Both Ears Wide Open: Towards Language-Driven Spatial Audio Generation

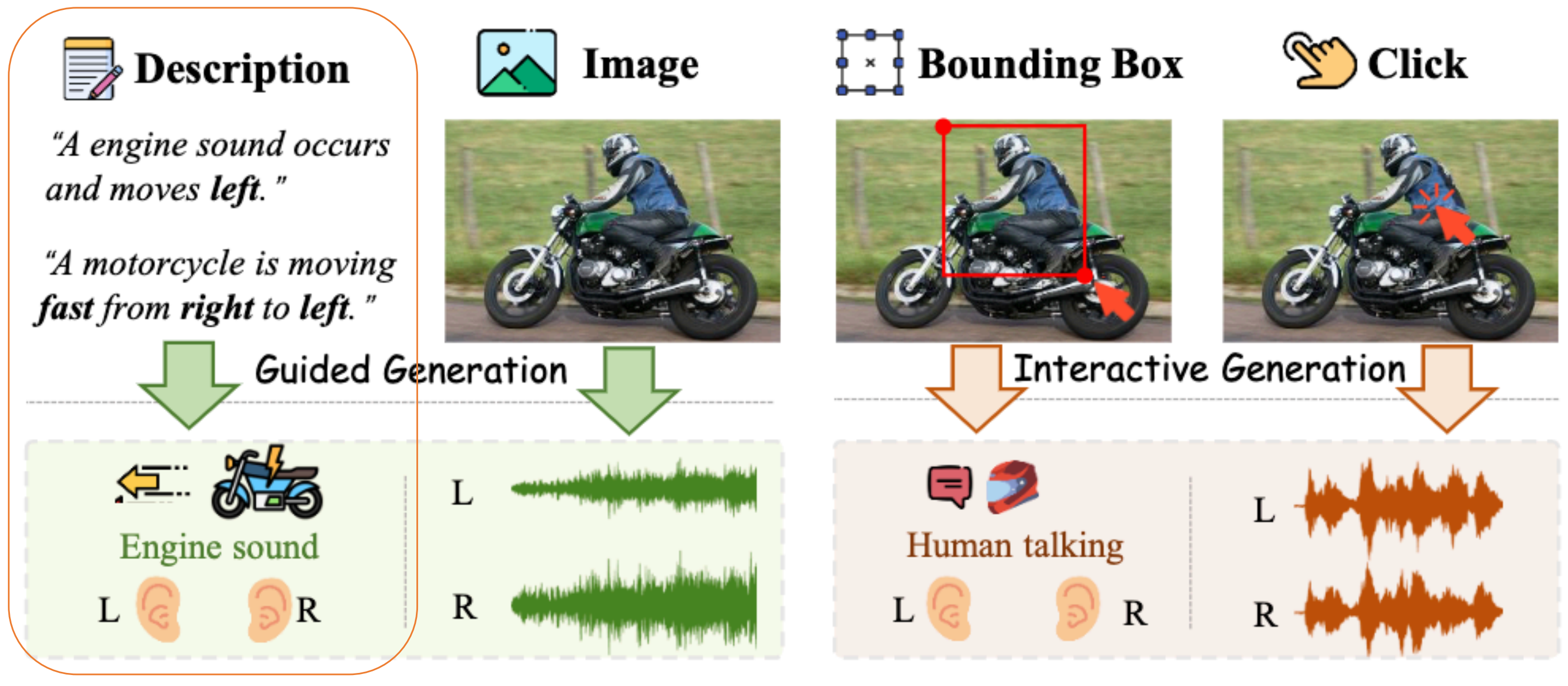
Peiwen Sun*[1,2], Sitong Cheng*[1], Xiangtai Li*[3], Zhen Ye[1], Huadai Liu[4], Honggang Zhang[2], Wei Xue[1], Yike Guo[1]

[1]Hong Kong University of Science and Technology,
[2]Beijing University of Posts and Telecommunications,
[3]Nanyang Technological University, [4]Zhejiang University

Demo Page

Code

Introduction



Why do this task?

Controlling stereo audio with spatial contexts to achieve immersive and attractive soundscapes that **adheres physical world**.

