VividAnimator: An End-to-End Audio and Pose-driven Half-Body Human Animation Framework

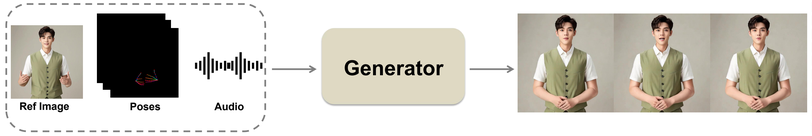
Donglin Huang[1*] Yongyuan Li[2*†] Tianhang Liu[1*] Junming Huang[1] Xiaoda Yang[1] Chi Wang[1] Weiwei Xu[1]
[1] Zhejiang University [2] idr.ai (* equal contribution, † project lead)

WACV 2026 TUCSON, AZ 3/6 - 3/10

Talking Human

Problem

- Existing audio-/pose-driven human animation methods often suffer from **stiff head motion and blurry hands**, due to weak audio–head correlation and the structural complexity of hands.



Goal

- Generate identity-preserving half-body videos with realistic gestures.
- Improve perceptual realism and motion fidelity (hands + head + lip sync).

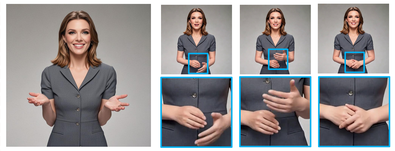
Key Contributions

HCC (Hand Clarity Codebook): a pre-trained hand codebook that provides rich hand texture priors.

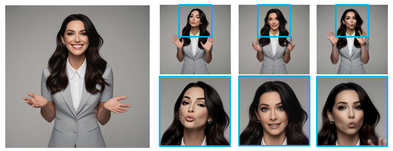
DSAA (Dual-Stream Audio-Aware Module): separately models lip synchronization and head pose dynamics .

PCT (Pose Calibration Trick): refines and aligns pose conditions

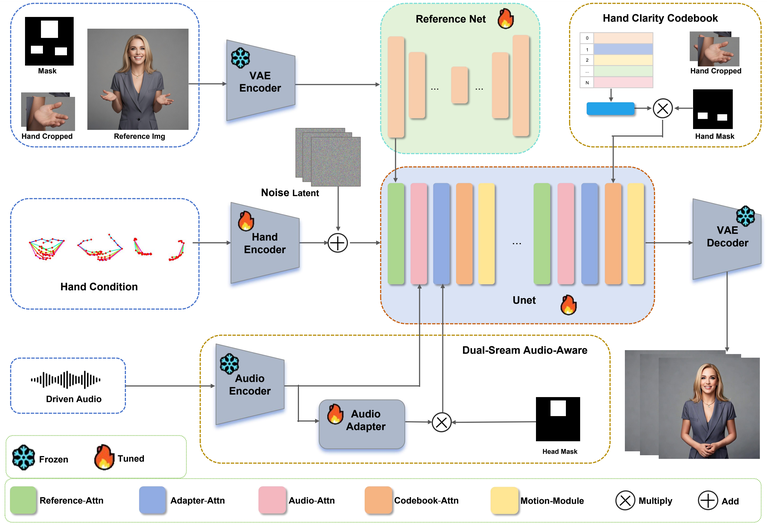
Hand Fidelity



Head Dynamics



VividAnimator



How it works

- **Inputs:** reference image + speech audio + sparse hand poses.
- **ReferenceNet** preserves identity.
- **U-Net** denoises with audio/hand/pose conditions.
- **HCC** and **DSAA** are injected via masked attention for localized control.

Dataset&Training

Two-stage training:

- single-frame training for reconstruction fidelity;
- multi-frame training for temporal consistency.

Dataset:

- 200 hours high-resolution videos with multi-stage filtering for quality.

Results

Quantitative (Table 1):

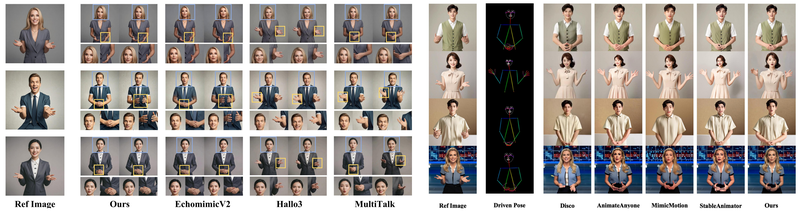
- **best** FID/FVD
- **highest** HKC & HyperIQA

Table 1. Quantitative comparison with existing half-body animation methods.

Methods	SSIM↑	PSNR↑	CSIM↑	FID↓	FVD↓	HKC↑	HyperIQA↑	Sync-C↑	Sync-D↓
Disco	0.616	16.93	0.912	152.88	2311.18	0.784	57.60	-	-
AnimateAnyone	0.671	20.29	0.968	61.45	563.96	0.868	64.39	3.136	11.592
MimicMotion	0.689	20.01	0.961	91.39	855.11	0.910	59.60	5.047	9.843
StableAnimator	**0.733**	**21.29**	**0.974**	60.92	334.12	0.896	61.23	5.184	9.433
Hallo3	0.679	18.64	0.933	106.01	642.54	0.861	55.82	2.687	12.226
MultiTalk	0.697	18.36	0.940	79.96	461.98	0.881	51.96	2.319	13.534
EchomimicV2	0.713	20.60	0.966	63.90	381.72	0.924	69.81	6.221	8.719
VividAnimator	0.711	<u>21.05</u>	<u>0.970</u>	**54.43**	**333.45**	**0.942**	**71.04**	**6.241**	**8.446**

Qualitative:

- clearer hands and more expressive head motion



Conclusion

VividAnimator: end-to-end co-speech half-body animation from image + audio + sparse hand poses.

HCC + DSAA + PCT jointly improve hand detail, lip/head dynamics, and pose consistency.

Delivers **state-of-the-art** perceptual realism and motion fidelity.

Contact
Donglin Huang
Zhejiang University
Email:hdl070607@gmail.com
Phone:+8618150582725