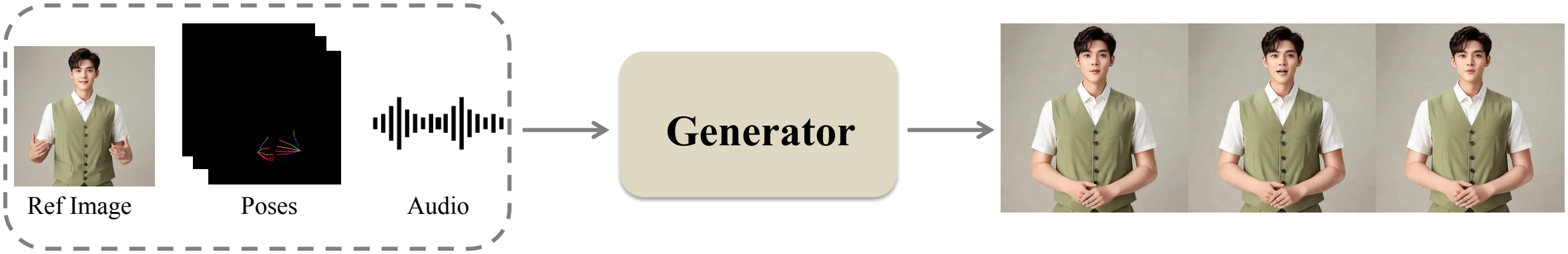
VividAnimator: An End-to-End Audio and Pose-driven Half-Body Human Animation Framework (WACV 2026)

Donglin Huang, Yongyuan Li, Tianhang Liu, Junming Huang, Xiaoda Yang, Chi Wang, Weiwei Xu

Motivation

Input: reference image + sparse pose (both hands) + speech audio

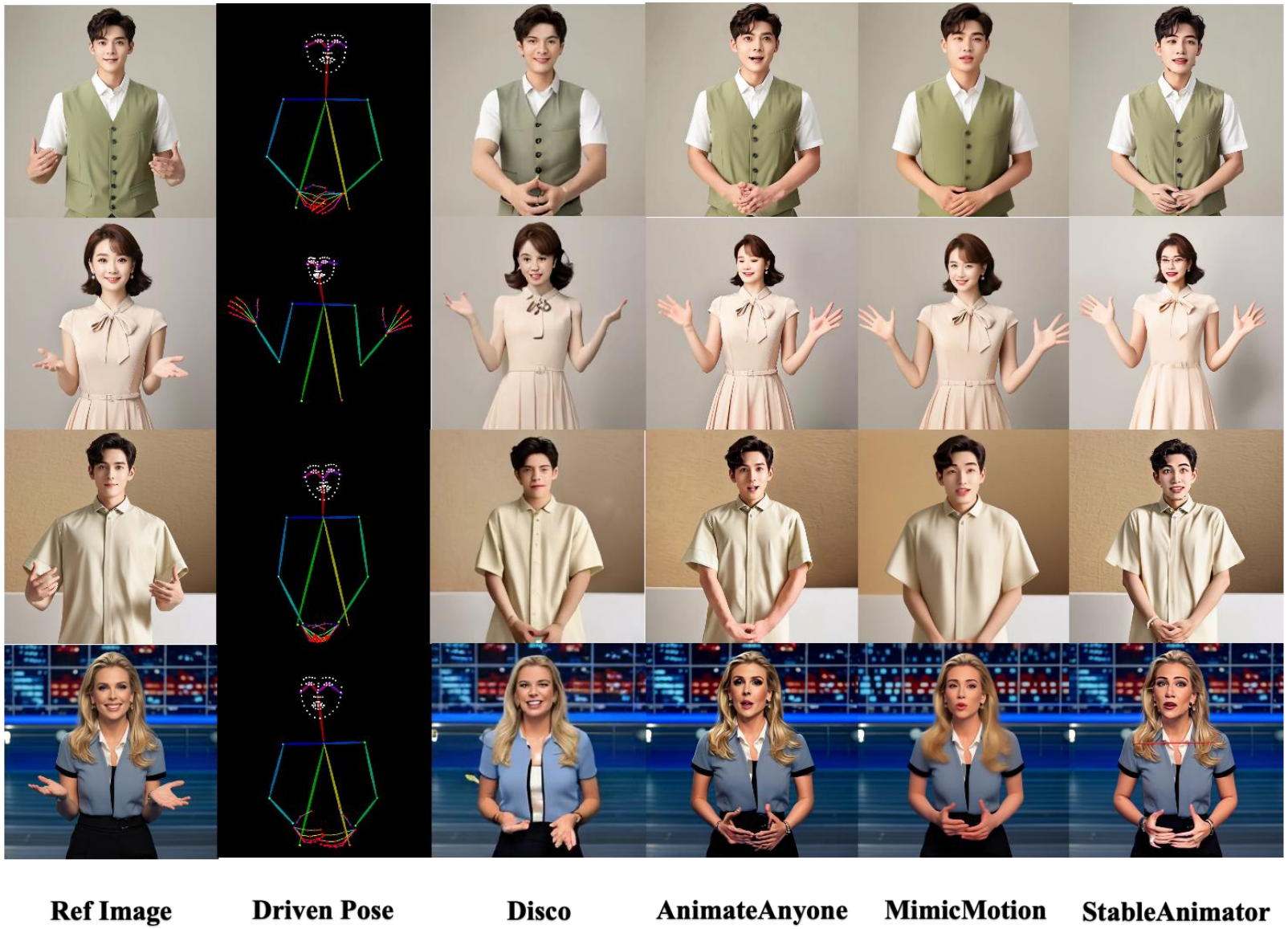
Output: identity-preserving half-body video with natural co-speech motion



