Talking Head Generation with Probabilistic Audio-to-Visual Diffusion Priors -Supplementary Material

Zhentao Yu[1*] Zixin Yin[1,2*†] Deyu Zhou[1,3*†] Duomin Wang[1]
Finn Wong[1] Baoyuan Wang[1‡]
[1]Xiaobing.AI.
[2]The Hong Kong University of Science and Technology.
[3]The Hong Kong University of Science and Technology (Guangzhou).
(yuzhentao,wangduomin,wangwenlan,wangbaoyuan)@xiaobing.ai
zyinaf@connect.ust.hk, dzhou861@connect.hkust-gz.edu.cn

I. Implementation Details

I.1. Pre-processing Details

In the training stage, all input images are cropped and resized to 224 × 224 following the data pre-processing pipeline of [6] with a face detector [11] and a landmark detector [3]. For the pre-processing of audio input, we follow [25] and convert the audio to mel-spectrogram with a sampling rate of 16kHz. Note that for each video frame, we extract an audio segment of 0.2s in the video centered at the video frame to construct an audio-video training sample.

I.2. Network Details

Our main framework consists of 6 modules:

- **Identity Encoder $\mathbf{E}_{id}$**, a pretrained ResNeXt50 [23].
- **Visual Encoder $\mathbf{E}_v$**, also named as non-identity encoder. It is a MobileNetV2 [15] following the pretraining scheme of LPD [4]. MLP$_{ol}$ and MLP$_{cl}$ are applied after E$_v$ to project the visual feature $\mathbf{f}^v$ into two subspace $\mathbf{f}^v_{ol}$ and $\mathbf{f}^v_{cl}$, respectively. The dimension of MLP$_{ol}$ and MLP$_{cl}$ are both 512×512. Then, they are separately mapped to two complementary features, $\mathbf{f}^v_{nl}$ for non-lip and $\mathbf{f}^a_l$ for lip features through MLP$_{nl}$ and MLP$_{a2l}$, respectively. The dimension of MLP$_{nl}$ is 512×42 and the dimension of MLP$_{a2l}$ is 512×470.
- **Audio Encoder $\mathbf{E}_a$**, a ResNet34 encoder from [5].
- **Prior Network $\mathbf{P}_{a2nl}$**, a 6-layer Transformer [19] encoder, with 512-d tokens and 1024-d fully forward layers. The positional embeddings are learnable. Note that our diffusion prior and auto-regressive prior networks share the same architecture but with different attention mechanisms, noted as bidirectional attention and causal attention, respectively. The max length of the input tokens of the encoder is set to 128.
- **Generator G**, borrowed from StyleGAN2 [10] and has the same modulated convolution as mentioned in [25].
- **Discriminator D**, same as the discriminator used in [25].

Three other pre-trained models are utilized during reconstruction learning and quantitative evaluation, including:

- **Gaze Encoder [1]**, the last 512-d feature of the encoder is used for the calculation of L_{gaze}.
- **VGG Network [16]** , for the calculation of VGG loss in [25].
- **Deep3DFace [7]**, a 3DMM model extracting the 3-d pose and 64-d expression coefficients for the evaluation of **FID$_{fm}$**, **FID$_{\Delta fm}$** and **SND**.

I.3. Loss Details

Lip & Non-lip Disentanglement

- **Audio-Visual Contrastive Learning** We use the same implementation as in CLIP [13] for the contrastive learning of audio encoder E$_a$ and a pre-trained visual encoder E$_v$ [4]. We utilize the audio and frames from the same video to construct a contrastive batch, where the corresponding pairs are positive pairs. Our models were trained on 4 A100 GPUs for 30 epochs with batch size of 288. The initial learning rate is set to $1e^{-5}$ with a decay rate of 0.93 for every 200,000 steps.
- **Reconstruction Learning for Non-Lip Space** The

---

[*] These authors have contributed equally to this work.

[†] This work was done when Zixin Yin and Deyu Zhou were interns at XiaoBing.AI.

[‡] Corresponding author.

loss formulas are listed as follows [4] [25]:

$$L_{\text{GAN}} = \min_{\mathrm{G}} \max_{\mathrm{D}} \sum_{n=1}^{N_D} \mathbb{E}_{I_{(i)}}[\log \mathrm{D}_n(I_{(i)})] \\ + \mathbb{E}_{\mathbf{f}_{cat(i)}}[\log(1 - \mathrm{D}_n(\mathrm{G}(\mathbf{f}_{cat(i)})))], \tag{I}$$

$$L_{L1} = \sum_{n=1}^{N_D} \left\| \mathrm{D}_n(I_{(i)}) - \mathrm{D}_n(\mathrm{G}(\mathbf{f}_{cat(i)})) \right\|_1, \tag{II}$$

