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

Huadai Liu[1,2], **Kaicheng Luo**[2], **Jialei Wang**[3], **Wen Wang**[2]
Qian Chen[2], **Zhou Zhao**[3], **Wei Xue**[1] *
[1]Hong Kong University of Science and Technology (HKUST)
[2]Tongyi Fun Team, Alibaba Group [3]Zhejiang University

Abstract

While end-to-end video-to-audio generation has greatly improved, producing high-fidelity audio that authentically captures the nuances of visual content remains challenging. Like professionals in the creative industries, this generation requires sophisticated reasoning about items such as visual dynamics, acoustic environments, and temporal relationships. We present **ThinkSound**, a novel framework that leverages Chain-of-Thought (CoT) reasoning to enable stepwise, interactive audio generation and editing for videos. Our approach decomposes the process into three complementary stages: foundational foley generation that creates semantically coherent soundscapes, interactive object-centric refinement through precise user interactions, and targeted editing guided by natural language instructions. At each stage, a multimodal large language model generates contextually aligned CoT reasoning that guides a unified audio foundation model. Furthermore, we introduce **AudioCoT**, a comprehensive dataset with structured reasoning annotations that establishes connections between visual content, textual descriptions, and sound synthesis. Experiments demonstrate that ThinkSound achieves state-of-the-art performance in video-to-audio generation across both audio metrics and CoT metrics, and excels in the out-of-distribution Movie Gen Audio benchmark. The project page is available at `https://ThinkSound-Project.github.io`.

1 Introduction

Generating realistic sound for video demands more than recognizing objects; it requires reasoning about complex visual dynamics and context – determining when an owl is chirping versus flapping its wings, identifying the subtle sway of tree branches, and synchronizing multiple sound events within a scene, as illustrated in Figure 1. Current end-to-end video-to-audio (V2A) generation systems (Luo et al., 2024; Xing et al., 2024; Zhang et al., 2024), while having largely improved, often struggle with this compositional complexity and contextual nuance. They may produce generic sounds or fail to synchronize precisely with subtle visual cues, thus limiting fidelity and user control.

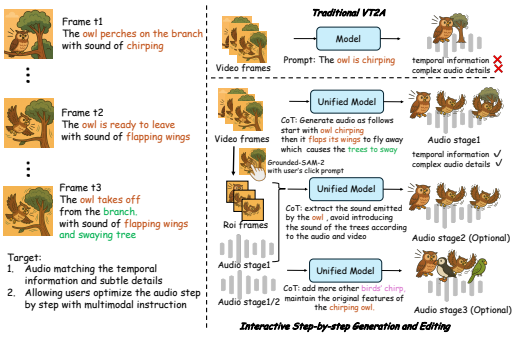

Figure 1: ThinkSound with CoT: (1) CoT-driven foley synthesis captures semantic and temporal details (2) interactive object-centric refinement for control (3) targeted editing.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recent breakthroughs in Multimodal Large Language Models (MLLMs) (Lu et al., 2024; Liu et al., 2023b; Wang et al., 2024b) offer powerful capabilities in understanding and reasoning across multiple modalities. Concurrently, Chain-of-Thought (CoT) prompting (Wei et al., 2022; Zhang et al., 2023; Wang et al., 2025) has proven effective at eliciting structured, step-by-step reasoning from large models. These advancements create great potential to fundamentally rethink V2A by decomposing complex sound design into explicit reasoning steps and actionable synthesis instructions. The necessity for CoT in V2A becomes evident when examining the process of sound designers. They employ a multi-stage approach: analyzing visual content, then reasoning about acoustic properties, and finally synthesizing and refining sounds. End-to-end approaches compress the process into a single black-box transformation, losing the nuanced reasoning required for audio generation.

Some approaches have attempted to adopt MLLMs for multi-stage V2A. SonicVisionLM (Xie et al., 2024) converts video to textual captions using MLLMs and employs a separate model for text-to-audio (T2A) generation, which inevitably loses critical visual details and motion dynamics essential for realistic sound synthesis. More recently, DeepSound-V1 (Liang et al., 2025), while introducing MLLM-generated CoT for V2A, fragments the process into three separate tasks (audio generation, vocal removal, and silence detection) using specialized models rather than a unified framework. This fragmentation fails to fully leverage the deep understanding and reasoning capabilities that MLLMs could bring to comprehensive audio design.

