Instruct-NeuralTalker: Editing Audio-Driven Talking Radiance Fields with Instructions

YUQI SUN, Fudan University, China
REIAN HE, Fudan University, China
WEIMIN TAN, Fudan University, China
BO YAN, Fudan University, China

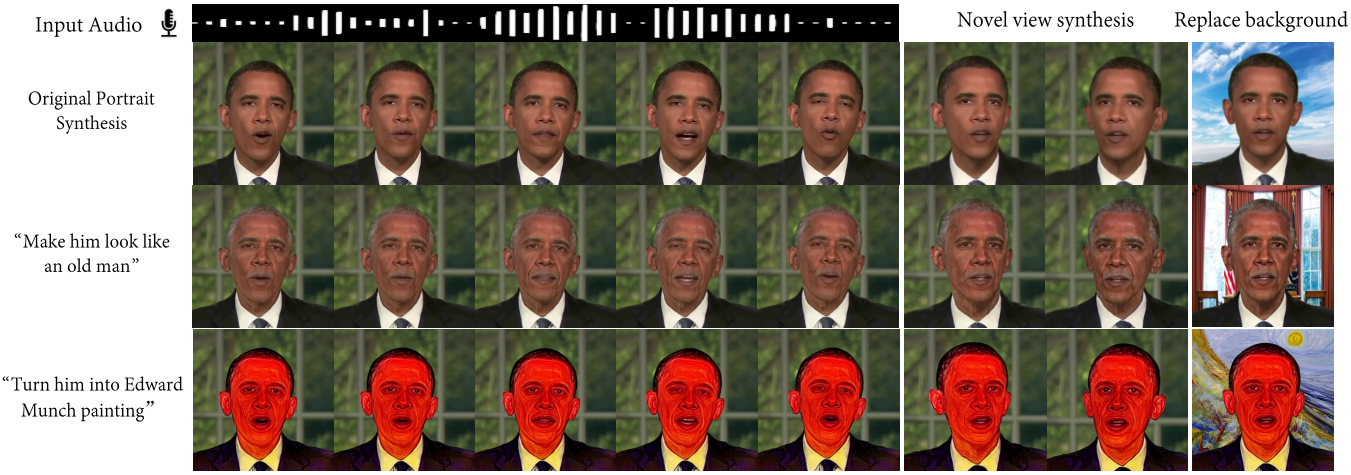

Fig. 1. Given a speech video, our method first builds an original audio-driven talking radiance field and edits it with human instructions. *Original Portrait Synthesis* shows the generated results of the original radiance field from the input audio. By giving editing instructions like *"Make him look like an old man"* and *"Turn him into Edward Munch painting"*, our methods can synthesize high-quality talking faces that meet the editing target well while maintaining audio-lip synchronization. The complete rendering pipeline can be done in real-time on consumer hardware. Benefiting from the 3D neural representation modeling, our method can easily perform other editing tasks like novel view synthesis and background replacement.

Recent neural talking radiance field methods have shown great success in photorealistic audio-driven talking face synthesis. In this paper, we propose a novel interactive framework that utilizes human instructions to edit such implicit neural representations to achieve real-time personalized talking face generation. Given a short speech video, we first build an efficient talking radiance field, and then apply the latest conditional diffusion model for image editing based on the given instructions and guiding implicit representation optimization towards the editing target. To ensure audio-lip synchronization during the editing process, we propose an iterative dataset updating strategy and utilize a lip-edge loss to constrain changes in the lip region. We also introduce a lightweight refinement network for complementing image details and achieving controllable detail generation in the final rendered image. Our method also enables real-time rendering at up to 30FPS on consumer hardware. Multiple metrics and user verification show that our approach provides a significant improvement in rendering quality compared to state-of-the-art methods.

CCS Concepts: • **Human-centered computing** → **Visualization**; **Interaction design**.

Additional Key Words and Phrases: Talking Face, Insturct-editing, Neural Radiance Field, Real Time, Audio Driven, View Synthesis

ACM Reference Format:
Yuqi Sun, Reian He, Weimin Tan, and Bo Yan. 2023. Instruct-NeuralTalker: Editing Audio-Driven Talking Radiance Fields with Instructions. *ACM Trans. Graph.* 1, 1 (June 2023), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Audio-driven talking face generation has been a long-standing task in computer vision and graphics, with widespread applications in digital humans, VR/AR, 3D telepresence, and virtual video conferencing. Recent methods [Guo et al. 2021; Tang et al. 2022] have made significant progress in achieving photorealistic and controllable talking face generation based on Neural Radiance Fields (NeRF) [Mildenhall et al. 2020]. By constructing an implicit 3D face model, their approaches allow impressive editing capabilities such as novel view synthesis and background replacement. However, achieving advanced editing tasks like semantic manipulation and style transfer

Authors' addresses: Yuqi Sun, Fudan University, China; Reian He, Fudan University, China; Weimin Tan, Fudan University, China; Bo Yan, Fudan University, China.

0730-0301/2023/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

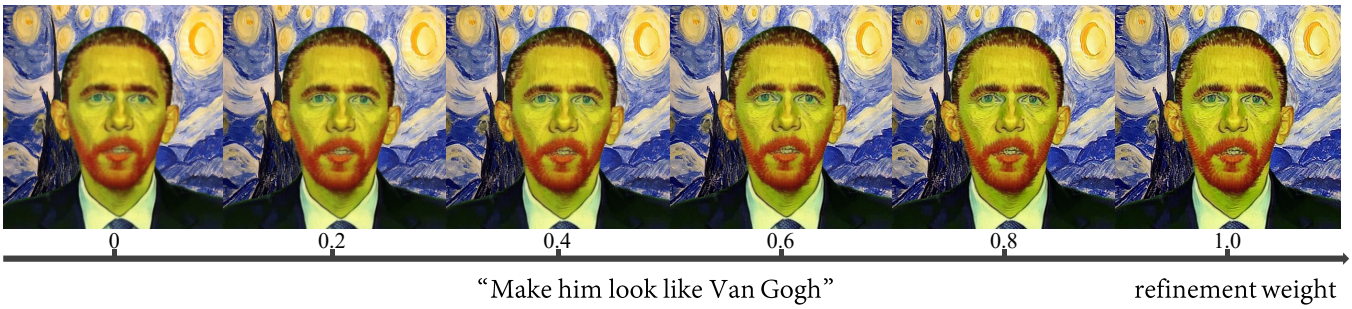

Fig. 2. **Illustration of controllable detail generation.** When we input instruction *"Make him look like Van Gogh"*, our method generates talking faces with the style of Van Gogh. By manipulating the added weight of the refinement network's output, we can generate controllable detail, where lower weight values yield smooth results, and higher weight produces rich details.

