintaining temporal consistency across frames, and preserving the original video's quality and subject's identity throughout the process. These challenges demonstrate why conventional video editing models are not suited for the precision required in facial video editing. Previous works in GAN-based [1, 46, 52] and Diffusion based [26] facial video editing faced similar hurdles. They often produced visible artifacts and also handled each frame independently, which significantly increased computation cost and editing time. Moreover, these techniques require end-to-end training tailored to specific facial videos and editing prompts, limiting their ability to generalize to diverse prompts and unseen head poses.

To tackle these challenges, we employ pre-trained Text-to-Image (T2I) diffusion models (Stable Diffusion 2.1 [35]), known for its generalization capabilities and high-quality image generation capability due to large-scale pre-training. Also using pretrained T2I models for Text2video tasks has proven to achieve significant reductions in editing time and compute latency [14, 49] .

While off-the-shelf (T2I) models provide strong generalization to unseen faces, they often struggle with precision when adhering to localized prompts and preserving identity [14, 49] Figure 1. To address this, we use directional-CLIP loss [13] to fine-tune one branch of our parallel T2I framework to enable it to make highly localized edits to the facial video. Additionally, we integrate an identity preservation loss [9], into one of our fine-tuned T2I branches. This explicitly guides the editing process during *inference*, focusing on maintaining the person's identity.

The combination of these two independently fine-tuned diffusers & careful loss setups enables our method to perform identity preserving localized edits to facial videos(Fig. 3).

In summary, we make the following contributions:

- We present a novel pipeline for facial video editing that leverages pre-trained text-to-image (T2I) diffusion models. This approach enables high quality prompt-consistent edits, generalizes effectively to challenging in-the-wild facial videos, and significantly reduces both computational cost and latency.
- We propose an inference-time guidance strategy that preserves the original identity throughout the editing process, ensuring global identity consistency across frames.
- We develop a fine-tuning strategy that allows for highly localized edits in facial videos and demonstrate that this framework significantly surpasses existing benchmarks for facial video editing in various quantitative and qualitative aspects.

2. Related Work

2.1. Text to Image Diffusion models

Text-guided diffusion models encode the rich latent features of an image coupled with the text guidance through cross-attention modules [11, 32, 37, 40]. ControlNet [54], LORA [20] and PNP-diffusion [45] delve into the controllability of visual generation by incorporating additional encoding layers, facilitating controlled generation under various conditions such as pose, mask, edge and depth. Notably [45] controls image generation by replacing the self-attention features of selective layers of the T2I network with another network that effectively preserves global features of the image and [54] provides guidance through zero convolutions. We take inspiration from these methods to facilitate additional identity guidance to our editing method.

2.2. Text guided video editing

Unlike advancements in image manipulation, progress in video manipulation methods has been slower, with a key challenge being the need to streamline editing processes for practical use, while ensuring temporal consistency across frames. Previous efforts [25, 33, 36] have tried to address these challenges but often at the expense of computational cost and impractical editing times.

To address this, recent advances [5, 14, 49] demonstrated using pre-trained T2I diffusion models results in faster video editing. In particular Geyer *et al*. [14] proposed using a T2I model with keyframe editing instead of editing every frame. Changes in these keyframes are propagated to other frames via interframe correspondences, significantly reducing editing time. Yao *et al*. [49] leveraged a pretrained (T2I) diffusion model and fine-tunes it with a small number of iterations using latent code initialization and temporal attention, significantly reducing video editing time.

2.3. Facial Video Editing

Facial video editing presents a unique challenge compared to traditional methods. Here, preserving the original person's identity and natural head movements is crucial, while maintaining temporal coherence throughout the edited video . Early methods [1–3, 13, 16, 46] leveraged StyleGAN's well-structured latent space for facial video editing. By manipulating specific codes, they could alter facial features like expressions, gender, and age. However, they were limited to broader edits and couldn't address more localized details like beards, glasses, or accessories.

To better preserve temporal consistency while allowing for more localized edits to the facial video, Kim *et al*. [26] proposed to use DDIM [43] to edit facial videos. This method first encodes the frames of a facial video into time-invariant (identity) and time-variant (motion) features using separate pre-trained encoders, then performs CLIP-guided editing with the text prompt to the time-invariant features. Then along with the noisy latents of the facial frames (obtained from the deterministic forward process of DDIM) and these edited features, they guide the reverse process of the conditional DDIM. But this involves a lot of individual preprocessing steps and processes every frame of the video during inference (Tab. 2). In contrast, our method tackles these challenges through a video editing framework that leverages off-the-shelf T2I models. To the best of our knowledge we are the first to explore using pretrained T2I models for facial video editing tasks.