To overcome these limitations and unlock the full reasoning potential of MLLMs for V2A, we present **ThinkSound** that harnesses CoT reasoning to enable stepwise, interactive generation and editing for V2A. ThinkSound decomposes audio generation into three intuitive and user-centric stages: (1) *foundational foley generation* to synthesize semantic and temporal matching soundscapes, (2) *interactive object-centric refinement* via user clicks, and (3) *targeted audio editing* guided by high-level natural language instructions. At each stage, an MLLM produces semantically and temporally aligned CoT instructions that guide a unified foundation model for audio generation, ensuring that the generated audio remains coherent, contextually grounded, and high quality throughout the workflow. Moreover, to support the training of ThinkSound and advance research in this area, we introduce **AudioCoT**, a comprehensive large-scale dataset with structured CoT reasoning annotations.

Technically, three key innovations are proposed: a) We fine-tune MLLMs on AudioCoT to generate structured, audio-specific reasoning chains that explicitly capture temporal dependencies, acoustic properties, and the decomposition of complex audio events. b) We design a unified audio foundation model based on flow matching that supports all three stages with the capabilities of synthesizing high-fidelity audio from arbitrary combinations of input modalities. This foundation model directly benefits from the detailed CoT reasoning provided by MLLMs, which effectively decomposes complex audio scenes into manageable components, enabling focused sound event generation while maintaining global coherence. c) We introduce a novel click-based interface that enables users to target specific visual objects for audio refinement, with CoT reasoning translating visual attention into contextually appropriate sound synthesis. The main contributions are summarized as follows:

- A novel three-stage interactive framework for V2A that progressively builds soundscapes through initial generation, object-centric refinement, and targeted editing, all unified by CoT reasoning from MLLMs.
- A unified multimodal foundation model capable of high-quality audio synthesis from an arbitrary combination of video, text, and audio inputs, leveraging CoT instructions to decompose complex scenes into manageable sound components.
- AudioCoT, a large-scale multimodal dataset with audio-specific CoT reasoning annotations that bridges visual content, textual descriptions, and sound synthesis.
- Experimental results demonstrate that ThinkSound achieves the state-of-the-art performance across objective metrics and subjective metrics, highlighting the effectiveness of our reasoning-guided generation.

2 Related Work

2.1 Video-to-Audio Generation

V2A (Cheng et al., 2024a; Wang et al., 2024c; Xu et al., 2024; Liu et al., 2025) focuses on synthesizing audio that aligns seamlessly with the visual content of a video clip. Earlier work (Luo et al., 2024;

Xing et al., 2024; Zhang et al., 2024; Viertola et al., 2024) try to generate audio samples based on silent video only using latent diffusion models (Rombach et al., 2022) or language models (Floridi & Chiriatti, 2020). For example, Diff-Foley (Luo et al., 2024) employs an audio-visual contrastive feature and latent diffusion to predict the spectrogram latent, while FoleyGen uses autoregressive techniques. However, recent research (Chen et al., 2024; Mo et al., 2024; You et al., 2025) pay more attention to generating audios using multimodal control, including videos, audios, and texts. For example, MMAudio (Polyak et al., 2024) uses flow matching conditioned on multi-modal inputs, including videos and texts. MultiFoley (Wu et al., 2024) further uses audio context as an additional input. Despite these advancements, existing work struggles to reason beyond simple object recognition and fails to conduct a step-by-step interactive process with users for audio generation. In this work, we propose ThinkSound, a V2A model with CoT reasoning in MLLMs, which supports step-by-step and interactive audio generation and editing.

2.2 Large Language Models and Reasoning

LLMs (Liu et al., 2024a; Guo et al., 2025; Hurst et al., 2024) have demonstrated remarkable reasoning capabilities through CoT prompting, enabling complex problem decomposition via intermediate reasoning steps. This paradigm, pioneered by (Wei et al., 2022), has been extended to MLLMs (Alayrac et al., 2022; Yang et al., 2023; Chu et al., 2024) that integrate visual, audio, and textual understanding through cross-modal alignment. Recent works (Rubenstein et al., 2023; Li et al., 2023; Wu et al., 2024) have explored MLLMs' potential in multimodal reasoning, particularly in visual-audio-textual grounding and cross-modal causal reasoning. Despite these, MLLMs remain under-explored for audio generation. While models like SonicVisionLM (Xie et al., 2024) and DeepSound-V1 (Liang et al., 2025) incorporate V2A, they lack mechanisms to decompose user intent into semantic and temporal reasoning steps. Our ThinkSound proposes a novel framework for audio generation and editing that decomposes the complex task into foundation Foley generation, interactive object-centric refinement, and targeted audio editing (Wang et al., 2023; Liu et al., 2024d).

