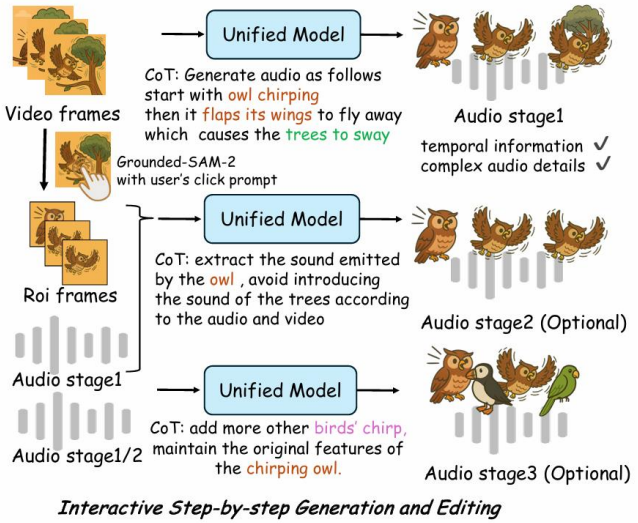
ThinkSound

Chain-of-Thought Reasoning in Multimodal Large Language Models for Audio Generation and Editing

Introduction

The Problem:

End-to-end V2A models often lack nuance and precise temporal synchronization. This is because they fail to perform sophisticated reasoning about visual dynamics and temporal relationships.

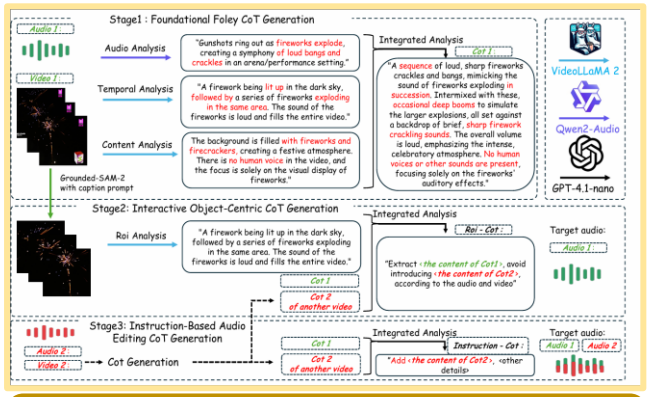


Interactive Step-by-step Generation and Editing

The Motivation: Using CoT as a Solution

To address this gap, we use Chain-of-Thought (CoT) reasoning to decompose the complex task into manageable, stepwise instructions, enabling models to reason like human sound designers for more coherent, controllable audio generation and precise editing.

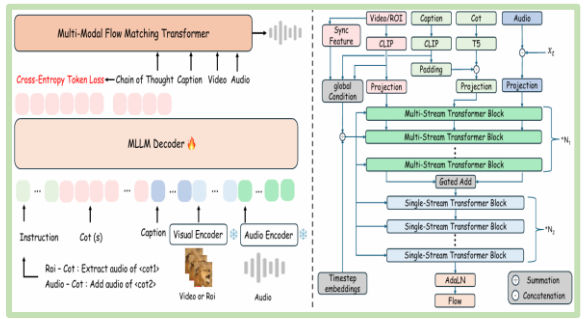
Dataset: AudioCoT



Quality Control:
● Audio-text alignment (CLAP) ● Object tracking consistency
● Human verification

● **Stage 1 Foundational Foley CoT**: Generates CoT reasoning chains by integrating video (via VideoLLaMA2) and audio (via Qwen2-Audio) information using GPT-4.1-nano.
● **Stage 2 Interactive Object-Centric CoT**: Uses Grounded SAM2 to identify sound-emitting objects (ROIs) and GPT-4.1-nano to create CoT for object-specific audio manipulations.
● **Stage 3 Instruction-Based Editing CoT**: Focuses on editing operations (e.g., inpainting, addition).

Method: ThinkSound



We propose **ThinkSound**, a framework enabling step-by-step, interactive audio generation and editing guided by CoT reasoning.

Our three-stage pipeline decomposes the task into:
● **Stage 1 – CoT-Guided Foley Generation**: synthesizing semantically and temporally matched soundscapes.
● **Stage 2 – Interactive Object-Focused Audio Generation**: refining sounds through user clicks on specific visual objects.
● **Stage 3 – Instruction-Based Audio Editing**: applying high-level, natural-language instructions for targeted modifications.

● **CoT Reasoning MLLM:** This component uses fine-tuned VideoLLaMA2 as its core reasoning engine. It understands complex audio-visual contexts, inferring acoustic properties and temporal relationships. It then decomposes complex tasks and interprets diverse user instructions, outputting them as structured CoT instructions.