➢ To achieve this with following contribution

a) An open-source, large-scale, stereo audio **dataset** with spatial captions.

b) An one-stage, controllable, spatial audio generation **framework**.

c) a series of subjective and objective **metrics** based on ITD and opinion score.

Dataset: BEWO-1M

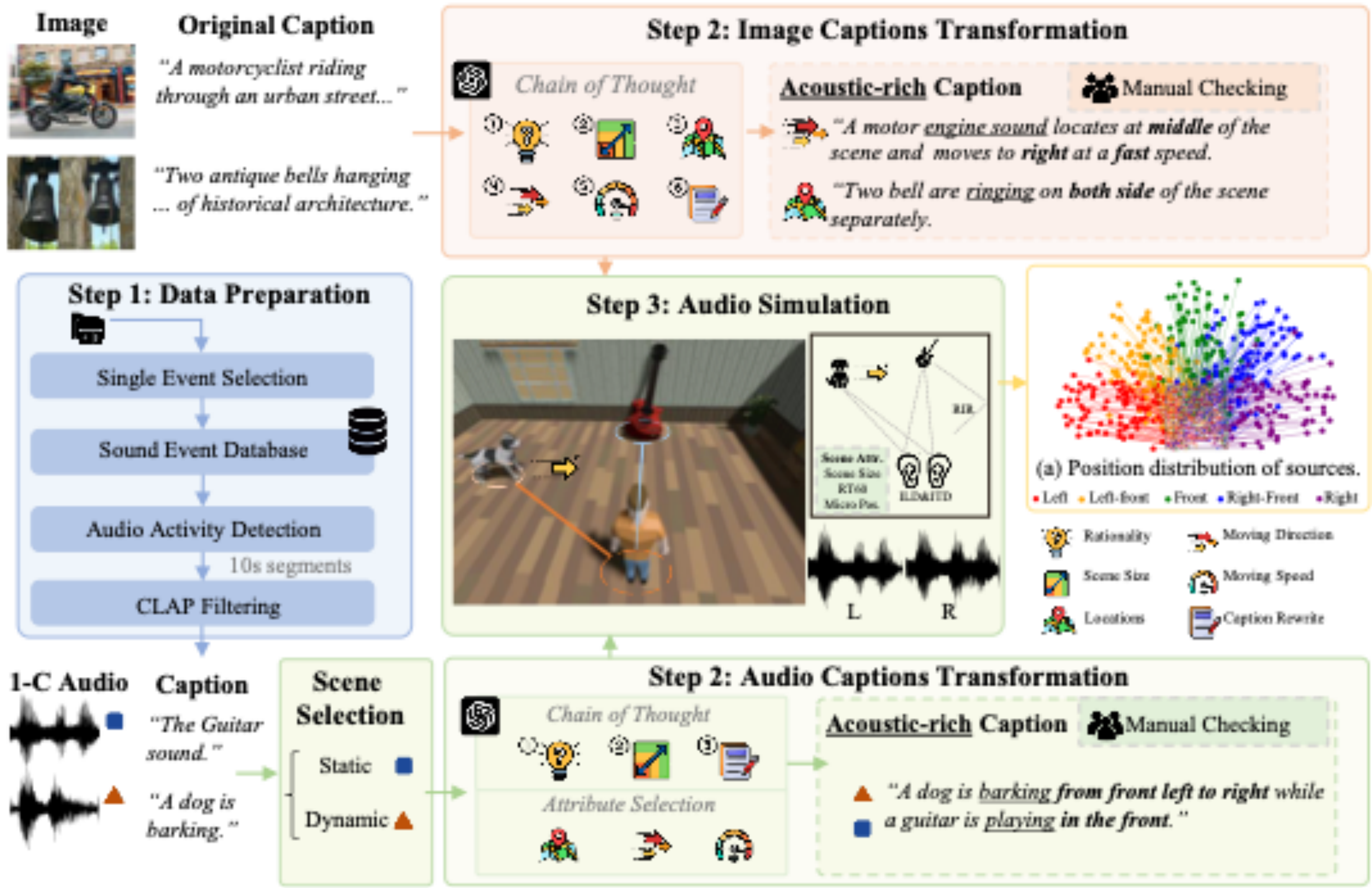
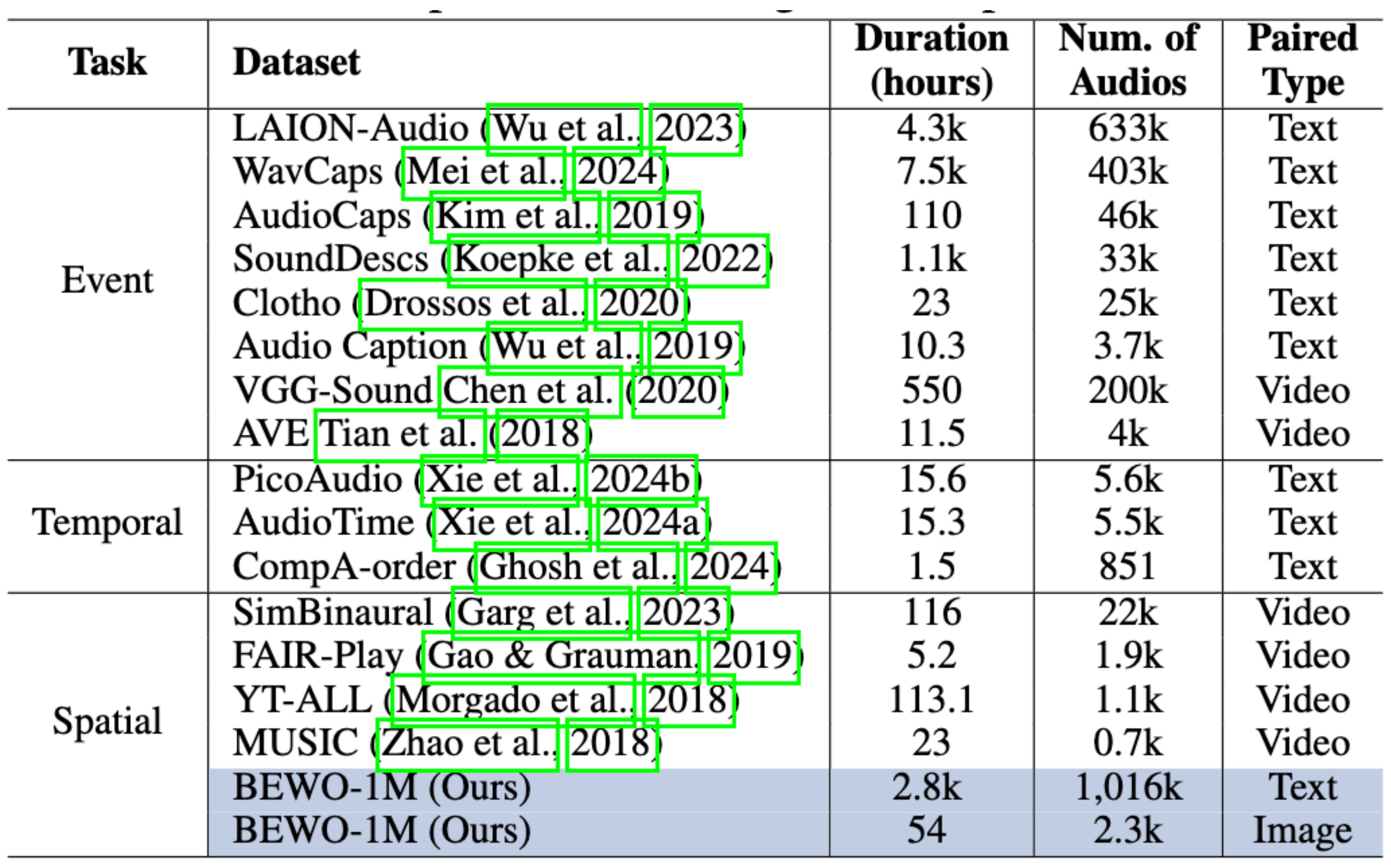


