IJCAI 2025 The 34th International Joint Conference on Artificial Intelligence
29-31, August, 2025, Guangzhou

Paper ID: 2522

SyncAnimation: A Real-Time End-to-End Framework for Audio-Driven Human Pose and Talking Head Animation

Yujian Liu[1,3], Shidang Xu[1], Jing Guo[2,3], Dingbin Wang[3], Zairan Wang[1], Xianfeng Tan[3], Xiaoli Liu[3]

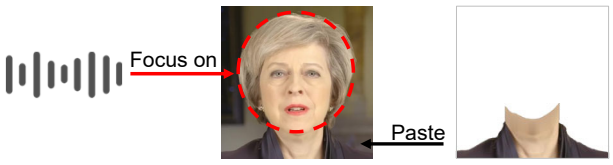
[1]AiShiWeiLai AI Research
[2]Beijing Institute of Technology
[3]South China University of Technology

Motivation

Can audio simultaneously drive and generate both the head and body when trained in a person-specific manner?

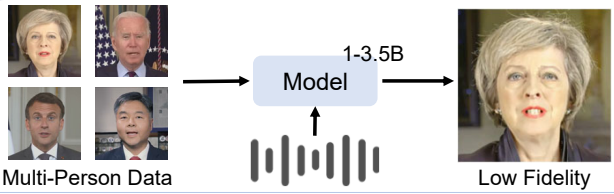
◆ Paste-back Generation: GAN- and Nerf-based Methods

While generating the full face, GAN- and Nerf-based methods often fail to synchronize audio with expressions, head poses, and upper-body movement, resulting in unnatural or desynchronized animations.



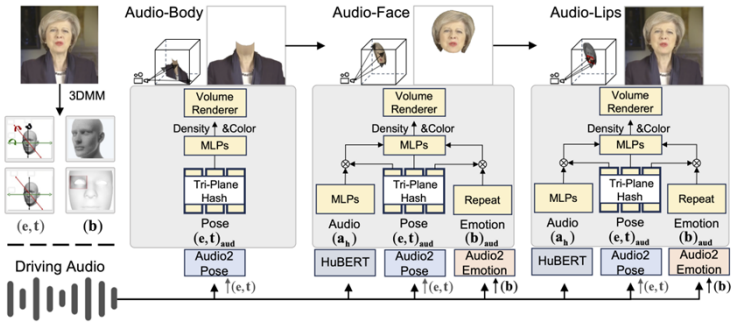
◆ Joint Generation: SD-based Method

Although SD-based methods can generate full-body avatars, they suffer from high computational cost, and poor identity consistency, making them unsuitable for real-time applications.



SyncAnimation

◆ Framework:



◆ Audio2pose:

Audio-Motion Ambiguity

Audio and motion lack a strict one-to-one mapping, often resulting in jitter or frame skipping in generated avatars.

Method: Two Conditional Vectors

1. Stability-Guided Conditional Vector S_{pose}:

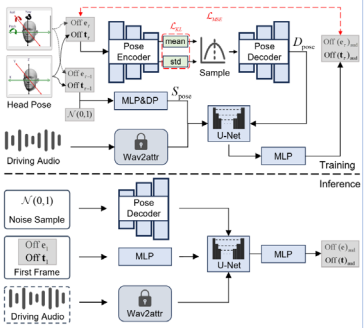
$$\mathbf{S}_{\text{pose}} = f_{\text{MLPs}}([\mathbf{Off}(\mathbf{e}_{\tau-1}), \mathbf{Off}(\mathbf{t}_{\tau-1})] + \mathcal{N}(\mu_S, \delta_S))$$

By leveraging the difference between averaged previous-frame poses and injecting Gaussian noise, the model reduces the one-to-many audio-to-pose mapping to a many-to-one form, enabling smoother and more stable motion.

2. Diversity-Guided Conditional Vector D_{pose} :

$$\mathbf{D}_{\text{pose}} = f_{\text{VAE}}([\text{Diff}(\mathbf{e}), \text{Diff}(\mathbf{t})])$$

A VAE-learned, diversity-guided latent vector that introduces stochasticity to avoid static poses, facilitating faster convergence and better accuracy compared to direct audio regression.



◆ Audio2Emotion:

Incomplete Audio-Facial Alignment

Most audio-driven models focus on lips, neglecting eyebrow and blink motions, leading to unnatural expressions.

Method: Two Conditional Vectors

1. Stability-Guided Conditional Vector S_{exp}:

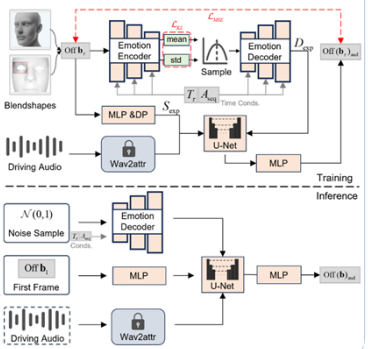
$$\mathbf{S}_{\text{exp}} = f_{\text{MLPs}}(\mathbf{Off}(b_\tau))$$

Enhances audio–non-lip facial region correlation by generating a stable expression reference from the current frame, enabling more lifelike and synchronized expressions.