Motivation

Pose-driven:

- rigid pose reliance
- blurry hands



Motivation

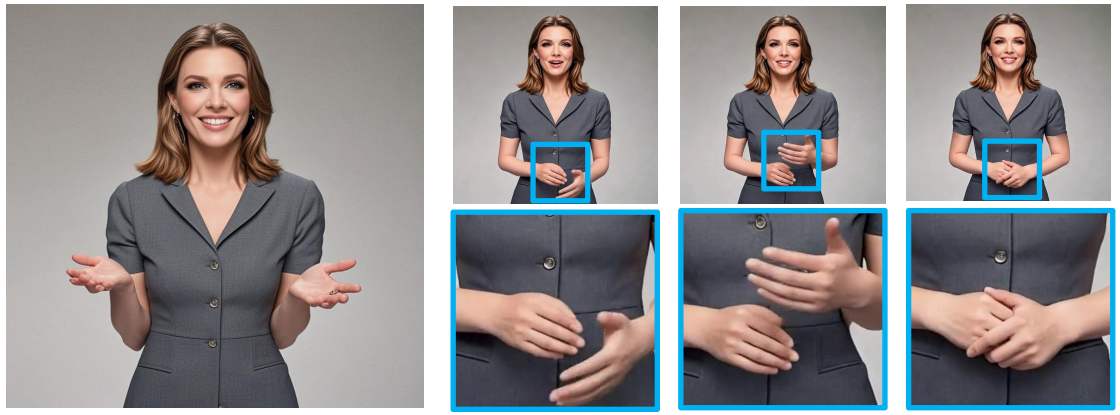
Audio-driven:

- head dynamics degrade
- uncontrollable hands

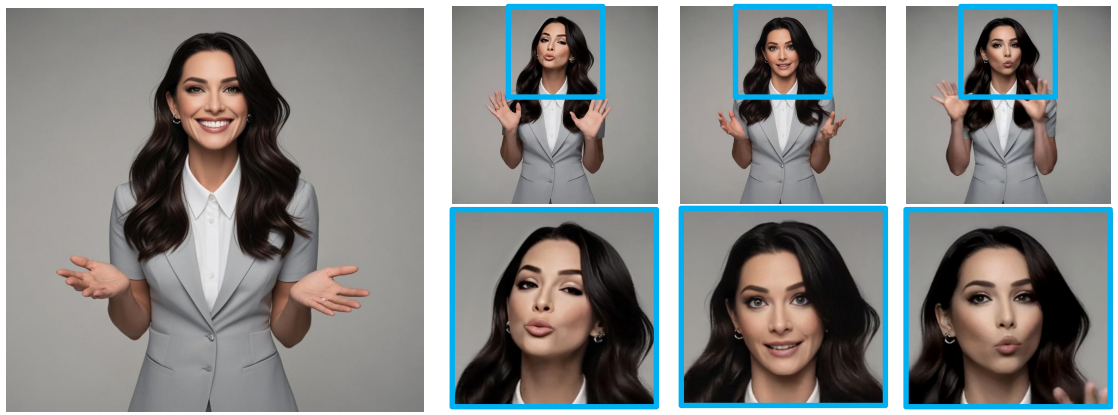
Contributions

- **HCC**: offline pretrained Hand Clarity Codebook injects rich hand texture priors
- **DSAA**: Dual-Stream Audio-Aware module separates lip-sync & head dynamics
- **PCT**: Pose Calibration Trick aligns driving poses to reference