on implicit representations still remain challenging and underexplored. Encouraged by the success of recent large language models (LLMs) such as ChatGPT, we propose the first interactive framework for semantically editing Talking Radiance Fields (TRF) with textual instructions, enabling highly personalized talking face generation.

We take inspiration from the recent instruction-based NeRF-editing method Instruct-NeRF2NeRF (in2n) [Haque et al. 2023]. It introduces a novel approach to edit NeRFs by manipulating training datasets during optimization. By incorporating a conditional diffusion model InstructPix2Pix (ip2p) [Brooks et al. 2022], in2n achieves impressive results in editing 3D scenes with text instructions. Nevertheless, their approach only focuses on static scenes. We extend this instruction-based NeRF editing method to dynamic neural radiance fields for editing TRF and synthesizing high-quality personalized talking faces in real-time.

We propose Instruct-NeuralTalker, an innovative interactive framework for semantically editing talking radiance fields using human instructions. Given a short speech video, our approach first builds an initial TRF for the original face. Then users can provide text instructions specifying personalized editing goals, such as "Make him look like an old man." They can also adjust the weighting of the text guidance to control the extent of the edits. Once the desired appearance is defined, we progressively modify the training dataset using ip2p during the optimization to guide the initial TRF toward the edited target. Ultimately, we obtain a TRF with an "old man" style that can generate corresponding talking faces driven by audio. Leveraging the latest efficient NeRF architecture, our method achieves real-time rendering on consumer hardware. We showcase some visual results in Figure 1 to demonstrate the editing ability of our method and an user interface demo in Figure 8.

In contrast to editing static scenes, it is essential to maintain motion semantically in editing dynamic radiance fields. For talking faces, It involves ensuring audio-lip synchronization in the edited results. We observed that using a large textual guidance scale in ip2p can lead to distortions in lip shape, while using a small scale may not achieve the desired editing goals. To address this, we propose a progressive dataset updating strategy that gradually increases the text-guided scale and steps of the reverse diffusion process during the optimization. This approach helps the model to preserve lip shape in a coarse-to-fine manner. We also propose a lip-edge loss to restrain the edges of the lips. Specifically, we use the lip parsing mask and the first-order gradient of the lip region of the original face image to constrain the edited lip shape to have sharp edges.

On the other hand, since the editing results of ip2p do not guarantee temporal consistency, the edited TRF tends to produce over smoothed results with a lack of image details. To solve this, we incorporate a refinement network to post-process the results and complement the image details. We obtain the final result by a weighted sum of the output of the refinement network with that of the TRF. It enables us to achieve controllable detail generation by adjusting the weighting factors. Figure 2 illustrates the results obtained with different refinement weights. When the weight is set to 0, the refinement network is disabled. The result directly rendered from the TRF is highly smooth. As the weight increases, details on the face become more prominent.

In summary, this paper mainly has the following contributions:

- We propose Instruct-NeuralTalker, the first interactive framework to semantically edit the audio-driven talking radiance fields with simple human instructions. It supports various taking face editing tasks, including instruction-based editing, novel view synthesis, and background replacement. In addition, Instruct-NeuralTalker enables real-time rendering on consumer hardware.
- In order to ensure audio-lip synchronization, we develop a progressive dataset updating strategy and a lip-edge loss to constrain the lips change of the editing results during the optimization process.
- We introduced a lightweight refinement network to complement high-frequency image details and achieve controllable detail generation in the rendering process.
- Extensive experiments demonstrate the superiority of our method in video quality compared to the state-of-the-art.

2 RELATED WORK

Audio-Driven Taking Face Generation. Audio-driven talking face generation is a long-standing problem in computer vision and graphics. It aims to reenact a specific person's speaking video based on arbitrary given audio input. Traditional methods require expertise in mapping from input waveforms to facial movements, such as phoneme-visual mapping [Fisher 2014]. With the emergence of deep learning techniques, some methods [Lele et al. 2019] learn the mapping between a deep audio feature and lip movements from

extensive data and generated realistic faces using GAN networks [Vougioukas et al. 2019]. More recently, many approaches introduce an intermediate 3D representation to achieve controllable face generation, such as 3D morphable face models [Thies et al. 2020] and facial landmarks [Zakharov et al. 2019]. Nevertheless, estimating explicit intermediate representations may result in information loss and error accumulation, resulting in mismatches between audio and lips. Our method aims to edit the implicit radiance field with human instructions. Although some recent one-shot methods, such as Wav2Lip[Prajwal et al. 2020], MakeItTalk[Zhou et al. 2020] and SadTalker[Zhang et al. 2022] can achieve this goal from an edited image, they can not support 3D-based editing tasks like novel view synthesis and background replacement.

Neural Talking Radiance Fileds. With the development of NeRF, recent methods propose to model talking faces with dynamic NeRFs [Pumarola et al. 2021]. AD-NeRF [Guo et al. 2021] was the first to directly map audio features to dynamic NeRFs for lips movement generation, achieving better audio-visual consistency results. DFA-NeRF[Yao et al. 2022] decouples lip movement features related to audio from personalized attributes unrelated to audio using a self-supervised learning approach. DFRF [Shen et al. 2022] can adapt to new identities with limited reference images through pretraining on a large dataset. RAD-NeRF [Tang et al. 2022] leverages recent successful grid-based NeRF approaches [Müller et al. 2022] to model dynamic head movements using efficient audio and spatial grids, which achieves real-time talking face generation and faster convergence. In our approach, we utilize RAD-NeRF as the backbone network and combine it with conditional diffusion models to achieve instruction-guided editing of talking radiance fields.

