Progressive Disentangled Representation Learning for Fine-Grained Controllable Talking Head Synthesis (Supplementary Material)

Duomin Wang Yu Deng Zixin Yin Heung-Yeung Shum Baoyuan Wang
Xiaobing.AI
{wangduomin, dengyu, yinzixin, harryshum, wangbaoyuan}@xiaobing.ai

I. More Implementation Details

I.1. Data Preparation

We train our method on all available videos in the training split of VoxCeleb2 [2] dataset. For evaluation, we use the test split of both VoxCeleb2 and Mead [8] dataset. We randomly sample 500 test video clips from VoxCeleb2, and 460 test clips from the Mead following the official setting.

All video frames are aligned following the official annotations [2], and then resized and center-cropped to 224×224. Corresponding audios are extracted from the original videos by ffmpeg, and then processed with a sample rate of $16,000$ and converted to Mel-spectrograms via FFT. The window size, hop size, and the number of Mel bands are set to $1,280$, 160 and 80, respectively.

I.2. More Training Details

Appearance and motion disentanglement. We follow [1] to learn the appearance encoder E_{app}, motion encoder E_{mot}, and the extra image generator G_0. Different from [1], for the appearance encoder, we send a single appearance reference as input to obtain the appearance latent feature during training, instead of taking the average latent feature of multiple appearance frames in a video clip. Apart from the original training losses proposed in [1], we further introduce a motion reconstruction loss as described in Sec. 3.1 in the main paper (*i.e.* Eq. (1)). We set the initial learning rates for E_{app}, E_{mot} to $5e^{-5}$. The initial learning rates for G_0 and an extra discriminator for computing the adversarial loss in [1] are set to $5e^{-5}$ and $5e^{-6}$, respectively. The learning rates of all networks are decayed by a rate of 0.5 for every $80,000$ iterations. We trained the whole pipeline with a batchsize of 24 for 50 epochs on 8 Tesla V100 GPUs with 32GB memory, which took around 2 weeks.

Lip motion disentanglement. We adopt the audio-visual contrastive learning scheme [9] to learn the lip motion encoder E_{lip} and the audio encoder E_{aud}. The two models are trained on audio-video pairs with the InfoNCE loss [7] as described in the main paper (*i.e.* Eq. (2) and (3)). The original training scheme in [9] utilizes frames from the videos different from those deriving the audio signals to construct the negative pairs, which we found can learn non-lip motion information in the obtained lip motion features. Therefore, we only use the unsynchronized frames and audio from the same video clip as the negative pairs during training. We set the initial learning rates of E_{lip} and E_{aud} to $1e^{-5}$, with a decay rate of 0.93 by every $200,000$ iterations. We train the two networks with a batchsize of 32 for 30 epochs. Each item in a batch contains 1 positive pairs and 8 negative pairs. The training took 2 days on 4 Tesla V100 GPUs.

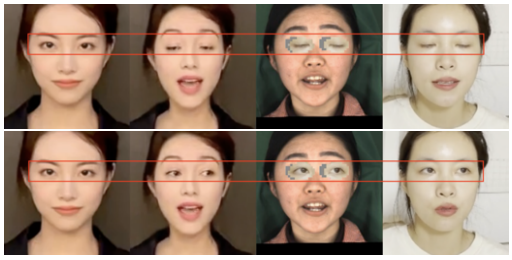

Figure I. Our observation on disentangled eye motion control in the face-reenactment setting in our appearance and motion disentanglement stage. The first column is the appearance reference, the second column is the reenactment result, the third column is the driving source where the eye region comes from the images in the last column. As shown in the figure, the eye motion can be controlled independently without affecting the lip motion in this scenario, which inspires us to design the eye-motion contrastive learning.

Eye motion disentanglement. The eye motion encoder E_{eye} is learned using our proposed eye-motion contrastive learning described in Sec. 3.2 in the main paper. We describe more details about the motivation behind. Specifically, since our first stage is based on the face reenactment

method of [1], we can already synthesize a talking face with the unified motion feature of a driving frame and a given appearance feature via the image generator G_0. We find that by simply replacing the eye region of the driving frame with a new one bearing different eye blink and gaze, we can achieve a disentangled control of eyes in the synthesized face without affecting other facial motions, as shown in Fig. I. Inspired by this, we formulate the eye-motion contrastive loss in the main paper.