Hand Fidelity



Head Dynamics



Method

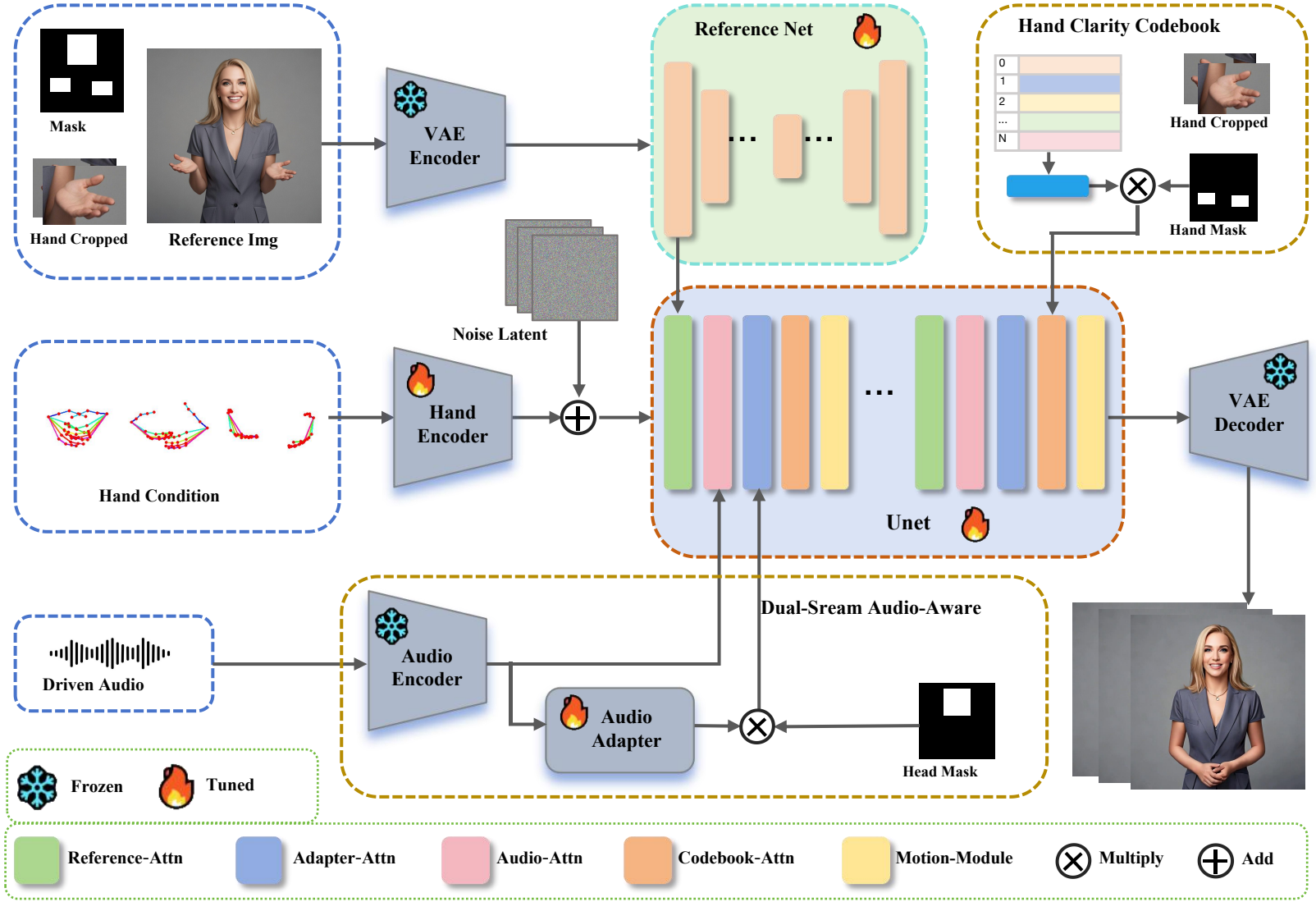
Core:

diffusion U-Net (latent diffusion)

