ICLR

CyberHost: A One-stage Diffusion Framework for Audio-driven Talking Body Generation

Gaojie Lin*[1], Jianwen Jiang*[1] *(equal contributions)*

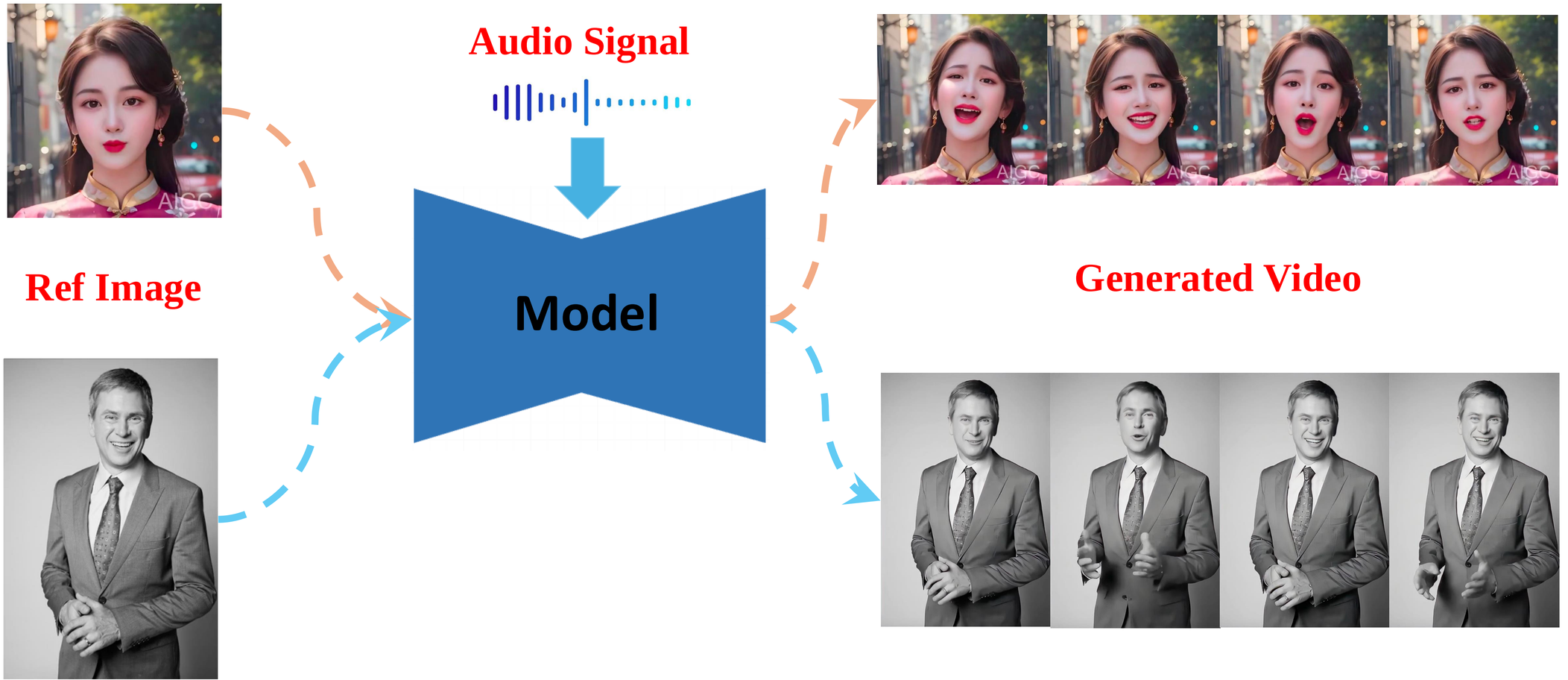
Chao Liang[1], Tianyun Zhong[2], Jiaqi Yang[1], Zerong Zheng[1], Yanbo Zheng[1]

[1]ByteDance [2]ZheJiang University

浙江大學 ZHEJIANG UNIVERSITY

Problem Definition

Audio-driven Human Animation



Background & Motivation

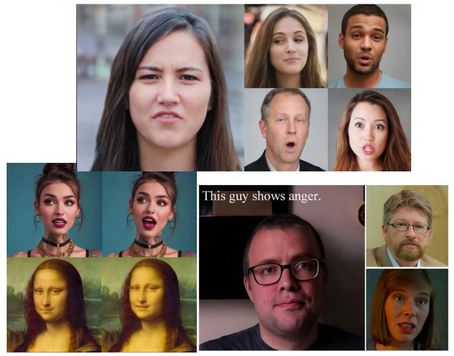
- Most previous work focused on Talking Head Generation.
- **Two-stage** Talking Body Generation methods suffer from:
  - Increased system complexity and reduced learning efficiency.
  - Limited by the capability of intermediate representations.
  - Effectiveness relies on pose/mesh annotation accuracy.