3. Preliminary

3.1. Denoising Diffusion Probabilistic Models:

DDPMs [18,43] are generative models based on iterative Markovian steps: forward noising and reverse denoising. Given an input image y_0, the forward process progressively adds noise using a Markov chain q over T steps:

$$q(y_{1:T} \mid y_0) = \prod_{t=1}^{T} q(y_t \mid y_{t-1}) = \mathcal{N}(y_t \mid \sqrt{\alpha_t} y_{t-1}, (1-\alpha_t) I) \tag{1}$$

where α_t controls the added noise level at each step. The reverse process uses a U-Net $\epsilon(y_t, t)$ to predict and remove the noise at each step, progressively denoising y_t back to y_0. This process can be guided by additional input signals, such as text descriptions or reference images, allowing the model to generate outputs that are more specific and aligned with the given conditions. By incorporating these inputs, the generative process becomes highly flexible and can be tailored to produce desired outputs that match particular requirements or constraints.

We leverage a pre-trained text-conditioned Latent Diffusion Model (LDM), like Stable Diffusion [35] for our task, which applies the diffusion process to a pre-trained image autoencoder's latent space. The U-Net architecture uses a residual block, a self-attention block to capture long-range dependencies and a cross-attention block for image features's interaction with text embeddings from the conditioning prompt.

3.2. Joint Keyframe Editing:

Traditional video editing techniques necessitate editing each frame separately and then ensuring temporal coherence, leading to significant computational overhead. To mitigate this, Geyer et al. [14] exploit the temporal redundancies within the feature space of T2I diffusion models in natural videos. This approach enables the attainment of temporal coherence in editing by altering only a selected subset of keyframes and propagating these modifications throughout the video.

The keyframe k_i generates a feature representation f_i, followed by the calculation of an attention weight a_{ij} for k_i, indicating the relevance of information with another keyframe k_j as follows:

$$a_{ij} = \text{softmax}(f_i^T W_a f_j) \tag{2}$$

where W_a is a learned weight matrix. After computing attention weights, each keyframe's feature is updated by aggregating information from all other keyframes:

$$f_i^{\text{updated}} = f_i + \sum_j a_{ij} W_c f_j \tag{3}$$

where W_c is another learned weight matrix. Then For each frame i they proceed to find the nearest neighbors in the original video's keyframe feature space. Let γ_{i^+} and γ_{i^-} denote the indices of the closest future and past keyframes (respectively) to frame i. Edited keyframes features (T_{base}) are propagated to f_i based on its distance to neighboring keyframes:

$$f_i^{\text{prop}} = w_i^+ T_{\text{base}}(\gamma_i^+) + w_i^- T_{\text{base}}(\gamma_i^-) \tag{4}$$

where $T_{\text{base}}(\gamma_{ij})$ refers to the edited feature vector of the keyframe with index γ_{ij} (either future or past), and w_{i^+} and w_{i^-} are weights determined by the distance between frame i and its corresponding neighbors. This technique ensures consistent video editing by integrating style information from all keyframes simultaneously, fostering global-editing coherence.

4. Method

This section details our facial video editing approach, consisting of two main stages: preprocessing and editing. In the preprocessing phase, identity features are extracted from video frames using the pre-trained T2I model ϵ_1. For each frame x_i in the input video (with i ranging from 1 to N frames), a DDIM inversion process is applied with ϵ_1 to extract self-attention features x_l^i at every network layer l during denoising, as depicted in Figure 2. In the editing step, we adopt joint keyframe editing(Sec. 3.2), focusing text-guided edits on sampled keyframes x_0^k (where k ranges from 1 to N) and then propagating these changes to the remaining frames using a nearest neighbor search [14] with network ϵ_2.

To ensure identity preservation, we introduce a novel pipeline where during editing we substitute the self-attention features of ϵ_2 with those precomputed by ϵ_1. Additionally, we've devised a framework to fine-tune ϵ_2 for more precise localized facial edits.

