Supplementary Materials AnyTalk: Multi-modal Driven Multi-domain Talking Head Generation

Anonymous Author(s)
Submission Id: 1373

APPENDIX

The appendix is organized as follows: First, we provide additional comparison experiments, ablation study in Sec. A. Next, we introduce the network and loss function of AnyTalk in details in Sec. B. Finally, we also present the experiment details and competing baselines in Sec. C.

A ADDITIONAL EXPERIMENTS

In this section, we provide additional experiment results to verify the superiority of AnyTalk. Beyond the commonly used metrics such as **FID**, **CSIM**, and **CPBD**, we employ the Average Euclidean Distance (**AED**) to measure the identity preservation the Average and the Average Keypoint Distance (**AKD**) to evaluate the motion preservation in the input driving image. We also adopt employ the Learned Perceptual Image Patch Similarity (**LPIPS**), and $\mathcal{L}_1$ distance to quantify the low-level similarity between the synthetic and driving images.

A.1 Additional Results on Within-Domain Face Reenactment

Beside the cross-domain evaluation, we also compare our AnyTalk with strong baselines in the task of face reenactment within the real human and Disney human domain.

Additional Results on Real Human Domain For the same-identity in real human domain evaluation, we follow the sampling strategy in previous methods [2] and use 200 test videos from VoxCeleb1 [3]. For the cross-identity in real human domain evaluation, we use 200 videos from VoxCeleb1 [3] to drive 2,100 images from CelebV [7] to versify the generalization ability.

We present the quantitative results in the real human domain in Tab. 1 and Tab. 2. For cross-identity evaluation, our method achieves the best performance across all metric. For same-identity evaluation, our method is overall comparable with DaGAN [2] and better than other baselines.

Additional Results on Disney Human Domain. For the same-identity in Disney human domain evaluation, we also randomly select 50 test videos with Disney human style from AniTalk to perform reconstruction tasks. For the cross-identity in Disney human domain evaluation, we use the 50 test videos from AniTalk to drive 2,100 images from different test videos with Disney human style in AniTalk.

	FID↓	CSIM↑	CPBD ↑
FOMM	16.504	0.860	0.2135
DaGAN	13.313	0.866	0.2928
ToonTalker	22.627	0.832	0.2617
Face-vid2vid	13.861	0.866	0.3244
AnyTalk	**13.197**	**0.867**	**0.3293**

Table 1: Cross-identity reenactment in VoxCeleb [3] and CelebV [7].

	L1↓	LPIPS↓	CPBD↑	AKD↓	AED↓
FOMM	0.0417	0.199	0.1946	1.386	0.130
DaGAN	**0.0383**	0.156	0.2277	**1.265**	**0.117**
ToonTalker	0.0554	0.238	0.2033	1.717	0.191
Face-vid2vid	0.0400	0.158	**0.2741**	1.540	0.129
AnyTalk	0.0387	**0.157**	0.2650	1.362	0.124

Table 2: Same-identity reconstruction in VoxCeleb [3].

We also present the results in Disney human domain in Tab. 4 and Tab. 3. For cross-identity evaluation, our method achieves the best performance across all metric. For same-identity evaluation, our method achieves the best **LPIPS**, **CPBD**, **AKD** and **AED**, indicating that the superiority of our method. Besides, we our method is comparable with DaGAN in term of **FID** and better than other baselines.

	FID↓	CSIM↑	CPBD ↑
FOMM	24.966	0.7959	0.2163
DaGAN	24.025	0.8042	0.2373
ToonTalker	26.818	0.8190	0.2592
Face-vid2vid	23.372	0.8133	0.2572
AnyTalk	**22.637**	**0.8196**	**0.2655**

