Supplemental Material: RADIO: Reference-Agnostic Dubbing Video Synthesis

Dongyeun Lee[1,*] Chaewon Kim[1,*] Sangjoon Yu[1] Jaejun Yoo[2,†] Gyeong-Moon Park[3,†]
[1]Klleon AI Research [2]UNIST [3]Kyung Hee University

A. Architectural Details

In this section, we provide a detailed description of the RADIO encoders and the sync discriminator.

Encoders. The two encoders (E_c, E_s) mentioned in the main paper have a similar structure consisting of residual-down blocks. The E_c and E_s receive RGB frames with 192×192 resolution and pass four numbers of a residual-down block along with two additional convolution blocks. A single residual-down block involves two convolutional layers (kernel size=3) and LeakyReLU activation with a skip connection that adds intermediate features passed through the additional convolutional layer (kernel size=1). The $f_c \in \mathbb{R}^{12\times12\times512}$, which is a output of E_c, is fed to generator G. The only difference between E_c and E_s is that E_s extracts $f_s \in \mathbb{R}^{1\times1\times512}$ via a spatial dimensional global average pooling operation and a fully-connected layer after four residual-down blocks.

The audio encoder E_a receives mel-spectrogram as inputs and encodes to $f_a \in \mathbb{R}^{1\times1\times512}$ through 2D convolutional layers. The E_a is implemented to follow [2] structure. [2] is considered one of the most effective architectures for speaker recognition using audio inputs and comprises SE layers and self-attention pooling with ResNet layers. We modified the activation function as LeakyReLU and normalization as instance normalization.

Sync Discriminator. The sync discriminator consists of an encoder that receives mel-spectrogram as input and an encoder that receives facial images as input. The audio encoder features follow exactly the structure of [2], while the visual encoder utilizes channel-attention and spatial-attention operations instead of self-attention of [2]. This is because it is important to concentrate on the mouth's shape within the face image or the local area around it. Finally, the sync discriminator is pre-trained with a loss function (eq. 4 in main paper) to increase the cosine similarity (eq. 5 in main paper) of the vision and audio features so that we can provide superior audio-video synchronization errors during the RADIO training scheme.

We pre-trained the sync discriminator on the LRW [3] dataset using tightly zoomed face images with a resolution of 144×144. This approach allowed the discriminator to focus specifically on the lip shape of the synthesized facial image. In addition, the images were converted to grayscale, making the sync discriminator color-agnostic and enabling it to focus solely on learning the sync accuracy of mouth shapes.

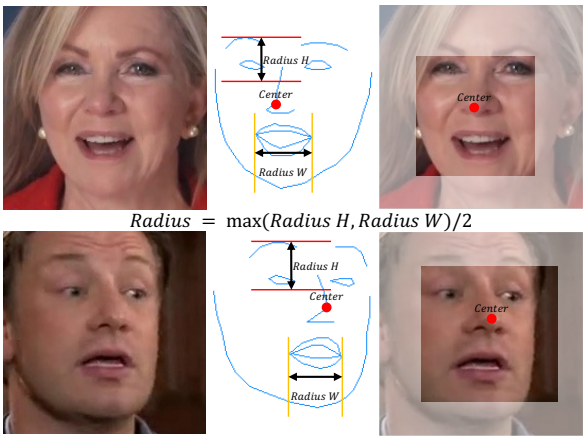

Figure 1. **Cropping method used for evaluation.** We applied this cropping method to all generated baseline results to evaluate the quantitative metrics with the same ground truth images.

B. Pre-processing algorithm for evaluation.

Including our proposed method, baselines generated different sizes of images with different alignments and different target regions for synthesis. In order to evaluate quantitative metrics fairly, we applied a pre-processing algorithm to all generated images before comparison. First, we aligned all baseline methods with FFHQ alignment [4]. Then, we applied a face cropping method based on the DINet [7] masking algorithm. Last, we resized tightly cropped faces to the same resolution.