One-Stage Talking Body Generation

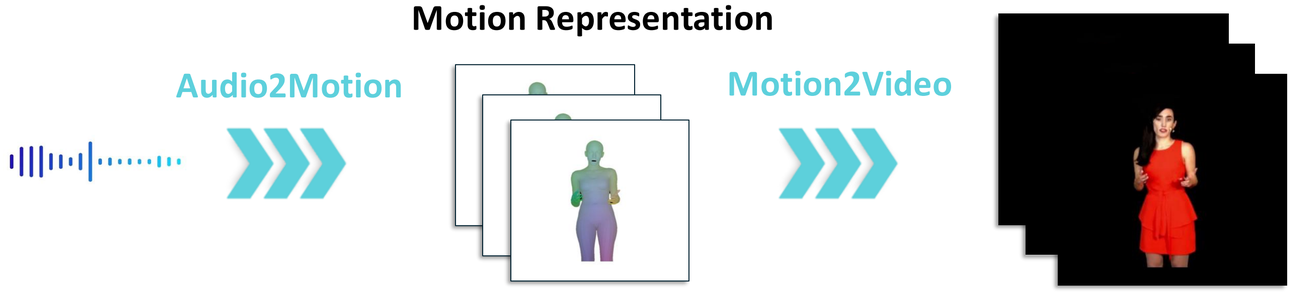
We aim to explore the upper limits of one-stage framework for Talking Body Generation.

Talking Head Generation

- VASA, NIPS 2024
- EMO, ECCV 2024
- Hallo2, ICLR 2025
- EchoMimic, AAAI 2025
- SadTalker, CVPR 2024
- ……



Two-stage Talking Body Generation



Comparsion with Two-Stage Method

Speech2Gesture

MoGlow

SDT

CyberHost *

Speech2Gesture: Learning Individual Styles of Conversational Gestures, CVPR 2019
MoGlow: Style-controllable speech-driven gesture synthesis using normalising flows, TOG 2020
SDT, Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates, ICCV 2021

Comparsion with Two-Stage Method

Reference Image

VLOGGER

CyberHost

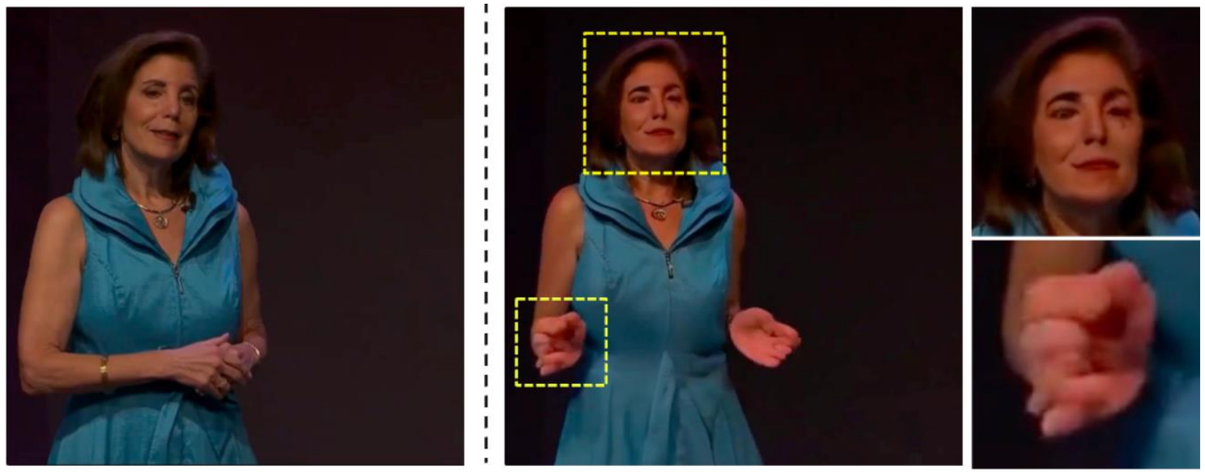
VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis, Arxiv 2024

Zero-shot Testing

Challenge

- **One-stage** Talking Body Generation faces two challanges:
  - **Details Underfitting**: Local structural priors missing & Small critical region coverage.
  - **Motion Uncertainty**: Higher motion freedom & Weak audio-limb correlation.

