Arbitrary Talking Face Generation via Attentional Audio-Visual Coherence Learning

Hao Zhu[1,2] , **Huaibo Huang**[2,3] , **Yi Li**[2,3] , **Aihua Zheng**[1] and **Ran He**[2,3*]
[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]NLPR&CEBSIT&CRIPAC, Institute of Automation, CAS, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
haozhu96@gmail.com, {huaibo.huang,yi.li}@cripac.ia.ac.cn, ahzheng214@ahu.edu.cn,
rhe@nlpr.ia.ac.cn

Abstract

Talking face generation aims to synthesize a face video with precise lip synchronization as well as a smooth transition of facial motion over the entire video via the given speech clip and facial image. Most existing methods mainly focus on either disentangling the information in a single image or learning temporal information between frames. However, cross-modality coherence between audio and video information has not been well addressed during synthesis. In this paper, we propose a novel arbitrary talking face generation framework by discovering the audio-visual coherence via the proposed Asymmetric Mutual Information Estimator (AMIE). In addition, we propose a Dynamic Attention (DA) block by selectively focusing the lip area of the input image during the training stage, to further enhance lip synchronization. Experimental results on benchmark LRW dataset and GRID dataset transcend the state-of-the-art methods on prevalent metrics with robust high-resolution synthesizing on gender and pose variations.

1 Introduction

Talking face generation, which aims to generate a realistic talking video for the given still face image and speech clip, has been an active research topic. It has wide potential applications such as movie animation, teleconferencing, talking agents, and enhancing speech comprehension while preserving privacy. Although recent efforts have achieved impressive talking face synthesis for arbitrary identities, it is still a huge challenge due to the heterogeneous between audio and video, together with the appearance diversity of arbitrary identities.

Existing state-of-the-art works [Zhou *et al.*, 2019; Chen *et al.*, 2019] either try to disentangle the content and identity features from the speech for the video generation step or leverage the landmark as a middle feature to bridge the gap between audio and video. Despite the great progress on feature representation for video generation, they usually leverage reconstruct loss between generated and real frames, while neglecting an essential problem in talking face generation: how to sufficiently express the audio feature into generated video? To reduce the uncertainty in the audio-to-video generation process and ensure the synchronized talking face transition, we argue that the giving audio and generated video should share maximum information, i.e., the minimized sharing entropy across modalities. Herein, we propose audio-visual coherence learning, which learns the shared entropy between audio and visual modalities to generate talking face video with precise lips shape.

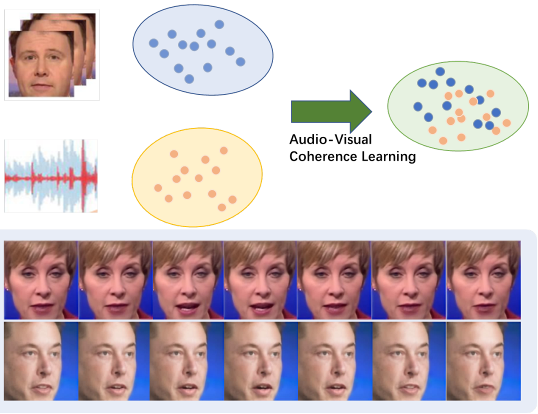


Figure 1: Illustration of proposed audio-visual coherence learning.

In the AI community, it is always active to exploit information theoretic measures to solve real-world problems [He *et al.*, 2009; Jun *et al.*, 2011; He *et al.*, 2014; Chen *et al.*, 2016]. Mutual information (MI), as a commonly used measure to explore the coherence between two distributions, has been successfully applied to feature selection [Sotoca and Pla, 2010; Liu *et al.*, 2009], face analysis [Jun *et al.*, 2011], and gene expression [Gupta and Aggarwal, 2010]. As explained in [Vincent *et al.*, 2010], mutual information can be utilized to learn a parametrized mapping from a given input to a higher-level representation while preserving information of the original input. This can be referred to as the infomax principle translating to maximize the mutual information between the output and input during the generative network. For instance, InfoGAN [Chen *et al.*, 2016] maximized the MI between generated samples and latent codes to learn more interpretable representations during generation.