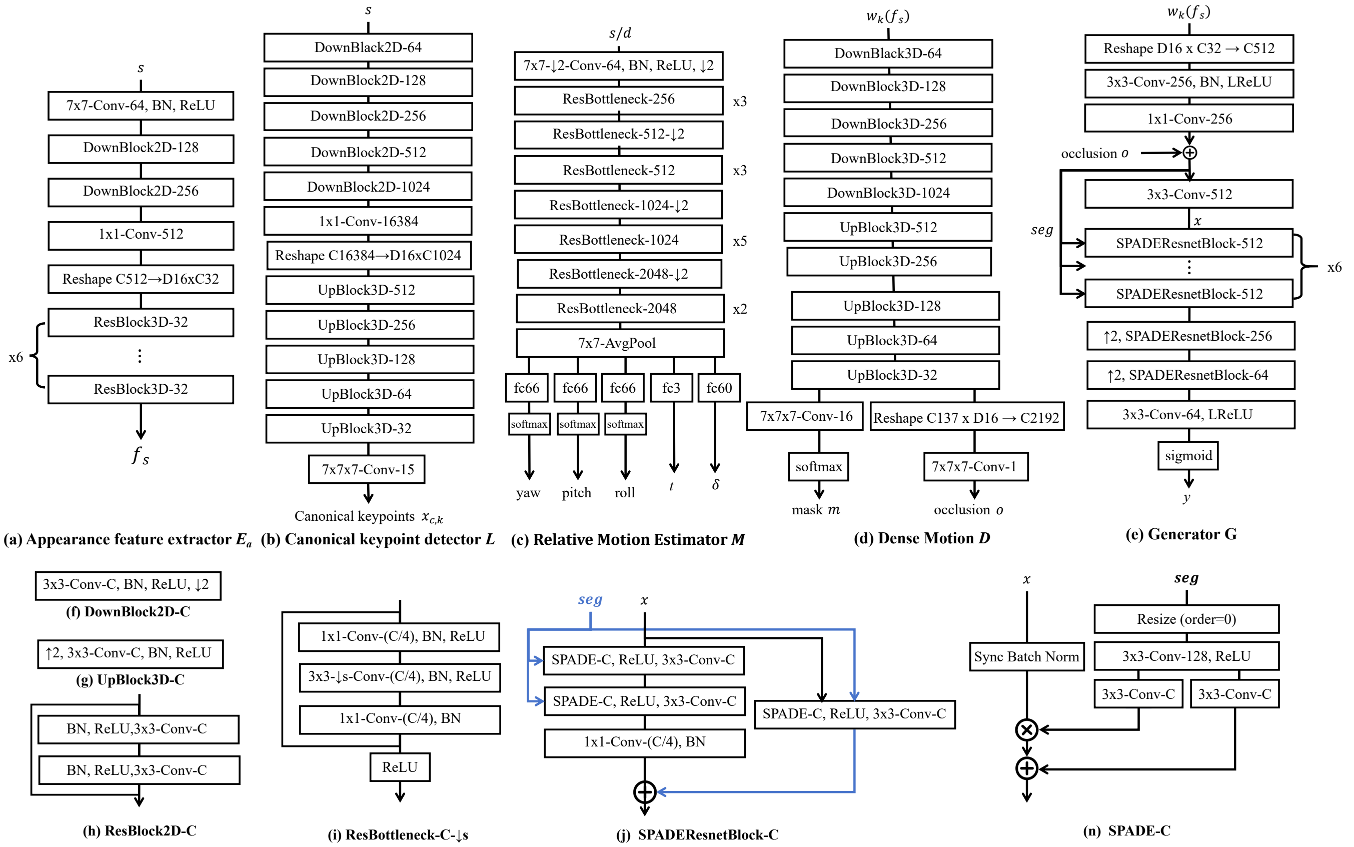
Table 3: The Cross-identity reenactment within Disney human domain in AniTalk.

A.2 Additional Ablation Study about the hyper-parameter λ_{exp}.

To further analyze the influence of expression consistent loss λ_{exp}, we perform an ablation study about the trade-off hyper-parameter λ_{exp} on within-domain face reenactment, *i.e.*the same-identity and cross-identity face reenactment in the real domain. We report the

Conference acronym '24, October 28 - November 01 2024, Melbourne, Australia

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX



(a) Appearance feature extractor E_a **(b) Canonical keypoint detector L** **(c) Relative Motion Estimator M** **(d) Dense Motion D** **(e) Generator G**

(f) DownBlock2D-C

(g) UpBlock3D-C

(h) ResBlock2D-C **(i) ResBottleneck-C-↓s** **(j) SPADEResnetBlock-C** **(n) SPADE-C**

Figure 1: The architecture of our AnyTalk.

	L1↓	LPIPS↓	CPBD↑	AKD↓	AED↓
FOMM	0.0431	0.1869	0.2140	2.787	0.220
DaGAN	**0.0423**	0.1832	0.2345	2.726	0.209
ToonTalker	0.0673	0.2509	0.2393	6.596	0.358
Face-vid2vid	0.0435	0.1819	0.2626	2.701	0.218
AnyTalk	0.0429	**0.1776**	**0.2663**	**2.597**	**0.205**

Table 4: The same-identity reconstruction task within Disney human domain in AniTalk.

L_1, FID and CPBD in Tab. 5. As we can see, it can be observed that λ_{exp} trades off the performance of the two cross-domain tasks. We empirically choose $\lambda_{exp} = 0.1$ in practical.