$$L_{\text{VGG}} = \sum_{n=1}^{N_G} \left\| \mathrm{VGG}_n(I_{(i)}) - \mathrm{VGG}_n(\mathrm{G}(\mathbf{f}_{cat(i)})) \right\|_1, \tag{III}$$

$$L_G = \lambda_{ol} \cdot L_{ol} + \lambda_{gaze} \cdot L_{gaze} + \lambda_{L1} \cdot L_{L1} \\ + \lambda_{\text{GAN}} \cdot L_{\text{GAN}} + \lambda_{\text{VGG}} \cdot L_{\text{VGG}}, \tag{IV}$$

where all λs are set to 1. $I_{(i)}$ and $\mathrm{G}(\mathbf{f}_{cat(i)})$ are the GT image and the generated image of the i-th sample, respectively. L_{GAN} is a multi-scale generative adversarial loss. N_D is the number of layers in discriminator D and the subscript n implies the n-th layer of D. L_{L1} is the L1 distance between the n-th layer features of GT and generated image extracted from D. Similar to L_{L1}, L_{VGG} is defined as the L1 distance between two features extracted from a pre-trained VGG network. L_{gaze} and L_{ol} are described in the paper.

In the disentanglement of lip and non-lip, the learning rate of two MLPs of E_v is initialized as $1e^{-5}$ with a decay rate as 0.5 for every 80,000 steps. The learning rates of G and D are initialized as $2e^{-5}$ and $3.5e^{-6}$, respectively, with the same decay rate. The batch size is set to 16 and the size of **MB**, K, is set to 32 which stores 512 features in total. After the 40k-th step, we freeze other modules and train only G and D for 5 epochs on 4 A100 GPUs.

Audio2Visual Prior After the disentanglement of lip & non-lip, we trained the prior network P_{a2nl} along with the pre-trained visual encoder E_v and audio encoder E_a, where E_v extracts the non-lip feature $\mathbf{f}^v_{nl}$ and E_a extracts the audio feature $\mathbf{f}^a$.

- **Diffusion Prior** The input of P_{a2nl} is a concatenated feature $cat(a_{1:L}, n^t_{1:L})$, where $a_{1:L}$ is an audio feature sequence and $n^t_{1:L}$ is a non-lip feature sequence with random noise added through the diffusion process. t is the time step and L is the length of the sequence. During training, t is uniformly sampled in [1, T] where T is set to 1000. MSE loss is calculated between the output of P_{a2nl} and the GT non-lip feature. At inference time, $n^t_{1:L}$ is replaced with random Gaussian noise. Note that bidirectional attention is used in P_{a2nl}, implying that denoised features are outputted at the same time. While trained with mask editing, $n^t_{1:L}$ is randomly replaced with GT values such that P_{a2nl} is trained to predict the unmasked region. As a result, the model is capable of interpolating non-lip sequences according to input audio and unmasked non-lip features, or generating non-lip sequences with input audio only at inference time. The model was trained on 1 A100 GPU for 50,000 steps with the batch size as 64 and L as 128. The learning rate is set to $1e^{-4}$.
- **Auto-Regressive (AR) Prior** The prior $\mathrm{P}_{a2nl}(\check{n}_i | \{\check{n}_{k<i}\}, a_{1:L})$ is modeled in an auto-regressive way where $k = [1, 2, ..., i-1]$ and $i = [1, 2, ..., L]$. Different from the diffusion prior, AR prior encodes the input with causal attention instead of bi-directional attention such that tokens can only refer to the previous tokens but not the succeeding ones. MSE loss is used to measure the prediction error of GT and the prediction. We use the same training setting as diffusion prior, *i.e.* batch size, etc.

I.4. Details about Mask Editing Mechanism

Figure I illustrates how our mask editing works during both the training and inference. As shown, our mask mechanism is very similar to the mask language modeling that is commonly used in Bert [8] and other pre-trained models, such as BeiTs [2, 20, 22]. Intuitively, we want to primarily count on the audio to infer the non-lip features, however, due to the weak correlation between audio and non-lip facial motions, additional conditions need to feed into the diffusion model to reduce the ambiguities of the mappings. As illustrated, during training, we empirically copy 10% frames of non-lip features from GT and ask the model to predict the rest 90% masked frames. This is inspired by the image inpainting works [14]. We leave a thorough study of the masking mechanism as future work.

I.5. Evaluation Details

Baselines Only methods that have pre-trained models released were chosen as baselines for fair comparisons. These methods are introduced as follows,

- **Wav2Lip [12]** generates the lower-half face given an identity image, an upper-half driving image, and an audio clip. Other facial regions remain unchanged.
- **MakeItTalk [26]** learns an identity-specific embedding and a speech-content embedding to predict facial landmarks. Face warping and image translation are applied afterward for face reenactment.
- **PC-AVS [25]** takes an identity image, a reference pose video, and an audio clip as inputs to generate a talking head video. It does not support other facial motions such as expression and eye blinks.