We set the initial learning rate for E_{eye} to $1e^{-5}$, decayed by a rate of 0.5 for every $80,000$ iterations. The network is trained with a batchsize of 128 for 30 epochs. The training took 2 days on 4 Tesla V100 GPUs.

Head pose disentanglement. The head pose encoder E_{pose} is learned by regressing the pseudo pose labels as depicted in Sec. 3.2 in the main paper. The learning rate of E_{pose} is also set to $1e^{-5}$ with a decay rate of 0.5 by every $80,000$ iterations. The network is trained with a batchsize of 128 for 30 epochs similar to the eye motion encoder. The training took 2 days on 4 Tesla V100 GPUs.

Expression disentanglement. The expression encoder E_{exp} and our final image generator G are learned via our proposed feature-level decorrelation and complementary self-reconstruction in Sec. 3.3 in the main paper. During this stage, all other networks are fixed, including E_{app}, E_{mot}, E_{lip}, E_{aud}, E_{eye}, and E_{pose}. For the in-window decorrelation, we set the window size to 13. For the lip-motion decorrelation, we set the memory bank size to 512 for an accurate estimation of the feature correlation. We set the initial learning rates to $1e^{-5}$ and $2e^{-5}$ for E_{exp} and G, respectively. The learning rate for an extra discriminator to compute the adversarial loss is set to $3.5e^{-6}$. The expression encoder is trained during the first $40,000$ iterations and frozen for the following steps. The learning rates for the generator and the discriminator are decayed with a rate of 0.5 by every $80,000$ iterations. We use a batchsize of 16 and train all networks for 50 epochs. It took 2 weeks on 8 Tesla V100 GPUs.

The weight of different losses. In the first stage, the training loss consists of GAN loss, VGG loss, and our proposed motion reconstruction loss. All of them are weighted by 1.0 except for the VGG loss weighted by 10.0. In the second stage, the lip-sync contrastive loss, the eye-motion contrastive loss, and the head pose regression loss are also weighted by 1.0. In the third stage, the VGG loss and the GAN loss are weighted by 10.0. The motion reconstruction term in the motion-level consistency loss is weighted by 10.0. All other terms in the motion-level consistency loss and the lip-motion decorrelation loss are weighted by 1.0. The above loss weights are empirically set without careful tuning. We will add this detail in the revision.

I.3. Quantitative Evaluation Details

In Sec. 4.1 in the main paper, we conducted multiple experiments for quantitative metrics calculation (*i.e.* Tab. 1 and 2 in the main paper) under two different settings, namely the *self-driving setting* and the *cross-video setting*.

In the self-driving setting, we use all test clips described in Sec. I.1 for evaluation. We set the first frame in each video as the appearance reference, and drive it using the video frames and the corresponding audio from the same video clip. The audio signals are used to drive the lip motion and the video frames for other motions. Since the source and the target are from the same video, we can easily use the driving frames as the ground truth to evaluate the performance of each method.

In the cross-video setting, we use the first frame from 100 randomly sampled test video clips as an appearance reference and use the first frame from another 100 random test video clip as the driving frame to control all non-lip motions. We still use the audio signals from the video clip of the corresponding appearance frame to control the lip motion. The cross-video setting is designed to evaluate the expression control performance, where we extract the expression parameters of the synthesized videos and their corresponding driving frames using a 3D face reconstructor [4], and compare the expression parameter difference. This helps us to evaluate if a method can precisely transfer the expression from a source to a target. By contrast, in the self-driving setting, since the source and the target are from the same video clip, their expressions are usually the same. Under this circumstance, if a method well mimics the expression motion of the appearance reference, it is difficult to judge whether it successfully transfers the source expression to the target or merely copies the expression from the appearance reference.