2.3 Flow Matching

Flow Matching has emerged as a powerful alternative to diffusion models for high-quality audio synthesis (Evans et al., 2024; Liu et al., 2024b, 2023a), directly learning continuous normalizing flows between distributions by training a time-dependent vector field. Recent works Lipman et al. (2022); Liu et al. (2022) establish its theoretical foundations, while Liu et al. (2024c) demonstrated its advantages for audio generation with improved quality and faster sampling. Polyak et al. (2024) further advances this approach by applying conditional flow matching to multimodal audio generation, showing its effectiveness in maintaining temporal coherence across complex audio signals. Our work extends these advances through CoT-guided generation and a novel multi-guidance strategy that integrates control signals from multiple modalities. Our modality-agnostic training approach with classifier-free guidance dropout allows the model to generate high-quality audio from flexible input combinations, addressing both control precision and data scarcity challenges that have limited previous approaches.

3 AudioCoT Dataset for CoT-Guided Generation and Editing

3.1 Multimodal Data Sources

The AudioCoT dataset comprises both video-audio and audio-text pairs. For video-audio data, we utilize VGGSound (Chen et al., 2020) and a curated, non-speech subset of AudioSet (Gemmeke et al., 2017) to ensure broad coverage of real-world audiovisual events. For audio-text data, we aggregate pairs from AudioSet-SL (Hershey et al., 2021), Freesound (Fonseca et al., 2017), AudioCaps (Kim et al., 2019), and BBC Sound Effects [2], resulting in a diverse and representative corpus for training multimodal models. Further details on data processing and statistics are provided in the Appendix A.1.

[2]https://sound-effects.bbcrewind.co.uk/

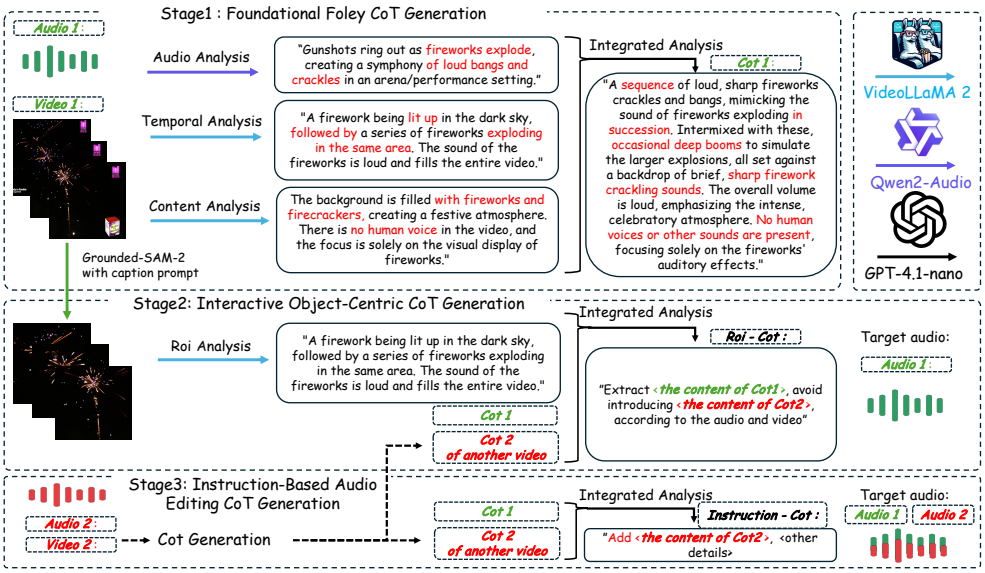

Figure 2: Overview of our AudioCoT dataset construction pipeline.

