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

Yu Wang, Yunfei Liu, Fa-Ting Hong, Meng Cao, Lijian Lin, Yu Li[*]

International Digital Economy Academy (IDEA)

Abstract

Cross-domain talking head generation, such as animating a static cartoon animal photo with real human video, is crucial for personalized content creation. However, prior works typically rely on domain-specific frameworks and paired videos, limiting its utility and complicating its architecture with additional motion alignment modules. Addressing these shortcomings, we propose AnyTalk, a unified framework that eliminates the need for paired data and learns a shared motion representation across different domains. The motion is represented by canonical 3D keypoints extracted using an unsupervised 3D keypoint detector. Further, we propose an expression consistency loss to improve the accuracy of facial dynamics in video generation. Additionally, we present AniTalk, a comprehensive dataset designed for advanced multimodal cross-domain generation. Our experiments demonstrate that AnyTalk excels at generating high-quality, multimodal talking head videos, showcasing remarkable generalization capabilities across diverse domains.

Introduction

With the advancement of mobile internet and the proliferation of short video platforms, individuals are increasingly appearing in videos. Speakers sometimes hope to present themselves by driving rich and vivid characters (*e.g.*, photorealistic person, Disney roles, cartoon animals, *etc.*) for entertainment consideration. However, animating these characters involves a complex 3D pipeline that requires extensive labor and significant time. Consequently, there is a burgeoning field of research focused on simplifying this animation process (Gong et al. 2023).

One-shot talking head generation aims to drive/animate a portrait image based on the motion provided by a driving video or an audio sequence. Previous methods mainly focus on one-shot talking head generation within the same domain, *i.e.*, driving a portrait with a video of a real person. These methods are usually trained with a large amount of real human talking videos and learn warping-based motion representations (Siarohin et al. 2019; Zhao and Zhang 2022), facial keypoints (Liu et al. 2023) or 3D Morphable Models (3DMMs) (Wu et al. 2021) to perform real person talking head generation. Although effective, these methods fail to generate pleasing talking heads for new domains, *e.g.*, Disney animals, cartoon characters, *etc.*. Only a few methods (Bansal et al. 2018; Xu et al. 2022; Gong et al. 2023; Kim et al. 2022) perform visually-driven cross-domain talking head generation. However, these methods only support videos from two different domains and usually require further model designs for different domains.

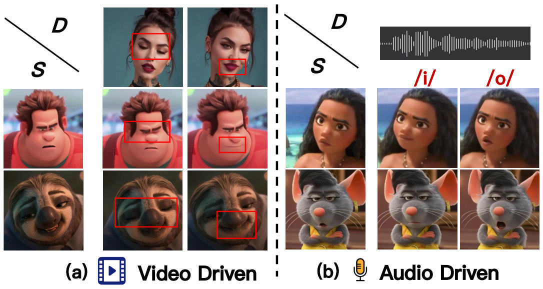

Figure 1: Cross-domain reenactment examples generated by our AnyTalk. Given a cartoon human or animal face as the source image, AnyTalk enables driving the source image with a video of a real person or an arbitrary audio speech.

In this paper, we introduce AnyTalk, a system designed to generate talking head videos across multiple domains. It does not require paired data or additional network modules for different domains. AnyTalk reveals two key insights: i) **a unified end-to-end cross-domain talking head generation framework is needed.** Recent works like ToonTalker (Gong et al. 2023) employ domain-specific motion estimators and cross-domain motion alignment modules to transfer motion across domains and perform cross-domain face reenactment (see Fig. 2.a). However, using separate networks for each domain leads to a cumbersome framework. Such a design requires more computational resources and limits their applications. Additionally, the absence of paired data poses significant challenges in optimizing cross-domain motion alignment, leading to inaccurate expression transfer and poor generated quality. Different from previous methods, we present a simple and unified framework, termed by AnyTalk, for cross-domain talking head generation (see Fig. 2.b). Every component in AnyTalk is shared across multiple domains. ii) **A large-scale and diverse cross-domain dataset**

[*]Corresponding Author.

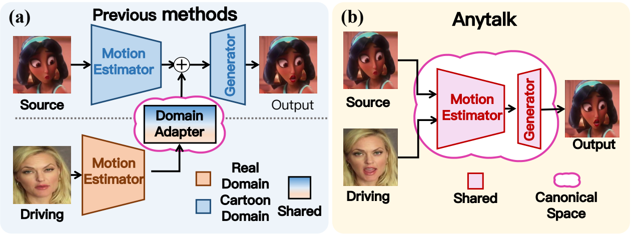

Figure 2: Comparison between previous methods and our AnyTalk. (a) Previous works (*e.g.* Toontalker (Gong et al. 2023)) employ domain-specific motion estimators and cross-domain motion alignment models to transfer motion across domains, which significantly limits their application. (b) Our AnyTalk is a unified cross-domain talking head generation framework, where each of its modules, e.g., such as the keypoint detector, are shared across multi-domains.

Methods	Video-driven	Audio-driven	3D Manipulation	Free View	Cross Domain
FOMM	✓				
Face-vid2vid	✓		✓	✓	
SadTalker		✓			
Wav2Lip		✓			
CVTHead	✓		✓		
Geneface			✓	✓	
ToonTalker	✓				✓
AnyTalk	✓	✓	✓	✓	✓

Table 1: The comparison between our AnyTalk and priors.

is needed. Apparently, a dataset in a single domain alone cannot facilitate the transfer of motions across multiple domains. However, videos from multiple domains exhibit significant diversity in expressions, head poses, and appearances. To equip the model with cross-domain capabilities, we enrich the existing data by collecting a high-quality talking head dataset (called AniTalk). This dataset consists of 1,250 videos from multiple domains, including 600 videos in Disney human style, 325 videos in Disney animals style, and 325 videos in mesh style (real humans).