ReferenceNet

Dual-Stream Audio-Aware

Hand Clarity Codebook

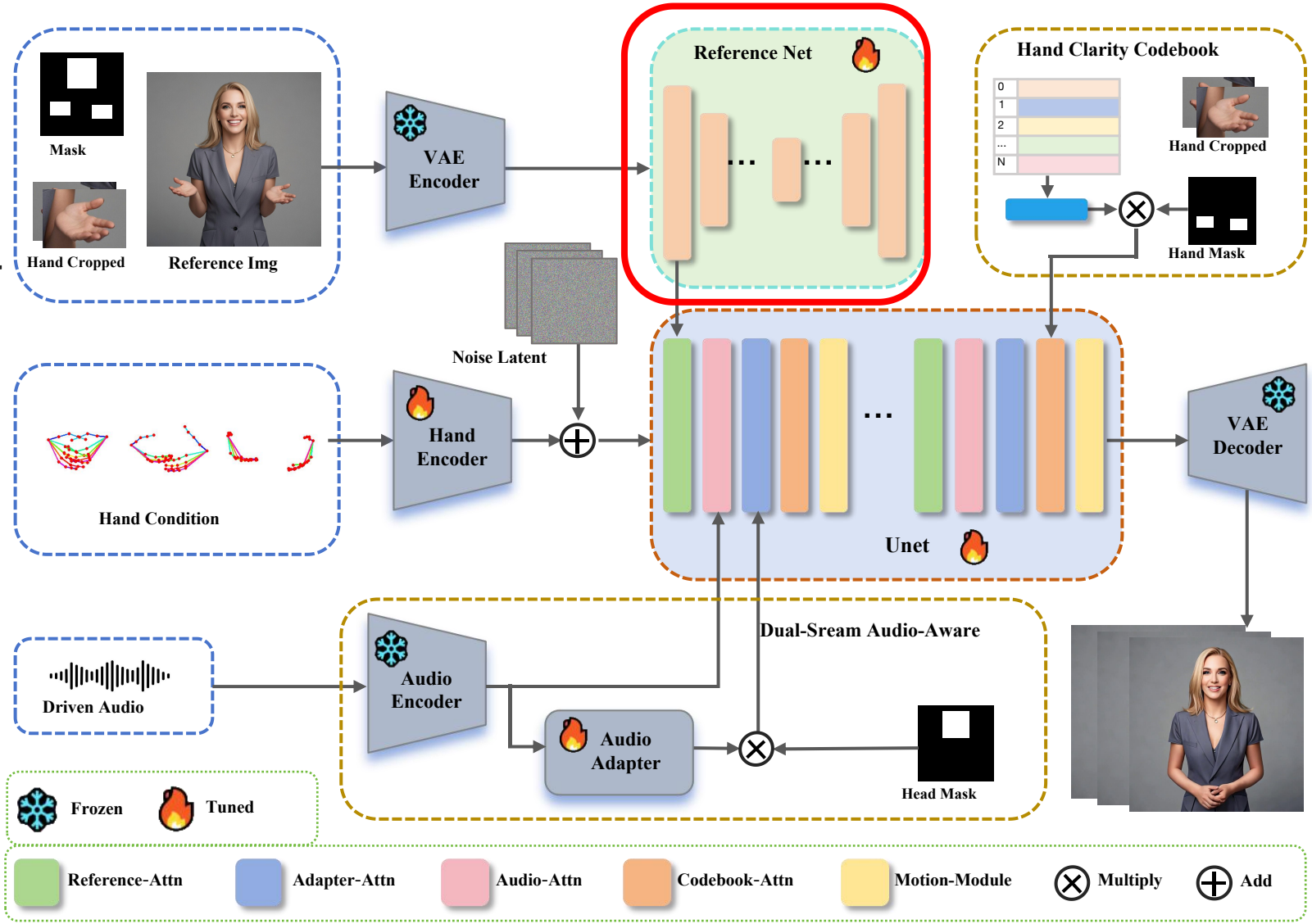


Method

ReferenceNet

injects identity features (via attn)