3.2 Automated CoT Generation Pipeline

Stage 1: Foundational Foley CoT Generation. To construct the Foley generation CoT data, we employ a multi-stage processing pipeline. For video-audio pairs, we utilize VideoLLaMA2 (Cheng et al., 2024b) to extract both temporal and semantic information from the videos through different prompting strategies, while for audio-text pairs (which lack video data), we employ a more streamlined approach without VideoLLaMA 2. In both cases, we generate audio descriptions using Qwen2-Audio (Chu et al., 2024), combining them with existing video captions or text annotations. The collected information is then integrated using GPT-4.1-nano[3] to synthesize comprehensive CoT reasoning chains that capture the complex relationships between content conditions and the corresponding audio elements, ensuring both data types contribute to a comprehensive understanding of audio generation reasoning.

Stage 2: Interactive Object-Centric CoT Generation. To facilitate object-focused audio generation, we develop a Region of Interest (ROI) extraction framework leveraging Grounded SAM2 (Ren et al., 2024; Ravi et al., 2024). Using audio captions as prompts, we identify and generate bounding boxes for potential sound-emitting objects. These coordinates are tracked temporally across video frames, while VideoLLaMA2 provides detailed semantic descriptions for each ROI segment. For complex audio manipulations (extraction/removal), we construct a hierarchical reasoning structure where the CoT of the target video is merged with another video's CoT to establish a global context. This composite representation is then integrated with ROI-specific generation information and processed by GPT-4.1-nano to formulate coherent manipulation rationales (detailed in Appendix A.2).

Stage 3: Instruction-Based Audio Editing CoT Generation. For instruction-guided audio editing, we analyze and integrate CoT information from Stage 1 based on four primary operations: extension, inpainting, addition, and removal. These operations address scenarios from extending sequences to eliminating unwanted segments. GPT-4.1-nano processes this integrated information to generate instruction-specific CoT reasoning chains while we perform corresponding audio operations, creating (Instruction-CoT, input audio, output audio) triplets for model training and evaluation.

4 ThinkSound

4.1 Overview

As depicted in Figure 3, ThinkSound introduces a novel step-by-step, interactive framework for audio generation and editing guided by CoT reasoning. Our approach decomposes the complex V2A task into three intuitive stages: (1) foundation Foley generation that creates a semantic and temporal

[3] https://openai.com/index/gpt-4-1/

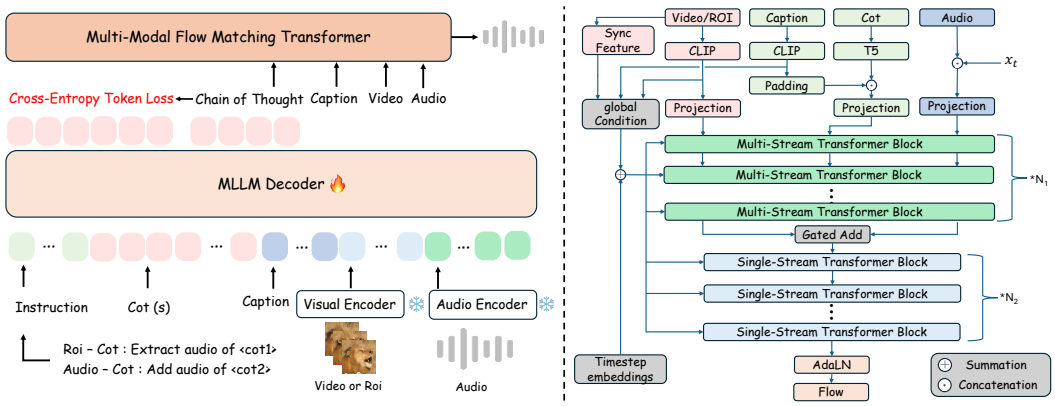

Figure 3: Overview of the ThinkSound architecture. **Left:** our Multimodal LLM framework, where a fine-tuned VideoLLaMA 2 model generates CoT reasoning for audio generation and editing. **Right:** our enhanced Multimodal Transformer architecture, which features an MM-DiT backbone with dedicated pathways for processing multimodal inputs and CoT-driven conditioning to enable high-fidelity, contextually grounded audio generation.

matching soundscape, (2) interactive region-based refinement through user clicks, and (3) targeted audio editing based on high-level instructions. At each stage, an MLLM generates CoT reasoning that guides a unified audio foundation model to produce and refine the soundtrack.

4.2 CoT Reasoning with Multimodal LLM

To enable stepwise, context-aware audio generation, we leverage VideoLLaMA2 (Cheng et al., 2024b) as the core multimodal reasoning engine. VideoLLaMA2 is selected for its state-of-the-art capability in fusing video, text, and audio modalities, and its advanced spatial-temporal modelling is essential for capturing the nuanced interplay between visual events and their corresponding acoustic manifestations.