I.4. User Study Details

We conduct two user studies to evaluate the motion control performance. In the first experiment, we ask participants to evaluate the accuracy of lip motion synchronization and expression control, as well as the naturalness of all facial motions. We generate 120 videos using 12 random appearance references and 10 random driving clips and randomly select 35 synthesized videos out of them for evaluation. Fifteen participants are asked to score from 1 to 5 for the quality of different properties in the synthesized videos (5 is the best). The corresponding results are in Tab. 3 in the main paper.

In the second experiment, we ask the same group of participants to evaluate the disentanglement controllability of our method. We generate 5 videos using an appearance ref-

Table I. Quantitative evaluation on factor disentanglement of our method. In each row, we compute the variance of a motion feature extracted from the synthesized videos when controlling different individual motion factors.

Variance	Control property				
	lip	pose	blink	gaze	exp
Speech lip motion	**11.24**	5.16	0.83	0.74	3.76
Head pose	0.0091	**0.1597**	0.0041	0.0045	0.0088
Eye blink	0.00038	0.00389	**0.06657**	0.00089	0.00225
Eye gaze	0.089	0.100	0.095	**0.105**	0.088
Expression	3.07	3.07	2.98	2.93	**3.59**

erence and 3 randomly selected driving videos for the head pose, expression, and eye motion, respectively. In each synthesized video, only one motion factor is controlled by the driving source and all other factors remain unchanged. The participants are asked to score from 1 to 5 for the variation level of each motion in the synthesized videos (5 indicates the largest variation, and 1 means nearly unchanged). The corresponding results are in Tab. II and discussed in Sec. II.2.

II. More Results

II.1. Fine-Grained Controllable Talking Heads

Figure V and VI show more talking head synthesis results by our method. Our method well mimics the motions from different driving sources and combines them to generate vivid talking heads. **Animations can be found in the accompanying video.**

II.2. Disentangled Controllability

We quantitatively evaluate the disentangled controllability of our method. To this end, we generate talking head images by only varying one motion factor and setting other factors to zeros (*i.e.* canonical positions). We then extract corresponding motion features from the synthesized results and compute the variance of each motion factor in a video clip. Ideally, if different motions are perfectly disentangled, the computed variances will be close to zero for all motions except the one being controlled.

In practice, we use off-the-shelf models to extract each motion feature from our synthesized images. For eye gaze and blink, we use the model of [6]. For expression and pose, we use a 3D face reconstruction model [5]. For lip motion, we use the model of [3]. The variance of each motion factor $\boldsymbol{f}$ is computed using the following equation:

$$var(\boldsymbol{f}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}\|\boldsymbol{f}_{ij} - \bar{\boldsymbol{f}}_i\|_2, \qquad \text{(I)}$$

where $\boldsymbol{f}_{ij}$ is the corresponding extracted motion feature of the j-th frame in the i-th video clip, $\bar{\boldsymbol{f}}_i$ is mean of $\boldsymbol{f}_{ij}$, N is the number of test videos, and M_i is the length of each video clip.

Table I shows the computed variance of each motion factor. Each row shows the variance of a single motion factor under different motion control. As shown, the variance of a factor reaches the maximum when the controlling factor is the same with it, and largely decreases when controlled under a different motion factor. This indicates that our method can disentangle different motion controls so that they have a minor influence on each other.

However, the computed variance can still be large in some cases (*e.g.* the left four columns in the last row in Tab. I). This is due to that the off-the-shelf motion feature extractors are not perfect and can be influenced by variations of other motions when extracting a certain motion feature. Therefore, we refer the readers to the accompanying video to examine the disentanglement ability of our method. We also conduct a user study to better evaluate the factor disentanglement. The results are in Tab. II (see Sec. I.4 for detailed description). As shown, the variance score is close to 5 when the factor for variance calculation and the factor to be controlled are the same, and close to 1 when they are different, which reveals the disentangled controllability of our method.

II.3. Expression Interpolation

We further investigate the expressive ability of our learned expression feature. We show expression interpolation results by linearly interpolating two expression features from different expression sources. As shown in Fig. II, our method can smoothly transfer between two different expressions. The synthesized images at interpolated points also have natural expressions. This indicates that our method learns a reasonable expression latent space that supports continuous expression control.