NeRF Editing from Instructions. Recently, the rise of large-scale language models LLMs (such as ChatGPT) has enabled complex task control through natural language, particularly instructions[Ouyang et al. 2022]. Recent works have introduced pre-trained language-visual models into the training process of NeRF to achieve text-based NeRF editing. ClipNeRF [Wang et al. 2021] and NeRF-Art [Wang et al. 2022] encourage similarity between the scene and text embeddings from CLIP [Radford et al. 2021] to enable text-based control. Instruct-Nerf2NeRF (in2n) [Haque et al. 2023] introduces an instruction-based image editing method ip2p to 3D scene editing and allows for intuitive and concise instruction-based editing for the first time. in2n can generate impressive results, showcasing the potential of instruction-based control in NeRF editing. Nevertheless, they focus on editing static scenes. Inspired by in2n, our method first introduces instruction-based editing into dynamic NeRFs to achieve a personalized generation of talking faces. It gives rise to an interface entirely controlled by language, facilitating a more comprehensive range of intuitive and content-aware 3D talking face editing.

3 METHOD

3.1 Preliminaries

RAD-NeRF. We employ RAD-NeRF[Tang et al. 2022] as the backbone network for constructing the radiance field of talking faces. RAD-NeRF achieves real-time rendering in TRFs by incorporating the latest advancements in the grid-based NeRF architecture. Grid-based NeRF stores 3D scene features in an explicit 3D feature grid structure and employs efficient data structures such as multi-resolution hash tables [Müller et al. 2022] and low-rank tensor components [Chen et al. 2022]. RAD-NeRF separately models the dynamic head with a 3D spatial grid and a 2D audio grid. To be specific, given a 3D coordinate $x \in \mathbb{R}^3$, it generates a 3D spatial code f from spatial encoder $E^3_{spatial} : f = E^3_{spatial}(x)$ for head reconstruction. Then RAD-NeRF compresses the high-dimensional audio features into a low-dimensional audio coordinate $x_a \in \mathbb{R}^D$ through an MLP: $x_a = MLP(a, f)$ and produces a 2D audio code g by $E^D_{audio} : g = E^D_{audio}(x_a)$. In the end, it uses a small MLP to estimate density and color:

$$c, \sigma = \mathbf{MLP}(f, g, e, i) \tag{1}$$

where e and i denote as an eye feature and a latent appearance embedding. The final color is rendered by volume rendering.

Instruct-NeRF2NeRF. Manipulating 3D scene representations with text instruction is an appealing task. Instruct-NeRF2NeRF (in2n) first employs ip2p to guide NeRF optimization through training set editing. Our method differs from in2n in two aspects: Firstly, we focus on editing dynamic radiance fields for personalized talking faces generation rather than static scenes. Secondly, in2n takes original input as additional conditions for editing NeRF-rendered output and edits a few images at a time. This strategy can be slow or even lead to unsuccessful edits. In contrast, our approach edits all training set images simultaneously and progressively increases the text guidance scale and diffusion steps to preserve lip shape. This strategy enables rapid convergence of the talking radiance fields towards the desired editing objectives.

InstructPix2Pix. Inspired by in2n, we leverage ip2p for dataset editing and update the TRF during optimization. InstructPix2Pix (ip2p) aims to achieve instruction-based image editing using the diffusion model. Given an image c_I, a text editing instruction c_T, and a purely noisy image z_t, the model uses a U-Net ϵ_θ to estimate the random noise added during the forward process at step t:

$$\epsilon = \epsilon_\theta(z_t, t, c_T, \xi(c_I)) \tag{2}$$

where ξ encodes the input image into a latent space, the noise prediction ϵ can be used to derive the edited image. ip2p also incorporates image and text guidance scales, allowing for a flexible trade-off between the fidelity to the input image and the adherence to the editing instructions.

3.2 Instruct-NeuralTalker

Our method, Instruct-NeuralTalker, aims to edit the talking radiance field based on text instructions, which enables more flexible and general talking face editing in addition to novel view synthesis and background replacement. Figure 3 illustrates our pipeline. Given an audio and an initial TRF R_o, we generate rendered frames $\{I^r_i\}^N$ from TRF for the first time as a training dataset. Then, using ip2p, we edit the training set to get $\{I^e_i\}^N$ based on text instructions, which we used to build the edited TRF R_e later. To preserve lip shape, we introduce a progressive dataset update strategy (Section 3.3) to update the training set iteratively. Since the images rendered by TRF lack texture details, we introduce a refinement network (Section 3.4)