Details Underfitting

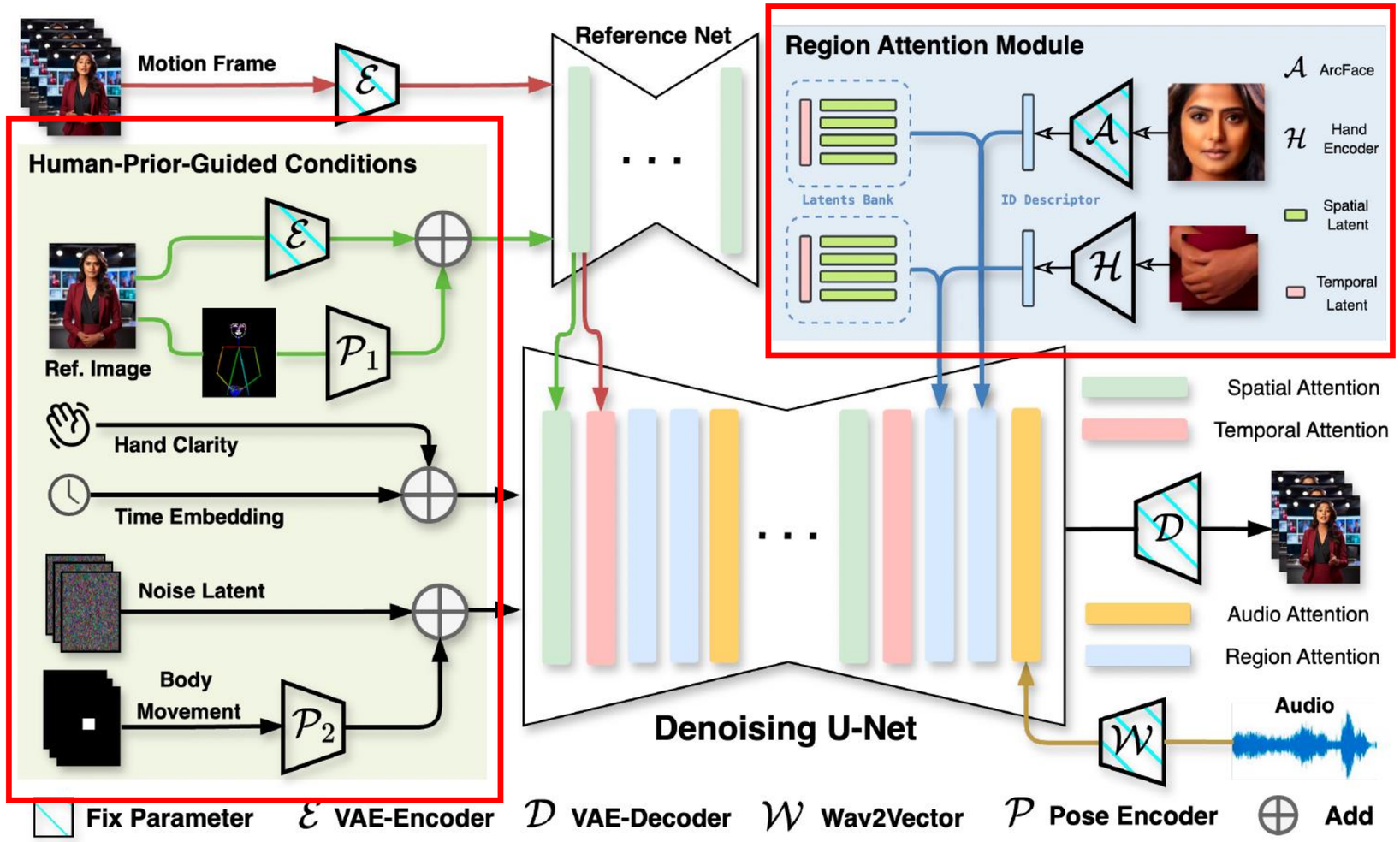


Ref. Image | Generated Result

Motion Uncertainty

Overall Framework

- To tackle Details Underfitting, we design **Region Attention Module (RAM)** to enhance the model's fitting ability for critical local regions.
- To tackle Motion Uncertaincy, we introduce **Human-Prior-Guided Conditions** to provide the model with prior knowledge of motion patterns and human structure.