We further adapt VideoLLaMA2 to the audio reasoning domain by fine-tuning it on our AudioCoT dataset, which contains rich, annotated reasoning chains tailored for audio-visual tasks. This fine-tuning process is designed to equip the model with three core competencies: (1) **audio-centric understanding**–the ability to infer acoustic properties, model sound propagation, and reason about audio-visual correspondences, including temporal and causal relationships among audio events (e.g., "footsteps occur before the door opens, then conversation ensues"); (2) **structured CoT decomposition**–the capacity to break down complex audio generation or editing tasks into explicit, actionable steps; and (3) **multimodal instruction following**–robustly interpreting and executing diverse generation or editing instructions across modalities. As illustrated in Figure 3, the fine-tuning objective is the standard cross-entropy loss for next-token prediction. Through this targeted adaptation, VideoLLaMA2 is transformed into a specialized audio reasoning module, capable of producing contextually precise CoT instructions that drive each stage of the ThinkSound pipeline.

4.3 CoT-Guided Unified Audio Foundation Model

The core of ThinkSound is our unified audio foundation model that seamlessly translates CoT reasoning into high-quality audio, whose detail is shown in the right part of Figure 3. We encode audio into latent representations using a pre-trained VAE (Pinheiro Cinelli et al., 2021) and train our model with conditional flow matching (Lipman et al., 2022), where the velocity field prediction is conditioned on multimodal context, including visual content, CoT reasoning, text captions, and audio context. To support any combination of input modalities, we adopt the integration of classifier-free guidance dropout during training. By randomly dropping different modality combinations with probability p_{drop}, we enable the model to handle arbitrary input configurations at inference time—essential for our interactive framework. We also incorporate strategic audio context masking to support advanced editing operations such as inpainting and extension.

For text processing, we employ a dual-pathway encoding strategy: MetaCLIP (Xu et al., 2023) encodes visual captions to provide scene-level context, while T5-v1-xl (Raffel et al., 2020) processes structured CoT reasoning to capture detailed temporal and causal relationships. These complementary

representations are effectively combined, with MetaCLIP's features serving as global conditioning signals while T5's outputs enable fine-grained reasoning-driven control.

Our enhanced MM-DiT architecture builds on recent advances in multimodal generative modeling (Esser et al., 2024; Labs, 2024; Cheng et al., 2024a) with three key components: (1) we implement a hybrid transformer backbone that alternates between modality-specific and shared processing. Multi-stream transformer blocks maintain separate parameters for each modality while sharing attention mechanisms, allowing efficient processing of diverse inputs without sacrificing cross-modal learning. (2) We design an adaptive fusion module that upsamples video features and fuses them with audio latents via a gated mechanism (Cho et al., 2014). This not only highlights salient visual cues and suppresses irrelevant information, but also ensures that video information is directly involved in subsequent single-stream transformer blocks. By integrating video into the audio latent space, the model can better capture subtle visual details and their nuanced effects on the soundscape, enabling richer cross-modal reasoning than using audio latents alone. (3) We implement global conditioning by mean-pooling CLIP features from both the caption and video, and, following MMAudio (Cheng et al., 2024a), incorporating sync features to improve audio-visual temporal alignment. The resulting global condition is added to the timestep embedding and injected via adaptive layer normalization layers (AdaLN) (Peebles & Xie, 2023) into both multi-stream and single-stream blocks.

4.4 Step-by-Step CoT-Guided Audio Generation and Editing

By enabling flexible combinations of input modalities along with CoT, ThinkSound supports decomposing audio generation into three intuitive stages shown in Figure 1. This three-stage pipeline enables progressively refined, highly customized audio generation through an intuitive interactive workflow, with CoT reasoning bridging user intent and audio synthesis at each step.

Stage 1: CoT-Guided Foley Generation. In the first stage, our framework analyzes the entire video to identify acoustic elements and their relationships. The fine-tuned MLLM generates detailed CoT reasoning that explicitly identifies primary sound events, ambient elements, acoustic properties, and their temporal dependencies - determining when objects make sounds and how these sounds interact. This structured reasoning guides the audio foundation model to synthesize high-fidelity audio that precisely matches both the semantic content and temporal dynamics of the visual scene. By decomposing complex audio scenes into explicit sound components through CoT reasoning, the model generates a diverse and coherent soundscape that captures subtle visual cues and motion dynamics essential for realistic audio synthesis.