[*]corresponding author

We argue that the audio and corresponding facial information contains potential mutual dependency, therefore, MI is able to calculate the shared information entropy between audio and video if we can maximize the MI between them, that means, we minimized the uncertainty in audio-to-video expression process, which can ensure more stable and precise lip motion generation. Therefore, we propose to explore the cross-modal audio-visual coherence via MI in this paper.

Belghazi *et al.* [2018] recently presented a Mutual Information Neural Estimator (MINE) to explore a KL-based MI lower bound and train this estimator on target distributions. Unfortunately, it cannot be directly applied to talking face generation. On the one hand, if MINE for GANs uses generated sample and input pairs for both training and estimating, the mutual information estimation may be misled due to the low quality of generation in the early stages. Thinking that GANs push the generated sample distribution to the real frame distribution, we propose to use the real sample and its corresponding input to update the mutual information estimator, followed by asymmetrically maximizing the mutual information between the generated sample and input for better GAN training. On the other hand, our objective is to maximize the MI between heterogeneous modalities while not concerning the accurate MI value [Belghazi *et al.*, 2018], we therefore replace the original formulation of MI by Jensen-Shannon represented MI estimator, which estimates the relative amplitude of MI rather than the exact value, and has been well studied in neural network optimization with empirically more stable learning [Hjelm *et al.*, 2018]. We refer to our strategy as Asymmetric Mutual Information Estimator (AMIE).

In addition, conventional approaches [Chung *et al.*, 2017; Chen *et al.*, 2018; Vougioukas *et al.*, 2018; Zhou *et al.*, 2019; Chen *et al.*, 2019] directly employ the whole global area on given face image during synthesis, which is difficult for neural networks to discover the relation between audio and local lips. We observe that a talking face video is mainly composited by the identity-related feature and the lip-related feature, where the former is more stable while the latter is more temporal dynamic. Therefore, it is essential to separate these two features for arbitrary identity generation. Herein, we propose to leverage the given face and previous generated frame to provide identity-related and lip-related information, respectively. We propose a dynamic attention block on lip area to preserve the identity information and leverage feature of lip motion, i.e., paying different attentions on given face image (identity-related) and the previous generated frame (lip-related) during different training stages.

Based on the above discussion, we propose a novel and robust method by exploring the coherence between audio and visual modalities for arbitrary talking face generation in this paper. The proposed model consists of three components: talking face generator, asymmetric mutual information estimator, and frame discriminator, as shown in Fig. 2. First, the talking face generator is designed to generate target frames from the given input: one audio clip, one still facial image, and the previous generated frame. Then, the asymmetric mutual information estimator is introduced to maximize the mutual information between generated video and audio distributions via the information learned from a neural network based on MI measure. Finally, we feed the generated frame and audio into frame discriminator to detect whether they are matched or not. In the training stage, the lip area will give different attention to make the network focus on the most important area by leveraging the Dynamic Attention block, as illustrated in Fig 4.

The main contributions of our paper can be summarized as follows:

- We propose to discover the audio-visual coherence via the Asymmetric Mutual Information Estimator (AMIE) to better express the audio information into generated video in talking face generation.
- We design a dynamic attention block to improve the transition of generated video for arbitrary identities by decoupling the lip-related and identity-related information.
- Extensive experiments yield a new state-of-the-art on benchmark datasets LRW [Chung and Zisserman, 2016] and GRID [Cooke *et al.*, 2006] with robust high-resolution synthesizing on gender and pose variations.

2 Related Works

In this section, we briefly review the related works on talking face generation and mutual information estimation.

2.1 Talking Face Generation

Earlier works on talking face generation mainly synthesized the specific identity from the dataset by given arbitrary speech audio. Rithesh *et al.* [2017] used a time-delayed LSTM to generate key points synced to the audio and use another network to generate the video frames conditioned on the key points. Supasorn *et al.* [2017] proposed a teeth proxy to improve the quality of the teeth during generation. Next, some works tried to synthesized the talking faces for the identities from the dataset [Chung *et al.*, 2017; Vougioukas *et al.*, 2018; Chen *et al.*, 2018]. Recently, the synthesis of the talking face for the arbitrary identities out of the dataset has drawn much attention. Zhou *et al.* [2019] proposed an adversarial learning method to disentangle the different information for one image during generation. Chen *et al.* [2019] proposed to leverage landmark as middle information to better guide face generation. However, they mainly focused on the inner-modal coherence, while lacking of discovering the cross-modal coherence.