II.4. Comparison with the prior methods

We show the lip motion synthesis comparison in Fig. III. The images are synthesized under the self-driving setting. As depicted, the lip motion generated by our method is nat-

Table II. User study on factor disentanglement of our method.

Variance	Control property				
	lip	pose	blink	gaze	exp
lip	**4.7**	1.1	1	1	1.1
pose	1.1	**4.6**	1	1.2	1.3
blink	1.1	1.1	**4.1**	1.5	1
gaze	1	1.1	1.4	**4.4**	1
exp	1.3	1.1	1	1	**3.7**

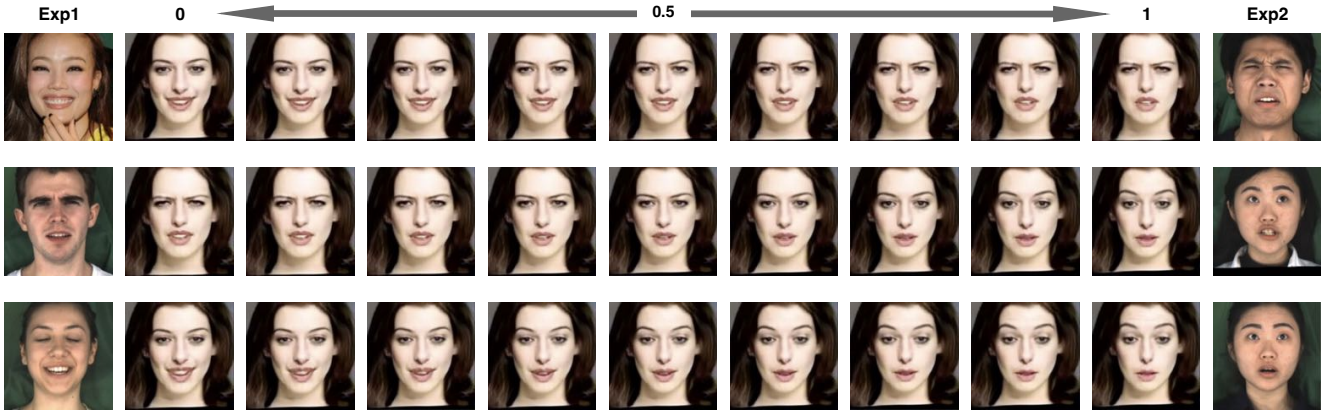

Figure II. Expression interpolation by our method.

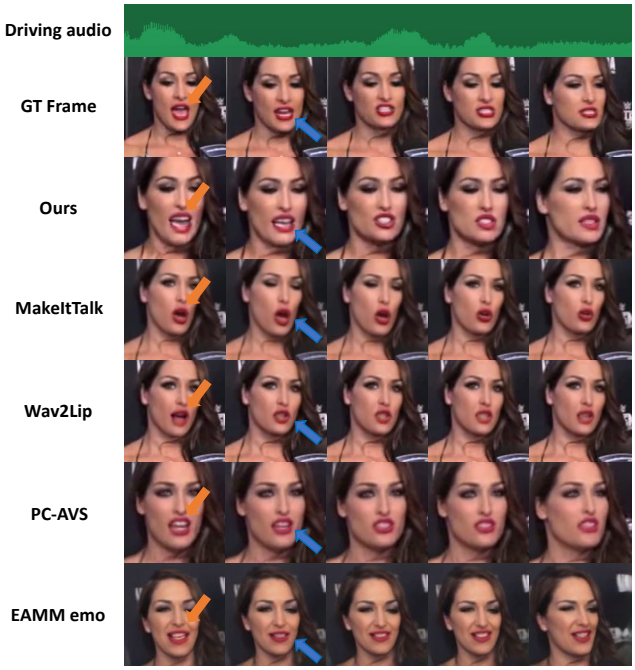

Figure III. Comparison on lip motion control. Images are synthesized under the self-driving setting where the lip motion is driven by the audio signal. Our method yields the best result.