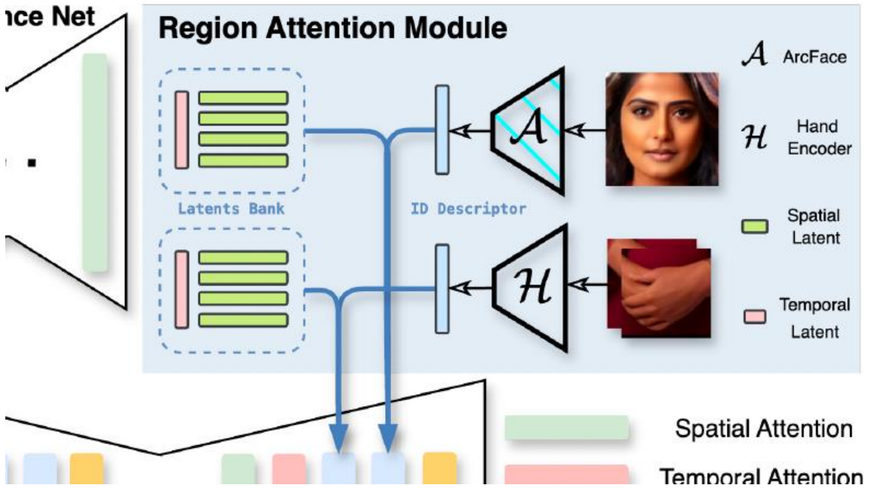


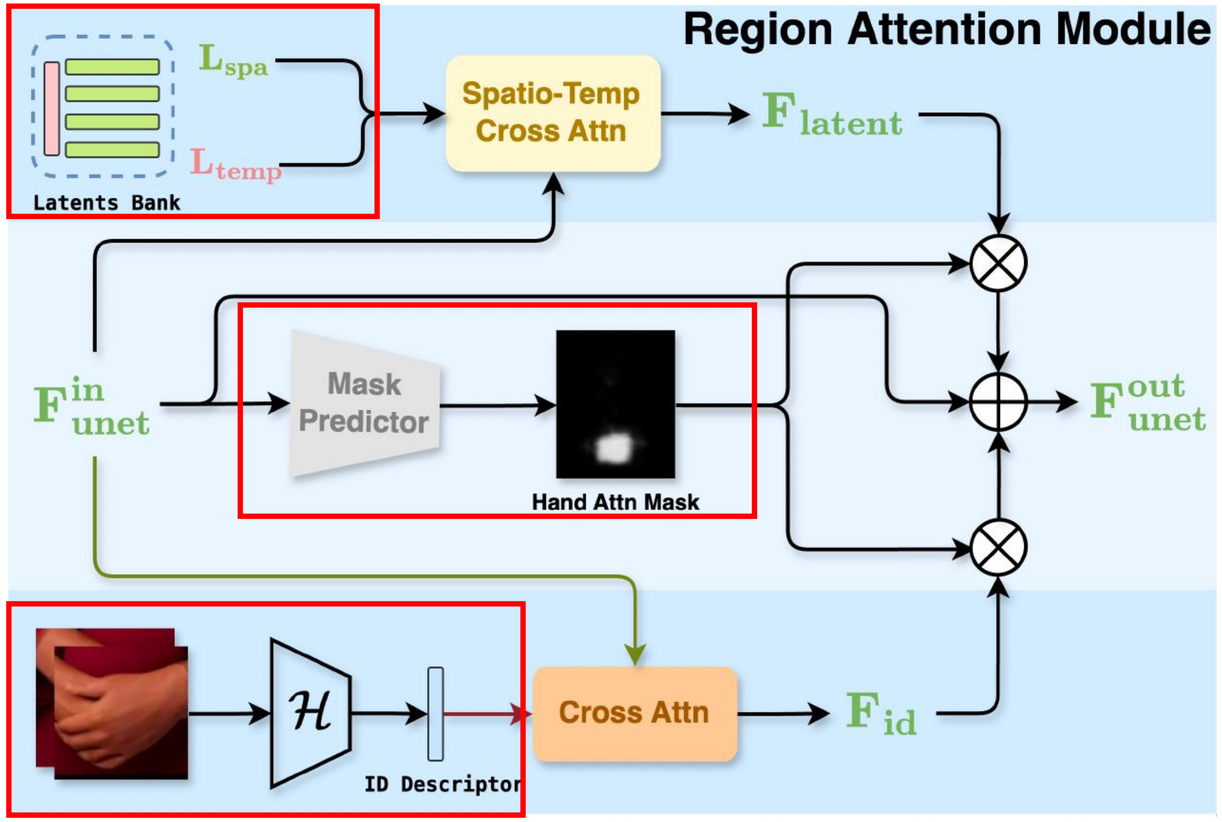
Region Attention Module

- **Spatio-temporal Latents Bank:** Additional learnable parameters, to learn shared priors of local details like structural topology and motion patterns.

- **Identity Descriptor:** Appearance features extracted from cropped images, to improve ID consistency.

- **Regional Mask Predictor:** Guide the learning process to focus on targeted local regions