Characters from different domains usually have very different styles, *e.g.*, the head movements and facial expressions of cartoon characters are often exaggerated compared to real people. Specifically, to bridge the domain gap, we propose to decompose head pose and expressions to ease the model learning. Inspired by Face-vid2vid (Wang, Mallya, and Liu 2021a), we utilize a canonical 3D keypoint detector to estimate the 3D canonical keypoints of the source image. Next, we estimate the relative motion, *i.e.*, head pose and keypoint perturbations, for both the source and driving images. Furthermore, to improve the facial expression details in generated videos, we propose a novel expression loss to constrain the expression consistency between the output videos and the driving videos. Benefiting from the proposed AniTalk dataset, our method exhibits high versatility, supporting various applications such as cross-domain synthesis and facial animation. Leveraging this decomposition approach, AnyTalk facilitates advanced functionalities like free-view control in talking head video editing. We summarize the differences between our AnyTalk and previous works in Tab. 1. It is important to emphasize that our objective is not to develop an intricate network architecture but rather to investigate a pioneering learning paradigm tailored for cross-domain face reenactment.

We conduct extensive experiments to evaluate our AnyTalk in the cross-domain setup. Our main contributions are summarized as follows:

- We propose a unified framework, **AnyTalk**, for cross-domain talking head generation that does not require paired data. AnyTalk leverages a cross-domain training scheme to extract precise canonical keypoints from images across different domains, facilitating accurate domain-agnostic motion transfer without requiring additional alignment modules. Additionally, we introduce a novel expression loss designed to enhance facial details during the cross-domain video generation process.
- We collect a high-quality and diverse talking video dataset, termed as AniTalk, for cross-domain talking head generation. AniTalk includes multiple domains like Disney human style, Disney animals, and mesh style.
- Experiments demonstrate that our AnyTalk can generate high-quality and diverse videos across multiple domains, including across species.

Related Works

Single-domain Talking Head Generation. Most talking head generation works focus on within-domain setup, *i.e.*, driving a real person image using another real person talking video/speech. These within-domain talking head generation works can generally be categorized into three classes. *(i) Video-driven Methods.* The video-driven methods focus on capturing the facial expressions from a driving video and blending them with the facial identity of a source image. Several approaches (Yao et al. 2020; Wu et al. 2021; Wang, Zhang, and Li 2021) employ a pretrained 3D Morphable Models (3DMMs) regressor (Tran and Liu 2018; Zhu et al. 2017) to decouple pose, expression, and identity to synthesize a new face. Additionally, many methods (Tripathy, Kannala, and Rahtu 2021; Ha et al. 2020; Zakharov et al. 2020, 2019; Zhao, Wu, and Guo 2021) utilize facial landmarks to represent facial dynamics, which are detected by a pre-trained face model (Guo et al. 2019). Recently, unsupervised learning techniques (Siarohin et al. 2019; Hong et al. 2022; Wang, Mallya, and Liu 2021b; Liu et al. 2021; Zhao and Zhang 2022; Hong and Xu 2023; Wang, Mallya, and Liu 2021a) are proposed to learn implicit keypoints to represent facial motion, which is used for modeling the motion transformations between two faces. *(ii) Audio-driven Methods.* Audio-driven talking head generation (He et al. 2024; Thies et al. 2020; Lu, Chai, and Cao 2021; Zhou et al. 2020; Lahiri et al. 2021; Wang et al. 2021; Zhou et al. 2021; Wang et al. 2022; Liu et al. 2023; Lu et al. 2023; Chen et al. 2024; Hong et al. 2024) is another popular direction, which aims to synthesize talking videos in sync with the input speech content.

Recently, zero-shot audio-driven methods have emerged, requiring only a single portrait image of the target avatar and the corresponding audio. Wav2Lip (Prajwal et al. 2020) directly learns the mapping between audio feature and mouth images. SadTalker (Zhang et al. 2022) leverages 3DMMs as an intermediate representation between speech and video. However, audio-driven talking head generation for different domain videos are not well explored. *(iii) 3D-based Methods.* Several common but increasingly explored categories include methods like NeRF-based, 3D Gaussian Splatting and other 3D-aware techniques (Wu et al. 2023; Guo et al. 2021; Shen et al. 2022; Ye et al. 2023; Chu et al. 2024; Li et al. 2024). These methods utilize an explicit 3D geometry and complex light interactions to create 3D head animations. For instance, AdNeRF (Guo et al. 2021) proposes model head and neck with two neural fields. Although effective, they exhibit limitations in cross-domain talking head generation, such as sub-par expression transfer, and insufficient decoupling of head poses.

Cross-domain Face Reenactment. Recently, significant efforts have been dedicated to designing and improving cross-domain Face Reenactment (Bansal et al. 2018; Xu et al. 2022; Gong et al. 2023; Kim et al. 2022; Song et al. 2021).

AnimeCeleb (Kim et al. 2022) introduces a novel dataset for cartoon talking-head videos, utilizing a 3D animation model to generate a vast collection of animated facial images with corresponding pose annotations. However, AnimeCeleb (Kim et al. 2022) relies on paired data restricts its applicability, and its inability to generalize to various cartoon styles with rich expressions is a limitation. ToonTalker (Gong et al. 2023) proposes a cross-domain face reenactment framework, which employs domain-specific motion estimators and generators for each domain, and needs cross-domain motion alignment models to transfer motion across domains. However, using separate networks for each domain leads to a cumbersome framework, requiring more computational resources and increasing inference time. Meanwhile, previous methods like (Gong et al. 2023) only support videos from two different domains at most and have poor generalization. In this paper, we propose a simple but effective unified end-to-end framework for cross-domain talking head generation. Our approach does not require paired data and performs well in generating talking heads across multiple domains.