Setup	λ_{exp}	0	0.1	0.2	0.3	0.4
Same Id	L_1↓	0.0400	**0.0387**	0.0396	0.0396	0.0440
	CPBD↑	**0.2741**	0.2650	0.2572	0.2565	0.2422
Cross Id↓	FID	13.861	**13.197**	13.342	13.528	13.629
	CPBD↑	0.3244	**0.3293**	0.3148	0.3011	0.2978

Table 5: The ablation study about the hyper-parameter λ_{exp}.

B ADDITIONAL NETWORK AND TRAINING DETAILS

B.1 Network architecture details of AnyTalk

The implementation details of several key model modules in our model are shown in Fig. 1.

Appearance feature extractor E_a. The network E_a (see Fig. 1.a) takes the source image s and extracts its 3D appearance features f_s. It incorporates several 2D downsampling blocks initially, followed by a convolutional layer that transforms the input 2D features into 3D features. Subsequently, we utilize six 3D residual blocks to calculate the final 3D feature representation, denoted as f_s.

Canonical keypoint detector L. The module L (see Fig. 1.b) leverages a U-Net styled encoder-decoder mechanism to estimate the canonical keypoints from source image s. To accommodate the extraction of 3D keypoints, the encoded features are projected into a three-dimensional space via a 1×1 convolution. This 1×1 convolution serves as the bottleneck of the U-Net structure. Subsequently, the decoder component of the U-Net is composed of 3D convolution layers and upsampling stages.

Relative Motion Estimator M. The relative motion estimator M (see Fig. 1.c) comprises multiple ResNet bottleneck blocks, succeeded by global pooling to eliminate the spatial dimension. Subsequently, various linear layers are employed to compute the rotation angles (*i.e.*yaw, pitch, roll), translation vector t, and expression deformations δ. The full angle range is divided into 66 bins for rotation angles, and the network predicts which bin the target angle is in. The computed head pose (*i.e.*yaw, pitch, roll) and deformations (t,δ) facilitate the transformation of canonical keypoints into the source or driving keypoints, essential for accurate alignment.

Dense Motion D. We utilize the dense motion D to predict an 2D occlusion map o to filter out the regions that should be inpainted, and a 2D motion flow mask m for weighting the motion field. As illustrated in Fig. 1.c, there are two heads at the end to predict these two parts.

Generator G. The generator G first processes the warped 3D appearance features, denoted as $w_k(f_s)$, projecting them into a 2D space. Subsequently, these features are element-wise multiplied by an occlusion mask, denoted as o, which is derived from the motion field estimator M. The next stage involves the application of a sequence of 2D residual blocks, coupled with upsampling layers, to synthesize the final image y.

B.2 Loss details

Expression consistency loss $\mathcal{L}_{exp}$. To improve facial expression detail in output images, We proposed a novel expression consistency loss ensures the consistency of driving frame expression ϕ_d and output expression ϕ_y, as follows:

$$\mathcal{L}_{exp}(\phi_d, \phi_y) = ||\phi_d, \phi_y||_2. \tag{1}$$

In fact, $\mathcal{L}_{exp}$ compute a perceptual difference between the driving frame expression ϕ_d and output expression ϕ_y. Optimizing $\mathcal{L}_{exp}$ improves facial expression details in output videos.

Perceptual Loss $\mathcal{L}_P$. Perceptual loss is a popular objective function in image generation tasks [2, 4, 6]. Following previous methods [4, 6], we use a multi-scale implementation. Specifically, the ground truth and output images are initially downsampled to four distinct resolutions: 256x256, 128x128, 64x64, and 32x32. We denote R_1, R_2, R_3, and R_4 as the images generated at these resolutions, and G_1, G_2, G_3, and G_4 as the corresponding ground truths. Subsequently, a pre-trained VGG network is employed to extract features from both the downsampled ground truths and the output images at these resolutions. We compute the $\mathcal{L}_1$ distance between the ground truth and output image in different resolutions:

$$\mathcal{L}_P = \sum_{i=1}^{4} \mathcal{L}_1(G_i, R_i) \tag{2}$$

GAN loss $\mathcal{L}_G$. We adopt the same patch GAN implementation as Face-vid2vid [6] and use the hinge loss. Feature matching [5] loss is also adopted to stabilize training. We use single-scale discriminators for training 256×256 images, and two-scale discriminators [5] for 512×512 images.