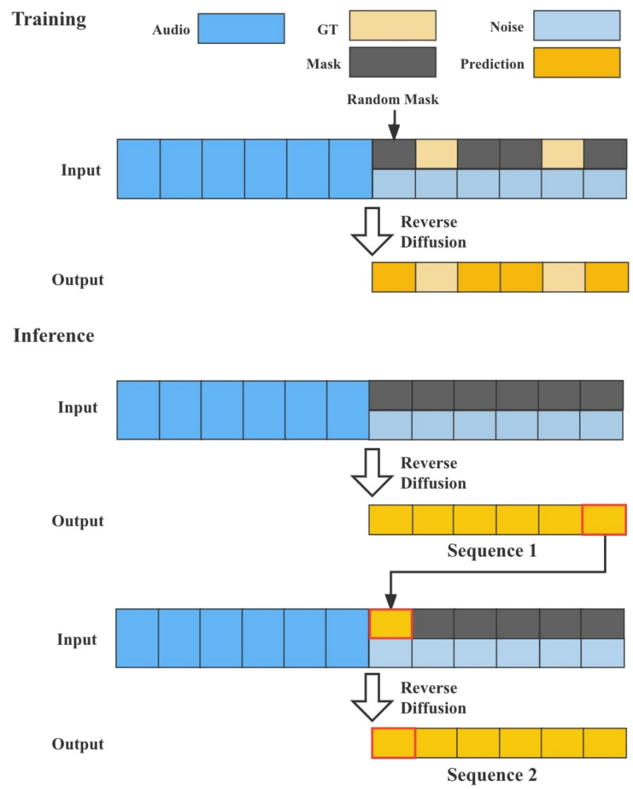

Figure I. Conceptual illustration of our mask editing technique during both training and inference. Here, "GT" denotes the non-lip feature from E_v. During training, we randomly masked 90% frames of the non-lip features and conditioned on the rest and audio input as well as noise to predict the masked non-lip features through the reverse diffusion process. During the inference, we feed the predicted non-lip feature of the last frame from the previous sequence as the unmasked non-lip features of the first frame in the next sequence, conditioned on which the subsequent masked non-lip features will be predicted, therefore smooth transition is achieved between two consecutive sequences.

- **Audio2head** [21] learns to predict the head pose auto-regressively given a reference image and an audio clip and then generates an audio-driven talking head accordingly.
- **EAMM** [9] generates a talking head video from an emotion video, a driving audio, an identity image, and a pose sequence.

Driving Settings Self-reenactment means driving an identity image with all signals from the same video clip of GT. Different from self-reenactment, cross-reenactment uses another video clip of GT as non-lip signals.

Lip & Non-Lip Disentanglement To evaluate the disentanglement between lip and non-lip on VoxCeleb2, we conduct two experiments to ensure that they do not interfere

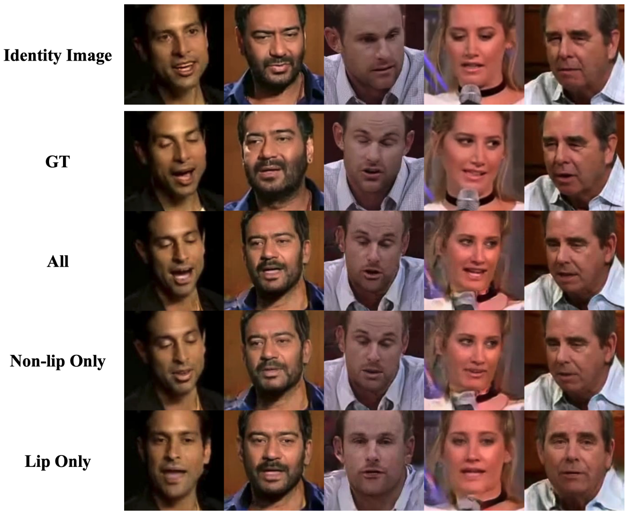

Figure II. Qualitative results on VoxCeleb2 [6] for non-lip signal only, lip signal only, and both of them, respectively.