ural and closer to the ground truth compare to the alternatives. **See the accompanying video for animations.**

II.5. Ablation Study

Motion reconstruction loss. We further conduct an ablation study to validate the efficacy of our motion reconstruction loss proposed in Sec. 3.1 in the main paper. As shown in Tab. III and Fig. VII, with the motion reconstruction loss, facial motions in the synthesized images contain more details and are closer to the driving sources. By contrast, removing the motion reconstruction loss leads to poor reenactment results for driving sources with rich expressions. As a result, the motion reconstruction loss is important for obtaining an informative unified motion feature to achieve accurate motion control.

Eye contrastive learning loss. We also conduct a visual ablation on the eye motion contrastive learning loss illustrated in Fig. IV. Without eye contrastive learning, blink&gaze are not controlled by the eye sources but kept the same as the reference.

III. Ethics Consideration

Our method enables precise and disentangled control over multiple facial motions for vivid talking head generation. While the major goal of it is to synthesize virtual

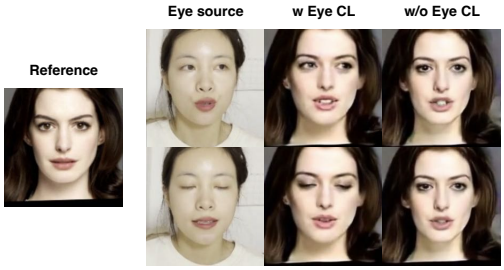

Figure IV. Visual effect of eye motion contrastive loss

Table III. Ablation of motion reconstruction loss on expression and pose control accuracy.

Method	Expression↓		Pose↓
	VoxCeleb2	Mead	VoxCeleb2
w/o mot loss	0.147	0.174	0.0021
w mot loss	**0.141**	**0.160**	**0.0017**

avatar for applications like live streaming, it can be misused to create deceptive and harmful content of real people. Especially, one may use it to synthesize fake videos of celebrities. We do not condone using our method for generating misleading information that could harm people's reputations. We also suggest investigating advanced forgery detection methods to identify the synthesized fake images and videos to prevent illegal usage.

References

[1] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 1, 2

[2] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 1

[3] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 3

[4] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2

[5] 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 IEEE Computer Vision and Pattern Recognition Workshop on Analysis and Modeling of Faces and Gestures*, 2019. 3

[6] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, pages 334–352, 2018. 3

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[8] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 1