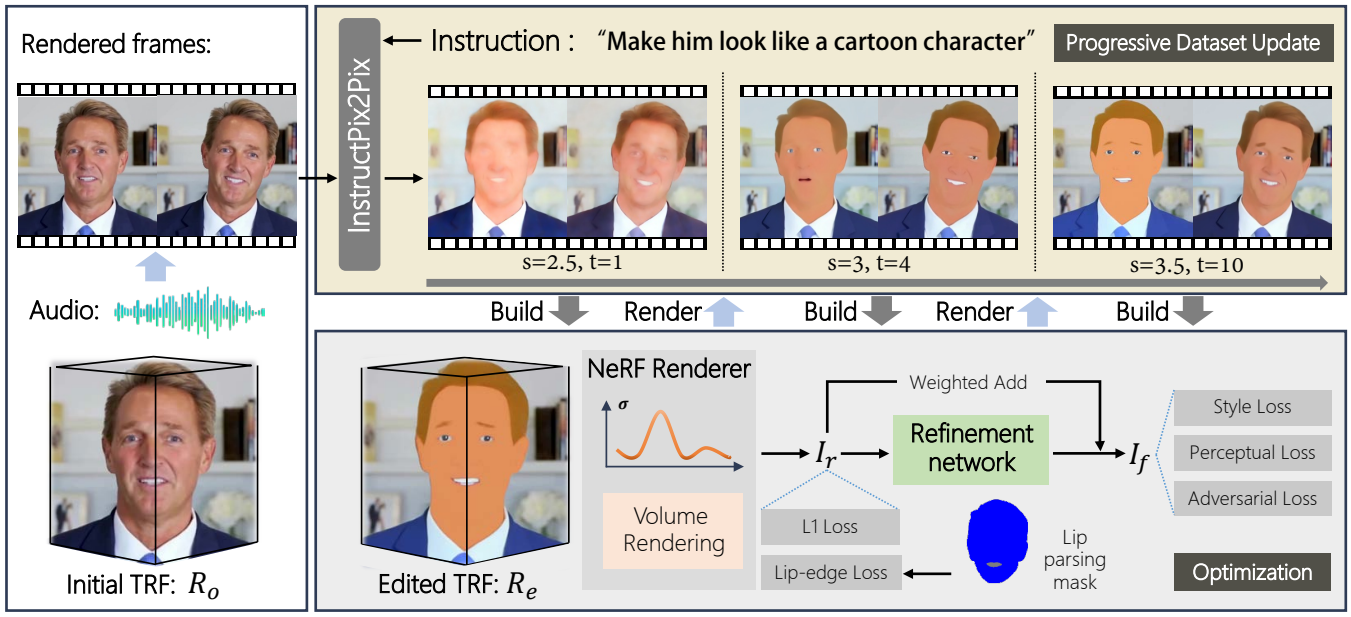

Fig. 3. **Overview of our pipeline.** We first use RAD-NeRF to initialize a talking radiance field R_o and render frames with the input audio as the training set. Then, given a text instruction such as "Make him look like a cartoon character," we use InstructPix2Pix to edit the rendered frames to get cartoon-style images, which are used to build the edited TRF. After D epochs, we stop the optimization and render the new frames for editing. Our progressive dataset update strategy keeps increasing the text guidance **s** and diffusion steps **t** to control the degree of editing. I_r denotes the image rendered from the edited talking radiance field R_e. A refinement network is proposed for compensating high-frequency image details. The final image I_f is obtained by weighted-add the output of our refinement network and I_r. We also adopt the lip parsing mask to calculate a lip-edge loss for preserving audio-lip synchronization.

to enhance image details and achieve controllable detail generation. In addition, we incorporate a lip-edge loss (Section 3.5) to control lip shape further and maintain audio-lip synchronization.

3.3 Progressive Dataset Update

Ip2p contains several hyperparameters that control the degree and quality of editing. Text guidance and image guidance scales control the edited images' alignment with the text description and the original image. The diffusion step can be used to control the image quality. In the case of facial editing, we find that a large text guidance scale may cause shape changes of the lips, resulting in lip flickering, and large diffusion steps require huge computational cost and significantly impact the efficiency of editing the taking radiance field. We use a small increasing text guidance scale and diffusion step to iteratively update the dataset to avoid this. With this coarse-to-fine strategy, we found that it maintains lip shape during editing.

Specifically, in the beginning, we manually set K parameter groups $\{(s_0, t_0), ..., (s_k, t_k)\}$with the gradual, incremental scales and steps for updating the dataset K times. We first generate rendered frames $\{I_i^r\}_1^N$ from initial TRF R_o and send them into ip2p for editing with text guidance scale $s = s_0$ and diffusion step $t = t_0$. Then we use the edited image $\{I_i^e\}_1^N$ as supervision in the optimization for building the edited talking radiance field R_e. After D epochs, we stop the optimization and render training frames $\{I_i^r\}_2^N$ from R_e, which are sent into ip2p for the second update with $s = s_2$ and $t = t_2$. After K times dataset updates and optimizations, we obtain the final talking radiance field that satisfies the instruction.

3.4 Refinement Network

Although the edited talking radiance field is able to render satisfactory results with great temporal consistency, the images appear overly smooth and lack fine details. We attribute this to the fact that the network capacity of MLP limits its ability to learn high-frequency signals. Recently, convolutional-based deep learning methods have achieved remarkable performance in image enhancement tasks. Therefore, we introduce a refinement network g that utilizes convolutional layers to complement details in the talking face images. Specifically, we draw inspiration from the Residual in Residual Dense Block (RRDB) in the super-resolution method ESRGAN [Wang et al. 2018], which consists of multi-level residual networks and dense skip connections. The multiple connections layers help the network capture the high-frequency details. To ensure a lightweight implementation, the refinement network utilizes only a single RRDB as the backbone network, resulting in a compact size of just 1MB. We send the result rendered from the talking radiance field I_r into the refinement network and add the output to I_r with the following equation:

$$I_f = \omega \cdot g(I_r) + I_r \tag{3}$$

This allows us to achieve controllable detail generation by choosing different weights ω. In addition, since the refinement network

Fig. 4. Qualitative results displaying the editing ability of our method. Given a simple text instruction, our method is able to edit the original sequence into a high-quality talking face that matches the editing target and maintains excellent audio-lip synchronization.

is lightweight, we can perform real-time rendering on consumer hardware.

3.5 Losses

Given an edited image I^e as supervision, we apply different losses for the rendered result from talking radiance field I_r and the final result I_f. For I_r, we expect that it has a similar facial structure to I_e while preserving the correct lip shape, so we employ a lip-edge loss and a reconstruction loss. For I_f, we hope it can learn high-frequency details such as texture from the edited image, so we adapt some losses at the feature level, including perceptual loss, style loss, and adversarial loss.

Lip-Edge Loss. We introduce constraints on the lip edges to further control the edited talking radiance field to generate the correct lip shape. For this sake, we borrow the depth smooth loss used in depth estimation [Yin and Shi 2018] and apply it to the lip constraint. We first parse the original face I_o to extract the lip mask M^{lip}, and then constrain the edges of I_r^{lip} inside the lip patch to be consistent with the correct edges in I_i^{lip}, with the following equation:

$$\mathcal{L}_{lip} = |\nabla I_r^{lip} * M^{lip}| \cdot exp(-|\nabla I_o^{lip} \cdot M^{lip}|) \tag{4}$$

where $|\cdot|$ means absolute value and ∇ means differential operator.