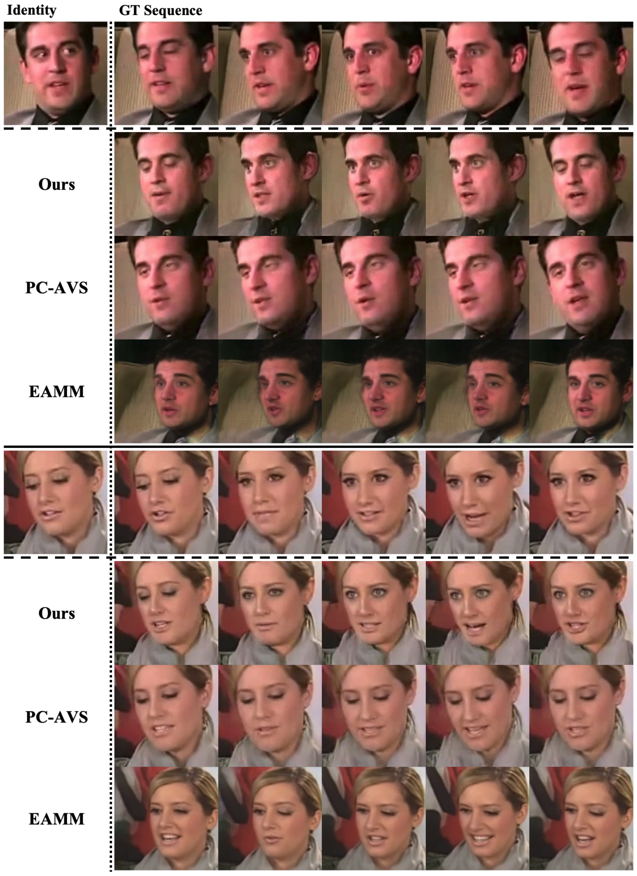

Figure III. Qualitative results on VoxCeleb2 with driving signals from the same video-audio pair as the identity image. Note that our model uses non-lip signals from video input, instead of the diffusion prior. Each row shows five uniformly sampled frames from videos.

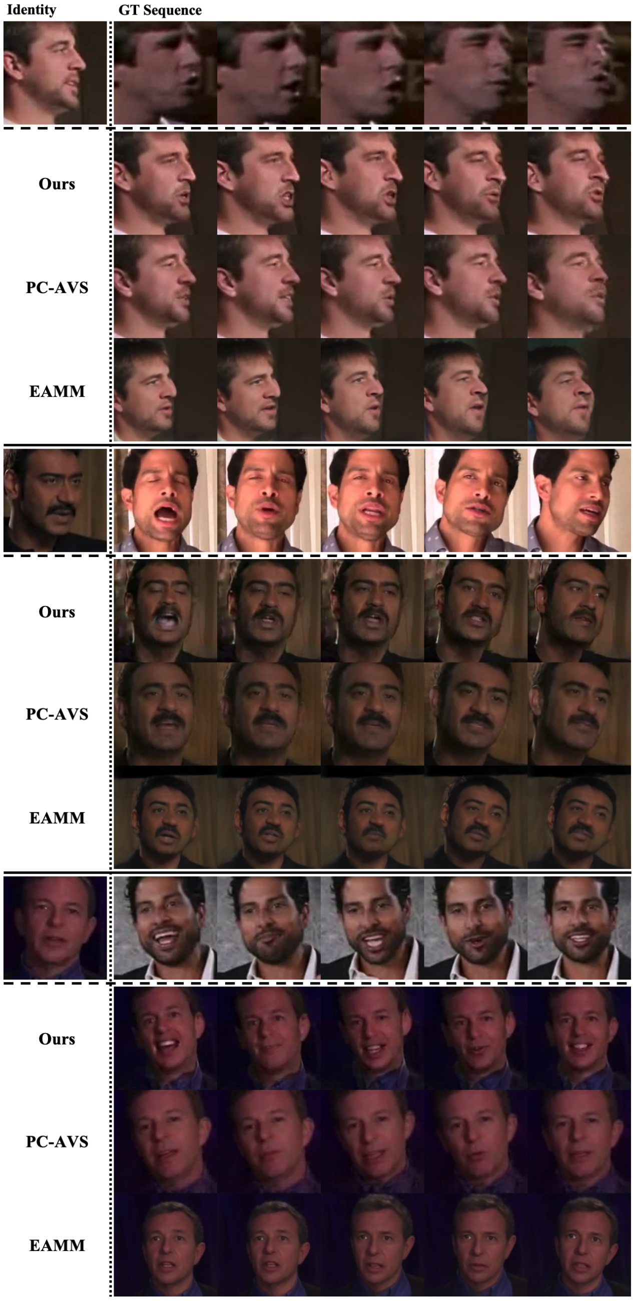

Figure IV. Qualitative results on VoxCeleb2 with driving signals from another video-audio pair. Note that our model uses non-lip signals from video input, instead of the diffusion prior. Each row shows five uniformly sampled frames from videos.

with each other. In the first experiment, we set non-lip features to all zeros when measuring lip accuracy, whereas in the second experiment, we set lip features to zeros when measuring non-lip accuracy. For the non-lip features, we use $\mathbf{f}^{v}_{nl}$ from the motion encoder rather than our diffusion model. We measure the Normalized Mean Error (**NME**) between the 2D landmarks [24] of GT images and generated images for blink and gaze, which we denote as B_d and G_d, respectively. For expression and pose, we calculate the L2 distances of the 3DMM coefficients using [7], which we denote as E_d and P_d, respectively.