Figure 1 depicts the face cropping method. We assumed that the tip of the nose (the thirty-fourth point of the facial landmark) is in the center of the human face. Then, we com-

*Equal Contribution. † Corresponding Authors.

puted two radius values: $Radius\ H$, which measures the distance between the highest point and the thirtieth point along the y-axis of the facial landmark, and $Radius\ W$, which measures the distance between the fifty-fifth point and the forty-ninth point along the x-axis of the facial landmark. We then set the final $Radius$ value as the maximum value between these two distances. Finally, we cropped the attached facial image with this $Radius$ value, starting from the thirty-fourth point on the facial landmark.

C. Baseline Models

In this section, we describe additional details about baselines mentioned in Section 4.1 of the main paper.

ATVGnet. ATVGnet [1] proposes constructing high-level representation (facial landmarks) from the audio signal and generating talking head videos conditioned on the facial landmark. ATVGnet leverages the pixel-wise loss with attention mechanisms to ensure temporal consistency and utilizes a regression-based discriminator to generate accurate facial shapes and realistic-looking images in the training scheme. Finally, ATVGnet can only generate 128×128 resolution videos and cannot keep up with the head motion of the target frames.

MakeItTalk. MakeItTalk [9] also proposes an audio-to-landmarks approach for controlling the motion of lips while determining the specifics of facial expressions and the rest of the talking-head dynamics from an audio signal. After that, MakeItTalk generates talking head animations (256×256 resolution) with a single image (cartoon or natural human) and predicted landmarks using image-to-image translation. MakeItTalk animates talking head videos based on facial landmarks extracted from audio signals. However, these facial landmarks are too sparse to describe lip motion details and do not represent significant head motion.

Wav2Lip. To the best of our knowledge, Wav2Lip [5] is the first approach to utilize a pre-trained Sync Discriminator in a training scheme and generate the lower half masked of the target frame. This method guarantees high audio-visual synchronization. However, it is highly dependent on the reference frame by feeding with concatenating the reference frame and masked input frame, and generates blurry results.

DINet. DINet [7] proposes a deformation inpainting network, which performs spatial deformation on feature maps of reference images to synthesize high-fidelity dubbing videos. DINet uses five reference facial images to create deformed features in order to align head poses and driving audio to preserve high-frequency details. In addition, DINet develops its masking algorithm around the lip, resulting in efficient inpainting synthesis of mouth shapes. Finally, DINet can generate high-resolution (416×320) videos. Although DINet utilizes multiple reference images, the synthesized results vary sensitively depending on the selected reference images. Also, their framework is restricted to frontalized head poses, and generates artifacts when the mouth region covers the background.

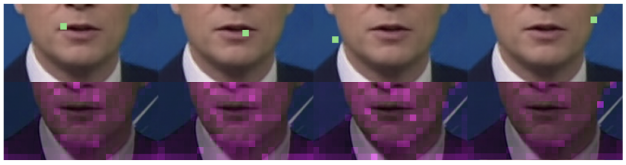

Figure 2. **Visualization of attention map in the last ViT block.** We visualized the green patches on generated frames (upper half) with the attention map on reference frames (lower half).

Configuration	PSNR↑	LPIPS↓	Sync↑
Baseline	33.089	0.072	0.576
+ ViT$(1,1)$	34.637	0.033	0.557
+ ViT$(1,2)$	34.757	0.032	0.559
+ ViT$(2,2)$	**34.938**	**0.031**	**0.609**

Table 1. **Ablation for the different number of attention layers.**

IP-LAP. IP-LAP [8] follows a two-stage training scheme like any other method of utilizing facial landmarks. IP-LAP is implemented to leverage facial sketch maps rather than face landmark coordinates so the framework can learn the driving face shapes clearly. Finally, IP-LAP aligns the twenty-five reference images using a warping-based alignment module and utilizes them to generate 128×128 resolutions while preserving the target head pose and expression. Despite the usage of a large number of reference images, the accuracy of the lip shape is insufficient to learn the audio-visual synchronization only with facial landmarks.

D. Ablation Study of ViT design

Attention Visualization of last ViT block. Figure 2 visualizes the attention map of $Att_{6,2}$, located in the second layer of the attention block within the last ($L = 6$) decoder layer. The upper half of the figure displays the green patches on the generated frames, while the lower half presents the corresponding attention map on the reference image. In contrast to Figure 5 in the main paper, each patch attended to the entire image for the last ViT block. We relate this phenomenon to the hierarchical nature of StyleGAN2 [4], which generates course-to-fine information for low-to-higher layers. While the intermediate attention layer, *i.e.*, $Att_{5,2}$, focused on the globally relevant features for each local patch, the last attention layer, *i.e.*, $Att_{5,2}$, captured the fine-grained textures and colors across the entire image.