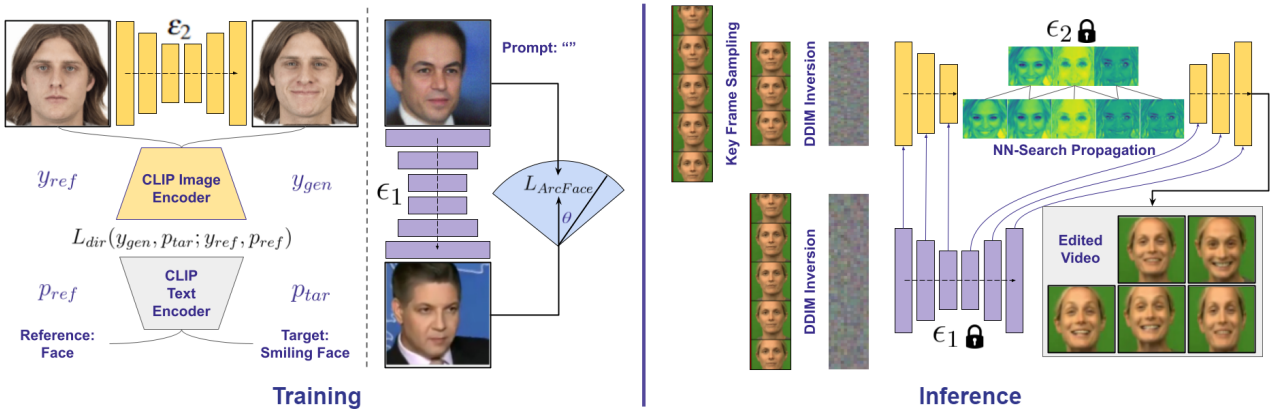

Figure 2. **Model Architecture.** Left: Pre-trained T2I models ϵ_1 and ϵ_2 are fine-tuned independently with ArcFace loss and directional CLIP loss for identity-preservation and prompt-adhering localization, respectively. Right: Video frames are inverted with DDIM and then processed through ϵ_1 to extract self-attention features at each timestep. Text-guided editing is applied to keyframes using ϵ_2, guided by identity features from ϵ_1. Edits are propagated to remaining frames using a nearest-neighbor search within the latent space.

4.1. Inference time Identity Preservation

Naively applying Text2Video models [14, 25, 33] for facial video editing often results in a loss of the original person's identity. These models are trained for generic video editing tasks such as changes in global style, scene of an image and thus when applied for facial videos struggle to maintain critical facial features.

To tackle this challenge, we introduce a method that incorporates additional structure and identity guidance during the editing process.

We fine-tuned an off-the-shelf T2I model ϵ_1 using a large facial image dataset (CelebA HQ [50]), employing a combined loss function that minimizes both pixel-wise reconstruction errors and identity differences (using ArcFace [9]) between generated and ground truth images. This approach aids in face reconstruction during editing while preserving identity. Mathematically we reduce the following loss:

$$L_{id} = \|\epsilon_1(x_0) - y_0\|_1 + D_{\cos}\left(f_{arc}(\epsilon_1(x_0)), f_{arc}(y_0)\right) \quad (5)$$

where y_0 is the ground truth face image and x_0 is the noisy latent obtained by DDIM inversion of y_0. The second term minimizes a cosine-similarity identity Loss using the pre-trained ArcFace [9] network (f_{arc}) where $D_{\cos}$ represents the cosine distance between the two features.

During inference, we guide the editing process of ϵ_2 by replacing its self-attention features with those from corresponding layers of the identity-preserving ϵ_1 network. This approach effectively incorporates identity information into the editing process, resulting in robust preservation of facial identity and head pose in the edited videos, as in Figure 2.

4.2. Directional Editing

T2I diffusion, pre-trained on generic generative tasks like style transfer and scene editing, are not well-suited for localized editing and often end up changing global semantics while increasing artifacts. TokenFlow [14], a non-faces geared (generic) diffusion-based video editor, suffers from this issue, see Fig. 1. To address this limitation, we fine-tune a pre-trained T2I model (SD 2.1 [35]) using a directional CLIP loss [13] on the large-scale face image dataset: CelebA-HQ [50]. We minimize the loss defined by:

$$\begin{aligned}\mathcal{L}_{dir}(\mathbf{y_{gen}}, \mathbf{p_{tar}}) = \lambda_1 \|f_I(\mathbf{y_{gen}}) - f_I(\mathbf{y_0})\|^2 \\ + \lambda_2 \left[1 - \left((f_I(\mathbf{y_{gen}}) - f_I(\mathbf{y_{ref}})\right) \cdot \right. \\ \left. \left(f_T(\mathbf{p_{tar}}) - f_T(\mathbf{p_{ref}})\right)\right)]\end{aligned} \quad (6)$$

where $\mathcal{L}_{dir}(\mathbf{y_{gen}}, \mathbf{p_{tar}})$ is the directional CLIP loss for image $\mathbf{y_{gen}}$ and target prompt $\mathbf{p_{tar}}$. $f_{I/T}(\cdot)$ denotes the CLIP Image/Text encoder mapping images and text into the same latent space. $\mathbf{y_{ref}}$ is original image. $\mathbf{p_{ref}}$ is the reference text embedding. λ_1 and λ_2 are scalar hyper-parameters.