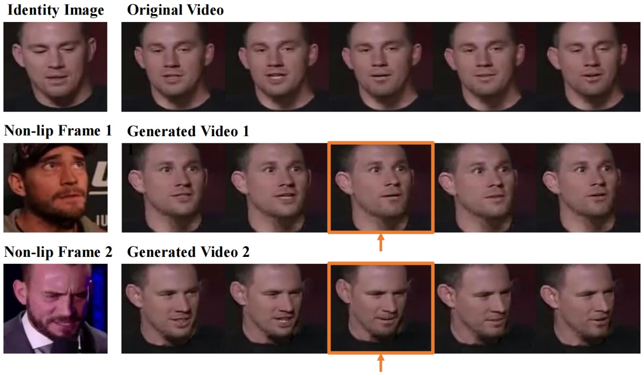

Figure V. Qualitative results of video editing. Each row shows five continuous frames in each video, where the frame with an orange box serves as the conditional non-lip features copied from the example frames shown on the left. Our model predicts smoothly transited sounding frames.

Multimodality To quantify the one-to-many diversity, we also investigated **Multimodality** from MDM [17] to measure the average distance of 3DMM parameters between different runs given the same inputs. However, due to the lack of one-to-many data, *i.e.*, multiple videos corresponding to the same audio, we only calculate it as 2.31 for future comparison.

II. Analysis and Results

II.1. Lip & Non-lip Disentanglement

Fig. II shows that non-lip signals can drive pose, expression, blink and gaze well with the mouth slightly opened. On the other hand, lip signals can only drive lip motions with others fixed.

We compare our proposed method with others in talking head reenactment, resulting in Fig. III. It can be observed that our method can control non-lip motions including pose, expression, blink, and gaze while lip motion is in-sync with the audio. It indicates that our method benefits from a well-disentangled motion space, which is a good foundation for our one-to-many diffusion prior.

Besides, we also showcase cross-id compared with other baselines as shown in Fig. IV. Our proposed method has more diverse motion than others and can control pose, expression, blink and gaze well.

II.2. Guidance Factor of Audio Condition

To study the impact of the guidance factor s on the diversity and performance of our method, we conduct ex-

s	Var $\rightarrow$	FID$_{fm}$ $\downarrow$	FID$_{\Delta fm}$ $\downarrow$	SND $\downarrow$
0.0	1.60	3.98	1.16	5.14
1.0	1.58	3.78	1.14	4.92
2.5	1.57	**3.60**	**1.08**	**4.68**
5.0	**1.85**	4.16	1.20	5.36
10.0	2.42	6.10	1.67	7.77

Table I. The quantitative results of variance and naturalness for different guidance factors on VoxCeleb2 under self-reenactment scenario.

periments and observe similar trends to prior works [17]. Specifically, as s increases, the performance of the generated results improves, but the diversity decreases. However, we also find that when choosing a larger value of s, the model will crash and fail to produce reasonable results. Therefore, we carefully consider all factors and choose 2.5 as the optimal value for s, as it strikes a balance between the indicators of diversity and performance.

It is worth noting that when the guidance factor s is set to 0, there is no audio condition, and all non-lip motions are sampled from **random noise**, as defined in Equation 10. However, we find that the performance of the model with $s = 0$ is poor, as reflected in the low value of **SND**. In contrast, our diffusion prior with a guidance factor of 2.5 generates more natural motions and achieves better performance. This proves that it is able to generate more reasonable non-lip motions under audio condition.

II.3. Video Editing Conditioned on Desired Non-lip Feature

Fig. V shows that our diffusion prior P_{a2nl} trained with mask editing technique can enable controlled video editing with conditional non-lip features. *i.e.* we can fix the non-lip feature of one particular frame while letting the diffusion prior model predict the non-lip features of the rest frames. In this example, we borrow two non-lip features (extracted by E_v) from a randomly chosen frame of different identities. Then, we condition the diffusion prior with these non-lip features and assign them to a particular frame (the 3rd frame in this example, marked with an orange box) to let the model predict the rest frames. As we can see, both resulting sequences (2nd and 3rd rows) respect the conditional input and perform smooth transitions surrounding the conditional frame. This indicates both the robustness and flexibility of our system.

II.4. Distribution Visualization

In order to visualize the distribution of poses and expressions generated by each method, we randomly sample 5000 samples for each method and employe t-SNE [18] for visualization, as described in the main paper. Specifically, Fig.VI,VII,VIII,IX show the results for poses of auto-regressive, MakeItTalk [26], Audio2Head [21] and our method with a diffusion prior, respectively. The corresponding results for expressions are shown in Fig.X,XI,XII,XIII.