Methodology

Given a source image s and a talking head driving video $\mathcal{D} = \{d^i\}_{i=1}^N$ (N is the number of frames), one-shot face reenactment aims to generate an output video $\mathcal{Y} = \{y^i\}_{i=1}^N$. Each frame y_i retains the identity of the source image s while following the motions from the corresponding driving frame d^i. In this work, we focus on the cross-domain video-driven talking head generation, also known as cross-domain face reenactment, where the source image s and driving video $\mathcal{D}$ come from different domains.

Overview

An overview of our proposed AnyTalk for cross-domain talking head generation is depicted in Fig. 3. It can be divided into three parts: (i) **Motion Estimation under Canonical Space.** Motivated by Face-vid2vid (Wang, Mallya, and Liu 2021a), we first utilize a canonical keypoint detector to extract the K canonical 3D keypoints $\mathcal{X}_c = \{x_{c,k}\}_{k=1}^K$, $x_{c,k} \in \mathbb{R}^3$ from s using a canonical 3D keypoint detection network L. Meanwhile, we adopt a relative motion estimator to extract the relative head pose and expression information from s. Then, we combine the canonical keypoints from L with the relative motion information from D to obtain the source 3D keypoints $\mathcal{X}_s = \{X_{s,k}\}_{k=1}^K$ and the driving 3D keypoints $\mathcal{X}_{d^i} = \{X_{d^i,k}\}_{k=1}^K$. With the 3D keypoints $\mathcal{X}_s, \mathcal{X}_{d^i}$ generated from the driving frame and the source image, we are able to calculate the motion flow $\mathcal{M} = \{m_j\}_{j=1}^K$ with the defense motion estimator D (Siarohin et al. 2019; Wang, Mallya, and Liu 2021a; Hong et al. 2022) between s and d^i. (ii) **Feature Wrapping and Image Generation.** Based on the motion flow $\mathcal{M}$, AnyTalk warps the source image s and then subsequently passes the warped results to the image generator G to produce the output image y^i. (iii) **Expression Consistency Learning.** We introduce a pretrained expression encoder E_{exp} to predict the facial expression and propose a novel expression consistency loss, namely $\mathcal{L}_{exp}$, designed to encourage similarity in facial expressions between the driving frames and the output.

Cross-domain Motion Estimation

To address the cross-domain facial reenactment problem, AnyTalk estimates motion under a canonical space, which naturally facilitates motion transfer across multiple domains without any requirement for training additional adapter modules. In this section, we introduce how to transfer motion across multiple domains under a canonical space without additional adapter module training.

Canonical Landmark Detection. Inspired by Face-vid2vid (Wang, Mallya, and Liu 2021a), we can decompose the top-K 3D facial keypoints $\mathcal{X}_j = \{x_{j,k}\}_{k=1}^K$ of image j into the canonical keypoints $\mathcal{X}_{c_j} = \{x_{c_j,k}\}_{k=1}^K$ and relative transformations, *i.e.*the relative head pose (R_j, t_j) and expression information $\delta_{j,k}$, as follows:

$$x_{j,k} = R_j x_{c_j,k} + t_j + \delta_{j,k}, \tag{1}$$

where the relative head pose is parameterized by a rotation matrix $R_j \in \mathbb{R}^{3\times3}$ and a translation vector $t_j \in \mathbb{R}^3$, the expression vector $\delta_{j,k} \in \mathbb{R}^3$. The generation of the face is controlled by canonical keypoints $\mathcal{X}_{c_j}$, with a neutral front view without expressions. Thus, the canonical keypoints predicted from images across multiple domains can form a well-aligned canonical space.

Thus, given a source image s and driving image d^i from different domains, we employ the canonical 3D keypoint detection network L to calculate the 3D canonical landmarks $\mathcal{X}_c = \{x_{c,k}\}_{k=1}^K$ of source image as reference, as follows:

$$\mathcal{X}_c = \{x_{c,k}\}_{k=1}^K = L(s), \tag{2}$$

where $x_{c,k} \in \mathbb{R}^3$ and K is the number of keypoints.