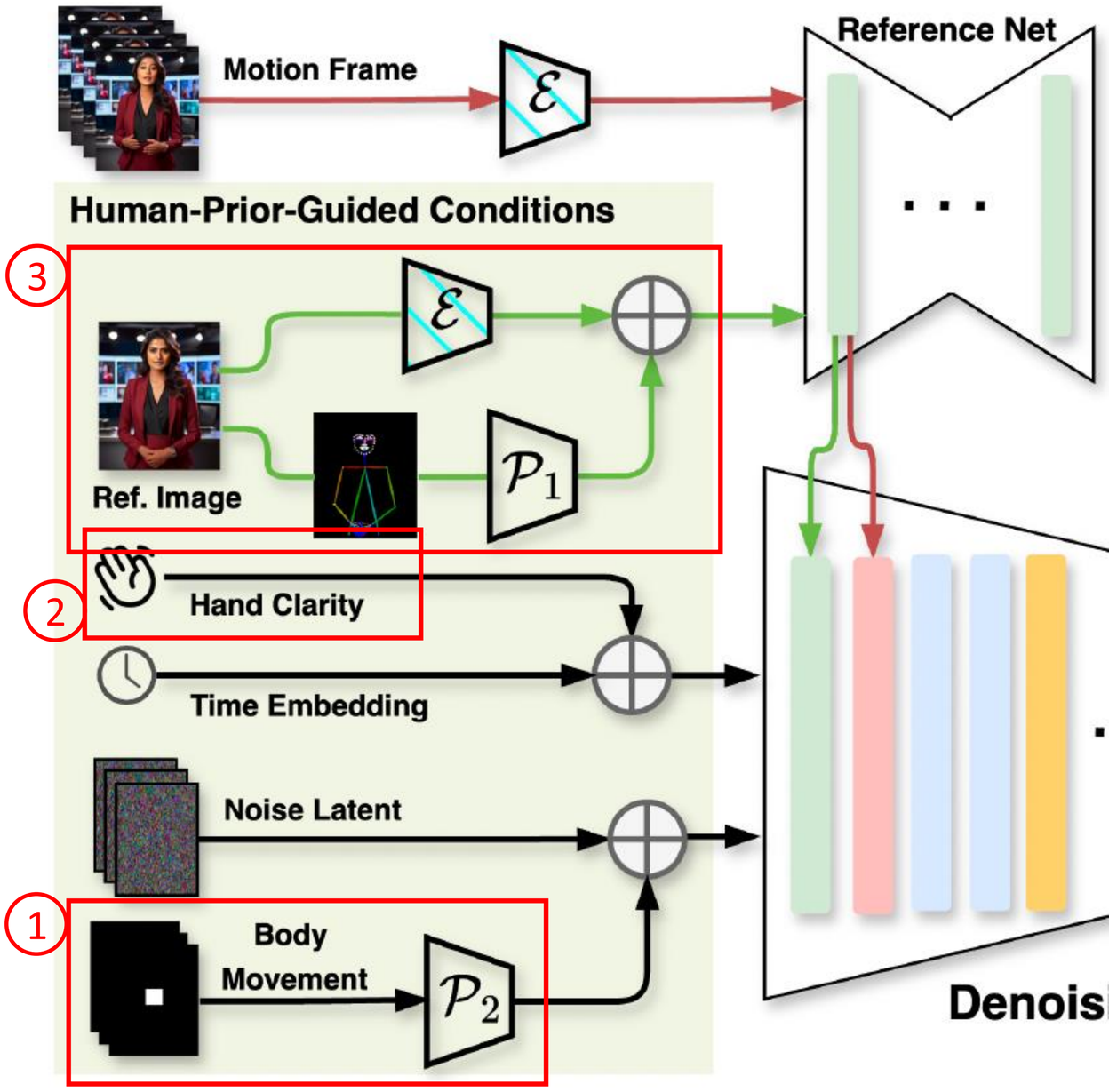


Human-Prior-Guided Condition

① **Body Movement Map**: A control signal for the movement amplitude of the body root

② **Hand Clarity Score**: Indicate the clarity of hand regions in the training video frames

③ **Pose-aligned Reference Feature**: Encode skeleton map along with reference image to incorporate its topological structure information.



Experiment Results

Hand Keypoint Variance

Hand Keypoint Confidence

Table 1: Quantitative comparison of audio-driven talking body. ∗ denotes evaluate on vlogger test set.

Methods	SSIM↑	PSNR↑	FID↓	FVD↓	CSIM↑	SyncC↑	SyncD↓	HKC↑	HKV↑
DiffTED	0.667	15.48	95.45	1185.8	0.185	0.925	12.543	0.769	-
DiffGest.+MimicMo.	0.656	14.97	58.95	1515.9	0.377	0.496	13.427	0.833	23.40
CyberHost (A2V-B)	**0.691**	**16.96**	**32.97**	**555.8**	**0.514**	**6.627**	**7.506**	**0.884**	**24.73**
Vlogger ∗	-	-	-	-	0.470	0.601	11.132	**0.923**	9.84
CyberHost (A2V-B) ∗	-	-	-	-	**0.522**	**7.897**	**7.532**	0.907	**18.75**

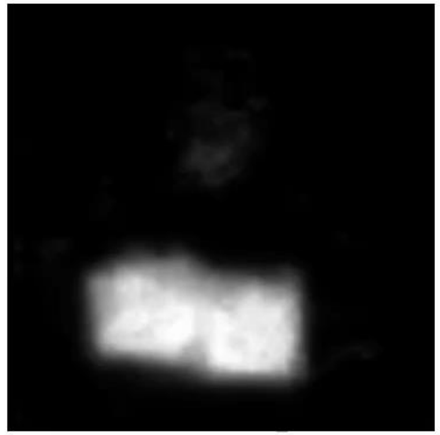
- Superior visual quality (SSIM, PSNR, FID, FVD)
- Enhanced ID consistency (CSIM)
- Improved lip-sync accuracy (SyncC, SyncD)
- High hand generation quality (HKC) and richer hand gesture diversity (HKV)

Ablation Study

Predicted **Hand Mask** at different timesteps within the Region Attention Module