Stage 2: Interactive Object-Focused Audio Generation. Stage 2 introduces an interactive framework that enables users to refine the initial soundscape by focusing on specific visual elements. Through a simple click-based interface, users can select objects of interest for audio enhancement. Unlike the holistic approach in Stage 1, this object-centric refinement leverages the segmented ROI to guide targeted audio synthesis. The fine-tuned MLLM analyzes the selected ROI and generates specialized CoT reasoning focused on the object's acoustic properties within the global context. This structured reasoning conditions the audio foundation model to synthesize object-specific sounds, seamlessly integrating them with the initial soundtrack produced in Stage 1. Notably, in this stage, the foundation model incorporates the existing audio context as an additional conditioning signal.

Stage 3: Instruction-based Audio Editing. In the final stage, users can provide high-level editing instructions to refine audio quality or modify specific elements. The MLLM translates these natural language instructions into precise audio processing operations through CoT reasoning, considering both visual content and current audio state. The foundation model, conditioned on both this reasoning and the existing audio context, applies targeted modifications while maintaining overall coherence. This natural language-driven approach enables non-technical users to perform sophisticated audio manipulation, including adding sounds, removing sounds, audio inpainting, and audio extension.

5 Experiments

5.1 Experiment Setup

Evaluation Metrics We conduct comprehensive evaluations using both objective and subjective metrics to assess audio quality, text-audio alignment, and video-audio synchronization. For objective

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_{PaNNs} ↓	DeSync↓	$CLAP_{cap}$ ↑	$CLAP_{CoT}$ ↑	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

evaluation, we compute the **Fréchet Distance (FD)** (Kilgour et al., 2018) in the OpenL3 feature space (Cramer et al., 2019; Evans et al., 2024) to measure distribution-level similarity, which we extend to evaluate stereo audio. We also use the **Kullback-Leibler (KL) Divergence** (Copet et al., 2024) based on predictions from PaSST model (KL_{PaSST}) and PaNNs (KL_{PaNNs}) to evaluate label-level consistency. Temporal alignment between audio and video is assessed using **DeSync** predicted by the Synchformer model (Iashin et al., 2024), while semantic alignment with text is measured using the **CLAP score** (Wu* et al., 2023; Chen et al., 2022), including both $CLAP_{cap}$ (caption) and $CLAP_{CoT}$ (CoT). For subjective evaluation, we collect human ratings using the **Mean Opinion Score (MOS)** to evaluate perceived audio quality (**MOS-Q**) and alignment with video and CoT (**MOS-A**). Further details are provided in the Appendix C.

Implementation Details For VAE training, we initialize our VAE using the VAE model weights trained on stereo data at 44.1kHz sample rate provided by Stability AI [4]. We employ mixed precision training with a batch size of 144 across 24 A800 GPUs for 500,000 steps. Subsequently, following Evans et al. (2024), we freeze the VAE encoder and train the VAE decoder with a latent mask ratio of 0.1 for an additional 500,000 steps. We use AdamW (Loshchilov & Hutter, 2019) as the optimizer, setting the generator learning rate to 3e-5 and the discriminator learning rate to 6e-5. In the foundation model training phase, we utilize an exponential moving average and automatic mixed precision for 100,000 steps on 8 A100 GPUs, with an effective batch size of 256. We adopt a cfg dropout of 0.2 for each modality with a learning rate of 1e-4. During the task-specific fine-tuning stage, we similarly apply exponential moving average and automatic mixed precision for 50,000 steps on 8 A100 GPUs, maintaining an effective batch size of 256. AdamW remains our optimizer of choice, with a learning rate set at 1e-4. We attach the benchmark details of VGGSound, Movie Gen Audio Bench, and AudioCoT test set for stages 2 and 3 into Appendix A.3.