Design of ViT attention layers. We additionally present quantitative results, evaluated on the LRW [3] validation dataset, for an ablation study of RADIO with varying num-

bers of ViT layers. Our evaluation metrics include PSNR, LPIPS, and the similarity score between audio and visual features. To obtain the similarity score, both the audio and visual features are encoded using the encoders of our sync discriminator described in Section A. We specifically used our pre-trained sync discriminator, which was trained with the LRW [3] training dataset, for accurate evaluation. For this experiment, all models were trained for 210K iterations with a batch size of 16, with resolution scaled down to 96×96, like the main ablation experiment. We only conducted the experiment with ViT on the last two layers of the decoder, because patches for earlier layers were too small to deliver semantically interpretable results. For example, the feature resolution is 12×12 for the fourth decoder layer, which is too small to divide into patches.

In Table 1, the baseline refers to the framework that generates audio-driven images by decoder layers modulated with style features, without additional components for fidelity mapping (method B in Table. 2 of the main paper). We denote ViT(n, m) as our RADIO framework with n ViT blocks, consisting of m attention layers. Note that in our main experiment, we applied ViT block to the last two decoder layers, with each block comprising two attention layers, *i.e.*, ViT$(2, 2)$. The results indicated that having two attention layers in a single ViT block was better than using only one layer. Additionally, employing ViT blocks in two decoder layers was more effective than placing them in a single decoder layer. Finally, ViT$(2, 2)$ achieved the best PSNR, LPIPS, and sync similarity scores compared to the baseline.

E. Additional Experimental Results

In this section, we show the additional experimental results of RADIO in Figure 3 and Figure 4. Throughout all examples, our results consistently generated the most natural and realistic mouth shapes, with high synchronization accuracy compared to the ground truth. ATVGNet, MakeItTalk, and PC-AVS commonly failed to generate identity-preserving details. Wav2Lip consistently created blurry images and generated artifacts for extreme face poses. Especially in harsh scenarios, IP-LAP and DINet struggled to generate realistic-looking mouth shapes, due to the significant distortion caused by warping and deformation. The mouth shapes generated by these methods were similar across all time steps, which also led to a degradation in synchronization quality. Especially, DINET failed to generate realistic faces with extreme poses, as their framework is limited to generate frontalized faces.

F. Limitation and Broader Societal Impact

While our model excels in producing high-quality images around the mouth region, it struggles to generate a natural-looking background. During our evaluation, we observed that frames significantly misaligned with the reference frame exhibited artifacts in the background. This limitation is observed across all baseline models [5, 8], but is more conspicuous for ours due to the alignment method that includes a larger portion of the background for generation. This issue can be easily fixed by borrowing a face-parsing model [6] to attach only the face region to the original video, thus improving the overall video quality.

Previous one-shot audio-driven frameworks have struggled to consistently generate realistic, high-fidelity frames. These challenges arise because they heavily rely on the reference image, which typically requires a frontalized pose with a neutral facial expression. In contrast, our reference-agnostic framework demonstrates exceptional capabilities in generating high-quality dubbed videos, even in the most challenging scenarios. This makes it suitable for a wide range of real-world industrial applications where diverse poses and expressions are encountered. We look forward to the application of our framework to generate realistic audio-driven faces for unseen speakers in real-time. Looking ahead, we aspire to enhance and extend our RADIO framework to support higher resolutions in the near future.