The directional CLIP loss [13] effectively guides the editing process by minimizing the directional change between the original facial image representation and the text prompt's influence in model's latent space. This loss allows for more localized edits compared to traditional clip losses.

We call this fine-tuned T2I model as ϵ_2 Figure 2. During fine-tuning, we focused on 2 edit categories: (1) *facial features* like age, gender and expressions and (2) *facial attributes* like hair color, spectacles, mustaches, or beards.

5. Experiments

5.1. Implementation Details

We use Stable Diffusion 2.1 [35] as our pre-trained Text-to-Image model (ϵ_1) for fine-tuning both our branches. We fine-tune ϵ_1 using the CelebAHQ dataset [27] comprising of 512×512 10,000 images. ϵ_2 was finetuned on the same

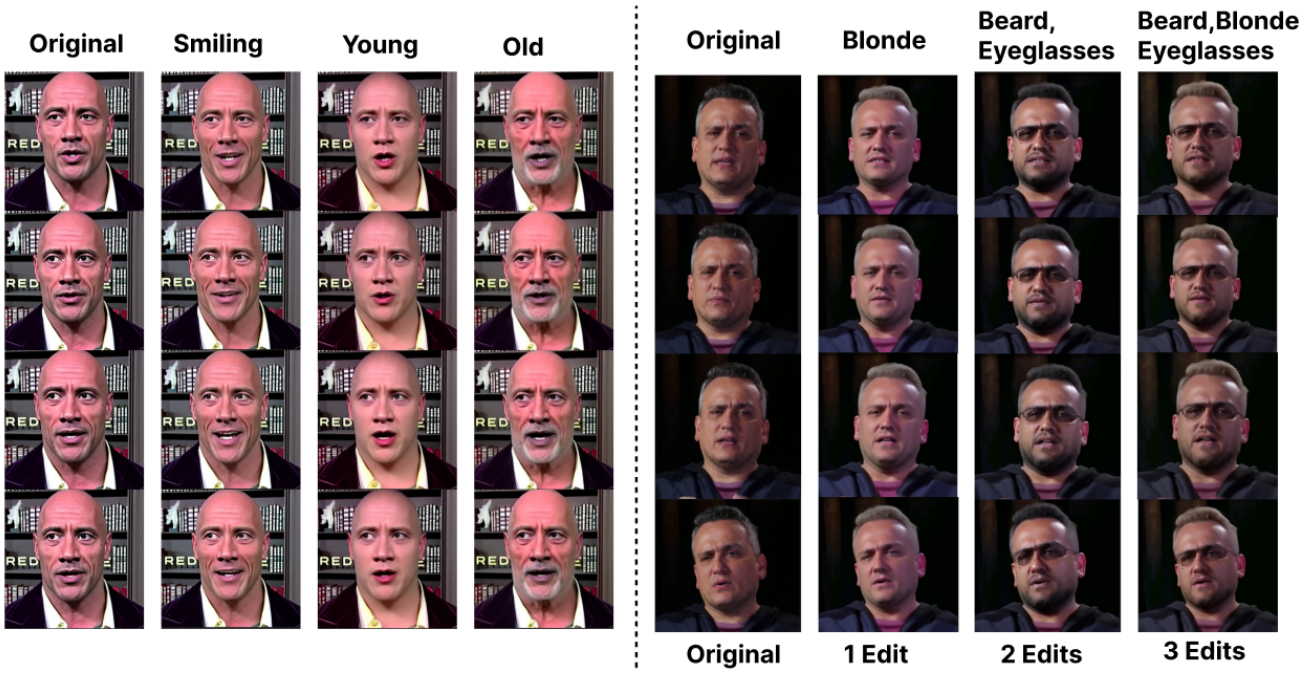

Figure 3. **Strong Prompt-Adhering Multiple Editing.** *Left*: Beyond local edits, our method manipulates facial expressions and age. *Right*: Our facial video editing method handles simultaneous edits to local facial features.