[9] 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

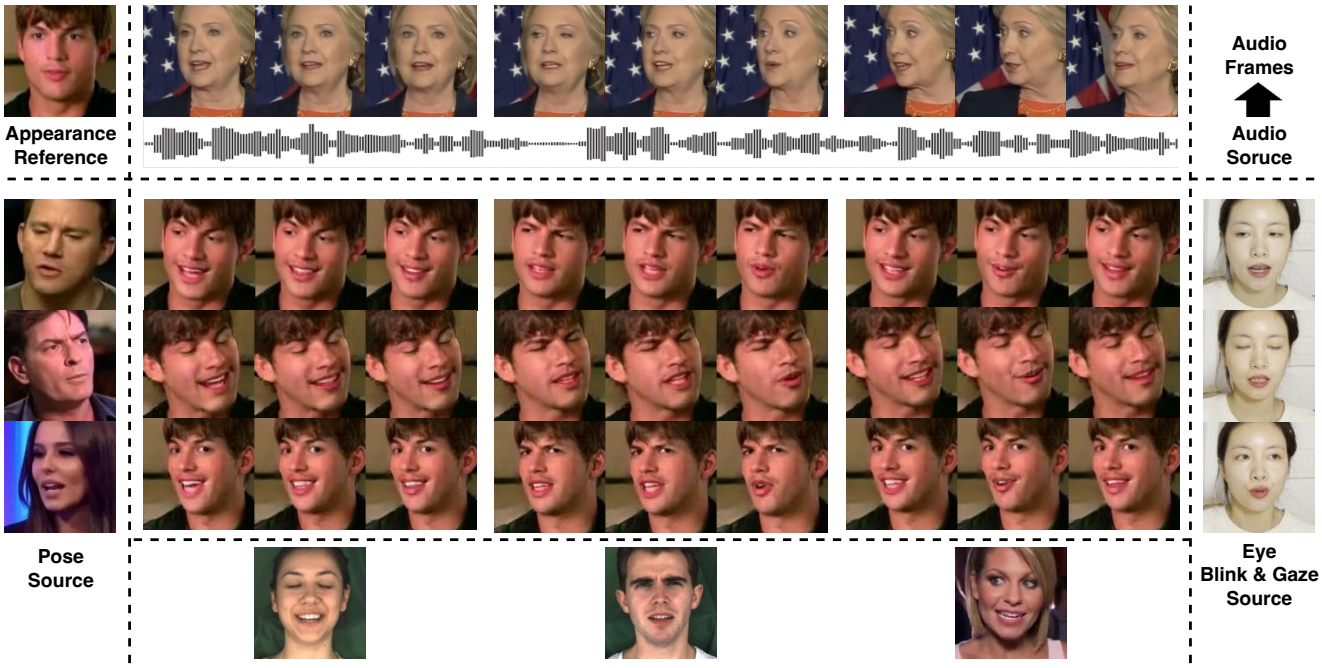

Figure V. Fine-grained controllable talking heads synthesized by our method.

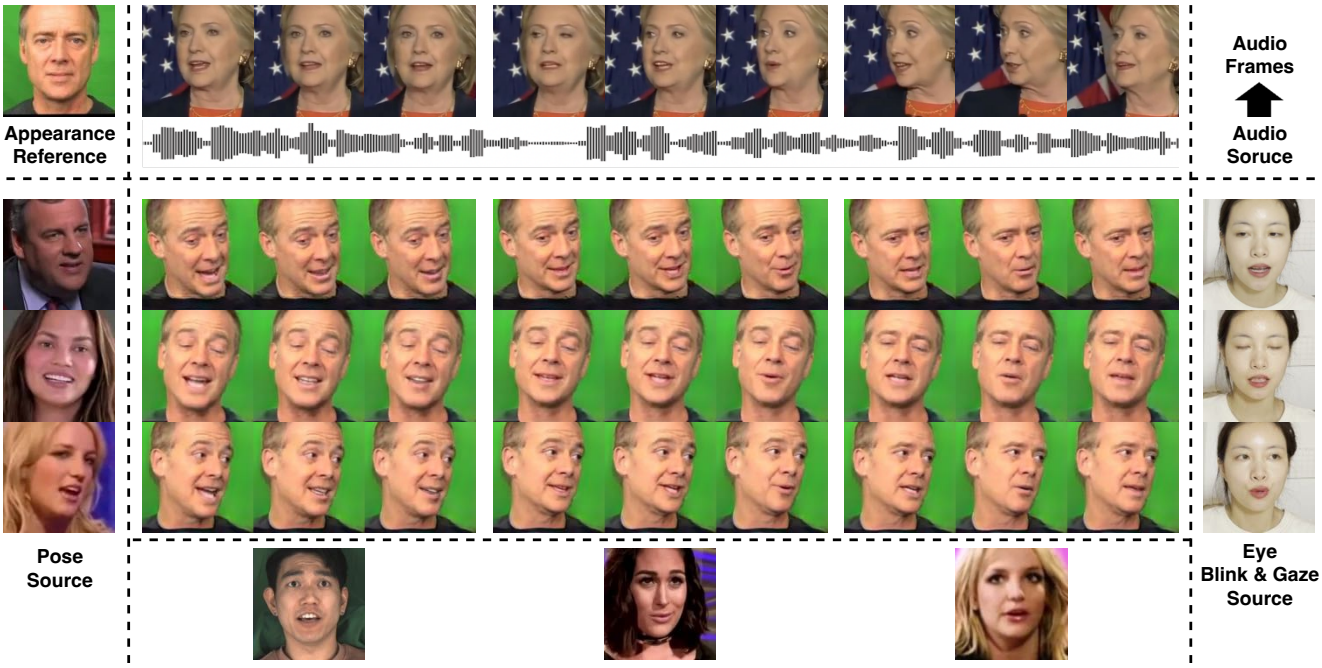

Figure VI. Fine-grained controllable talking heads synthesized by our method.