Generated Video

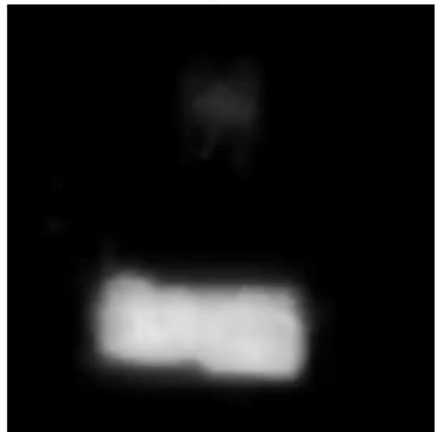
Step 1



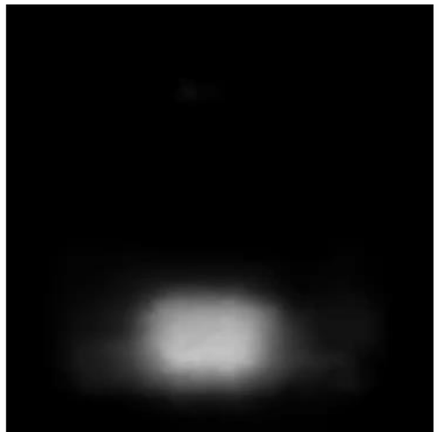
Step 201

Step 401

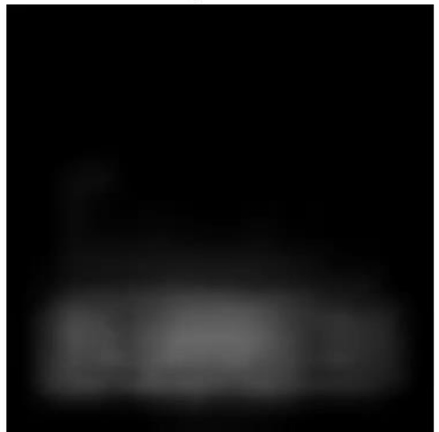
Step 601



Step 801



Step 961



Ablation Study

Visual ablation studies on the Latents Bank and ID Descriptor:

- **w/o Latents Bank:** decrease in the local structural stability.
- **w/o ID Descriptor:** decrease in ID consistency.

Ref. Image	w/o Latents Bank	w/o ID Descriptor	CyberHost (Full)

Ablation Study

Hand Clarity Score: sacrifice some hand movement richness, greatly reduces hand blur occurrences.

Low **Medium** **High**

Ablation Study

Pose-Align Referece Feature: Preventing the generation of results with evident limb ambiguities.