2. Diversity-Guided Conditional Vector D_{exp} :

$$\mathbf{D}_{\text{exp}} = f_{\text{CVAE}}(\mathbf{Off}(\mathbf{b}) \mid T_\tau, \mathbf{A}_{\text{seq}})$$

A CVAE-based expression representation that incorporates time encoding and sequential audio features to capture temporal dynamics, provide diverse expression templates, and ensure natural blinking and continuous facial motion.

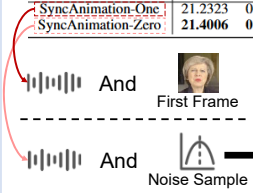


Results

◆ Comparison with Other Methods:

Table 1: Quantitative comparisons with state-of-the-art methods. "SyncAnimation-One" refers to one-shot inference, while "SyncAnimation-Zero" indicates zero-shot inference using noise of the same dimension. We achieve state-of-the-art performance on most metrics.

	Method	Image Quality				Lip Sync			Head Motion	
		PSNR↑	LPIPS↓	SSIM↑	FID↓	LMD↓	AUE↓	SyncScore↑	Diversity↑	EAR↓
GAN	Wav2Lip	18.8071	0.2523	0.6571	40.3604	5.2161	4.5499	**9.2332**	0.0782	0.0535
	DINet	18.6620	0.2547	0.6515	34.0215	5.2784	4.4457	7.3138	0.0560	0.0568
	IP-LAP	18.7763	0.2438	0.6531	34.8322	5.4905	4.7474	3.4331	0.0444	0.0545
	EDTalk	18.7482	0.2896	0.6670	53.7573	5.0925	4.8511	7.3092	0.1046	0.0441
Nerf	ER-NeRF	18.8610	0.2323	0.6799	37.5604	3.7192	4.8899	6.1909	0.0935	0.0439
	SyncTalk	18.8035	0.2287	0.6766	32.9779	3.6106	3.6198	7.0865	0.0822	0.0512
	GeneFace++	19.0344	0.2443	0.6819	40.5045	5.3823	4.0530	7.1118	0.0726	0.0504
SD	Hallo	17.9117	0.2678	0.6305	35.8346	6.0491	4.3952	7.2458	0.2166	0.0471
	V-Express	17.3918	0.2842	0.6121	46.6975	6.9958	5.4740	7.4739	0.1039	0.0490
	EchoMimic	14.4399	0.4524	0.4562	59.3261	8.0643	5.1434	6.0574	0.1864	0.0468
	SyncAnimation-One	21.2323	0.1543	0.7343	**20.0567**	3.2215	**3.5485**	7.1389	0.2570	**0.0357**
	SyncAnimation-Zero	**21.4006**	**0.1532**	**0.7411**	21.7353	**3.1878**	3.5515	7.1499	**0.2652**	0.0386

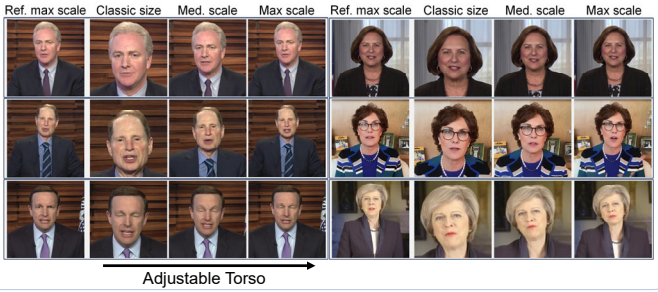


SyncAnimation-Zero, which replaces identity input with noise. Results show that our model can generate digital humans **from audio alone** — no identity input required.

◆ Qualitative Experimental Results:

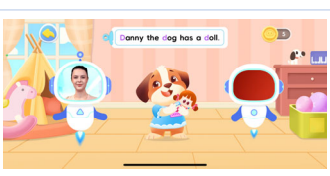
Fixed pose and expression, independent of audio

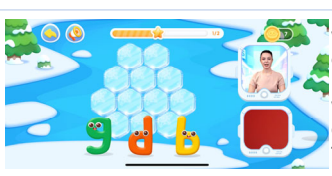
◆ Upper-body Scaling Expansion Results



Real-world Scenario

IN 与爱为舞爱学





We proudly introduce **"爱学" APP**, the first AI-powered educational product featuring the digital human **Teacher Lucy**, designed to support early childhood English learning.

With our continued efforts, we hope she will become a **dedicated, ever-present, and all-knowing companion teacher** for every learner.