From Fig.VI and Fig.X, we observe that the distribution of auto-regressive results appears comprehensive, but most of the samples are concentrated at the edge of the distribution (**Best viewed with zoom-in** to observe the contour line that represents the degree of data aggregation). Moreover, the most prominent problem of auto-regressive, rapid and unreasonable motion changes, cannot be discerned from the frame-level distribution alone, but is visible in the accompanying demo video. For MakeItTalk [26], as shown in Fig.VII for poses and Fig.XI for expressions, the generated samples are concentrated in small clusters, indicating that the method is limited in its ability to generate motions beyond those of the identity image. This same conclusion also applies to Audio2Head [21], as shown in Fig.VIII and Fig.XII, which is only able to generate images facing the camera.

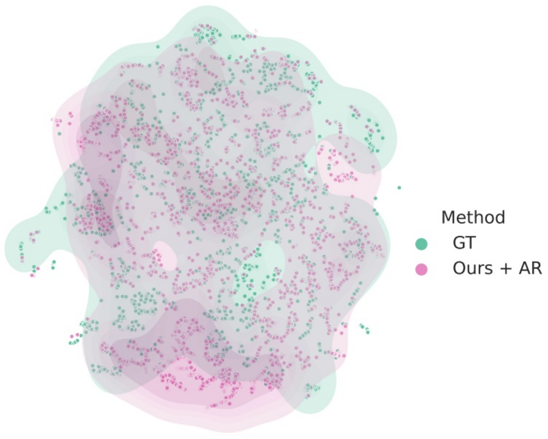

Figure VI. Distribution visualization of AR's poses.

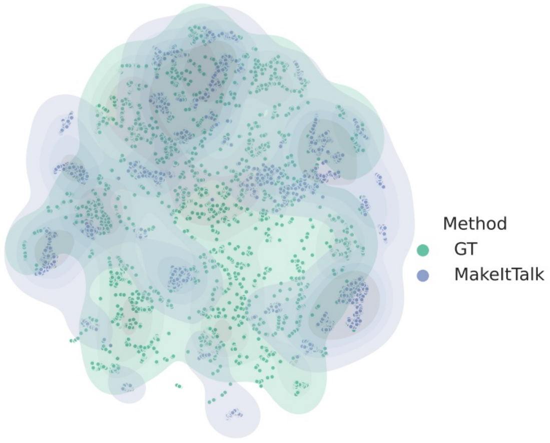

Figure VII. Distribution visualization of MakeItTalk's poses.

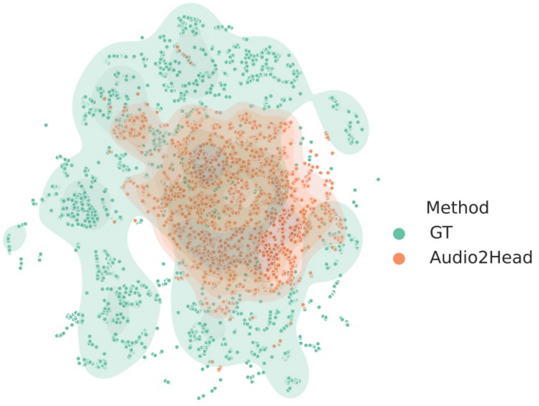

Figure VIII. Distribution visualization of Audio2Head's poses.

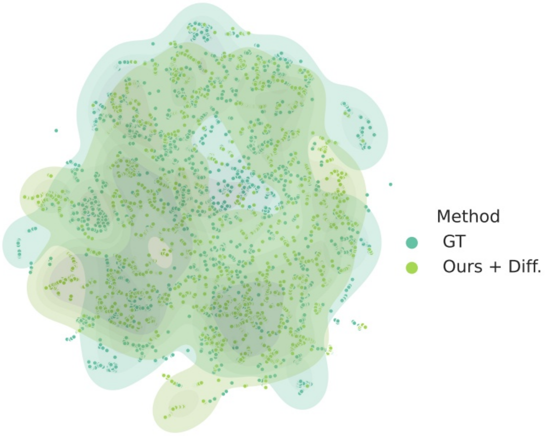

Figure IX. Distribution visualization of diffusion's poses.

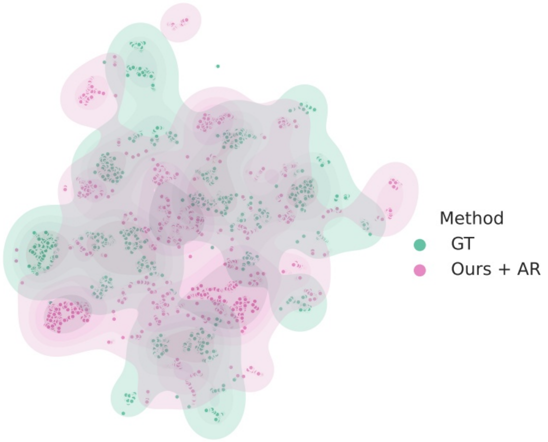

Figure X. Distribution visualization of AR's expressions.