2.2 Mutual Information Estimation

It is historically difficult to measure the mutual dependence between two random variables in information theory. Some past solutions are only suitable for discrete variables or limited known probability distributions. Previous works tried to employ parameters-free approaches [Kraskov *et al.*, 2004], or relied on approximate Gaussianity of data distribution [Hulle, 2005] to estimate the mutual information. Recently, Belghazi *et al.* [2018] proposed a backpropagation MI estimator that exploited a dual optimization based on dual representations of the KL-divergence [Ruderman *et al.*, 2012] to estimate divergences beyond the minimax objective as formalized in GANs.

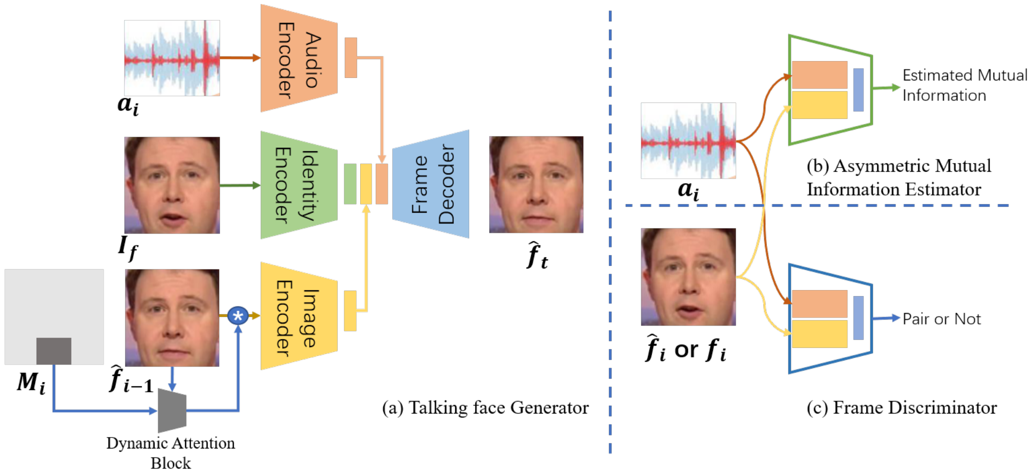


Figure 2: Pipeline of our proposed method.

3 Proposed Method

In this paper, we propose a novel model for arbitrary talking face generation by attentively discovering the cross-modal coherence, as shown in Fig. 2. We first overview the architecture of our model, followed by the elaboration of two novel components: Asymmetric Mutual Information Estimator (AMIE) and Dynamic Attention (DA) modules. The training details are provided at the end of this section.

3.1 Overview

Our model consists of three parts: a Talking Face Generator, a Frame Discriminator, and a Mutual Information Estimator.

Talking Face Generator. There are three inputs of the generator: 1) the input face I_f which ensures the texture information of the output frame, 2) the speech audio clip $A = \{a_1, a_2, ..., a_n\}$, working as the condition to supervise the lip changing, and 3) the previously generated frame $\hat{f}_{i-1} \in \hat{F} = \{\hat{f}_1, \hat{f}_2, ..., \hat{f}_n\}$ where the i-th frame guarantees the smoothness of the image generation by feeding more temporal information. The three inputs are fed to the Identity Encoder, Audio Encoder, and Image Encoder respectively, while the target video frame $\hat{f}_i$ is generated by the Frame Decoder. The generated frames $\hat{F}$ should be similar to real frames $F = \{f_1, f_2, ..., f_n\}$.

Asymmetric Mutual Information Estimator (AMIE). AMIE aims to approximate the mutual information between the generated video frame and audio. It is trained using $\{f_i, a_i\}$ and $\{f_j, a_k\}$, where j and k are random sampled from $1 \sim n$. These two pairs are serve as samples that the sampled from joint and marginal distributions. While in the estimating stage, we asymmetrically estimate mutual information using $\{\hat{f}_i, a_i\}$, and $\{\hat{f}_j, a_k\}$, i.e., we use the real pairs for training while the generated samples to estimate. It consists of an Image Encoder and an Audio Encoder having the same architecture as defined in the generator, followed by a 3-layer classifier with the output as a 1-dimension scalar. We shall elaborate on the details of the proposed AMIE in the following section.