dataset using captions generated by BLIP-2 [28] as reference prompts. These were paired with extensive set of editing prompts that covered both facial attributes (e.g., beard, glasses) and facial features (e.g., gender, age). We ensured prompt diversity by incorporating a wide range of facial editing categories present in the large-scale facial attribute dataset CelebA [29]. The hyperparameters for the diffusion CLIP loss, λ_1 is 0.3 and λ_2 is 0.7. For both ϵ_1 and ϵ_2, we trained at a resolution of 512x512, batch size of 16 with a learning rate of 10^{-5} for 30,000 iterations using 2 NVIDIA A100's. Inference can be performed on a single one.

5.2. Baselines and Dataset

We conducted both quantitative and qualitative experiments, comparing our method against established Facial Video Editing benchmarks:

1. *DVA*: Kim *et al.*'s [26] diffusion-based method for facial video editing that enables editing a set of predefined facial features (e.g., expressions, age) and local facial attributes (e.g., beard, accessories).

2. *StitchGAN (STIT)* [46]: This method exclusively edits predefined global facial features (e.g., age, gender, expressions) but struggles with localization.

3. *Latent Transformer for Facial Video Editing (LTFE)* [52]: A StyleGAN-based [24] facial video editing method limited to a narrow set of attributes.

Other recent advancements in facial video editing leveraging StyleGAN inversion have been reported, but experimental comparisons with these models were not feasible due to the lack of publicly available implementations or pretrained models [23, 51, 53]. Consequently, we focused on widely recognized benchmarks that remain competitive.

Evaluation Dataset For our experiments, we selected a subset of 50 original videos from the CelebV-HQ dataset [8]. From these, we created an evaluation dataset consisting of 25 edited videos of each of the baselines and our method. Each video had a duration of 4 seconds (10 fps) and a frame resolution of 512x512.

5.3. Identity Preservation Metric

To quantitatively assess identity preservation in edited videos, we use three face recognition models: CosFace [48], VGGFace [34], and FaceNet [41]. These models generate high-dimensional embeddings that capture facial features, enabling identity evaluation.

For each video, we compute embeddings by averaging frame-wise embeddings for the original and edited frames. Identity preservation is assessed using retrieval metrics: 1) Recall at Rank 1 (R@1) [47], 2) Mean Reciprocal Rank (MRR) [31], and 3) Cosine Similarity. Using the CelebV-HQ database [8] (35,000 videos), we rank the embedding of the original video among the database embeddings based on its similarity to the edited video embedding.

Figure 4. **Editing Faces in the Wild.** We successfully overcome a previous hurdle of out-of-domain adaptation for facial video editing methods.

Recall at Rank 1 (R@1): Measures the proportion of queries where the original embedding is ranked first based on average Euclidean distance of the frames:

$$\text{R@1} = \frac{\text{Number of correct top-1 queries}}{\text{Total queries}}$$

Mean Reciprocal Rank (MRR): Quantifies retrieval performance by averaging reciprocal ranks of ground-truth embeddings, prioritizing higher ranks:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$$

where rank_i is the position of the ground-truth embedding for the i-th query.

Cosine Similarity: Evaluates similarity between embeddings of original (x_0) and edited frames (x) to assess identity preservation:

$$d_{\text{cosine}}(x_0, x) = 1 - \frac{x_0 \cdot x}{\|x_0\|\|x\|}$$

Lower values of d_{cosine} indicate better identity preservation.

Model	VGGFace [9]			CosFace [48]			FaceNet [48]		
	Cosine ↓	R@1 ↑	MRR ↑	Cosine ↓	R@1 ↑	MRR ↑	Cosine ↓	R@1 ↑	MRR ↑
DVA [26]	0.361	0.76	0.794	0.318	0.76	0.794	0.256	0.76	0.794
STIT [46]	0.447	0.70	0.748	0.471	0.70	0.748	0.273	0.72	0.773
LTFE [52]	0.506	0.64	0.662	0.531	0.64	0.673	0.488	0.66	0.683
Ours (w/o ID guidance)	0.762	0.32	0.352	0.722	0.34	0.396	0.706	0.34	0.396
Ours (with ID guidance)	**0.221**	**0.96**	**0.97**	**0.206**	**0.96**	**0.97**	**0.185**	**0.96**	**0.97**

Table 1. **Quantitative Comparison of Identity Preservation.** We evaluate identity preservation using Cosine Distance, R@1, and MRR metrics averaged across videos in our evaluation dataset, using embeddings from VGGFace, CosFace, and FaceNet.