Reconstruction Loss. We calculate a MSE loss between the result I_r and the edited image I_i^e as the reconstruction loss:

$$\mathcal{L}_{rec} = ||I_r - I^e||_2 \tag{5}$$

Perceptual and Style Loss. Perceptual loss [Chen and Koltun 2017]calculates the feature distance between the input image and the target image in a pre-trained VGG network[Simonyan and Zisserman 2014]. Style loss[Johnson et al. 2016] adds a Gram matrix [Gatys et al. 2015] in perceptual loss to penalize differences in style: color, texture, common patterns, etc. The formula is as follows:

$$\begin{aligned} \mathcal{L}_{pcp} &= \sum_l \lambda_l \|\Phi_l(I_f) - \Phi_l(I^e)\|_1 \\ \mathcal{L}_{style} &= \sum_l \lambda_l \|gram(\Phi_l(I_f)) - gram(\Phi_l(I^e))\|_1 \end{aligned} \tag{6}$$

Φ_l denotes the outputs of middle layers of a pretrained VGG network, and the weights λ_l determine which layers of the network are used. $gram(\cdot)$ is a matrix of inner products of a set of vectors in an inner product space.

Adversarial Loss. We also use adversarial loss [Mirza and Osindero 2014] to increase image detail. The adversarial loss $\mathcal{L}_{adv}$ is calculated on the final result I_f and the edited image I^e through a learnable discriminator, which aims to distinguish the training data distribution and predicts one.

We first train the network using all losses other than lip-edge loss, the total loss is calculated by:

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{pcp}\mathcal{L}_{pcp} + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv} \tag{7}$$

After that, we finetune our method by adding lip-edge loss.

$$\mathcal{L}_{ft} = \mathcal{L}_{total} + \lambda_{lip}\mathcal{L}_{lip} \tag{8}$$

We set $\lambda_{rec} = 1, \lambda_{pcp} = 0.001, \lambda_{style} = 10, \lambda_{adv} = 0.01, \lambda_{lip} = 0.1$.

4 EXPERIMENTS

4.1 Implementation Details and Metrics

Datasets. We borrow the Obama video from [Suwajanakorn et al. 2017] and another nine videos from HDTF [Zhang et al. 2021a] for our experiments. Each video has a duration of approximately 5 minutes. For initializing the TRF, we use the first 90% of frames from each video as the training set and reserve the remaining 10% of frames for testing. During the editing process, we only utilize 200 images from the original training set for training.

Fig. 5. Qualitative comparisons with the state-of-the-art methods. Wav2Lip only generates static heads. The lip shapes of MakeitTalk and SadTalker are far from that of the origin sequence. The two post-processing based methods ip2p, ip2p+deflicker suffer from significant flicker and distortion. Our method was able to generate the best results.

Implementation Details. Since our method introduces convolution-based refinement work, the random ray sampling strategy is unsuitable here. Instead, we apply a patch-based ray sampling to facilitate spatial dependencies between rays. We randomly crop a 256×256 patch from the edited images as the supervision of the TRF. When finetune our method with lig-edge loss, we only use a 64×64 patch of lip regions for training. Since the talking radiance field has already been initialized, we only need a few images to participate in the optimization during editing, further speeding up the model training. Typically, we set the number of images participating in training N to 200. We first train 300 epochs on a large patch and then finetune another 100 on a lip patch. We take Adam as the optimizer, and the learning rate is set as 5e-4 for MLP and 2e-4 for our refinement network. All experiments are conducted on an NVIDIA RTX 3090.

Evaluation Metrics. We evaluate our method for editing equality on four main aspects. To assess the audio-lip synchronization, we follow previous works [Guo et al. 2021] and employ the identity-agnostic SyncNet confidence (**Sync score** [Chung and Zisserman 2017]). We evaluate the instruction editing quality with the directional similarity in CLIP space (**CLIP Direction** [Gal et al. 2021]). This metric measures the consistency of the change between the two images (in CLIP space) with the change between the two image captions. We manually provide pre- and post-editing image captions corresponding to the editing instructions. For example, we provide "A photograph of a man" and "A photograph of a young man" for the instruction "Make him look like a young man". Additionally, we employ a face identity distance loss (**ArcFace** [Deng et al. 2018]) to measure face identity preserving. Finally, we measure temporal inconsistency based on a warping error ($\mathbf{E_w}$ [Lei et al. 2023]) that considers both short-term and long-term warping errors.

Table 1. Quantitative comparison results. Our method achieves top rankings in all four metrics. We show the best in bold and the second with underline.

Methods	↑ Sync score	↑ CLIP Direction	↓ ArcFace	↓ E_w
Wav2Lip [Prajwal et al. 2020]	**8.82**	0.0477	2.54	-
MakeitTalk [Zhou et al. 2020]	4.48	0.0518	2.36	0.0053
SadTalker [Zhang et al. 2022]	5.36	0.0442	2.86	<u>0.0045</u>
ip2p [Brooks et al. 2022]	5.17	<u>0.0597</u>	**1.81**	0.0120
ip2p+deflicker [Lei et al. 2023]	6.61	0.0435	2.16	0.0054
Ours	<u>6.66</u>	**0.0642**	<u>1.99</u>	**0.0044**

Table 2. User studies. Participants prefer our approach on all four aspects.