Frame Discriminator. Frame discriminator is fed by the pairs of the real frame and audio clip $\{f_i, a_i\}$, or the pairs of the generated frame and corresponding audio clip $\{\hat{f}_i, a_i\}$. The output of the discriminator is the probability of whether the inputs (audio and frame) are matched. The discriminator consists of an Image CNN, an Audio FC, and a classifier. We flatten the output of Image CNN, and the Audio FC then feed concatenated features to the final classifier to produce 1-dimensional output.

3.2 Asymmetric Mutual Information Estimator

We first introduce the theory of mutual information neural estimation (MINE), followed by the mutual information estimation in talking face generation task which leverages the proposed AMIE to learn the cross-modal coherence.

Preliminary Theory of 'MINE'

Mutual information is a measurement of mutual dependency between two probability distributions,

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}, \tag{1}$$

where $p(x,y)$ is the joint probability function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

Clearly, mutual information is equivalent to the Kullback-Leibler (KL-) divergence between the joint $p(x,y)$ and the product of the marginal distributions $p(x)$ and $p(y)$:

$$I(X,Y) = D_{KL}(p(x,y) \parallel p(x)p(y)), \tag{2}$$

where D_{KL} is defined as,

$$D_{KL}(p \parallel q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)}. \tag{3}$$

Furthermore, the KL divergence admits the following Donsker-Varadhan (DV) representation [Donsker and Varadhan, 1983; Belghazi *et al.*, 2018] :

$$D_{KL}(p \parallel q) = \sup_{T:\omega \to \mathbb{R}} \mathbb{E}_p[T] - log(\mathbb{E}_q[e^T]), \tag{4}$$

where the supremum is taken over all functions T and $\omega \subset \mathbb{R}^d$ so that the two expectations are finite. Therefore, we leverage the bound:

$$I(X, Y) \geq I_\Theta^{DV}(X, Y), \tag{5}$$

where $I_\Theta^{DV}(X, Y)$ denotes the neural information measure,

$$\begin{aligned} I_\Theta^{DV}(X, Y) = \sup_{\theta \in \Theta} \mathbb{E}_{p(x,y)}[T_\theta(x, y)] - \\ log(\mathbb{E}_{p(x)p(y)}[e^{T_\theta(x,y)}]), \end{aligned} \tag{6}$$

and T_θ denotes a neural network trained by maximizing the mutual information [Belghazi *et al.*, 2018].

'AMIE' in Talking Face Generation

In this cross-modal talking face generation task, we propose to explore the mutual information between the audio and the visual modality via an Asymmetric Mutual Information Estimator. Given $\hat{F}$ and A as the generated video frames and the given audios respectively, the neural network T_θ is fed by $\{\hat{f}_i, a_i\}$ and $\{\hat{f}_j, a_k\}$ to output a scalar which will be used to estimate the MI. Furthermore, as we do not concern the accurate value of MI while maximizing it, inspired by [Hjelm *et al.*, 2018], we replace Donsker-Varadhan (DV) representation by Jensen-Shannon representation (JS), which has been noted with more stable learning in neural network optimization. Therefore, the estimated mutual information between $\hat{F}$ and A is:

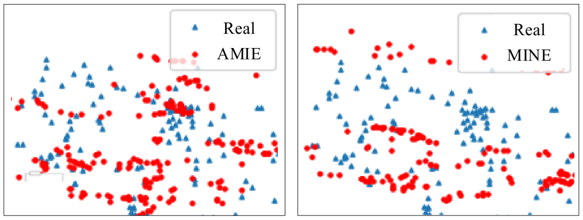
$$\begin{aligned} I_\Theta^{JS}(\hat{F}, A) = \sup_{\theta \in \Theta} \mathbb{E}_{p(\hat{f},a)}[-\varphi(-T_\theta(\hat{f}_i, a_i))] - \\ \mathbb{E}_{p(\hat{f})p(a)}[\varphi(T_\theta(\hat{f}_j, a_k))], \end{aligned} \tag{7}$$