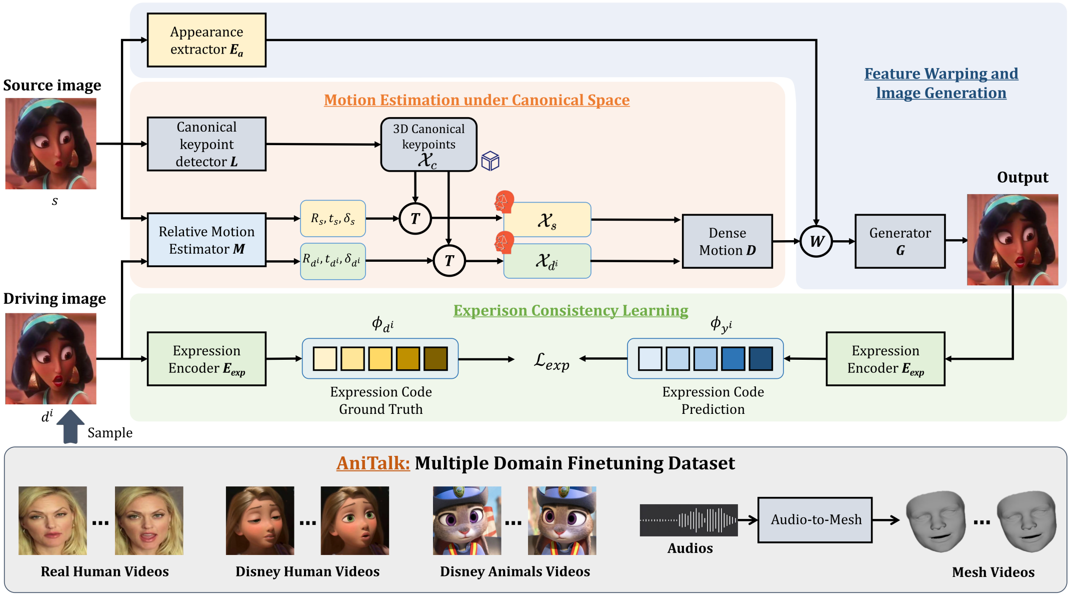
Appearance extractor E_a
Feature Warping and Image Generation
Source image
s
Motion Estimation under Canonical Space
Canonical keypoint detector L
3D Canonical keypoints $\mathcal{X}_c$
Output
R_s, t_s, δ_s
T
$\mathcal{X}_s$
Relative Motion Estimator M
$R_{d^i}, t_{d^i}, \delta_{d^i}$
T
$\mathcal{X}_{d^i}$
Dense Motion D
W
Generator G
Driving image
Experison Consistency Learning
ϕ_{d^i}
ϕ_{y^i}
Expression Encoder E_{exp}
Expression Code Ground Truth
$\mathcal{L}_{exp}$
Expression Code Prediction
Expression Encoder E_{exp}
d^i
Sample
AniTalk: Multiple Domain Finetuning Dataset
Real Human Videos
Disney Human Videos
Disney Animals Videos
Audios
Audio-to-Mesh
Mesh Videos

Figure 3: The illustration of our AnyTalk, which contains three parts: i) Motion Estimation under Canonical Space, ii) Feature Warping and Image Generation, and 3) Expression Consistency Learning. Powered by the proposed AniTalk dataset, AnyTalk learns a general and unified network for cross-domain talking head video generation.

Relative Motion Transfer. In addition, the relative motion estimator M predicts the relative head poses (R_s,t_s), (R_{d^i},t_{d^i}) and expression vectors δ_s,δ_{d^i} of source image s and driving frame d^i, respectively. We combine the canonical keypoints extracted by L with the motion-related information extracted by D to obtain the source 3D keypoints $\mathcal{X}_s = \{x_{s,k}\}_{k=1}^{K}$ and the driving 3D keypoints $\mathcal{X}_{d^i} = \{x_{d^i,k}\}_{k=1}^{K}$, as follows:

$$\begin{aligned} x_{s,k} &= R_s x_{c,k} + t_s + \delta_{s,k} \\ x_{d^i,k} &= R_{d^i} x_{c,k} + t_{d^i} + \delta_{d^i,k} \end{aligned} \tag{3}$$

By applying Eq. 3, we can easily transfer relative motion information across different domains based on $\mathcal{X}_c$, without the need for training additional adapter modules.

Video Generation. Given $\mathcal{X}_s$ and $\mathcal{X}_{d^i}$, the dense motion network D estimates the 3D motion flow $\mathcal{M}_{s \leftarrow d^i} = \{m_j\}_{j=1}^{K}$ between s and d^i. AnyTalk warps the appearance features of the source extracted by the appearance encoder E_a, and our generator G uses the warped feature to output y^i.

Expression Consistency Learning

Because AnyTalk estimates sparse facial keypoints, it is challenging to thoroughly recognize facial expression by sparse keypoints residual $\delta_{s,k}, \delta_{d^i,k}$. To enhance the learning of facial expression details (Danecek, Black, and Bolkart 2022), we apply a pre-trained emotion recognition network E_{exp} to accurately predict the expression labels, as follows:

$$\phi_i = E_{exp}(I_i), \tag{4}$$

where $\phi_i \in \mathbb{R}^8$ and I_i is a talking head image. E_{exp} uses ResNet-50 as backbone, followed by a fully connected prediction head for expression classification. Additionally, E_{exp} is pretrained on AffectNet (Mollahosseini, Hasani, and Mahoor 2019), a large-scale annotated emotion dataset.

Expression Consistency Loss. Furthermore, we introduce a new expression consistency loss to ensure that the expression in the driving frame, ϕ_{d^i}, closely aligns with the expression in the output frame, ϕ_{y^i}, as follows:

$$\mathcal{L}_{exp}(\phi_{d^i}, \phi_y) = ||\phi_{d^i} - \phi_{y^i}||_2. \tag{5}$$

Here, $\mathcal{L}_{exp}$ computes a perceptual difference between the driving frame expression ϕ_{d^i} and output expression ϕ_{y^i}. Optimizing $\mathcal{L}_{exp}$ improves facial expression details in output.

Network Learning

Directly training AnyTalk, to achieve talking head generation across multiple domains, is challenging due to the diverse styles present and the imbalance in the number of videos for different styles. Therefore, we propose a two-stage learning approach for our AnyTalk. Unlike ToonTalker (Gong et al. 2023), we employ a unified model across all domains without the need for employing the same architecture on each domain and training additional adapter modules to transfer motion.