Methods	Lip Sync.	Instruction Editing Quality	Face Identity Preserving	Overall Video Quality
Wav2Lip [Prajwal et al. 2020]	1.75%	1.75%	0.75%	1.25%
MakeitTalk [Zhou et al. 2020]	1%	2.75%	1.0%	1.25%
SadTalker [Zhang et al. 2022]	<u>13.5%</u>	<u>17%</u>	<u>16.75%</u>	<u>27%</u>
ip2p [Brooks et al. 2022]	5%	4.75%	6.25%	2.0%
ip2p+deflicker [Lei et al. 2023]	8%	4.0%	8.5%	4.75%
Ours	**70.75%**	**69.75%**	**66.75%**	**63.75%**

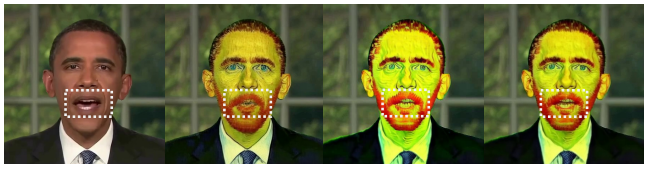

Fig. 6. Illustration of ablation study with instruction "Make him look like Van Gogh". We show the results of baseline and after adding progressive dataset update (PDU) and lip-edge loss.

Table 3. Ablation study comparing our full method with and without progressive dataset update (PDU) and lip-edge loss.

Methods	↑ Sync score	↑ CLIP Direction	↓ ArcFace	↓ E_w
Baseline	3.604	**0.1070**	3.0580	0.0045
Baseline+PDU	4.597	0.0708	2.9642	**0.0042**
Baseline+PDU+L_{lip}	**4.954**	0.0442	**2.6490**	0.0063

4.2 Comparisons

Comparison settings. Since there is no method to consider instruction-based talking face editing, we compare our method with some two-stage methods. These methods consist of two main categories: (1) editing an image first using ip2p and then applying it to one-shot talking face generation methods (Wav2Lip[Prajwal et al. 2020], MakeitTalk[Zhou et al. 2020], SadTalker[Zhang et al. 2022]) (2) using RAD-NeRF to render the target image first, and then post-process it using ip2p. The first type of method does not enable editing tasks such as novel view synthesis and background replacement, and the second type of method may result in significant flicker artifacts. We use the latest deflickering method[Lei et al. 2023] to post-process the second type of method (ip2p+deflicker).

Qualitative results. We first illustrate some results of our method, as shown in Figure 4 and Figure 7. Our method enables complex 3D talking face editing using simple instructions. Leveraging dynamic neural radiance fields, we can generate high-quality, temporally consistent, and audio-visual lip-synchronized talking faces based on the audio input. Figure 5 shows two editing results for comparisons. Wav2Lip can only produce a static head result. MakeitTalk generates a lip shape that differs significantly from the original result. SadTalker, the current SOTA one-shot-based talking face generation method, can produce high-quality results but does not match the lip shape's origin well. The post-processing-based network can restore a more consistent lip shape but suffers from flicker and distortion, while our method produces the best editing results. For a more immersive experience, please refer to our supplementary video.

Quantitative Results. We show quantitative results in Table 1. Wav2Lip only focuses on generating lip movements and achieving the highest audio-lip synchronization. We disregard evaluating its video consistency as it only generates static heads. MakeitTalke performs poorly in audio-lip consistency, and SadTalker can generate high-quality talking faces but fails to preserve the original identity and guarantee instruction editing direction. The method utilizing ip2p for post-processing exhibits poor temporal consistency. After deflickering, there is some improvement in generated quality, but it compromises the original identity and editing direction. Our method achieves top rankings in all four metrics, either first or second.

User Studies. We conduct user studies to evaluate better the quality of generated talking faces. We generate 20 videos from different characters and text instructions. We invite 20 participants and let them choose the best method for lip synchronization, instruction editing quality, face identity preserving, and overall naturalness. We show the result in Table 2, where the participants prefer our method mostly in all four evaluation terms. We attribute this to the superiority of our method in both audio-lip consistency and video quality.

4.3 Ablation study

We conduct our ablation study on the Obama video. Firstly, we finetune RAD-NeRF in edited images by ip2p with a large text guidance scale and take this as our baseline method. Then we add the progressive dataset update strategy (PDU) and lip-edge loss on it step by step. Figure 6 show the results of our methods with the instruction "Make him look like Van Gogh." Large text guidance will destroy the lip's shape, shown in a white box. Adding PDU makes the lip structure more defined after several iterative edits with small text guidance. Our lip-edge loss further aligns the lip shape with the origin.

We show the quantitative results in Table 3. Both PDUs and lip-edge loss bring a significant improvement in audio-lip synchronization. On the other hand, both methods lead to reduced image details and make the results closer to the original image (ArcFace) and away from the editing target (CLIP Direction). We find it a trade-off between audio-lip synchronization and image details, but enhancing the former is more important for editing the talking face.

Figure 2 shows the ablation of the refinement network. Since we can control the role of the refinenet's output on the final result with the weight, it is equivalent to disabling the refinement network when the weight is set to 0. We can see from Figure 2 that the final result lacks detail. We also conduct experiments on the efficiency of rendering. On NVIDIA 2080 Ti, disabling the refinement network

achieves a rendering speed of about 30FPS. When the refinement network is enabled, the rendering speed is about 25FPS, which also supports real-time rendering.

5 CONCLUSION

We present Instruct-NeuralTalker, the first novel interactive framework to edit talking radiance fields using instructions. Instruct-NeuralTalker greatly expands the ability to edit 3D talking faces. It enables users to generate personalized talking faces with their instructions. In order to keep the audio-lip synchronization, we introduce a progressive dataset update strategy and lip-edge loss to constrain the lip shape. We also introduce a refinement network to overcome over-smoothed results and support controllable detail generation. In addition, our approach can achieve real-time rendering on consumer-grade hardware. Finally, multiple evaluation metrics and user studies demonstrate that our method achieves higher quality editing than the state-of-the-art.

REFERENCES

Matthew Brand. 2005. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99*. https://doi.org/10.1145/311535.311537

Christoph Bregler, Michele Covell, and Malcolm Slaney. 2005. Video Rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*. https://doi.org/10.1145/258734.258880

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2022. InstructPix2Pix: Learning to Follow Image Editing Instructions. *ArXiv* abs/2211.09800 (2022).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *ArXiv* abs/2005.14165 (2020).

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 9630–9640.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. In *European Conference on Computer Vision*.

