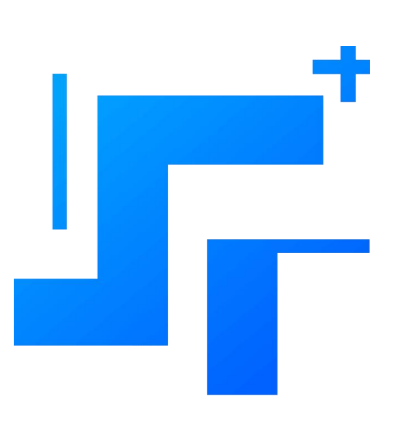




SpeakerVid-5M: A Large-Scale High-Quality Dataset for Audio-Visual Dyadic Interactive Human Generation

Youliang Zhang^{1,2}, Zhaoyang Li², Duomin Wang², jiahe zhang, Deyu Zhou^{2,3}, Zixin Yin^{2,3}, Xili Dai³, Gang YU², Xiu Li¹
¹Tsinghua University; ²StepFun; ³The Hong Kong University of Science and Technology.



Motivation

Embodied Multimodal Dialogue Video Generation

Input

Hey, Tom. How's the weather today ?

Output

What a wonderful day! How about...

Motion Degree: [2]
Entities List: [man, sofa, handrail...]
Camera Movement: [static]
Human Body: [portrait]
Motion Cap: man crossed his hands...
Expression Cap: man looks serious...
Topic: [entertainment...]
Speaker ID: A
Visual ID: person1
ASR: "Let me think about how to...."
Audio: [audio waveform]