Pretraining Stage. At the pretraining stage, we train our AnyTalk using a real human talking-head dataset, where each video contains a single person. For each video, we sample two frames: one as the source image s and the other as the driving image d, respectively. We train the networks M, L, E_a, D and G by minimizing the following loss:

$$\begin{aligned} \mathcal{L}_{total} = \lambda_{exp}\mathcal{L}_{exp} + \underbrace{\lambda_P \mathcal{L}_P + \lambda_G \mathcal{L}_G}_{\text{Perceptual and GAN loss}} + \\ \underbrace{\lambda_E \mathcal{L}_E + \lambda_{dist}\mathcal{L}_{dist}}_{\text{Equivalence and keypoint dist loss}} + \underbrace{\lambda_M \mathcal{L}_M + \lambda_\Delta \mathcal{L}_\Delta}_{\text{Relative motion loss}}, \end{aligned} \tag{6}$$

where the λ_P, λ_G, λ_E, λ_{dist}, λ_M, λ_Δ and λ_{exp} are hyperparameters to balance these losses.

Perceptual loss and GAN loss. Similar to Face-vid2vid (Wang, Mallya, and Liu 2021a), we leverage the perceptual loss $\mathcal{L}_P$ and GAN loss $\mathcal{L}_G$ to minimize the gap between the model output and the driving frame.

Equivalence loss and keypoint dist loss. The equivalence loss $\mathcal{L}_E$ and the keypoint dist loss $\mathcal{L}_{dist}$ help AnyTalk to learn more stable keypoints in an unsupervised way.

Relative motion loss. Additionally, the relative motion loss includes a head pose loss $\mathcal{L}_H$ and a deformation priors loss $\mathcal{L}_\Delta$, which constrain the estimated head pose and expression deformation, respectively. More details refer to Appendix.

Fine-tuning Stage. At the pre-training stage, we train AnyTalk to perform reconstruction tasks using real person videos. This process allows our AnyTalk to learn general knowledge about facial motion and expression details from real human data. However, directly transferring facial motion across multiple domains is challenging for AnyTalk due to domain shifts. Thus, we further fine-tune the pre-trained model to achieve talking head generations across multiple domains. To mitigate the risks of catastrophic forgetting and data imbalance across domains, we randomly select 300 talking head videos from the real domain and incorporate them into our proposed cross-domain dataset,AniTalk, for fine-tuning stage. This addition of live talk videos contributes to preserving the performance of AnyTalk in the real human domain. The losses used in the fine-tuning stage are the same as in the pre-training stage.

AniTalk Dataset

Diverse and high quality talking head videos from different domains are essential to train a one-shot face re-enactment framework across multiple domains. However, most existing open-source reenactment datasets only contain real human talking head videos. Even ToonTalker (Gong et al. 2023) only collects cartoon videos with Disney style without making them open-source.

To address these challenges, we introduce a novel talking head videos dataset, called AniTalk. This dataset comprises 1,250 talking head videos with multiple styles and multiple species, present in high-definition MP4 format. Specifically, we collect high-resolution (1080p) videos with Disney human style, Disney animals style, and mesh style from YouTube or other websites. Due to the complex scene transitions present in the original videos, which pose challenges for models in talking head generations, we perform a manual screening process to split all original videos into single-subject talking videos. Finally, we obtain a total of 1,250 videos, including 600 videos in Disney human style, 325 videos in Disney animals style, and 325 videos in mesh style (i.e. real humans).

Given AniTalk, AnyTalk can generate talking head videos across styles, even across species. Meanwhile, we extend the video-driving method AnyTalk to encompass text/audio-driving tasks, utilizing mesh data as an intermediary. For example, we first generate mesh videos through off-the-shelf text/audio-to-mesh methods (Xing et al. 2023), followed by face reenactment using AnyTalk.

Experiments

Experiment Setup

Dataset. We first pretrain our AnyTalk on VoxCeleb1 (Nagrani, Chung, and Zisserman 2017), a popular talking head generation dataset. Then, we further finetune our AnyTalk on AniTalk. For more details, refer to the Appendix.

Evaluation Metrics. We utilize the Frechet Inception Distance (**FID**) to measure the realism of our generated outputs. To assess the identity preservation, we follow the previous works (Gong et al. 2023; Hong et al. 2022) and utilize the cosine similarity (**CSIM**) between synthetic and source images through ArcFace (Deng et al. 2019). Meanwhile, we use the cosine similarity of expression embedding (**CEIM**) to quantify the subtle yet significant facial expression between the driving and generated images. For more details and results, refer to the Appendix.

Cross-Domain Face Reenactment

We compare our AnyTalk under cross-domain face reenactment setting with several state-of-the-art face reenactment methods: FOMM (Siarohin et al. 2019), DaGAN (Hong et al. 2022), ToonTalker (Gong et al. 2023), Face-vid2vid (Wang, Mallya, and Liu 2021a). It is worth noting that FOMM, DaGAN, and Face-vid2vid are trained on a single domain (*i.e.*'Real') in their original papers. For fair comparisons, we train them with the proposed training pipeline on the proposed AniTalk dataset. Specifically, we conduct four transfer tasks for evaluation: *'Disney Human → Real', 'Real → Disney Human', 'Disney Animals → Real', and 'Real → Disney Animals'*. Here, the notation *D* → *S* denotes the task of using a video from domain *D* to drive a source image from domain *S*, where *D* and *S* can be either 'Real', 'Disney Human', or 'Disney Animals'. For more details, refer to the Appendix.

Quantitative Evaluation. The quantitative results of four cross-domain reenactment tasks are reported in Tab. 2. Our AnyTalk outperforms all baselines in terms of FID across the four tasks, indicating that our synthetic results are most consistent with the source distribution. Additionally, our AnyTalk also performs the best in identity preservation and expression consistency, *i.e.*, the highest CSIM and CEIM.