identity features from VAE Encoder



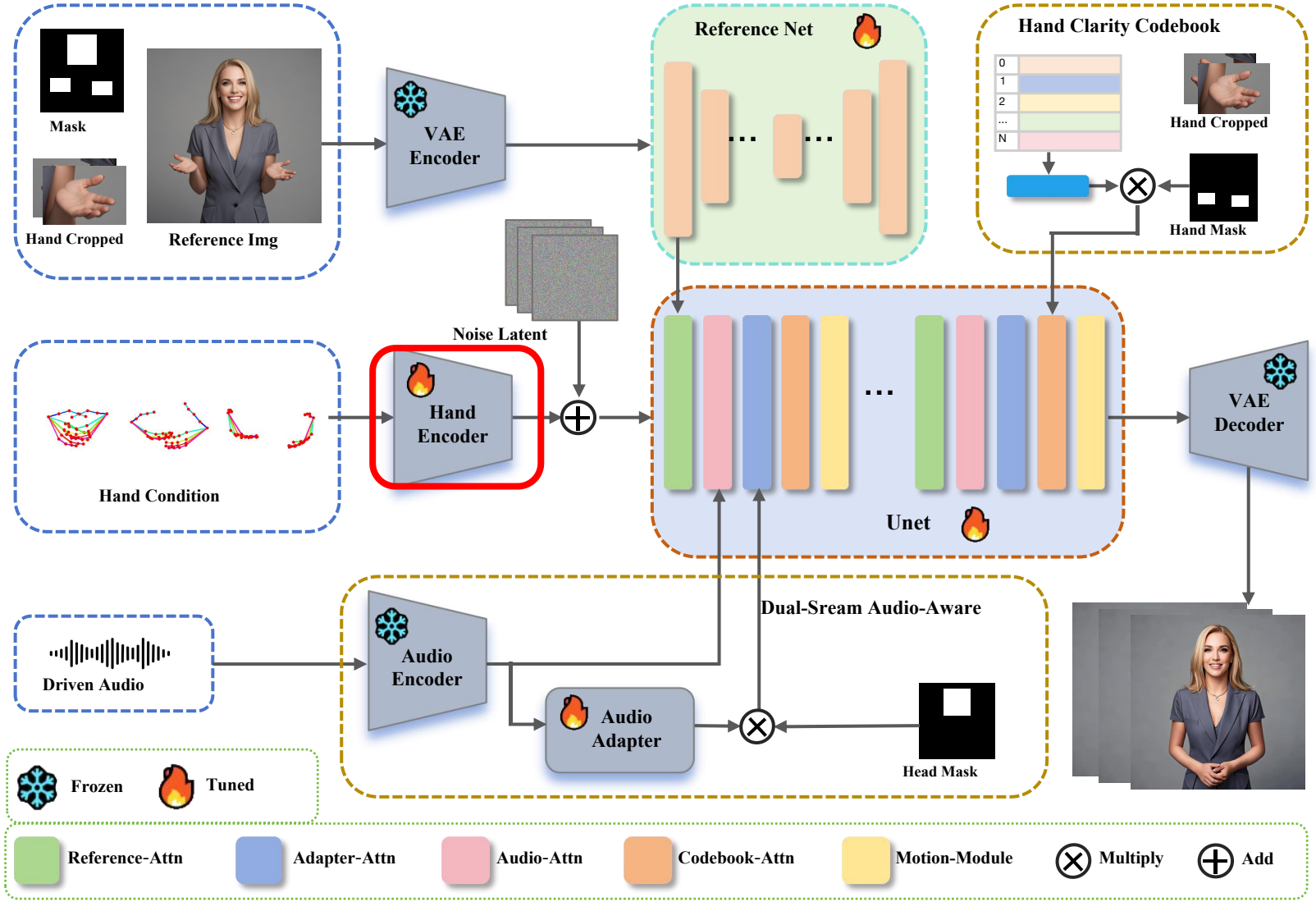
Method

Hand Encoder

lightweight conv encoder

injects sparse hand poses

hand poses from DWPose

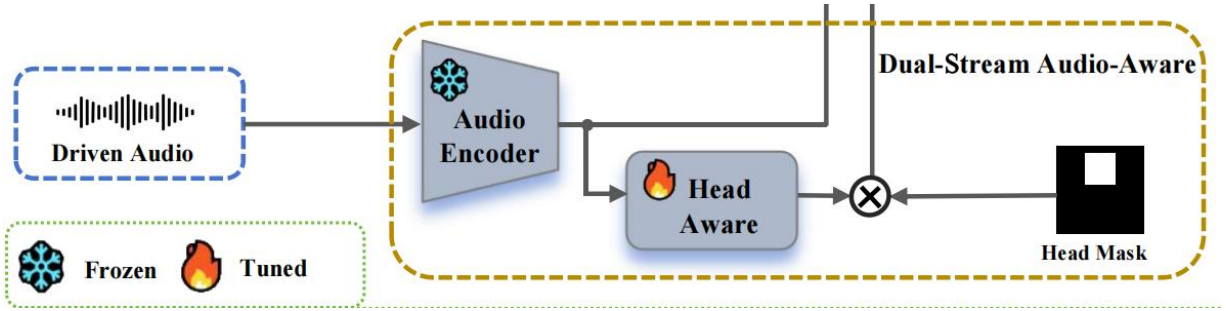


DSAA: Dual-Stream Audio-Aware Module

Problem:

audio→lip strong;

audio→head weak → stiff head in prior work



Solution: two complementary pathways:

Head-aware stream: global rhythm → vivid head dynamics

Local lip-sync stream: fine-grained phoneme cues → accurate lip motion

Streams interact but keep specialization

DSAA Head-Aware Stream (Global Rhythm)

audio features:

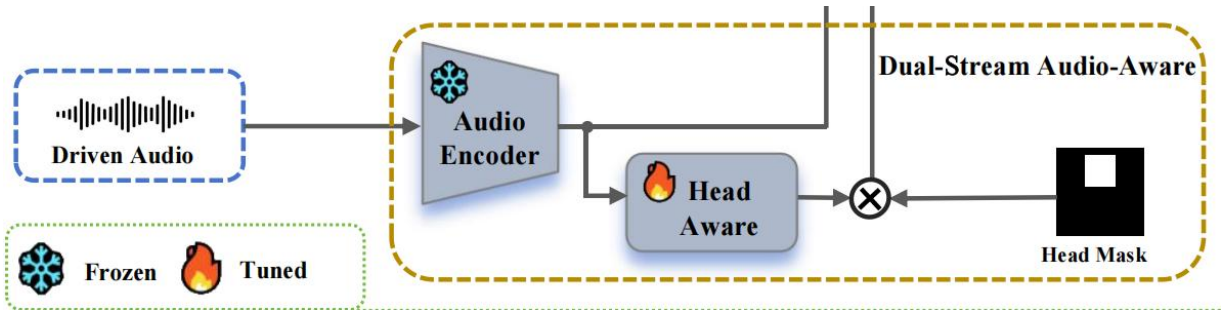
$$W_a \in R^{B\times F\times T\times D}, F = 24, T = 50, D = 384$$

Average pooling along window dimension:

$$F_r \in R^{B\times T\times D}$$

Audio Adapter projects F_r then:

injected via cross-attn localized to head region using M_{head}

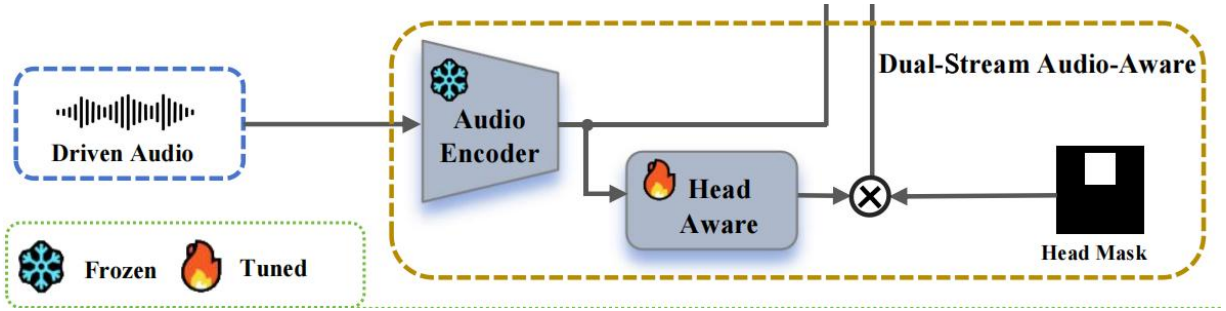


DSAA Local Lip-Sync Stream (Fine Alignment)