5.4. Frames vs. Inference Time.

Existing facial video editing benchmarks require end-to-end training for each facial video prior to inference, resulting in high computational costs and significant time consumption. In contrast, our method leverages our pre-trained T2I models to directly perform edits, eliminating the need for additional per-video training. Also by focusing on editing only keyframes, our approach reduces inference time by approximately 80% compared to state-of-the-art methods, offering a robust & practical solution for facial editing.

For instance, editing a 30-second, 10 fps facial video (300 frames) takes just ∼6 minutes on a single NVIDIA A100 GPU, demonstrating significant gains over the nearest baselines: STIT [46] (32 minutes) and DVA [26] (28 minutes). Tab. 2 shows a quantitative comparison.

Video Duration	Ours	Kim [26]	Tzaban [46]	Yao [52]
30 sec	**6 mins**	28 mins	32 mins	35 mins
60 sec	**15 mins**	45 mins	60 mins	68 mins
90 sec	**32 mins**	90 mins	110 mins	132 mins
120 sec	**98 mins**	180 mins	194 mins	206 mins

Table 2. **Inference Speed Comparison** on **10 fps** videos, averaged across video clips from the evaluation dataset and reported in minutes using a single NVIDIA A100

5.5. Temporal Consistency Evaluation

Our method incorporates Geyer *et al.*'s [14] algorithm, which enforces temporal consistency by aligning and propagating latent tokens across frames within the latent space of pre-trained T2I diffusion models.

In contrast to existing benchmark approaches that rely on per-frame encoders and process each frame independently,

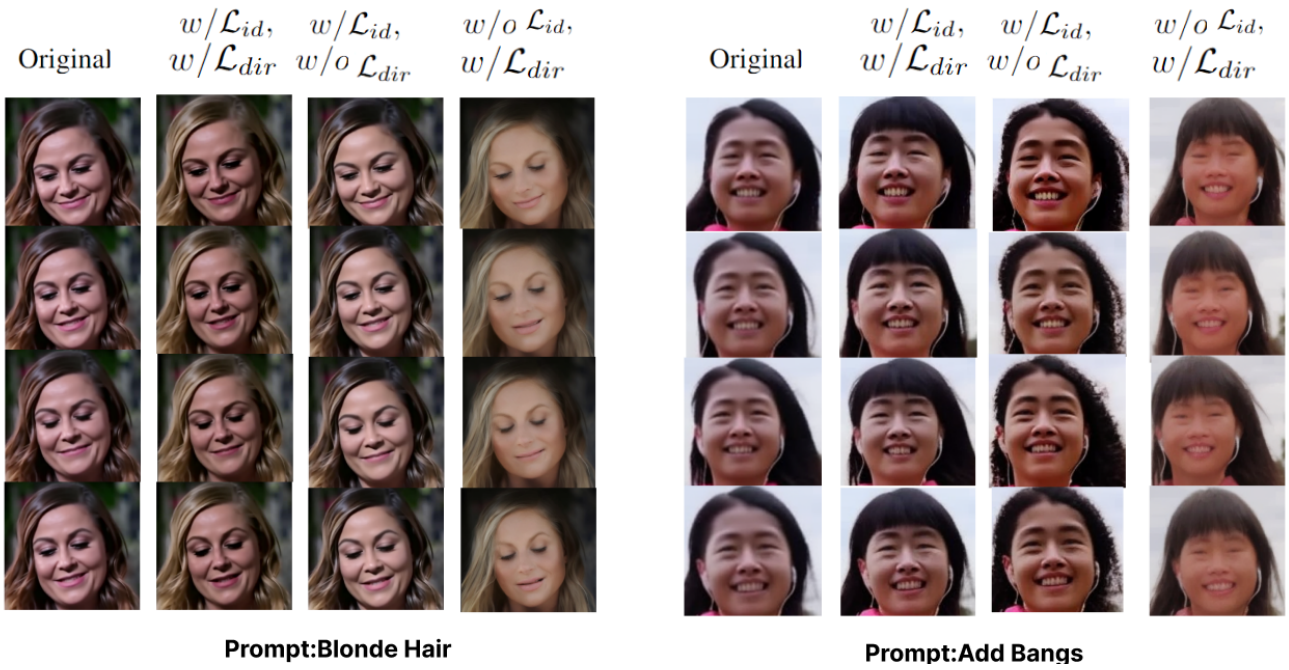

Figure 5. Ablation: fine-tuning ϵ_1 and ϵ_2 with Arc-Face and directional-Clip Loss for identity preservation and performing localized edits in facial videos.

our method inherently enforces frame-to-frame coherence at the latent representation level. This helps in reducing perceptual flickering and improving temporal stability Tab. 3.