Qualitative Evaluation. The qualitative results of cross-domain face reenactment are in Fig. 4 and Fig. 5. Compared to baselines, AnyTalk achieves superior image quality in terms of image sharpness, reduced distortion. Furthermore, our model outperforms others in motion consistency and facial expression preservation (row 2,3 in Fig. 5). AnyTalk generates videos with higher naturalness and motion consistency under challenging scenarios, *e.g.*, large poses (row 4 in Fig. 4 and Fig. 5).

Ablation Study

To verify the effectiveness of our proposed expression consistency loss $\mathcal{L}_{exp}$, we perform an ablation study by removing $\mathcal{L}_{exp}$ in our method. As shown in Tab. 3, $\mathcal{L}_{exp}$ largely improves the performance of our AnyTalk. In addition, we also apply the expression loss to a different method, *i.e.*, FOMM, which brings a large performance gain. Specifically,

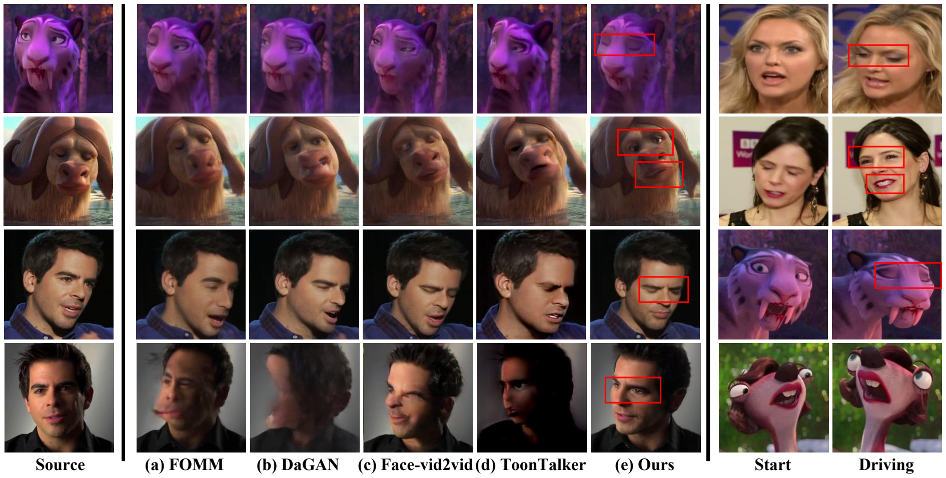

Figure 4: Qualitative comparisons with state-of-the-art methods on *'Real → Disney Animals'* (row 1,2) and *'Disney Animals → Real'* (row 3,4) tasks. Our AnyTalk outperforms existing approaches in terms of image sharpness, distortion, and artifacts, even in challenging cross-species scenarios.

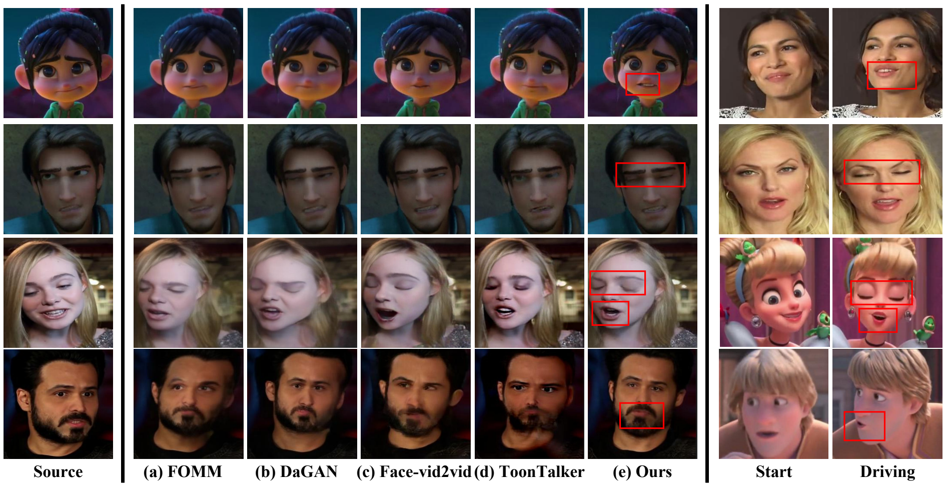

Figure 5: Qualitative comparisons with state-of-the-art methods on the *'Real → Disney Human'* (row 1,2) and *'Disney Human → Real'* (row 3,4) tasks. AnyTalk achieves superior naturalness, visual quality, motion consistency and facial expression details.

	Disney Human→ Real			Real → Disney Human			Disney Animals → Real			Real → Disney Animals		
	FID↓	CSIM↑	CEIM↑	FID↓	CSIM↑	CEIM↑	FID↓	CSIM↑	CEIM↑	FID↓	CSIM↑	CEIM↑
FOMM	17.666	0.816	0.8881	24.547	0.807	0.8813	24.258	0.672	0.9048	47.595	0.736	0.9035
DaGAN	17.541	0.804	0.8861	23.665	0.821	0.8842	26.404	0.675	0.8998	45.531	0.768	0.8924
ToonTalker	17.598	0.824	0.8910	28.901	0.803	0.8872	16.233	0.835	0.9047	55.577	0.748	0.9103
Face-vid2vid	15.224	0.827	0.8923	24.921	0.825	0.8899	15.977	0.804	0.9036	50.038	0.756	0.9131
AnyTalk	**13.039**	**0.848**	**0.8927**	**24.144**	**0.830**	**0.8936**	**13.312**	**0.848**	**0.9065**	**44.481**	**0.770**	**0.9139**

Table 2: Quantitative comparisons on cross-domain reenactment. These methods are trained under the same setup on the proposed AniTalk dataset for fair comparisons.

the methods with $\mathcal{L}_{exp}$ yield better performance in terms of CEIM metric. These results indicate that the proposed expression consistency loss is effective in expression consistency, and can be easily generalized to different methods. Note that all methods are trained using the proposed training pipeline on the AniTalk dataset.