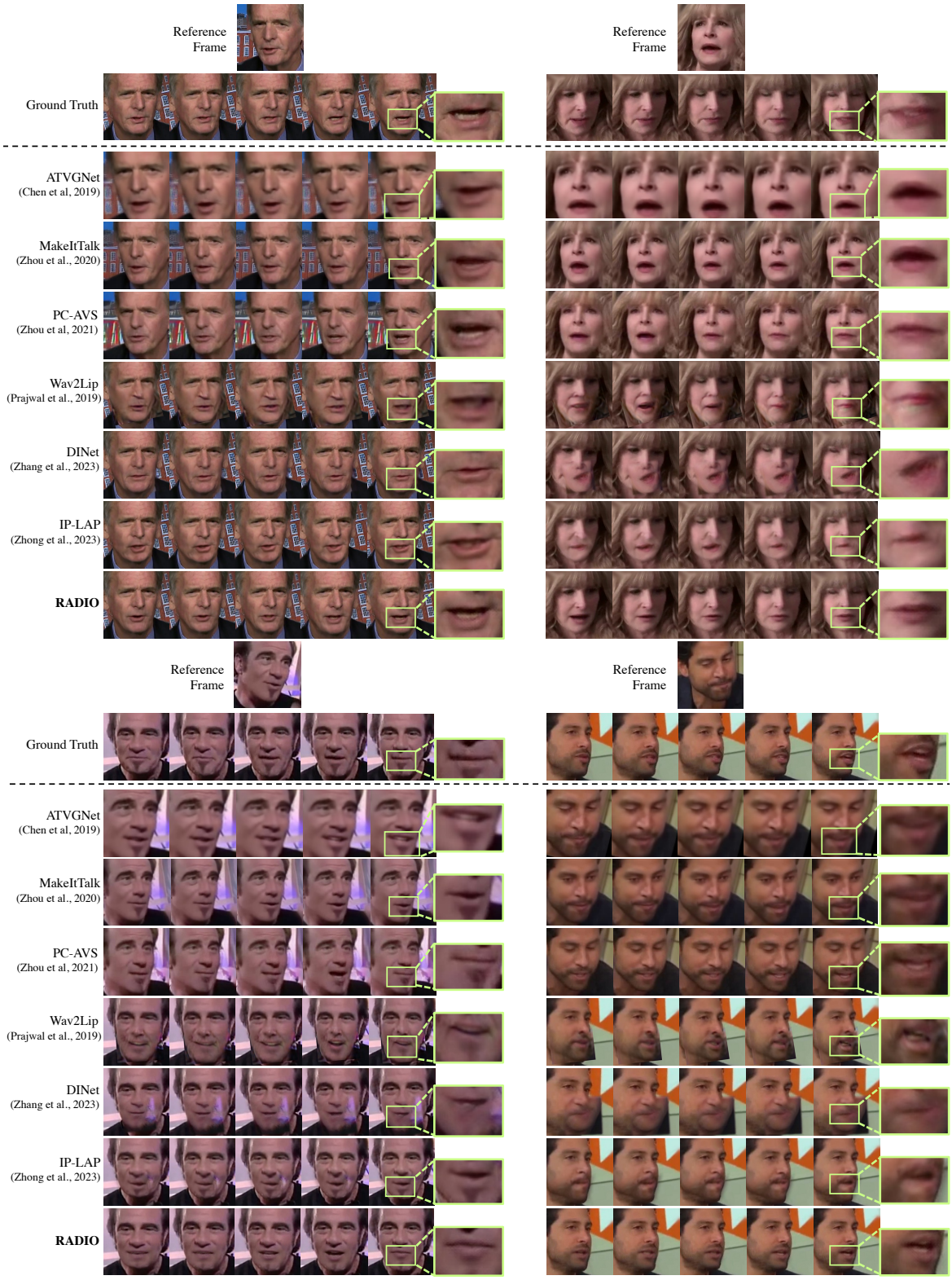

Figure 3. **Qualitative comparison with baselines.** We visualized the dubbed results for challenging scenarios where the ground truth pose and expression significantly differ from the reference frame.

Figure 4. **Qualitative comparison with baselines.** We visualized the dubbed results for challenging scenarios where the ground truth pose and expression significantly differ from the reference frame.

References

[1] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 2

[2] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. In *Proc. Interspeech*, 2020. 1

[3] A. Zisserman J. S. Chung. Lip reading in the wild. In *Asian Conference on Computer Vision (ACCV)*, 2016. 1, 2, 3

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[5] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 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*, page 484–492, 2020. 2, 3

[6] Xuansong Xie Tao Yang, Peiran Ren and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[7] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *AAAI*, 2023. 1, 2

[8] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2023. 2, 3

[9] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 2