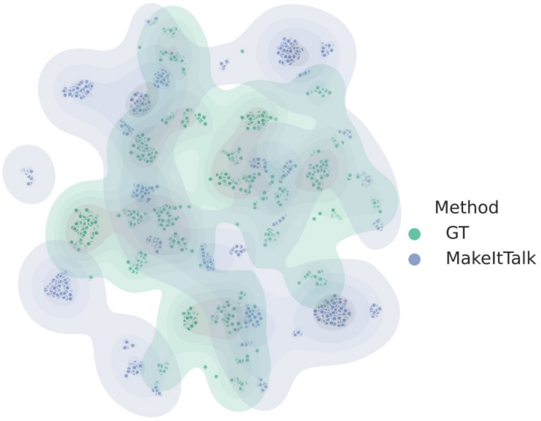

Figure XI. Distribution visualization of MakeItTalk's expressions.

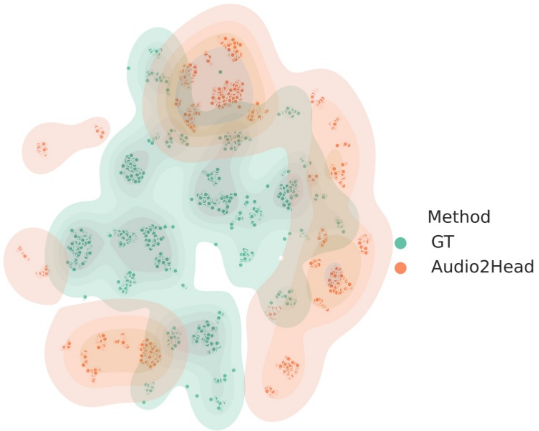

Figure XII. Distribution visualization of Audio2Head's expressions.

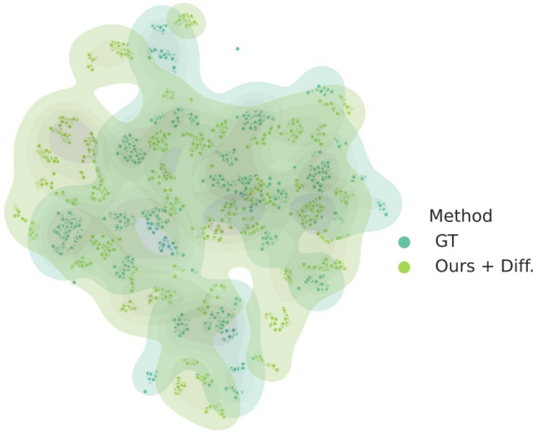

Figure XIII. Distribution visualization of diffusion's expressions.

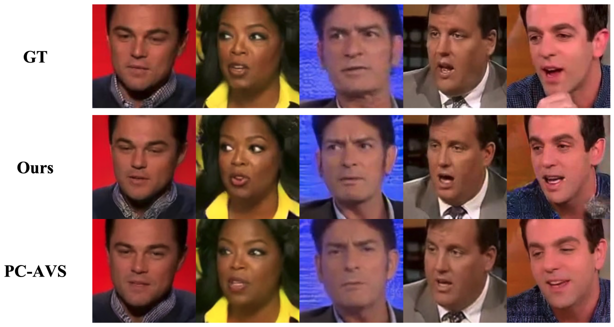

Figure XIV. Rendering cases compared with PC-AVS on VoxCeleb2 with driving signals from the same video-audio pair as the identity image. Note that our model uses non-lip signals from video input, instead of the diffusion prior.

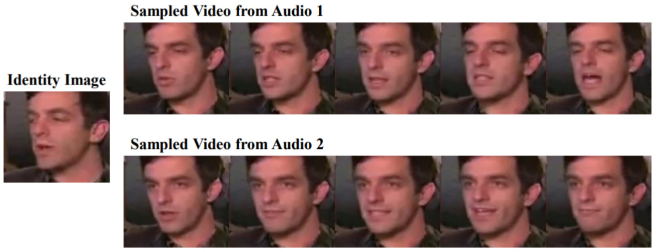

Figure XV. Qualitative results of generating videos with the different audios. Each row shows five continuous frames in each video.

II.5. Limitations of Rendering

Although we train the generator G similarly to that in PC-AVS [25] and our **FID** is the best, sometimes there will be artifacts in non-face area. As shown in Fig. XIV, our generator may produce stripe artifacts in regions with high-frequency details, *i.e.*, clothes and background, thus reducing **FID**. This is likely caused by our rendered images having a larger proportion on the face than PC-AVS's. Another possible reason is the imperfect disentanglement within the pipeline. Nevertheless, our major focus is to predict the non-lip facial motions based on the audio sequence with the help of a probabilistic diffusion prior model. We leave the improvement of rendering to future works.