	Disney Human→ Real			Real → Disney Human		
	FID↓	CSIM↑	CEIM↑	FID↓	CSIM↑	CEIM↑
FOMM	17.666	0.816	0.8881	24.547	0.807	0.8813
FOMM w/ $\mathcal{L}_{exp}$	**17.447**	**0.819**	**0.8924**	**20.530**	**0.822**	**0.8913**
Ours w/o $\mathcal{L}_{exp}$	15.224	0.827	0.8923	24.921	0.825	0.8899
AnyTalk	**13.039**	**0.848**	**0.8927**	**24.144**	**0.830**	**0.8936**

Table 3: Ablation study on the proposed expression consistency loss $\mathcal{L}_{exp}$. These methods are trained under the same setup on the proposed AniTalk dataset.

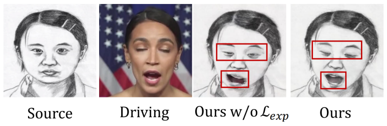

Figure 6: Ablation study about L_{exp}.

User Study

We conduct a user study to further evaluate the performance of all the methods. We invite twenty-five participants and let them answer fifteen single-choice questions. In each question, a rater chooses the best from 5 synthetic cartoon videos generated by four competing methods and our method, based on video sharpness, motion consistency and identity preservation, respectively. Our method is the most favorable with a selection rate of 59.20%. In contrast, the selection rates for Face-vid2vid (Wang, Mallya, and Liu 2021a), Toontalker (Gong et al. 2023), DaGAN (Hong et al. 2022) and FOMM (Siarohin et al. 2019) are 10.67%, 18.13%, 8.80%, and 3.20%, respectively.

Generalization and Application

Generalization on unseen domain. In this section, we compare AnyTalk with several state-of-the-art face reenactment methods on unseen domains. Specifically, we use real human videos from VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) to drive the sketch-style images from CUFS (Zhang, Wang, and Tang 2011). We report the FID, CSIM, and CEIM scores in Tab. 4. Additionally, we show some samples from various unseen domains animated by AnyTalk in Fig. 7 to demonstrate its generalization.

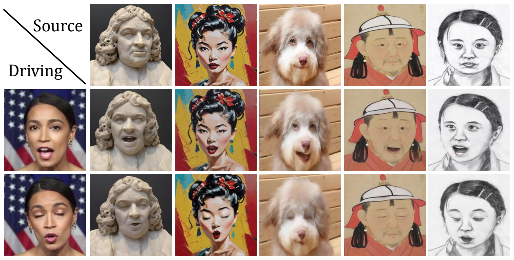

Figure 7: AnyTalk results on various unseen domains.

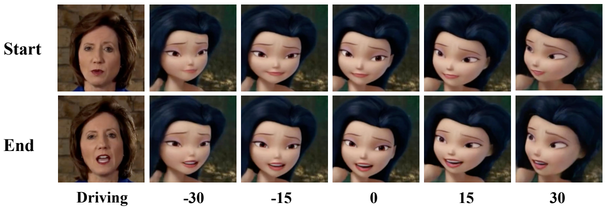

Figure 8: AnyTalk animates cartoon characters with various head poses based on different yaw angles.

	FOMM	DaGAN	ToonTalker	Face-vid2vid	**Ours**
FID↓	41.289	32.142	40.030	62.202	**28.446**
CSIM↑	0.862	0.870	0.843	0.821	**0.871**
CEIM ↑	0.3144	0.3719	0.3954	0.4206	**0.4486**

Table 4: Face reenactment results on unseen-domain.

	NIQE↓	LSE-D↓
SadTalker	8.444	13.661
Ours	**6.293**	**13.493**

Table 5: Audio-driven comparison with SOTA method.

Application on audio-based animation. In this section, we introduce the superiority of our AnyTalk in cross-domain audio-driven talking head generation. Considering that few audio-driven frameworks drive cartoon images with audio without keypoint priors, we compare our AnyTalk with a strong audio-driven baseline, SadTalker (Zhang et al. 2023). We report the NIQE and LSE-D in Tab. 5, where our AnyTalk outperforms SadTalker across all metrics. These underscore the superiority of Anytalk. More results are provided in the Appendix.

Cross-domain face reenactment with free view. In this section, we show that the superiority of our AnyTalk in local free-view control of the output video. We present the visualization examples in Fig. 8. As we can see, our model allows changing the viewpoint of the talking-head during synthesis, without the need for a 3D graphics model. But current state-of-the-art cross-domain face reenactment method, i.e., ToonTalker (Gong et al. 2023) fails to do this.

Real-time performance and computational efficiency. We evaluate the computational efficiency of Anytalk and report frame per second (fps). Anytalk runs at **42** fps using naive PyTorch implementation, being much faster than ToonTalker (17 fps) and suitable for realtime applications.

Conclusion

In this paper, we present a unified framework for cross-domain talking head generation. Without the need for paired data and introducing additional modules, AnyTalk employs a novel canonical 3D keypoint detector to address the domain shift issue, while improves facial detail accuracy through a new expression loss. Additionally, we contribute a high-quality, large-scale dataset for this task. Experimental results demonstrate the superiority, generalization, and flexibility of our framework.