where $p(\hat{f}, a)$, $p(\hat{f})$ and $p(a)$ denote the joint distribution of the generated frame and audio, the marginal distributions of generated frame and the marginal distributions of generated audio respectively, in addition, $\varphi(\cdot)$ represents softplus operation:

$$\varphi(x) = log(1 + e^x). \tag{8}$$

As proved in [Belghazi *et al.*, 2018], one can estimate the mutual information between the generated video frame and audio by directly training T_θ on them. However, this may disturb the MI estimation since the generated frames are blurry and not accurate at earlier training epochs. Furthermore, GANs are usually used to learn the probability distribution consistent with the real data, and mutual information is used to estimate the amount of shared information between the two distributions. Therefore, our solution is to use mutual information in three distributions, we use the real frame and audio distribution during the training while the generated frame and audio distribution.

Specifically, let F represent real video frames, we update the mutual information estimator by maximizing mutual information between the real frames and the audio ($I_\Theta^{JS}(F, A)$), then use the updated estimator to maximize the mutual information between the generated frames and audio ($I_\Theta^{JS}(\hat{F}, A)$) for better GAN training. We refer to this as Asymmetric Mutual Information Estimator (AMIE) in this paper which can improve the quality of generation as verified in our experiments. As shown in Fig. 3, comparing our AMIE with conventional MINE, we see that our generated samples are closer to the real samples than those from MINE.



(a) AMIE (b) MINE

Figure 3: Visualization of distributions of real and generated frames. We reduce the dimension of frames into two-dimension via PCA for better demonstration. It is obvious that the generated samples are closer to the real samples than that with original MINE.

3.3 Dynamic Attention on Lip Area

As we discussed in Sec. 1, we observe that a talking face video is mainly composited by the identity-related and lip-related features. Directly leveraging the whole area of the given face tends to generate face images with a slight jitter problem. In order to capture different information among the given faces and disentangle the identity-related and lip-related information, we introduce a dynamic attention block. When generating a video sequence, we assign an initial attention rate to the first frame, then predict the fine-grained attention mask for the following video frames. When given low attention on the lip, the input image mainly contains identity-related information (facial texture without the lip-related information), and the previous frame and audio clip together provide the information about the shape of the lip. Therefore, our model can divide the feature of one identity into two parts: identity-related feature and lip-related feature. The lower attention, the information is separated more completely.

However, it may significantly affect the quality of generation in the early stage if we directly assign a low initial attention mask M_i on the lip since lacking supervision of lip

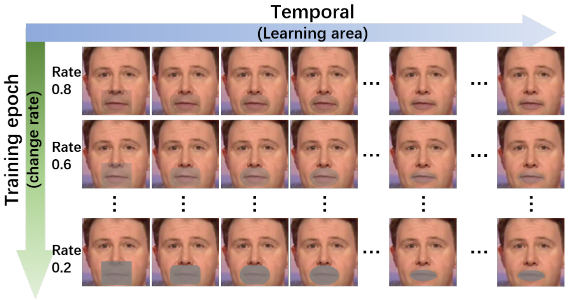


Figure 4: The illustration of the proposed dynamic attention.

information. Therefore, we progressively decrease the initial attention after several training epochs, which will enforce the visual information of the lip deriving from the previous frame $\hat{f}_{i-1}$. Specifically, in the training stage, we start from relatively high attention (rate = $0.7 \sim 0.9$), and progressively decrease it to relatively low attention (rate = $0.1 \sim 0.3$), then we fix the rate to 1 for the last few epochs. The attention masks of the following generated frame are predicted by the previous attention mask and the generated faces. In practice, we give a coarse mask area and an attention rate, the following frames generation only predict a fine-grained mask area which is close to the lip shape, while leaving the rate unchanged.