8421 Hours
4839K Clips
81295 IDs

Blur
1.13
0.97

DWpose

Method

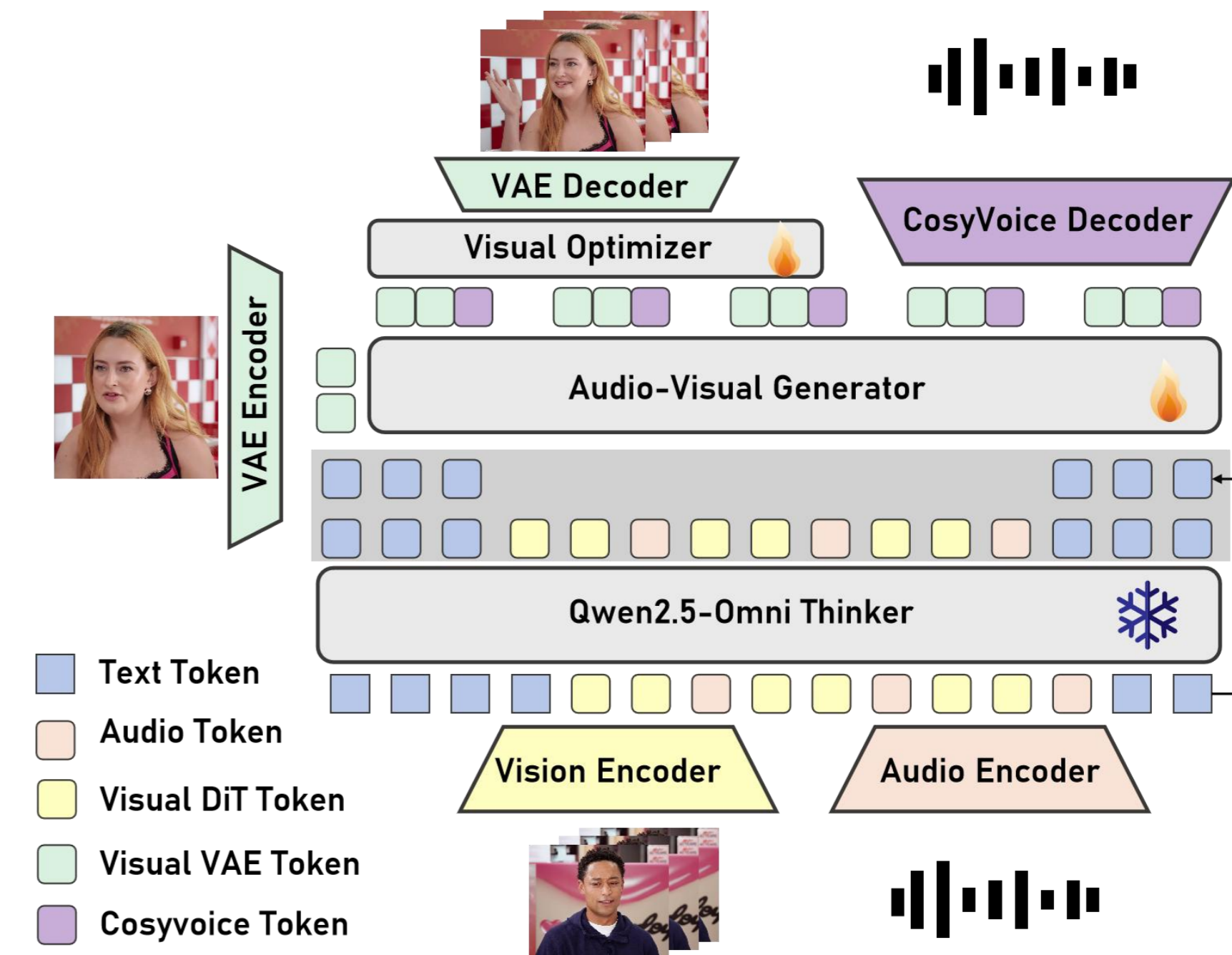


Illustration of our proposed autoregressive audio-visual generation method.

[Core Contributions]

- The first large-scale dataset designed specifically for the audiovisual dyadic interactive virtual human task. It includes 770K high-quality dialogue audiovisual pairs, with support for multi-turn conversations..
- Contains 5M single-speaker audiovisual clips, making it the largest talking human dataset.
- We open-source the entire dataset, including the raw data, annotations, and data processing pipeline.

Data Collection

1. Source Data Collection

YouTube 140K Videos, 57K Hours

Manual search, Original Videos, Highest resolution, Audio Check

4. Data Quality Filter

Luminance Filtering (Exclude Extreme Luminance)

Video Quality Filtering (DOVER, Bitrate, $\sqrt{H * W}$, Mean Blur: 0.15)

Blur Filtering (Mean Blur: 0.15)

Audio Filtering (Whisper ASR Confidence, No Speech Probability, Compression Ratio)

2. Audio-Visual Pre-processing

Scene Splitting (SceneDetect, 7M Clips, 3-15s)

Speaker Diarization (3D-Speaker, Time: [2,7], ID: 1)

Human Detection (YOLO, person1: [x1,y1,x2,y2])

Lip Sync (SyncNet, person1: 1.3 < person2: 6.4)

ID Correction (ArcFace)