● **CoT-Guided Unified Audio Foundation Model:** This component translates the MLLM's CoT reasoning into high-quality audio using conditional flow matching. It employs a dual-path text encoding strategy: MetaCLIP encodes scene-level context, while T5 processes the structured CoT for fine-grained control.

Experiments

Table 1: Comparison of our ThinkSound foundation model with existing video-to-audio baselines on the VGGSound test set. ↓ indicates lower is better, ↑ indicates higher is better. For MOS, we show the mean and variance of the MOS scores. † indicates that the method does **not use text** for inference.

Method	Objective Metrics						Subjective Metrics		Efficiency	
	FD↓	KL_{PASST} ↓	KL_{PaSST_S} ↓	DeSync↓	$CLAP_{ap}$ ↑	$CLAP_{CoT}$ ↑	MOS-Q↑	MOS-A↑	Params	Time(s)↓
GT	-	-	-	0.55	0.28	0.45	4.37±0.21	4.56±0.19	-	-
See&Hear	118.95	2.26	2.30	1.20	0.32	0.35	2.75±1.08	2.87±0.99	415M	19.42
V-AURA†	46.99	2.23	1.83	0.65	0.23	0.37	3.42±1.03	3.20±1.17	695M	14.00
FoleyCrafter	39.15	2.06	1.89	1.21	**0.41**	0.34	3.08±1.21	2.63±0.88	1.20B	3.84
Frieren†	74.96	2.55	2.64	1.00	0.37	0.34	3.27±1.11	2.95±1.09	159M	-
V2A-Mapper†	48.10	2.50	2.34	1.23	0.38	0.32	3.31±1.02	3.16±1.04	229M	-
MMAudio	43.26	1.65	1.40	**0.44**	0.31	0.40	3.84±0.89	3.97±0.82	1.03B	3.01
ThinkSound	**34.56**	**1.52**	**1.32**	0.46	0.33	**0.46**	**4.02±0.73**	**4.18±0.79**	1.30B	1.07
w/o CoT Reasoning	39.84	1.59	1.40	0.48	0.29	0.41	3.91±0.83	4.04±0.75	1.30B	0.98

Table 2: Out-of-distribution evaluation on MovieGen Audio Bench. This benchmark does not provide the GT audios, so we cannot compare FD and KL.

Method	$CLAP_{CoT}$ ↑	DeSync↓	MOS-Q↑	MOS-A↑
MMAudio	0.45	0.77	3.95±0.87	3.62±1.03
MovieGen	0.47	1.00	3.98±0.77	3.70±0.96
ThinkSound	**0.51**	**0.76**	**4.11±0.74**	**3.87±0.82**

Table 3: Object-focused generation performance.

Method	FD↓	KL_{PASST} ↓	CLAP↑	MOS-Q↑	MOS-A↑
MMAudio	44.46	1.38	0.41	3.61±0.63	3.64±0.69
ThinkSound	**43.27**	**1.32**	**0.48**	**3.89±0.52**	**3.91±0.53**
w/o CoT	45.28	1.34	0.43	3.77±0.64	3.81±0.59

Table 4: Audio editing results on AudioCoT test set (MOS-A: alignment between audio and text; DDPM: DDPM-Friendly).

Method	FD↓	KL_{PASST} ↓	CLAP↑	MOS-Q↑	MOS-A↑
AudioLDM-2	61.28	1.94	0.35	3.28±0.59	3.48±0.82
DDPM	55.56	1.75	0.39	3.34±0.28	3.67±0.56
ThinkSound	**34.78**	**1.45**	**0.51**	**3.92±0.82**	**3.85±0.82**
w/o CoT	45.78	1.58	0.44	3.53±0.45	3.52±0.62

Quantitative Results

● ThinkSound achieves **state-of-the-art** video-to-audio generation, surpassing all baselines on VGGSound (FD 34.56 ↓, MOS-Q 4.02 ↑).
● **CoT reasoning** notably enhances temporal and semantic alignment (CLAP 0.46 vs 0.41 w/o CoT).
● Demonstrates strong **generalization** on MovieGen Audio Bench (CLAP 0.51, MOS-Q 4.11).
● **Object-focused generation** shows excellent control (MOS-A 3.91).
● **Instruction-based editing** outperforms AudioLDM-2 / DDPM baselines (FD 34.78, CLAP 0.51).

Scan for Demos

- Project Page: https://thinksound-project.github.io/
- Contact Us: huadai.liu@connect.ust.hk