3.4 Training Details

In the training stage, frame discriminator predicts the probability of whether frame and audio are paired or not, resulting in the following loss of our GAN:

$$\begin{aligned} \mathcal{L}_{GAN}(D,G) &= \mathbb{E}_{f \sim P_d}[log(D(f_i,a_i))] + \\ &\mathbb{E}_{z \sim P_z}[log(1 - D(G(a_i,z),a_i))]. \end{aligned} \tag{9}$$

In order to synchronize the lip movements more accurately, we employ perceptual loss [Johnson *et al.*, 2016] to capture high-level features differences between generated images and ground-truth. The perceptual loss is defined as:

$$\mathcal{L}_{perc}(f_i,\hat{f}_i) = \parallel \phi(f_i) - \phi(\hat{f}_i) \parallel_2^2, \tag{10}$$

where ϕ is a feature extraction network.

To focus on the lip movement, we only utilize the lip area of the frame for L_1 reconstruction loss,

$$\mathcal{L}_{lip}(f_i,\hat{f}_i) = \parallel corp(f_i) - corp(\hat{f}_i) \parallel_1, \tag{11}$$

where $corp(\cdot)$ means the cropped lip area from image.

While $\mathcal{L}_{perc}$ and $\mathcal{L}_{lip}$ measure the distance between visual concept, it is also important to shorten the distance between audio and visual modalities in high-level representation.

We implement mutual information as described before. We try to maximize it between generated frames and audios,

$$\mathcal{L}_{mi}(\hat{F},A) = -I_{\Theta}^{JS}(\hat{F},A). \tag{12}$$

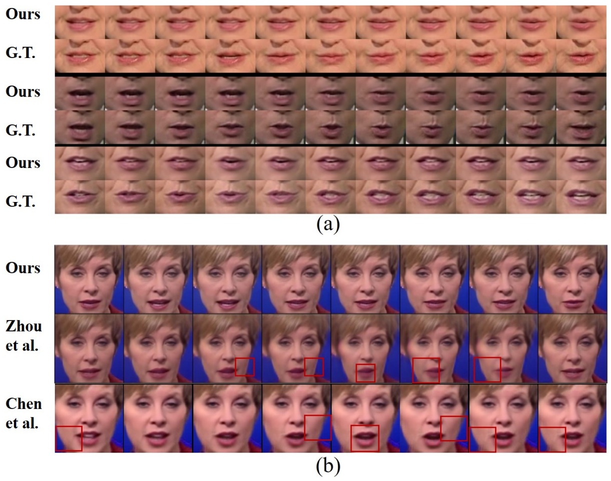
Our full model is optimized according to the following objective function:

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda_1\mathcal{L}_{perc} + \lambda_2\mathcal{L}_{lip} + \lambda_3\mathcal{L}_{mi}. \tag{13}$$

4 Experiments

4.1 Dataset and Metrics

We evaluate our method on prevalent benchmark datasets LRW [Chung and Zisserman, 2016] and GRID [Cooke *et al.*, 2006]. Frames are aligned into 256×256 faces and audios are processed into (Mel Frequency Cepstrum Coefficient) MFCC features at the sampling rate of 5000Hz. Then we match each frame with an MFCC audio input with a size of 20×13. We use common reconstruction metrics such as PSNR and SSIM [Wang *et al.*, 2004] to evaluate the quality of the synthesized talking faces. Furthermore, we use Landmark Distance (LMD) to evaluate the accuracy of the generated lip by calculating the landmark distance between the generated video and the original video. The lower LMD, the better of the generation.



(a)

(b)

Figure 5: Generation examples of our method comparing with Ground Truth (G.T.) (a), and Zhou *et al.* and Chen *et al.* (b). (Better zoom in to see the detail).

Methods	Evaluation on LRW		
	PSNR ↑	SSIM ↑	LMD ↓
Chung *et al.* [2017]	28.06	0.46	2.23
Chen *et al.* [2018]	28.65	0.53	1.92
Zhou *et al.* [2019]	26.80	0.88	—
Chen *et al.* [2019]	30.91	0.81	1.37
AMIE (Ours)	**32.08**	**0.92**	**1.21**

Table 1: Quantitative results.

4.2 Quantitative Results

We compare our model with four recent state-of-the-art methods, including Chung *et al.* [2017], Chen *et al.* [2018] Zhou *et al.* [2019], and Chen *et al.* [2019]. Table 1 shows the quantitative results of our method and its competitors with higher PSNR, SSIM and lower LMD, suggesting the best quality of the generated video frames of the talking faces. Although Zhou *et al.* [Zhou *et al.*, 2019] obtains the lowest PSNR, it obtains the second highest SSIM and its SSIM is significantly better than Chung *et al.* [Chung *et al.*, 2017] and Chen *et al.* [Chen *et al.*, 2018].