Image
Original Caption
"A motorcyclist riding through an urban street..."
"Two antique bells hanging ... of historical architecture."
Step 2: Image Captions Transformation
Chain of Thought
Acoustic-rich Caption
Manual Checking
"A motor engine sound locates at middle of the scene and moves to right at a fast speed."
"Two bell are ringing on both side of the scene separately."
Step 1: Data Preparation
Single Event Selection
Sound Event Database
Audio Activity Detection
10s segments
CLAP Filtering
Step 3: Audio Simulation
(a) Position distribution of sources.
Left
Left-front
Front
Right-Front
Right
Rationality
Moving Direction
Scene Size
Moving Speed
Location
Caption Rewrite
1-C Audio
Caption
"The Guitar sound."
"A dog is barking."
Scene Selection
Static
Dynamic
Step 2: Audio Captions Transformation
Attribute Selection
"A dog is barking from front left to right while a guitar is playing in the front."

Pipeline of our proposed dataset.



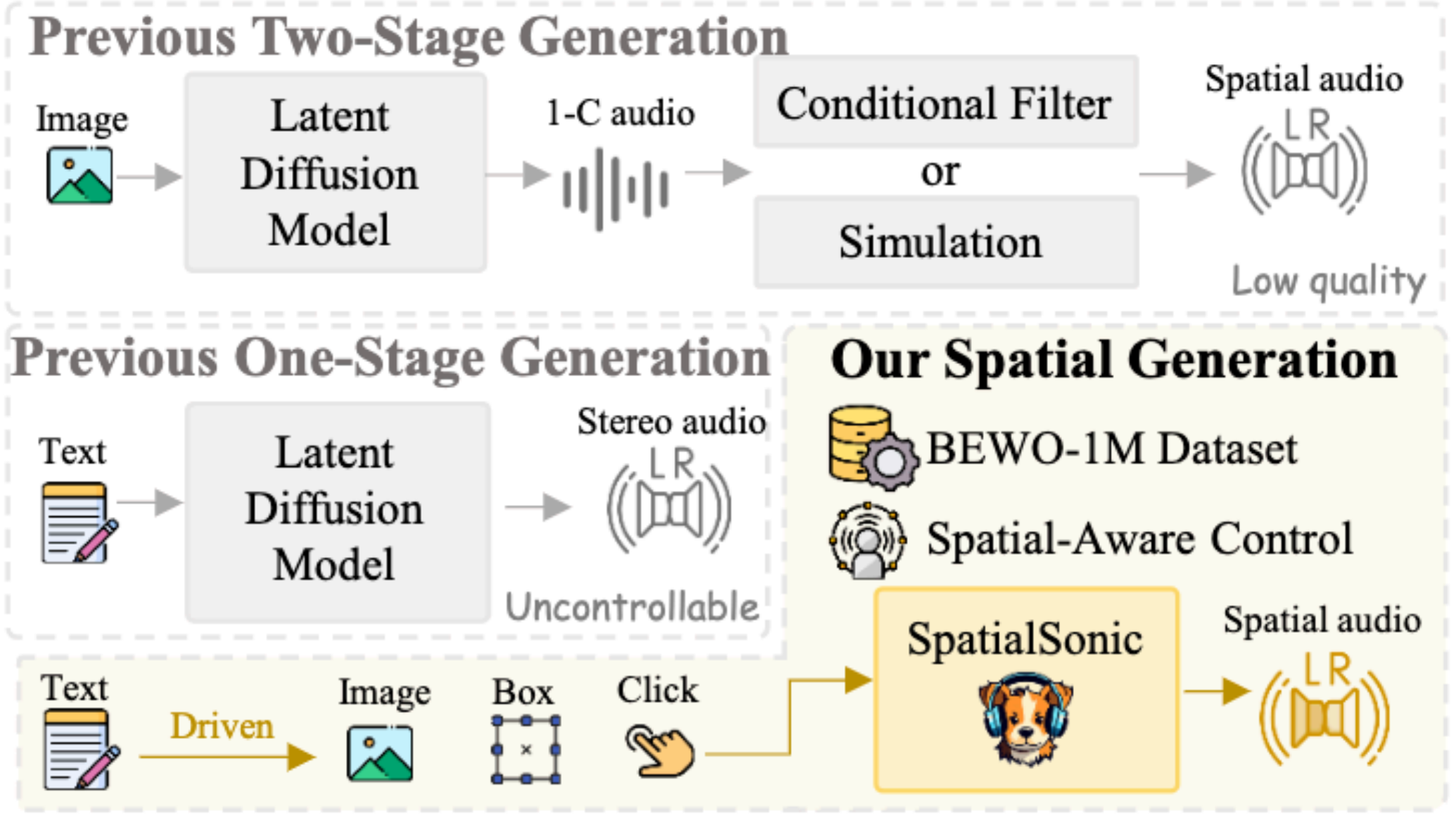
Task	Dataset	Duration (hours)	Num. of Audios	Paired Type
Event	LAION-Audio (Wu et al., 2023)	4.3k	633k	Text
	WavCaps (Mei et al., 2024)	7.5k	403k	Text
	AudioCaps (Kim et al., 2019)	110	46k	Text
	SoundDescs (Koepke et al., 2022)	1.1k	33k	Text
	Clotho (Drossos et al., 2020)	23	25k	Text
	Audio Caption (Wu et al., 2019)	10.3	3.7k	Text
	VGG-Sound Chen et al. (2020)	550	200k	Video
	AVE Tian et al. (2018)	11.5	4k	Video
Temporal	PicoAudio (Xie et al., 2024b)	15.6	5.6k	Text
	AudioTime (Xie et al., 2024a)	15.3	5.5k	Text
	CompA-order (Ghosh et al., 2024)	1.5	851	Text
Spatial	SimBinaural (Garg et al., 2023)	116	22k	Video
	FAIR-Play (Gao & Grauman, 2019)	5.2	1.9k	Video
	YT-ALL (Morgado et al., 2018)	113.1	1.1k	Video
	MUSIC (Zhao et al., 2018)	23	0.7k	Video
	BEWO-1M (Ours)	2.8k	1,016k	Text
	BEWO-1M (Ours)	54	2.3k	Image