Condition each U-Net block on audio tokens

$$F_{lip} \in R^{B \times T \times D}$$

Audio-attention modulates hidden states with phoneme-level cues



Complements head stream:

- head stream sets rhythm;
- lip stream enforces precise sync

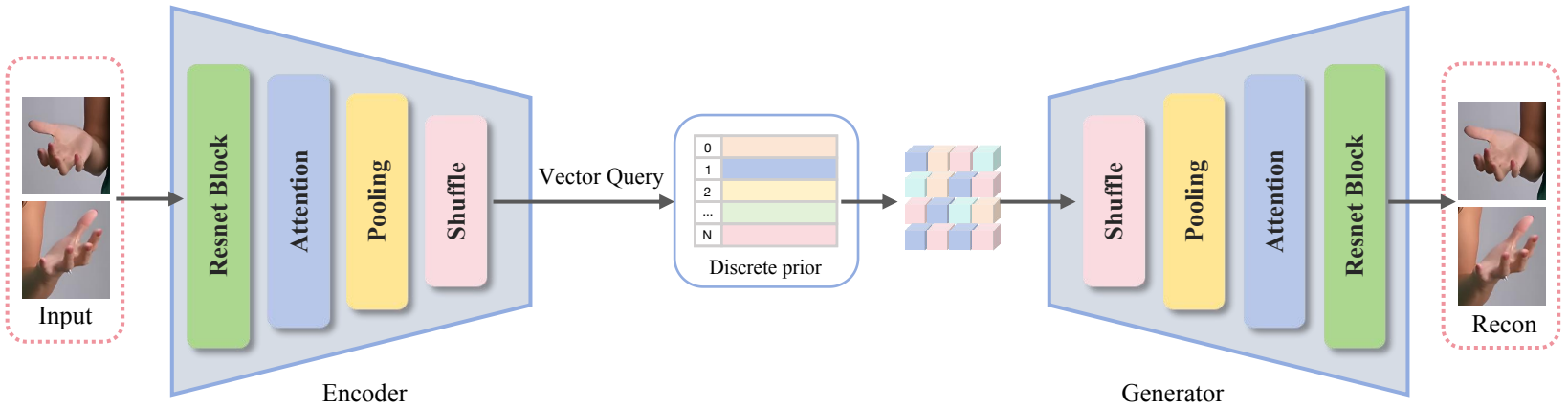
HCC: Hand Clarity Codebook

Problem:

- Hands are small
- high-frequency texture
- self-occlusion

existing solution:

- entangled gradients
- unstable convergence
- higher compute



Our solution:

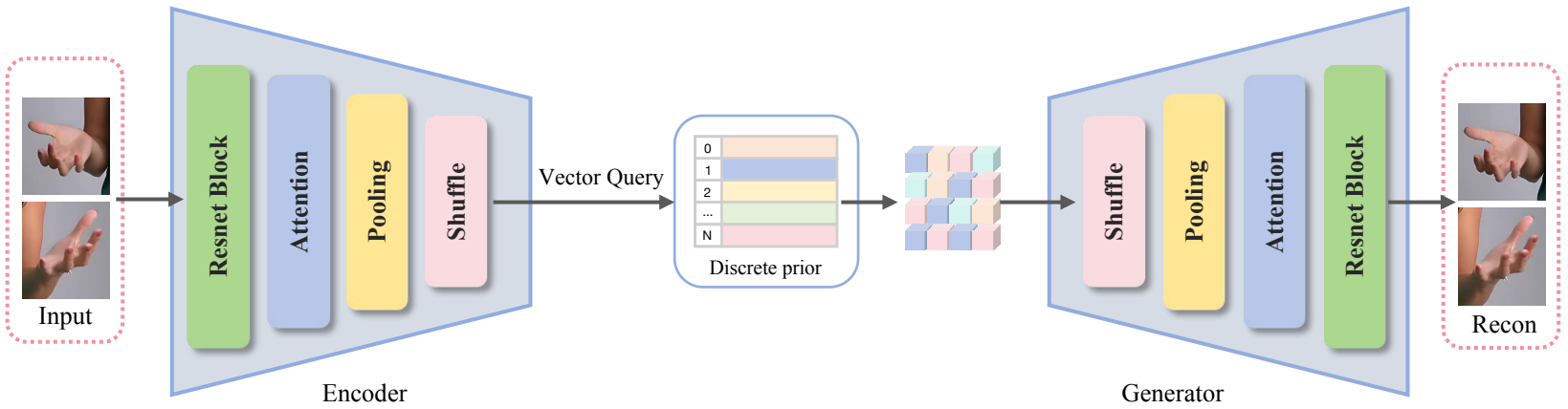
offline pre-train a VQ-VAE hand codebook → stable discrete priors

HCC: Hand Clarity Codebook

- codebook size: 1024
- embed dim: 256
- latent grids: 16×16
- traning dataset:

  260k hand images

  resolutin: 256x256



Method

HCC: Hand Clarity Codebook

- codebook size: 1024
- embed dim: 256
- latent grids: 16×16
- traning dataset:

  260k hand images

  resolutin: 256x256

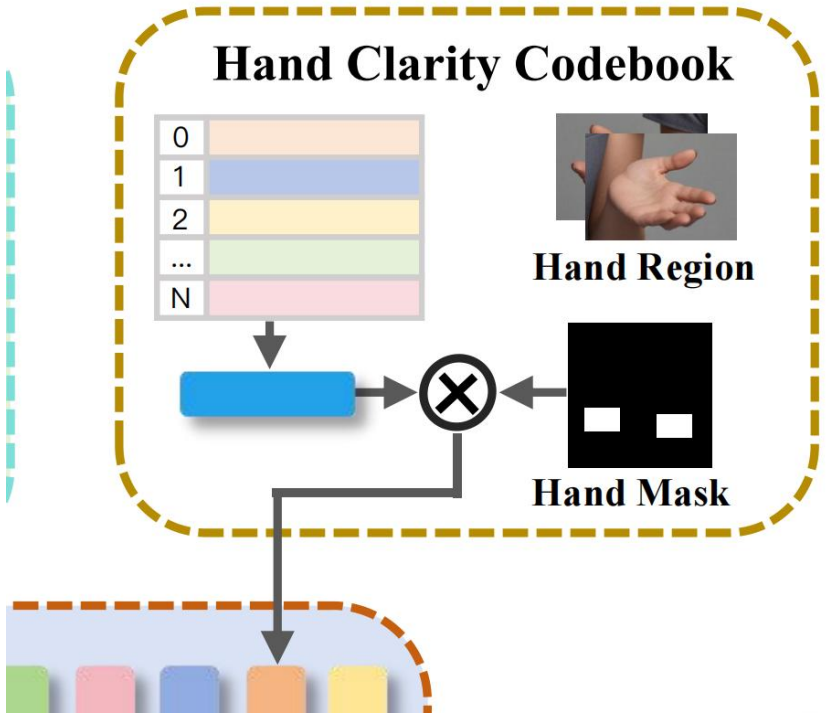
HCC: Hand Clarity Codebook

During denoising:

- crop left&right hand regions from reference image
- Encode+quantize → concatenate → F_h
- injected via Codebook Attention using M_{hand}

Effect:

- sharper contours
- richer hand texture
- better HKC/HKV



Method

PCT: Pose Calibration Trick

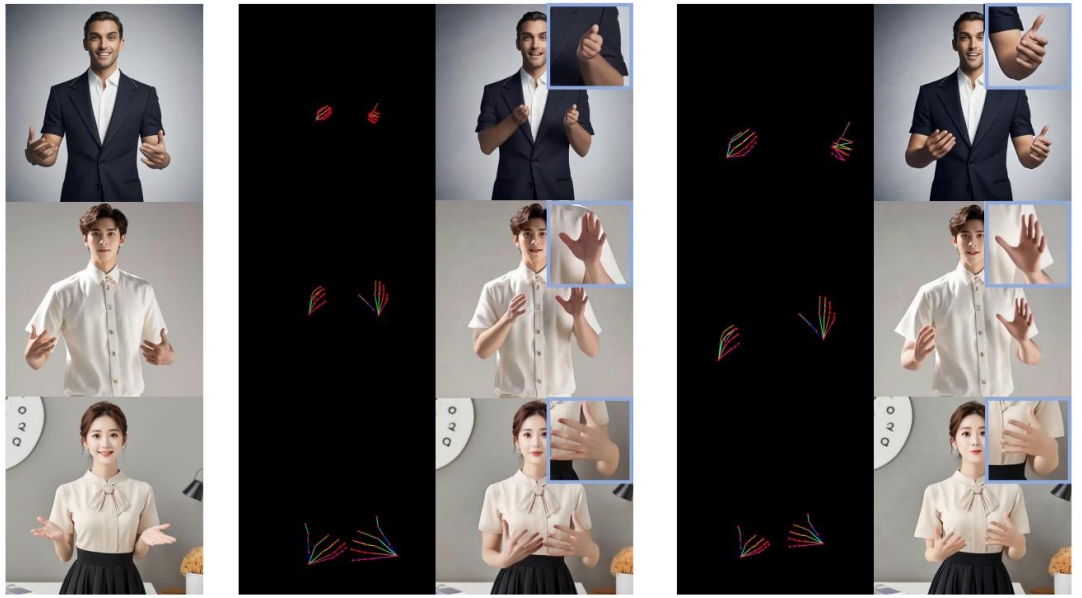
Raw driving keypoints often misaligned:

- scale mismatch
- proportion distortion
- translation offsets

PCT: training-free alignment in 3 steps:

- Global scale normalization via torso anchors (r_x, r_y)
- Segment-wise proportion adjustment $\rho_{ij} = \frac{l_{ij}^{ref}}{l_{ij}}$
- Anchor-based translation using torso center difference