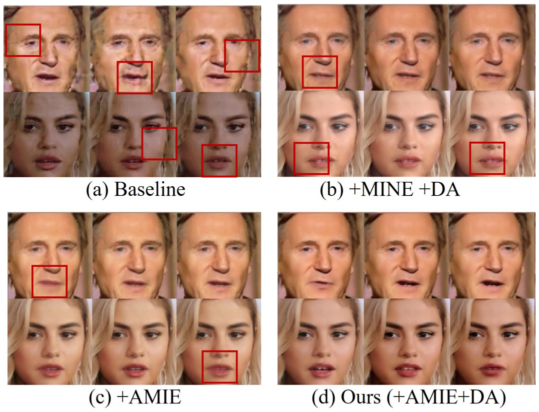
To further verify the robustness for arbitrary person generation, we evaluate our method on another benchmark dataset GRID [Cooke *et al.*, 2006] and report the comparison results in Table 2. From Table 2, we observe that our model achieves the highest SSIM and the lowest LMD, demonstrating the effectiveness and robustness of our method. Although the PSNR of our method is a little lower than that of Chen *et al.* [Chen *et al.*, 2019], our method surpass [Chen *et al.*, 2019] on the metrics of both SSIM and LMD. Therefore, our method always achieves the highest PSNR and SSIM, demonstrating our method is able to generate high-quality videos.

4.3 Qualitative Results

To present the superiority of our method, we provide generated samples compared with Zhou *et al.* [2019], and Chen *et*

Methods	Evaluation on GRID		
	PSNR ↑	SSIM ↑	LMD ↓
Chung *et al.* [2017]	29.36	0.74	1.35
Chen *et al.* [2018]	29.89	0.73	1.18
Chen *et al.* [2019]	**32.15**	0.83	1.29
AMIE (Ours)	31.01	**0.97**	**0.78**

Table 2: Cross-dataset evaluation of our method on GRID dataset pre-trained on LRW dataset.


(a) Baseline

(b) +MINE +DA

(c) +AMIE

(d) Ours (+AMIE+DA)

Figure 6: Qualitative results of ablation.

al. [2019]. As shown in Fig. 5 (a), the identity face image is obtained from the testing set in LRW [Chung and Zisserman, 2016] dataset, while in Fig. 5 (b), the arbitrary identity face image is downloaded from internet. It is clear to see that, Zhou *et al.* suffer from a ”zoom-in-and-out” effect, while the lip shapes of Chen *et al.* appear differences from the real one. In general, our model can generate more realistic and synchronous frames.

4.4 Ablation Study

In order to quantify the effect of each component of our method, we conduct ablation study experiments to verify the contributions of the two key components in our model: Asymmetric Mutual Information Estimator (AMIE) and Dynamic Attention (DA), and the two important strategies in AMIE: the asymmetric training strategy ($Asy.$) and JS represented estimator. From Table 3, it is clear to see that, (1) Simply adopting MI into our baseline did not significantly improve the results, comparing Table 3 (a) to baseline. (2) DA plays an important role in talking face generation, comparing Table 3 (b) to baseline. (3) Introducing either $Asy.$ or JS strategy shows the great improvement on almost metrics, comparing Table 3 (e) and (f) to Table 3 (c). (4) After integrating the asymmetric training strategy ($Asy.$) and JS represented estimator into the MINE, our AMIE can further improve the performance on all the metrics, as shown in Table 3 (g). (5) Integrating AMIE (comparing Table 3 ’Ours’ to Table (b)) can further boost the performance, while simply integrating MINE (comparing Table 3 (c) to (b)) the performance declined, that verifies the contribution of proposed AMIE. Fig. 6 demonstrates the visualized examples.