Comparison of datasets.

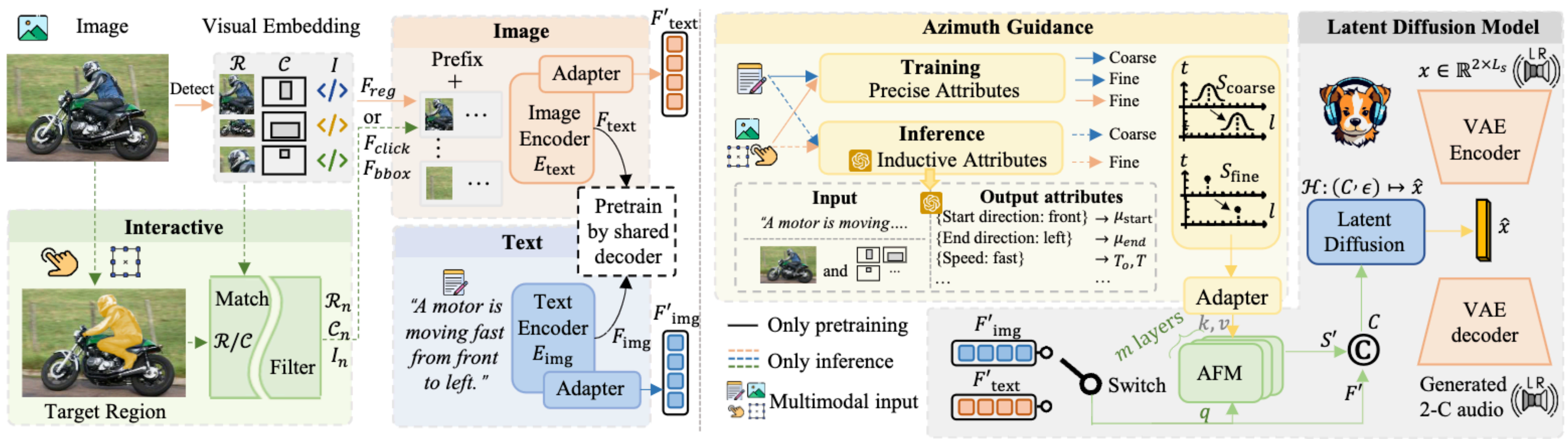
➢ Developing a semi-automated pipeline to create an open-source, large-scale, stereo audio dataset with spatial captions, **BEWO-1M** and supporting both large-scale training and precise evaluation.