Result: better hand position and fewer artifacts

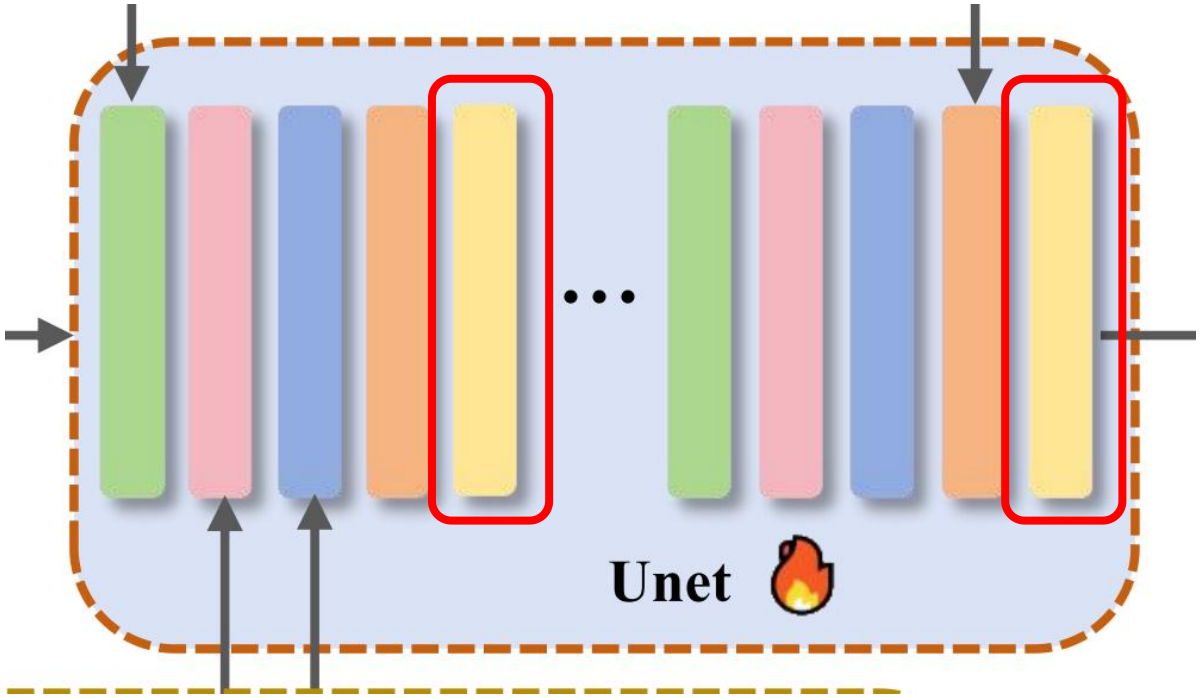


Ref Image | a) w/o PCT | b) with PCT

Method

Temporal Consistency

- Use temporal transformer on latent sequence
- Multi-frame training on continuous 24 frames
- temporal smoothness & cross-frame consistency



Dataset & Filtering Pipeline

Dataset:

about 200 hours high-res videos (40h subtle-motion + 160h large-motion).

Filtering:

5–10s clips → remove subtitles → remove poor A/V sync → remove motion blur → apply super-resolution to low-quality.

Test set:

130 clips (planned public release).

Reference images: 230 synthesized by FLUX (balanced age/gender/skin tone).

Experiments

Training Setup

Gpus:

8×NVIDIA A800

HCC pretrain:

200k iters, Adam lr=1e-4, batch=256, β(commit)=0.25.

Full training:

single-frame bs=8, multi-frame bs=4

AdamW lr=1e-5

audio & image CFG=2.5

multi-frame sequence length=24 frames.

Comparison: Quantitative

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**

Comparison: Qualitative (videos)

Ablation Studies

HCC training scheme:

online vs offline pretraining

HCC latent grid:

8×8 vs 16×16

Remove HCC: w/o HCC

Remove head-aware audio stream:

w/o Head-Aware

Table 2. Ablation study on VividAnimator across different metrics.

Methods	HKC↑	HKV↑	HMV↑	HyperIQA↑
HCC(online 8x8)	0.911	44.657	4.063	63.04
HCC(online 16x16)	0.914	44.435	4.157	63.37
HCC(offline 8x8)	0.920	44.495	4.095	65.24
w/o HCC(offline 16x16)	0.907	43.749	3.636	69.40
w/o Head Aware	0.922	44.641	3.295	69.65
EchomimicV2	0.922	42.221	3.061	69.81
VividAnimator	**0.942**	**46.452**	**4.435**	**71.04**

Limitations

- Ambiguous inputs / rare failure cases: extreme occlusions or noisy poses can cause artifacts.
- Fast hand motions: motion blur and self-occlusion remain challenging.
- Scope: currently optimized for half-body animation (not full-body).

Future work

- Stronger temporal hand modeling for fast and complex gestures.
- Extend to full-body co-speech animation with richer motion priors.

Conclusion

- **VividAnimator** generates end-to-end half-body co-speech animation from reference image + audio + sparse hand poses.
- **HCC** injects discrete high-frequency hand priors → crisp hands.
- **DSAA** decouples global head rhythm and local lip-sync → vivid & synchronized motion.
- **PCT** aligns driving pose to the reference → smoother gestures, fewer artifacts.
- Achieves **state-of-the-art** perceptual & motion results.

THANK YOU!

Contact:
Donglin Huang (Zhejiang University)
hdl070607@gmail.com