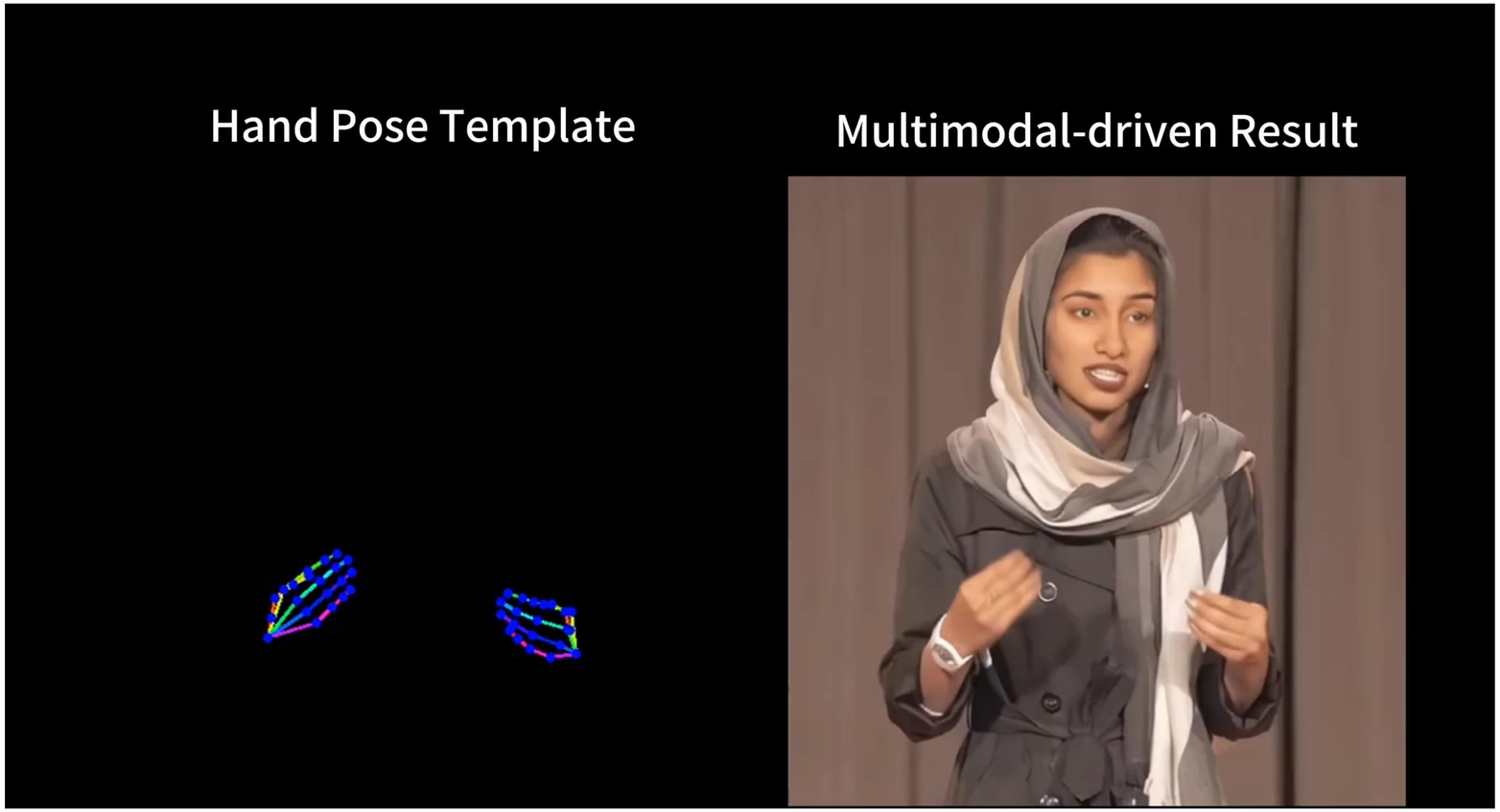
Ref. Image

w/o Pose-Align

w Pose-Align

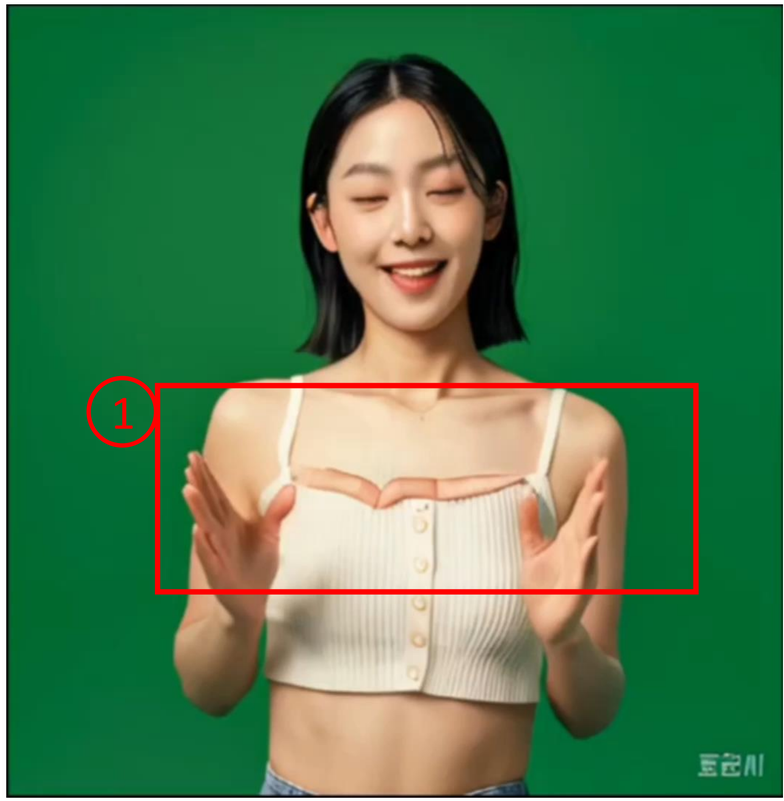
Multimodal-driven Results

Combining hand pose templates with audio signals to support Multimodal-driven setting.



Failure Cases & Limitations

- **Challenging reference images**
- Exaggerated limb proportions
- Challenging audio signals



Failure Cases & Limitations

- Challenging reference images
- **Exaggerated limb proportions**
- Challenging audio signals

Failure Cases & Limitations

- Challenging reference images
- Exaggerated limb proportions
- **Challenging audio signals**

Conclusion

- The first **One-stage** audio-driven talking body framework without relying any intermediate representations.
- A **Regional Attention Module** to address the issue of ***Details Underfitting*** and enhance the generation quality of critical local regions.
- A suite of **Human-Prior-Guided Conditions** to mitigate the ***Motion Uncertainty*** in audio-driven settings.
- Comprehensive visualization results to validate its effectiveness and superiority.

THANK YOU

Poster Location: Hall 3 + Hall 2B #71

Paper

Proj Page