3. Audio-Visual Annotation

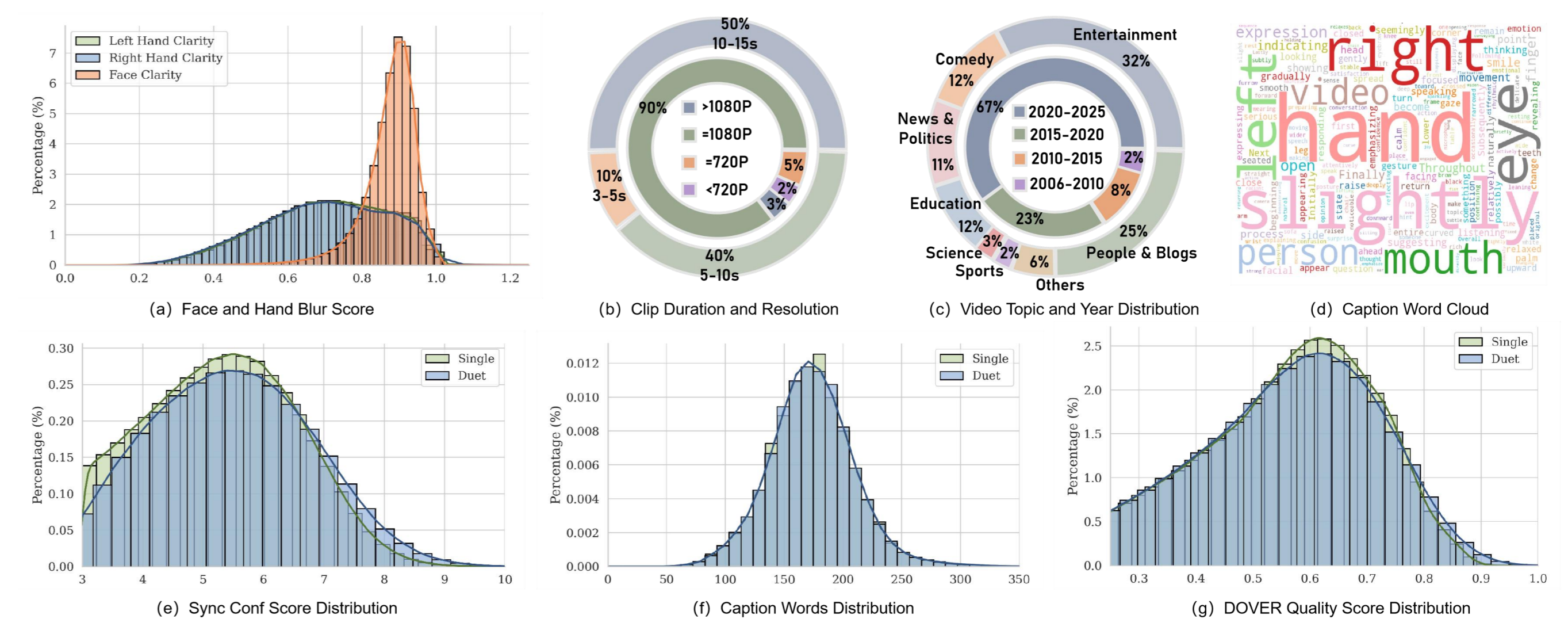
Structured Textual Caption (Qwen2.5-VL)

Audio Annotation (OpenAI Whisper, ASR: "I used to enjoy cooking very much...", 3D Speaker, ID: person1:ID2)

Face and Hand Blur (DWpose, Laplacian Variance, Blur Score: 0.07, 0.32, 0.75, 1.02)

Skeleton Sequence (DWpose)

Statistics



Datasets	Domain	Clips	Duration (hours)	Generation	Person num	audio	pose	Speaker ID	Blur anno	Body composition	Caption type	IDs	Resolution
UCF-101	Human	13.3K	26.7	-	N/A						Text	N/A	240P
ActivityNet Caba	Human	100K	849	-	N/A						Text	N/A	N/A
NTU RGB+D	Human	114K	3.7	Conditioned	single		✓				-	N/A	1080P
TikTok-v4	Human	350	1	Conditioned	single		✓				-	N/A	N/A
Openhumanvid	Human	13.4M	16.7K	Conditioned	multi	✓	✓				Structured	N/A	720P
VoxCeleb Nagrani et al. (2017)	Head	21.2K	352	Conditioned	single	✓					-	1.2k	224P
VoxCeleb2 Chung et al. (2018)	Head	150.4K	2.4K	Conditioned	single	✓					-	6.1K	224P
MEAD Wang et al. (2020)	Head	281.4K	39	Conditioned	single	✓					-	60	1080P
CelebV-HQ Zhu et al. (2022)	Head	35.6K	68	Conditioned	single	✓					Structured	15.6K	512P
CelebV-Text Yu et al. (2023)	Head	70K	279	Conditioned	single	✓					Structured	N/A	512P
SpeakerVid-5M	Human	5.2M	8.7K	Conditioned	single	✓	✓	✓	✓	✓	Structured	83K	1080P
SpeakerVid-5M (Dialogue)	Human	770K	1.8K	Dyadic	single	✓	✓	✓	✓	✓	Structured	16K	1080P

The SpeakerVid-5M curation pipeline: (1) Source data collection; (2) Multi-step audio-visual pre-processing; (3) Rich multi-modal annotation; (4) Rigorous quality filtering stage for data fidelity