To quantitatively assess temporal consistency, we utilize the pre-trained RAFT model [44] to compute optical flow o_t between consecutive frames I_t and I_{t+1}. Using the computed optical flow, we forward-warp frame I_t to frame I_{t+1} with the bilinear inverse warping operator [22], denoted as $\mathcal{W}$. This process ensures precise alignment between consecutive frames, enabling an accurate evaluation of temporal coherence.

The temporal loss, calculated as the average ℓ_1 distance between the warped frame $\mathcal{W}(I_t, o_t)$ and the actual frame I_{t+1} across all video frames, is defined as:

$$\mathcal{L}_{\text{temp}} = \frac{1}{T}\sum_{t=1}^{T} \|\mathcal{W}(I_t, o_t) - I_{t+1}\|_1, \tag{7}$$

where T is the total number of frames in the video. A lower temporal loss indicates superior temporal consistency, with smoother transitions and fewer artifacts. Table 3 presents the temporal loss values, comparing our approach against state-of-the-art methods, demonstrating the effectiveness of our method in maintaining temporal coherence across frames.

Method	Temporal Consistency ↓	MOS Score ↑
DVA [26]	0.256	4.7
STIT [46]	0.273	5.6
LTFE [52]	0.463	3.4
Ours	**0.238**	**7.7**

Table 3. **Quantitative Fidelity Comparison.** The table compares identity preservation, RAFT-based temporal consistency, and Mean Opinion Score (MOS) for different methods. Evaluations are conducted on 20 ten-second videos at 512×512 resolution, highlighting the superior temporal consistency and perceived quality of our approach compared to existing methods.

5.6. Ablation Study

We conducted an ablation study to assess the contributions of key components in our model: the effect of identity guidance from the fine-tuned model ϵ_1, optimized with $\mathcal{L}_{\text{id}}$, and the impact of fine-tuning ϵ_2 with $\mathcal{L}_{\text{dir}}$ to achieve temporally consistent and localized facial edits.

First, we retained the identity guidance from ϵ_1 and edited facial videos with ϵ_2 before it was fine-tuned with the directional CLIP loss $\mathcal{L}_{\text{dir}}$. As shown in Column 3 of Fig. 5, the results demonstrated that while the T2I model struggled with highly localized edits, validating the challenge of using pre-trained T2I models for precise facial modifications.

Figure 6. **More Identity Preserving Localized Editing in the Wild.** Left: Hair color change and accessory addition are performed from a randomly scraped online music video. Right: Facial hair and facial expressions of a consenting volunteer are edited using our method.

This emphasizes the necessity of $\mathcal{L}_{\text{dir}}$ for enabling effective, localized edits in our framework.

Despite the limitations in performing precise edits, the model successfully preserved identity, attributed to the guidance provided by ϵ_1.

Next, we removed the identity guidance from ϵ_1, which was fine-tuned with $\mathcal{L}_{\text{id}}$, and tested the model using only ϵ_2, fine-tuned with $\mathcal{L}_{\text{dir}}$. The results, as seen in Column 3 of Fig. 5, showed a substantial failure to preserve the original identity, which was corroborated by identity preservation metrics such as CosFace, VGGFace, and FaceNet. These metrics recorded significantly suboptimal values in this configuration (Tab. 1).

This ablation study highlights the critical importance of fine-tuning both ϵ_1 and ϵ_2 with $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{dir}}$, respectively, to achieve accurate, identity-preserving localized edits.

5.7. Mean Opinion Score Study

We conducted a mean opinion score (MOS) study with 50 participants to qualitatively evaluate the realism and believability of our facial video edits, using a 1 to 10 rating scale (Tab. 3). A double-blind setup, with anonymized methods and randomized video order, ensured unbiased and robust assessment of our approach's perceived fidelity.