Methods	Evaluation on LRW		
	PSNR ↑	SSIM ↑	LMD ↓
Baseline	28.88	0.89	1.36
(a) + MINE	29.05	0.89	1.38
(b) + DA	29.19	0.90	1.37
(c) + MINE + DA	29.08	0.89	1.32
(d) + MINE + JS	29.33	0.90	1.50
(e) + MINE + JS + DA	29.12	0.91	1.21
(f) + MINE + $Asy.$ + DA	29.30	0.90	1.33
(g) + AMIE	29.41	**0.92**	1.22
Ours (+ AMIE + DA)	**29.64**	**0.92**	**1.18**

Table 3: Ablation study of the key components AMIE and DA in our method as well as two strategies applied in AMIE: Asymmetric Training ($Asy.$) and JS represented estimator (JS). Ours = Baseline + AMIE + DA, and AMIE = MINE + $Asy.$ + JS.

Methods	Realistic	Synchronization
Zhou *et al.* [2019]	15.22%	18.17%
Chen *et al.* [2019]	28.37%	32.92%
AMIE (Ours)	**56.41%**	**48.91%**

Table 4: Results of user study.

4.5 User Study

We conduct a user study on the LRW dataset with 42 volunteers in both realistic and synchronization of generation. We randomly select the samples generated by our method, Zhou *et al.* [2019] and Chen *et al.* [2019]. Then, the volunteers were asked to answer the following two questions: (1) which one appears more realistic and (2) which one provides more temporal synchronism referring to the ground truth. Table 4 clearly demonstrates that our model achieves the highest rating in both realistic and synchronization.

5 Conclusion

We have proposed a novel model of talking face generation for arbitrary identities via exploring the cross-modality coherence in this paper. Our model leverages the asymmetric mutual information estimator to learn the correlation of audio and facial image features and utilizes dynamic attention to simulate the process of disentangling. Extensive experimental results on benchmark datasets demonstrate the promising performance of our method.

Acknowledgments

This work is partially funded by the Beijing Natural Science Foundation (JQ18017), Youth Innovation Promotion Association CAS (Grant No. Y201929), the National Nature Science Foundation of China (61976002), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046).

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[Chen *et al.*, 2018] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *European Conference on Computer Vision (ECCV)*, 2018.

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

[Chung and Zisserman, 2016] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[Chung *et al.*, 2017] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *British Machine Vision Conference (BMVC)*, 2017.

[Cooke *et al.*, 2006] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[Donsker and Varadhan, 1983] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[Gupta and Aggarwal, 2010] Neelima Gupta and Seema Aggarwal. Mib: Using mutual information for biclustering gene expression data. *Pattern Recognition*, 43(8):2692–2697, 2010.

[He *et al.*, 2009] Ran He, Bao-Gang Hu, and Xiao-Tong Yuan. Robust discriminant analysis based on nonparametric maximum entropy. In *Asian Conference on Machine Learning*, pages 120–134. Springer, 2009.

[He *et al.*, 2014] Ran He, Baogang Hu, Xiaotong Yuan, and Liang Wang. *Robust recognition via information theoretic learning*. Springer, 2014.

[Hjelm *et al.*, 2018] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[Hulle, 2005] Marc M Van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910, 2005.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[Jun *et al.*, 2011] Bongjin Jun, Taewan Kim, and Daijin Kim. A compact local binary pattern using maximization of mutual information for face analysis. *Pattern Recognition*, 44(3):532–543, 2011.

[Kraskov *et al.*, 2004] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[Kumar *et al.*, 2017] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017.

[Liu *et al.*, 2009] Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang. Feature selection with dynamic mutual information. *Pattern Recognition*, 42(7):1330–1339, 2009.

[Ruderman *et al.*, 2012] Avraham Ruderman, Mark Reid, Darío García-García, and James Petterson. Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*, 2012.

[Sotoca and Pla, 2010] José Martínez Sotoca and Filiberto Pla. Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition*, 43(6):2068–2081, 2010.

[Suwajanakorn *et al.*, 2017] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95:1–95:13, 2017.

[Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research (JMLR)*, 11(12):3371–3408, 2010.

[Vougioukas *et al.*, 2018] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. In *British Machine Vision Conference (BMVC)*, 2018.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004.

[Zhou *et al.*, 2019] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *International AAAI Conference on Artificial Intelligence (AAAI)*, 2019.