References

Bansal, A.; Ma, S.; Ramanan, D.; and Sheikh, Y. 2018. Recycle-GAN: Unsupervised Video Retargeting. In *ECCV*.

Chen, J.; Liu, Y.; Wang, J.; Zeng, A.; Li, Y.; and Chen, Q. 2024. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *CVPR*.

Chu, X.; Li, Y.; Zeng, A.; Yang, T.; Lin, L.; Liu, Y.; and Harada, T. 2024. GPAvatar: Generalizable and Precise Head Avatar from Image(s). In *ICLR*.

Danecek, R.; Black, M. J.; and Bolkart, T. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *CVPR*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

Gong, Y.; Pang, Y.; Cun, X.; Xia, M.; He, Y.; Chen, H.; Wang, L.; Zhang, Y.; Wang, X.; Shan, Y.; et al. 2023. Interactive Story Visualization with Multiple Characters. In *SIGGRAPH Asia*.

Guo, X.; Li, S.; Yu, J.; Zhang, J.; Ma, J.; Ma, L.; Liu, W.; and Ling, H. 2019. PFLD: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.; Bao, H.; and Zhang, J. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *ICCV*.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*.

He, T.; Guo, J.; Yu, R.; Wang, Y.; Zhu, J.; An, K.; Li, L.; Tan, X.; Wang, C.; Hu, H.; Wu, H.; Zhao, S.; and Bian, J. 2024. GAIA: Zero-shot Talking Avatar Generation. arXiv:2311.15230.

Hong, F.-T.; Liu, Y.; Li, Y.; Zhou, C.; Yu, F.; and Xu, D. 2024. DreamHead: Learning Spatial-Temporal Correspondence via Hierarchical Diffusion for Audio-driven Talking Head Synthesis. *arXiv preprint arXiv:2409.10281*.

Hong, F.-T.; and Xu, D. 2023. Implicit Identity Representation Conditioned Memory Compensation Network for Talking Head video Generation. In *ICCV*.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-Aware Generative Adversarial Network for Talking Head Video Generation. In *CVPR*.

Kim, K.; Park, S.; Lee, J.; Chung, S.; Lee, J.; and Choo, J. 2022. AnimeCeleb: Large-Scale Animation CelebHeads Dataset for Head Reenactment. In *ECCV*.

Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; and Bregler, C. 2021. LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video using Pose and Lighting Normalization. In *CVPR*.

Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting. In *ECCV*.

Liu, P.; Wang, R.; Cao, X.; Zhou, Y.; Shah, A.; Oquab, M.; Couprie, C.; and Lim, S.-N. 2021. Self-appearance-aided Differential Evolution for Motion Transfer. *arXiv preprint arXiv:2110.04658*.

Liu, Y.; Lin, L.; Yu, F.; Zhou, C.; and Li, Y. 2023. Moda: Mapping-once audio-driven portrait animation with dual attentions. In *CVPR*.

Lu, L.; Zhang, T.; Liu, Y.; Chu, X.; and Li, Y. 2023. Audio-Driven 3D Facial Animation from In-the-Wild Videos. *arXiv preprint arXiv:2306.11541*.

Lu, Y.; Chai, J.; and Cao, X. 2021. Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation. *arXiv*.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE T AFFECT COMPUT*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *INTERSPEECH*.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *ACM MM*.

Shen, S.; Li, W.; Zhu, Z.; Duan, Y.; Zhou, J.; and Lu, J. 2022. Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis. In *ECCV*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. In *NeurIPS*.

Song, L.; Wu, W.; Fu, C.; Qian, C.; Loy, C. C.; and He, R. 2021. Pareidolia Face Reenactment. In *CVPR*.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural Voice Puppetry: Audio-driven Facial Reenactment. In *ECCV*.

Tran, L.; and Liu, X. 2018. Nonlinear 3d face morphable model. In *CVPR*.

Tripathy, S.; Kannala, J.; and Rahtu, E. 2021. Facegan: Facial attribute controllable reenactment gan. In *WACV*.

Wang, Q.; Zhang, L.; and Li, B. 2021. SAFA: Structure Aware Face Animation. In *3DV*.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. In *IJCAI*.

Wang, S.; Li, L.; Ding, Y.; and Yu, X. 2022. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021a. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *CVPR*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021b. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*.

Wu, X.; Zhang, Q.; Wu, Y.; Wang, H.; Li, S.; Sun, L.; and Li, X. 2021. F^{3}A-GAN: Facial Flow for Face Animation With Generative Adversarial Networks. *TIP*.

Wu, Y.; Xu, S.; Xiang, J.; Wei, F.; Chen, Q.; Yang, J.; and Tong, X. 2023. AniPortraitGAN: Animatable 3D Portrait Generation from 2D Image Collections. In *SIGGRAPH Asia*.

Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *CVPR*.

Xu, B.; Wang, B.; Deng, J.; Tao, J.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Motion and appearance adaptation for cross-domain motion transfer. In *ECCV*.

Yao, G.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Mesh guided one-shot face reenactment using graph convolutional networks. In *ACM MM*.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. *arXiv preprint arXiv:2301.13430*.

Zakharov, E.; Ivakhnenko, A.; Shysheya, A.; and Lempitsky, V. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2022. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *CVPR*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*.

Zhao, J.; and Zhang, H. 2022. Thin-Plate Spline Motion Model for Image Animation. In *CVPR*.

Zhao, R.; Wu, T.; and Guo, G. 2021. Sparse to dense motion transfer for face image animation. In *ICCV*.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In *CVPR*.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM TOG*.

Zhu, X.; Liu, X.; Lei, Z.; and Li, S. Z. 2017. Face alignment in full pose range: A 3d total solution. *TPAMI*.