5.8. Multiple Prompt-Localized Editing.

Existing methods are constrained to performing one edit at a time on facial videos, necessitating successive processing for multiple edits and significantly increasing inference time. In contrast, our approach leverages the robust latent space of pre-trained SDE models. By embedding text-guided editing directly into the cross-attention layers of the fine-tuned U-Net [39] ϵ_2, we enable simultaneous application of multiple edits without added computational overhead. For instance, our method can process complex prompts such as "[Add bangs, sunglasses, and a French beard to the face]" in a single step Fig. 3.

5.9. Generalization: Editing Faces in the Wild.

Prior facial video editing works, trained end-to-end on curated datasets, tends to struggle when applied to in-the-wild facial videos(Fig. 1 and Fig. 4 the results produced by DVA [26] and STIT [46]). Our approach, however, harnesses the robustness of pre-trained SDE models trained on extensive data (e.g., LAION 2B dataset [42]), enabling effective editing of wild facial videos while preserving temporal consistency, even amidst challenging head poses and varied viewing angles as shown in Fig. 1 and Fig. 4 .

6. Conclusion

To address challenges such as high editing time, identity preservation, and achieving diverse edits in facial video editing, we proposed a novel pipeline that leverages pre-trained text-to-image (T2I) diffusion models, fine-tuned to maintain global identity while enabling directional edits.

Furthermore, although stable diffusion is effective, it may introduce slight shifts in image texture, lighting,and background. These shifts were observed in some of our results and should be considered in future work. Also further exploration for enhanced flexibility in editing facial videos through various guidance attributes, as well as a focus on both micro and macro expressions, presents promising advancements for creative applications.

7. Acknowledgement

We would like to acknowledge partial support from IITM Pravartak Technologies Foundation. KM would like to acknowledge support from Qualcomm Faculty Award 2024.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4431–4440, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8293–8302, 2020. 2

[3] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics*, 38(4):1–11, July 2019. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2022. 2

[5] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy Jyoti Mitra. Pix2video: Video editing using image diffusion. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23149–23160, 2023. 2

[6] Wenhao Chai, Xun Guo, Gaoang Wang, and Yang Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22983–22993, 2023. 2

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42:1 – 10, 2023. 1

[8] Hao clip, Wayne Wu, Wentao clip, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, page 650–667, Berlin, Heidelberg, 2022. Springer-Verlag. 5

[9] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, Oct. 2022. 2, 4, 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 1, 2

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, page 89–106, Berlin, Heidelberg, 2022. Springer-Verlag. 1, 2

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2022. 2

[13] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada. *ACM Transactions on Graphics (TOG)*, 41:1 – 13, 2021. 2, 4

[14] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, 2023. 2, 3, 4, 6

[15] Alexandros Graikos, Srikar Yellapragada, and Dimitris Samaras. Conditional generation from pre-trained diffusion models using denoiser representations. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 2

[16] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3018, 2019. 2

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2022. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 1

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-clip, Yuanzhi Li, Shean Wang, Lu Wang, and Weiclip Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[21] Inwoo Hwang, Hyeonwoo Kim, and Young Min Kim. Text2scene: Text-driven indoor scene stylization with part-aware details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1890–1899, June 2023. 2

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 7

[23] Yue-Ren Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. Identity-aware and shape-aware propagation of face editing in videos. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3444–3456, 2024. 5

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018. 5

[25] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15908–15918, 2023. 2, 4

[26] Gyeongman Kim, Hajin Shim, Hyunsung Kim, Yunjey Choi, Junho Kim, and Eunho Yang. Diffusion video autoencoders:

Toward temporally consistent face video editing via disentangled video encoding. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6091–6100, 2022. 2, 3, 5, 6, 7, 8

[27] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[30] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: a fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. 1

[31] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 5

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan clip, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2

[33] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2, 4

[34] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, 2015. 5

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2023. 2, 3, 4

[36] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15886–15896, 2023. 2

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[38] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 8

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. 1, 2

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 5

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. 8

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1, 3

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

[45] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2022. 2

[46] 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*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery. 2, 5, 6, 7, 8

[47] Ellen M Voorhees. The trec-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*. National Institute of Standards and Technology (NIST), 1999. 5

[48] H. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 5, 6

[49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7589–7599, 2022. 2

[50] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[51] Yangyang Xu, Shengfeng He, Kwan-Yee K. Wong, and Ping Luo. Rigid: Recurrent gan inversion and editing of real face videos. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13645–13655, 2023. 5

[52] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13769–13778, 2021. 2, 5, 6, 7

[53] 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*, abs/2401.02616, 2024. 5

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, October 2023. 2