Qifeng Chen and Vladlen Koltun. 2017. Photographic Image Synthesis with Cascaded Refinement Networks. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 1520–1529.

Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. 2021. Stylizing 3D Scene via Implicit Representation and HyperNetwork. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2021), 215–224.

Joon Son Chung and Andrew Zisserman. 2017. *Out of Time: Automated Lip Sync in the Wild*. 251–263. https://doi.org/10.1007/978-3-319-54427-4_19

Jiankang Deng, J. Guo, and Stefanos Zafeiriou. 2018. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 4685–4694.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2020. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw.2019.00038

Cletus G. Fisher. 2014. Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research* (Jul 2014), 796–804. https://doi.org/10.1044/jshr.1104.796

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ArXiv* abs/2108.00946 (2021).

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. Texture Synthesis Using Convolutional Neural Networks. In *NIPS*.

Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas A. Funkhouser. 2020. Object-Centric Neural Scene Rendering. *ArXiv* abs/2012.08503 (2020).

Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 5764–5774.

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. *ArXiv* abs/2303.12789 (2023).

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM Trans. Graph.* 41 (2022), 161:1–161:19.

Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2021. Learning to Stylize Novel Views. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 13849–13858.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. 2022. Decomposing NeRF for Editing via Feature Field Distillation. *ArXiv* abs/2205.15585 (2022).

Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. 2023. Blind Video Deflickering by Neural Filtering with a Flawed Atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen Lele, KMaddox Ross, Duan Zhiyao, and Xu Chenliang. 2019. Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss.

Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. 2022. Language-driven Semantic Segmentation. *ArXiv* abs/2201.03546 (2022).

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *ArXiv* abs/1411.1784 (2014).

Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. 2021. Extracting Triangular 3D Models, Materials, and Lighting From Images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 8270–8280.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics* (Jul 2022), 1–15. https://doi.org/10.1145/3528223.3530127

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv* abs/2203.02155 (2022).

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.01018

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr52688.2022.01042

Shuai Shen, Wanhua Li, Zhengbiao Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. 2022. Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis. In *European Conference on Computer Vision*.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. 2020. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 7491–7500.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama. *ACM Transactions on Graphics (TOG)* 36 (2017), 1 – 13.

Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. 2022. Real-time Neural Radiance Talking Portrait Synthesis via Audio-spatial Decomposition. *ArXiv* abs/2211.12368 (2022).

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. *Neural Voice Puppetry: Audio-driven Facial Reenactment*. 716–731. https://doi.org/10.1007/978-3-030-58517-4_42

Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. 2022. Neural Feature Fusion Fields: 3D Distillation of Self-Supervised 2D Image Representations. *2022 International Conference on 3D Vision (3DV)* (2022), 443–453.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. 2021. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 5481–5490.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. Realistic Speech-Driven Facial Animation with GANs.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2021. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 3825–3834.

Can Wang, Ruixia Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022. NeRF-Art: Text-Driven Neural Radiance Fields Stylization. *ArXiv* abs/2212.08070 (2022).

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. *ArXiv* abs/1809.00219 (2018).

Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. 2022. NeuMesh: Learning Disentangled Neural Mesh-based Implicit Field for Geometry and Texture Editing. *ArXiv* abs/2207.11911 (2022).

Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. 2021. Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 13759–13768.

Shunyu Yao, Ruizhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. 2022. DFA-NeRF: Personalized Talking Head Generation via Disentangled Face Attributes Neural Rendering. *ArXiv* abs/2201.00791 (2022).

Zhichao Yin and Jianping Shi. 2018. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 1983–1992.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models.

Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. 2021b. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)* 40 (2021), 1 – 18.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2022. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *arXiv preprint arXiv:2211.12194* (2022).

Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul E. Debevec, William T. Freeman, and Jonathan T. Barron. 2021c. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. *ACM Trans. Graph.* 40 (2021), 237:1–237:18.

Zhimeng Zhang, Lincheng Li, and Yu Ding. 2021a. Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 3660–3669.

Yang Zhou, Dingzeyu Li, Xintong Han, Evangelos Kalogerakis, Eli Shechtman, and Jose I. Echevarria. 2020. MakeItTalk: Speaker-Aware Talking Head Animation. *ArXiv* abs/2004.12992 (2020).