II.6. More Results

Fig. XV shows the generated videos sampled by different audios with the same identity image. Fig. XVI shows more comparing results of our model and other baselines. Please refer to our supplementary video for more visual results.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022. 1

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 2

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 1

[4] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[5] 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. *arXiv preprint arXiv:2003.11982*, 2020. 1

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1, 3

[7] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 4

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[9] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022. 3

[10] 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/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[11] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 1

[12] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted

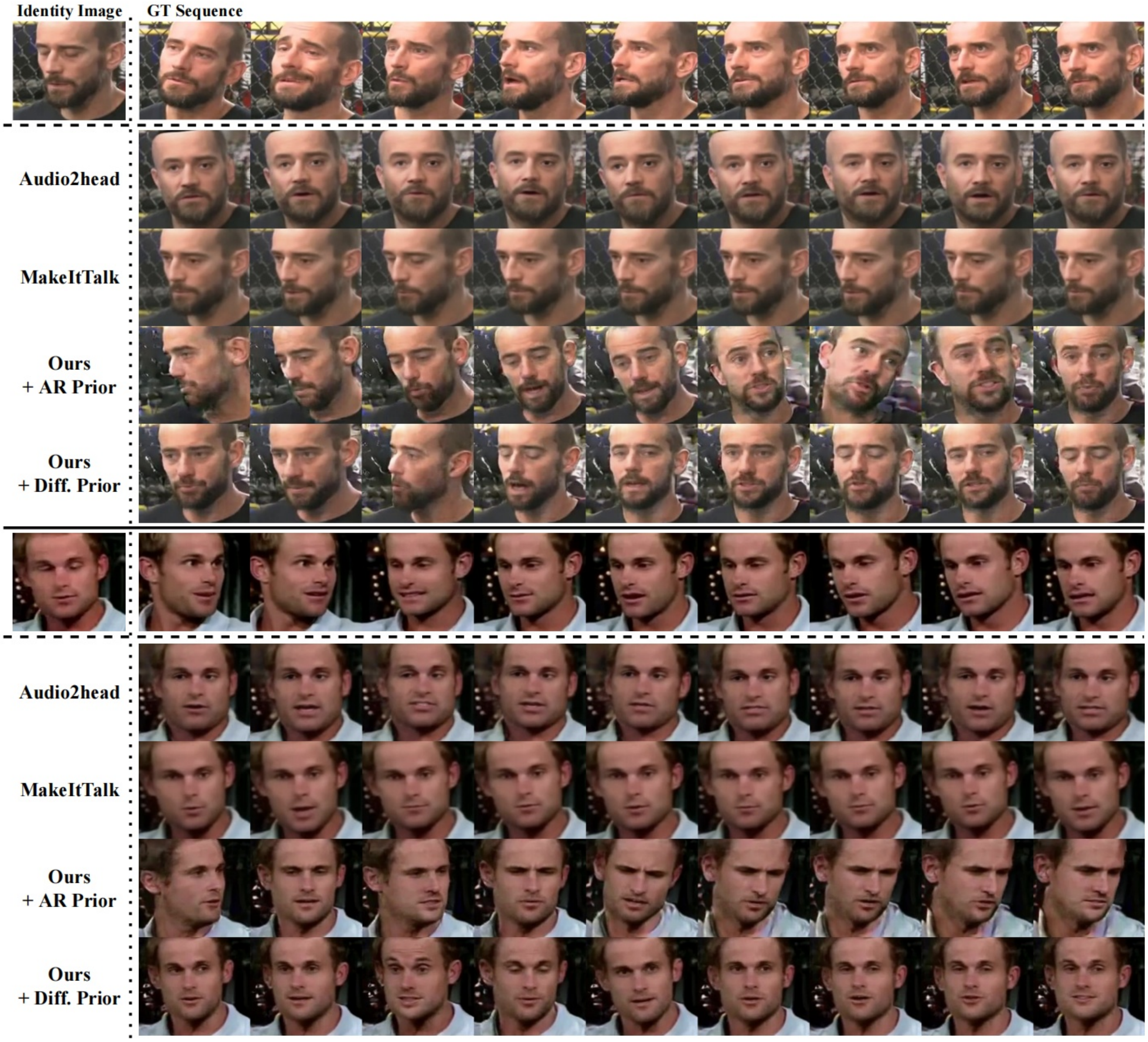

Figure XVI. Qualitative results of our method as compared to other baselines. Each row shows nine uniformly sampled video frames. Here we use two different audio sources to drive the identities. Our method shows accurate lip-audio synchronization with diverse and natural facial motions.

residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[17] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 4, 5

[18] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 5

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[20] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022. 2

[21] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-

head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 3, 5

[22] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442, 2022. 2

[23] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[24] Baosheng Yu and Dacheng Tao. Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[25] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 1, 2, 7

[26] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 2, 5