Equivariance loss $\mathcal{L}_E$. This loss function is employed to maintain the consistency of the estimated keypoints, similar to the approach used by FOMM [4]. Consider an image I and a detected keypoint within it. A predefined spatial transformation T is applied to I, generating a transformed image I_T. Consequently, the corresponding keypoints $X_{T(k)}$ in I_T should undergo the same transformation. Thus, for the K keypoints detected in I, the following condition must hold:

$$\mathcal{L}_E = \sum_{i=1}^{K} ||x_k - T^{-1}(x_{T(x)})||_1 \tag{3}$$

Keypoint prior loss $\mathcal{L}_{dist}$. To make the detected facial keypoints much less crowded around a small neighbourhood, we apply a keypoint distance loss $\mathcal{L}_{dist}$ [2]. This loss penalizes the model when the distance between any two corresponding keypoints, x_i and x_j, is below some threshold D_t, or if the mean depth value deviates from a preset target value z. For every pair of keypoints in an image, the following constraint is applied:

$$\mathcal{L}_{dist} = \sum_{i=1}^{K}\sum_{j=1}^{K} max(0, D_t - ||x_i, x_j||_2^2) + ||Z(x_d) - z_t||, i \neq j, \tag{4}$$

where where $Z(\cdot)$ extracts the mean depth value of the keypoints, and the D_t and z_t is the threshold of distance. We set D_t to 0.1 and z_t to 0.33 in our work, which shows good performance in our practice.

Head pose loss $\mathcal{L}_M$. We compute the $\mathcal{L}_1$ distance between the estimated head pose R_d and the estimated head pose $\hat{R}_d$ by a pre-trained pose estimator [6], which is considered as the ground truth. Specifically, the head pose loss $\mathcal{L}_M$ is defensed as

$$\mathcal{L}_M = ||R_d - \hat{R}_d||, \tag{5}$$

where the distance is calculated as the sum of the absolute differences across the corresponding Euler angles.

Deformation priors loss $\mathcal{L}_\Delta$. Given that the expression deformation Δ represents the deviation from the canonical keypoints, its magnitude should be reasonably constrained. To enforce this constraint, we impose a loss on the L1 norm of these deviations:

$$\mathcal{L}_\Delta = ||\Delta_{d,k}||, \tag{6}$$

Total loss. The total loss is given by

$$\begin{aligned}\mathcal{L}_{total} = \lambda_{exp}\mathcal{L}_{exp} + \underbrace{\lambda_P\mathcal{L}_P + \lambda_G\mathcal{L}_G}_{\text{Perceptual and GAN loss}} + \\ \underbrace{\lambda_E\mathcal{L}_E + \lambda_{dist}\mathcal{L}_{dist}}_{\text{Equivalence and keypoint dist loss}} + \underbrace{\lambda_M\mathcal{L}_M + \lambda_\Delta\mathcal{L}_\Delta}_{\text{Relative motion loss}},\end{aligned} \tag{7}$$

where the $\lambda_P, \lambda_G, \lambda_E, \lambda_{dist}, \lambda_M, \lambda_\Delta$ and λ_{exp} are the hyper-parameters that facilitate balanced learning from these losses.

C ADDITIONAL EXPERIMENT DETAILS

C.1 Implementation Details.

The 3D keypoints detector, dense motion estimator and generator in AnyTalk follows Face-vid2vid [6]. We train AnyTalk in two stages to enhance the generality. To optimize the training objectives, we set λ_P = 10, λ_G = 1, λ_E = 10, λ_M = 20 , λ_Δ = 5, λ_{dist} = 10 and λ_{exp} = 0.1. The number of keypoints is set to 15, which is the same as that of Face-vid2vid [6]. In the first stage, we employ 8 RTX A6000 GPUs to trained AnyTalk for 200 epochs in an end-to-end manner. Then, we further finetune our AnyTalk on AniTalk and a subset of VoxCeleb1 [3] for 100 epochs. For a fair comparison, all competing methods are fine-tuned on the same datasets.[1]. For

[1] Because without the open-source training code, we use the checkpoint of ToonTalker to evaluate it.

the evaluation on cross-domain face reenactment, we use 78 videos with Disney human/animal style from AniTalk to drive 2,100 images from VoxCeleb1 [3].

C.2 Compare methods.

FOMM [4]. FOMM propose a paradigm that aims to detect the keypoints of the face image and model the motion between two images using detected keypoints.

DaGAN [2]. DAGAN introduces depth into the estimation of key points and has a similar formulation of motion transfer as FOMM.

Facevid2vid [6]. Face-vid2vid proposes a pure neural rendering approach, which renders a talking-head video using a deep network in the one-shot setting without using a graphics model of the 3D human head.

ToonTalker [1]. ToonTalker proposes a unified cross-domain face reenactment framework, which employs domain-specific motion estimators and generators for each domain and cross-domain motion alignment models to transfer motion across domains.

REFERENCES

[1] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, et al. 2023. Interactive Story Visualization with Multiple Characters. (2023).

[2] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. 2022. Depth-Aware Generative Adversarial Network for Talking Head Video Generation. In *CVPR*.

[3] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *INTERSPEECH*.

[4] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *NeurIPS* (2019).

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*.

[6] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. (2021).

[7] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *ECCV*.