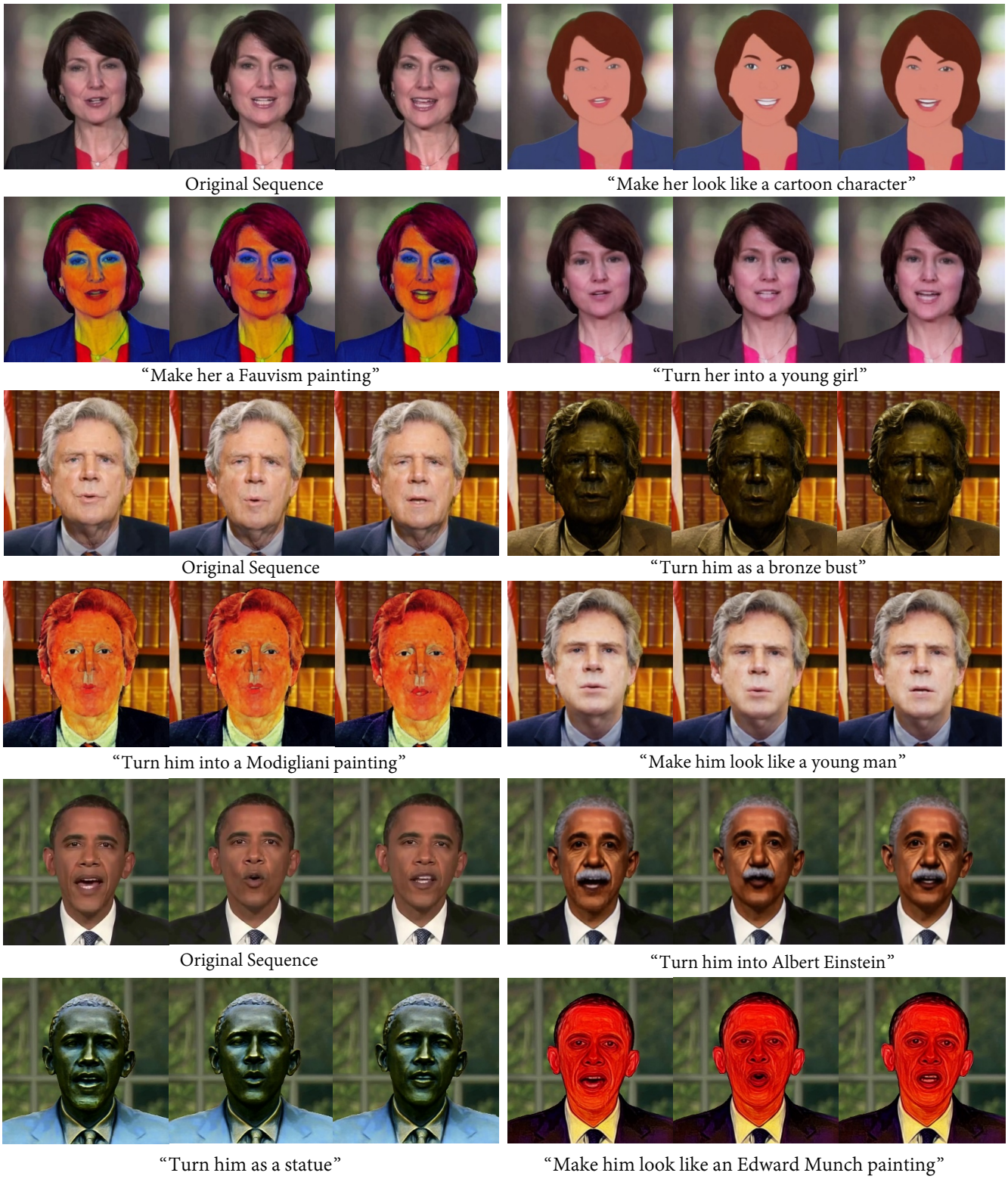

Fig. 7. More editing results of Instruct-NeuralTalker.

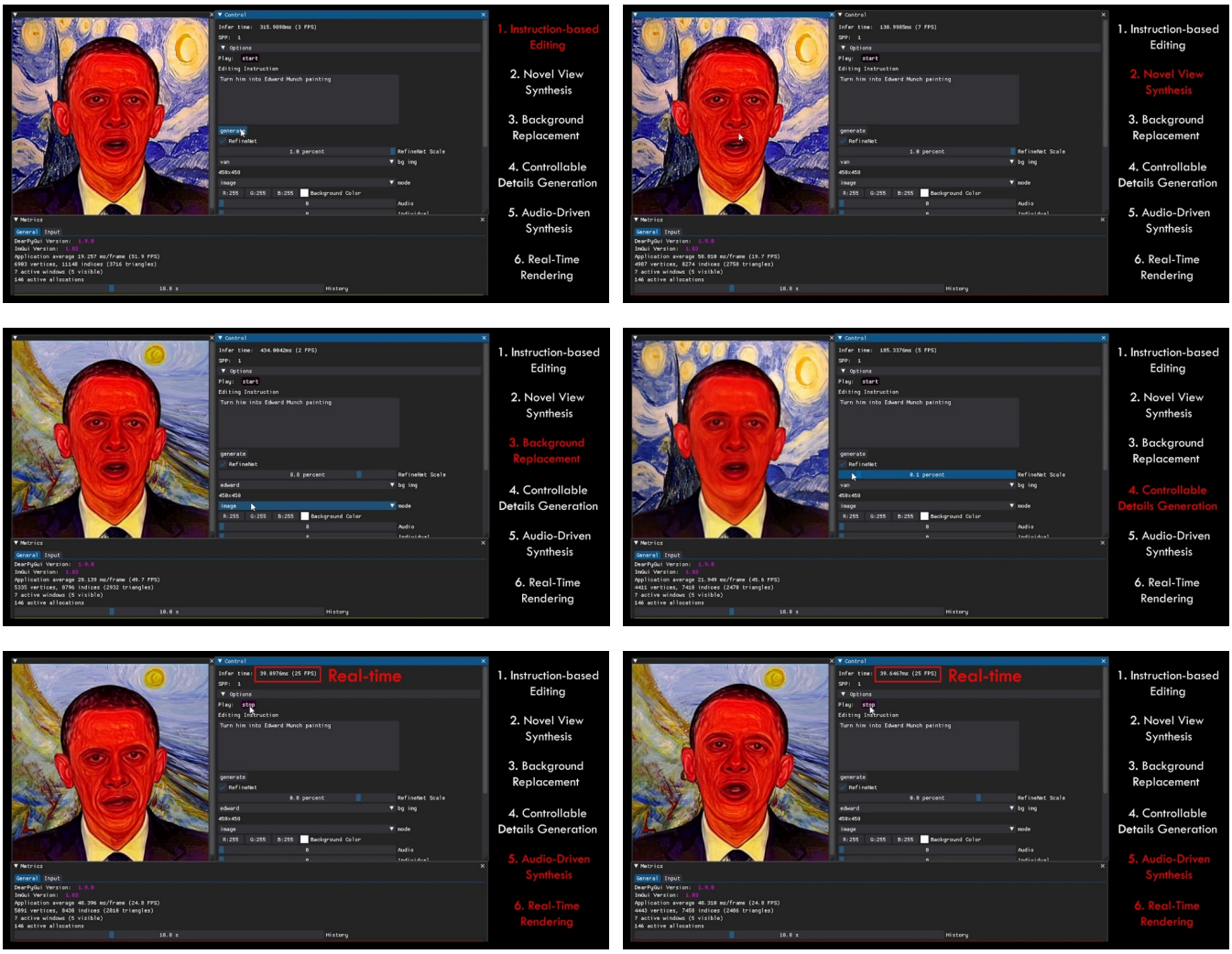

Fig. 8. User interface. Please refer to our demo video for more details.