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

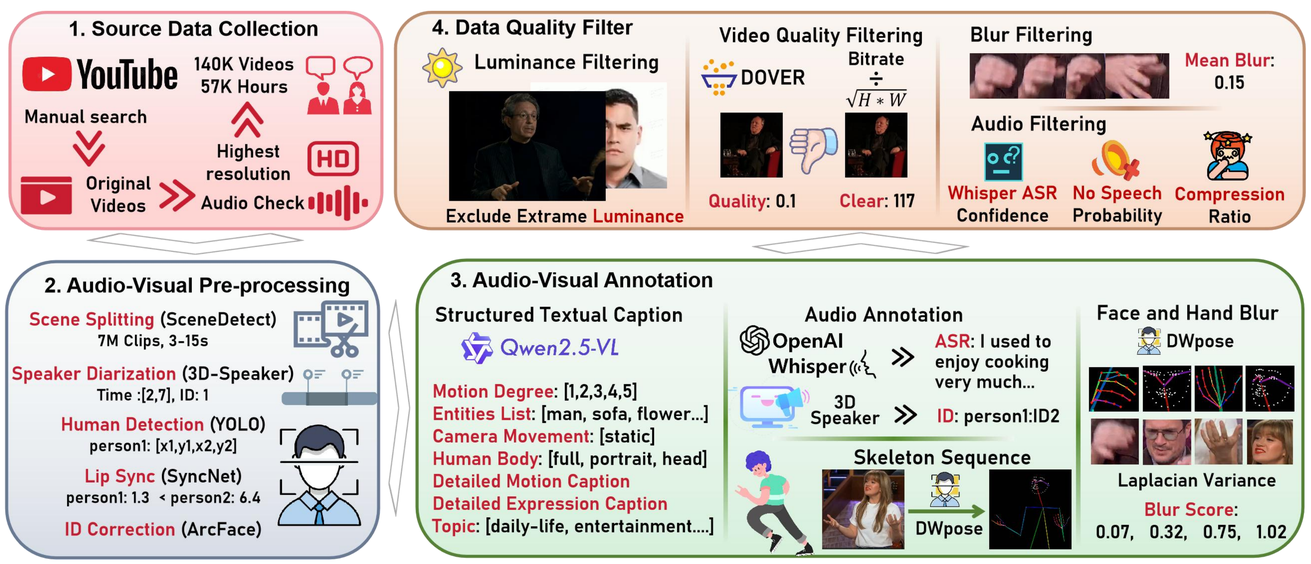
Youliang Zhang[1,2], Zhaoyang Li[2], Duomin Wang[2], jiahe zhang, Deyu Zhou[2,3], Zixin Yin[2,3], Xili Dai[3], Gang YU[2], Xiu Li[1]
[1]Tsinghua University; [2]StepFun; [3]The Hong Kong University of Science and Technology.

Motivation

[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



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

Method

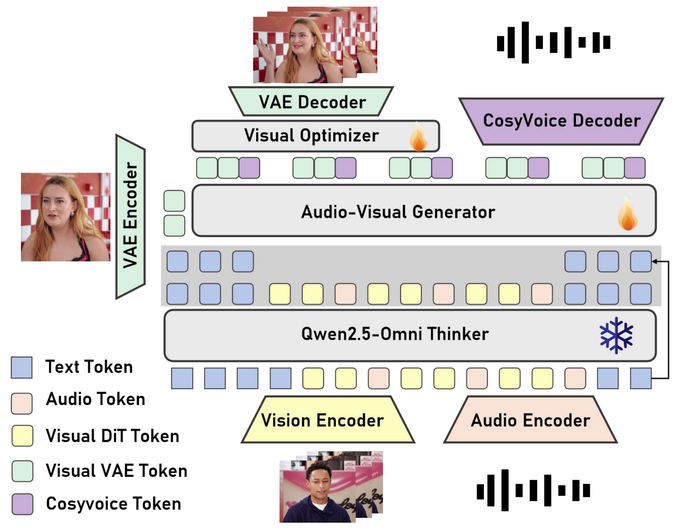
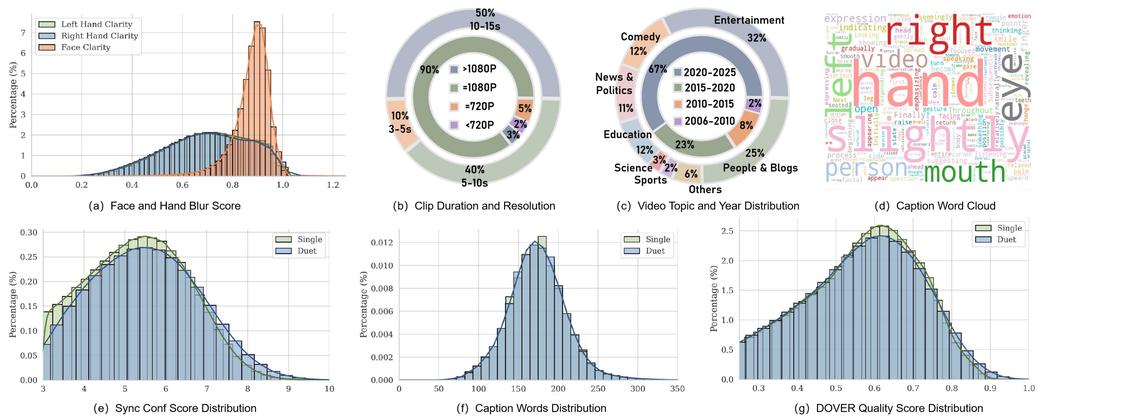


Illustration of our proposed autoregressive audio-visual generation method.

Statistics



Datasets	Domain	Clips	Duration (hours)	Generation	Person num	audio	pose	Speaker ID	Blur anno	Body composition	Caption type	IDs	Resolution
UCF-101 Soomro et al. (2012)	Human	13.3K	26.7	–	N/A						Text	N/A	240P
ActivityNet Caba Heilbron et al. (2015)	Human	100K	849	–	N/A						Text	N/A	N/A
NTU RGB+D Shahroudy et al. (2016)	Human	114K	3.7	Conditioned	single		✓				-	N/A	1080P
TikTok-v4 Chang et al. (2023)	Human	350	1	Conditioned	single		✓				-	N/A	N/A
Openhumanvid Li et al. (2025)	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