➢ Totally, we constructed **2.8k hours** of training audio with more than **1M** audio-text pairs and approximately 17 hours of validation data with 6.2k pairs. To the best of our knowledge, this work represents the first attempt to address these issues.

Model: SpatialSonic



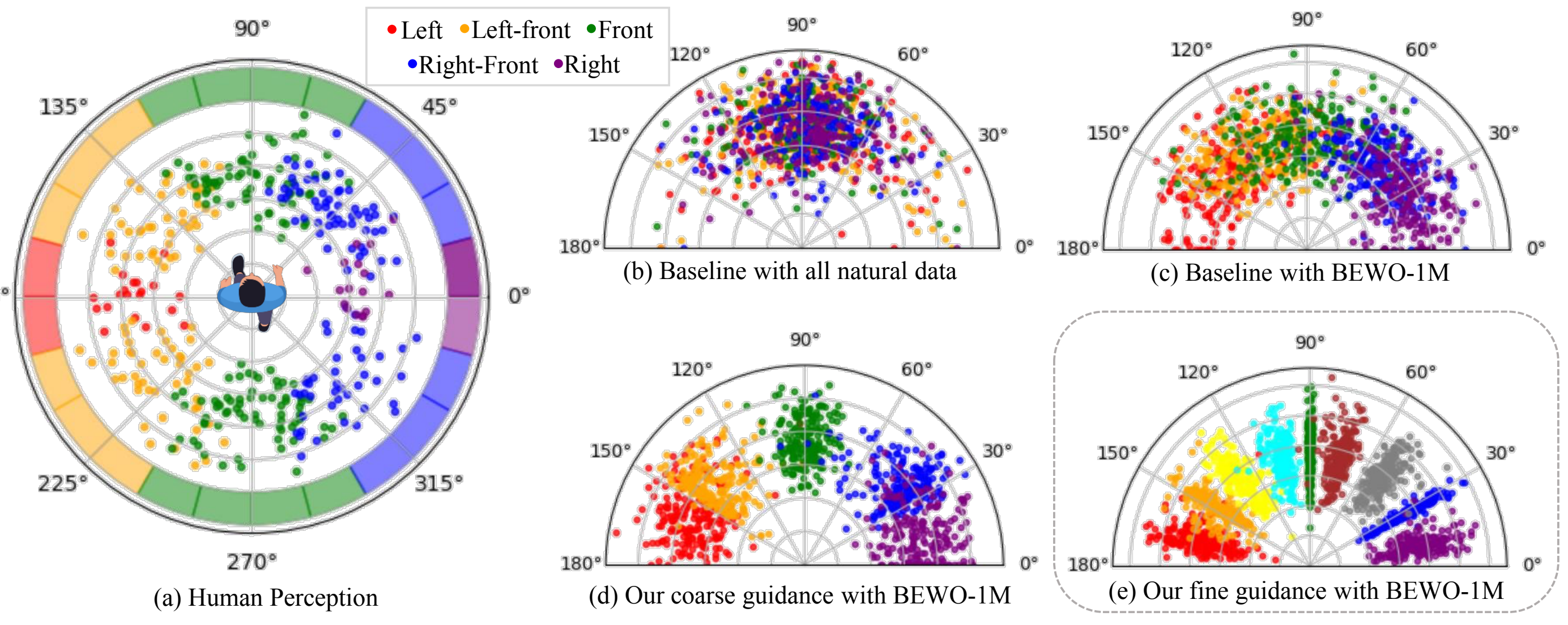
One stage generation model



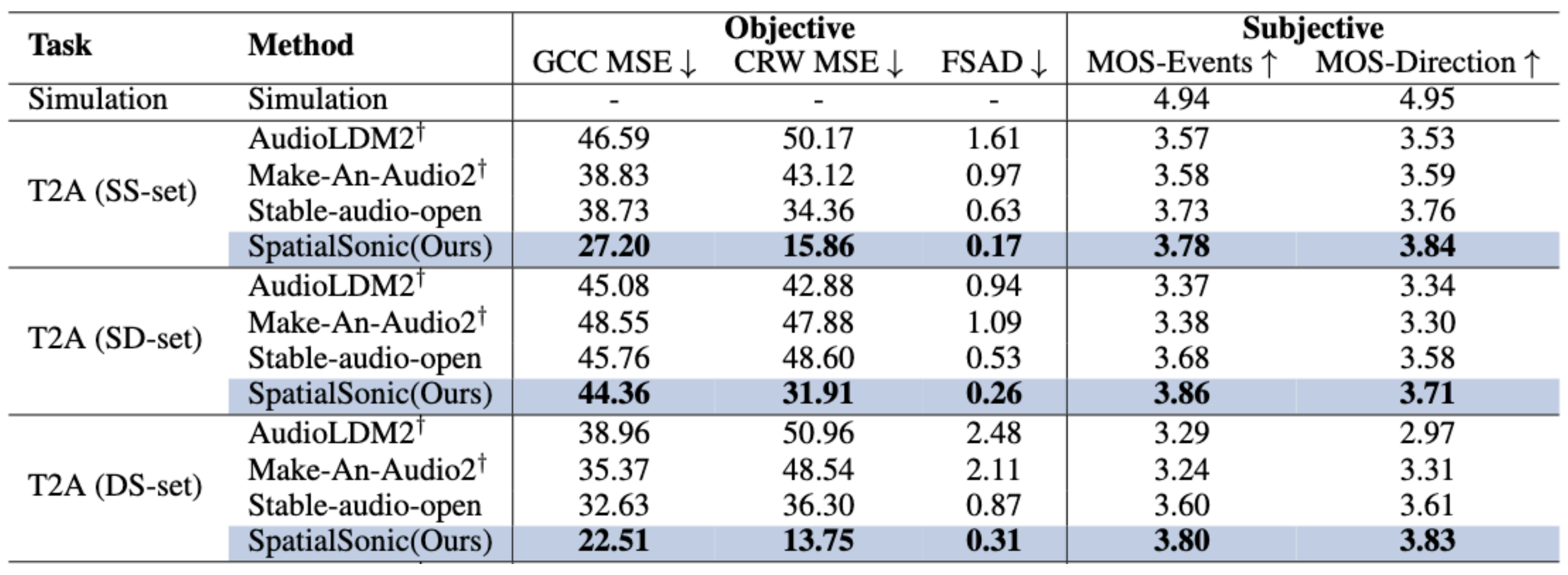
Spatial audio generation pipeline based on DiT

➢ The proposed SpatialSonic takes azimuth as precise guidance and generate spatial audio with one stage behaviour.

Evaluation: ITD and MOS



Visualization of the angle of generated samples



Task	Method	Objective			Subjective	
		GCC MSE ↓	CRW MSE ↓	FSAD ↓	MOS-Events ↑	MOS-Direction ↑
Simulation	Simulation	-	-	-	4.94	4.95
T2A (SS-set)	AudioLDM2†	46.59	50.17	1.61	3.57	3.53
	Make-An-Audio2†	38.83	43.12	0.97	3.58	3.59
	Stable-audio-open	38.73	34.36	0.63	3.73	3.76
	SpatialSonic(Ours)	**27.20**	**15.86**	**0.17**	**3.78**	**3.84**
T2A (SD-set)	AudioLDM2†	45.08	42.88	0.94	3.37	3.34
	Make-An-Audio2†	48.55	47.88	1.09	3.38	3.30
	Stable-audio-open	45.76	48.60	0.53	3.68	3.58
	SpatialSonic(Ours)	**44.36**	**31.91**	**0.26**	**3.86**	**3.71**
T2A (DS-set)	AudioLDM2†	38.96	50.96	2.48	3.29	2.97
	Make-An-Audio2†	35.37	48.54	2.11	3.24	3.31
	Stable-audio-open	32.63	36.30	0.87	3.60	3.61
	SpatialSonic(Ours)	**22.51**	**13.75**	**0.31**	**3.80**	**3.83**

Some important results of the subset (3/5)