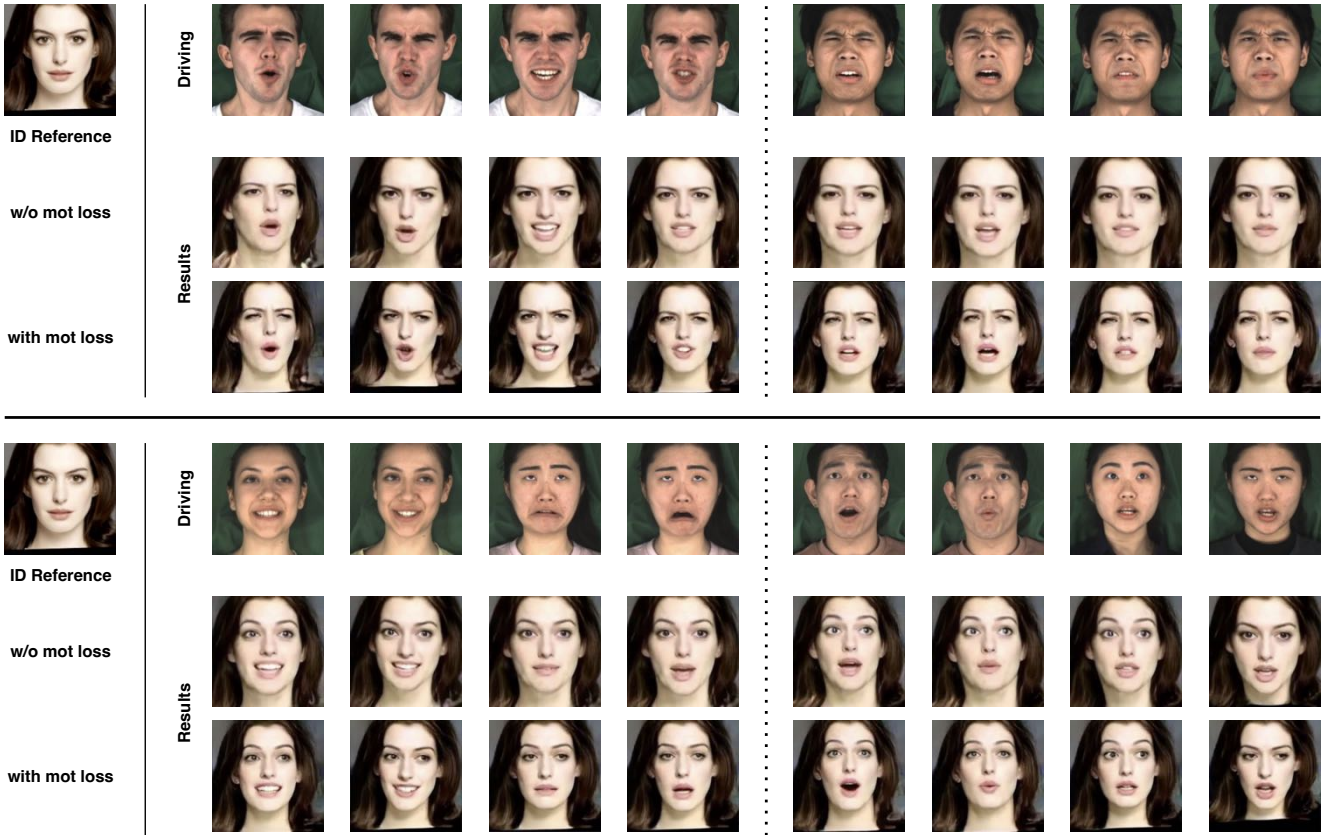

Figure VII. Ablation study on the motion reconstruction loss in the appearance&motion disentanglement learning.