5.2 Main Results

Video-to-Audio Generation Results We compare our foundation model against existing video-to-audio baselines including Seeing and Hearing (Xing et al., 2024), V-AURA (Viertola et al., 2024), FoleyCrafter (Zhang et al., 2024), Frieren (Wang et al., 2024d), V2A-Mapper (Wang et al., 2024a), and MMAudio (Cheng et al., 2024a). From Table 1, we observe three key findings: (1) GT audio exhibits low $CLAP_{cap}$ scores (0.28), revealing that original VGGSound captions inadequately capture the semantic content and temporal relationships needed for high-fidelity audio generation. (2) ThinkSound outperforms all baselines across most objective metrics and all subjective metrics. Compared to the strongest baseline (MMAudio), our model achieves substantial improvements in audio quality (KL_{PaSST}: 1.52 vs. 1.65) and semantic alignment ($CLAP_{CoT}$: 0.46 vs. 0.40), while maintaining comparable temporal synchronization. Subjective evaluations further confirm these improvements (MOS-Q: 4.02 vs. 3.84, MOS-A: 4.18 vs. 3.97). (3) Removing CoT reasoning notably degrades both audio quality and alignment metrics, especially for $CLAP_{CoT}$, decreasing from 0.46 to 0.41, confirming that CoT provides crucial information about sound events, their temporal relationships, and acoustic characteristics.

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**

[4] https://github.com/Stability-AI/stable-audio-tools

Furthermore, we conduct an out-of-distribution evaluation on the MovieGen Audio Bench (Polyak et al., 2024) to assess the generalization capability of our model. The results in Table 2 show that our ThinkSound model still achieves the best CLAP_{CoT} score of 0.51 while DeSync is on par with the best baseline. For subjective evaluation, ThinkSound achieves the best performance in both alignment and fidelity metrics. This demonstrates that ThinkSound exhibits strong generalization capability across different scenarios.

Object-Focused Audio Generation and CoT-Guided Audio Editing For object-focused generation, we compare two approaches: (1) MMAudio, which does not have the ROI design, and (2) ThinkSound w/o CoT reasoning. Results in Table 3 show that our interactive approach with CoT reasoning achieves significantly better results, demonstrating superior object-specific sound quality and integration with the foundation audio.

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

For text-guided audio editing, we compare ThinkSound with AudioLDM-2 (Liu et al., 2024e) and Edit Friendly DDPM (Huberman-Spiegelglas et al., 2024), adapting these models to our experimental setup. As shown in Table 4, ThinkSound outperforms the baselines across all objective and subjective metrics. Specifically, ThinkSound achieves the lowest FD (34.78) and KL_{PaSST} (1.45), and the highest CLAP score (0.51), indicating better audio fidelity and semantic relevance. In human evaluation, ThinkSound also gets the best MOS-Q and MOS-A. Removing CoT reasoning clearly drops all metrics, showing the importance of CoT-guided reasoning for text-based audio editing.

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

5.3 Ablation Studies

To better understand the contribution of each component in ThinkSound and to validate the effectiveness of our design choices, we conduct comprehensive ablation studies on the VGGSound test set. Mainly focus on: 1) text encoding strategies and 2) multi-modal integration mechanisms. For more ablation and exploratory results, we attach them to Appendix D.

Text Encoding Strategies. We evaluate different text encoding strategies with or without CoT reasoning, and the results are compiled in Table 5. The results show that (1) CoT reasoning substantially improves audio fidelity, with the FD score improving from 39.84 to 37.65 when comparing CLIP-only to T5-based CoT approaches. (2) The integration of contrastive features from CLIP with contextual features from T5 further enhances performance, reducing KL divergence metrics from 1.54 to 1.52 (KL_{PaSST}) and from 1.35 to 1.32 (KL_{PaNNs}).

Table 5: Comparison of text encoder fusion strategies. The input of the CLIP text encoder is the caption. The CLAP score means the value of CLAP_{CoT}.

Method	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	CLAP ↑
CLIP	39.84	1.59	1.40	0.48	0.41
T5 (CoT)	37.65	1.54	1.35	0.46	0.44
CLIP+T5	**34.56**	**1.52**	**1.32**	**0.46**	**0.46**

Multi-Modal Integration Mechanisms. We investigate different multi-modal integration mechanisms for video and audio before feeding them into the single-stream transformer, and the results are displayed in Table 6. We find that (1) the element-wise addition of video and audio features performs better than audio-only input with weak-supervision global conditioning, especially for synchronization metric DeSync,

Table 6: Comparison of different multi-modal integration mechanisms between video and audio features.

Integration	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	CLAP ↑
audio only	37.13	1.58	1.37	0.50	0.43
linear video	38.96	1.58	1.38	0.46	0.45
gated video	**34.56**	**1.52**	**1.32**	**0.46**	**0.46**

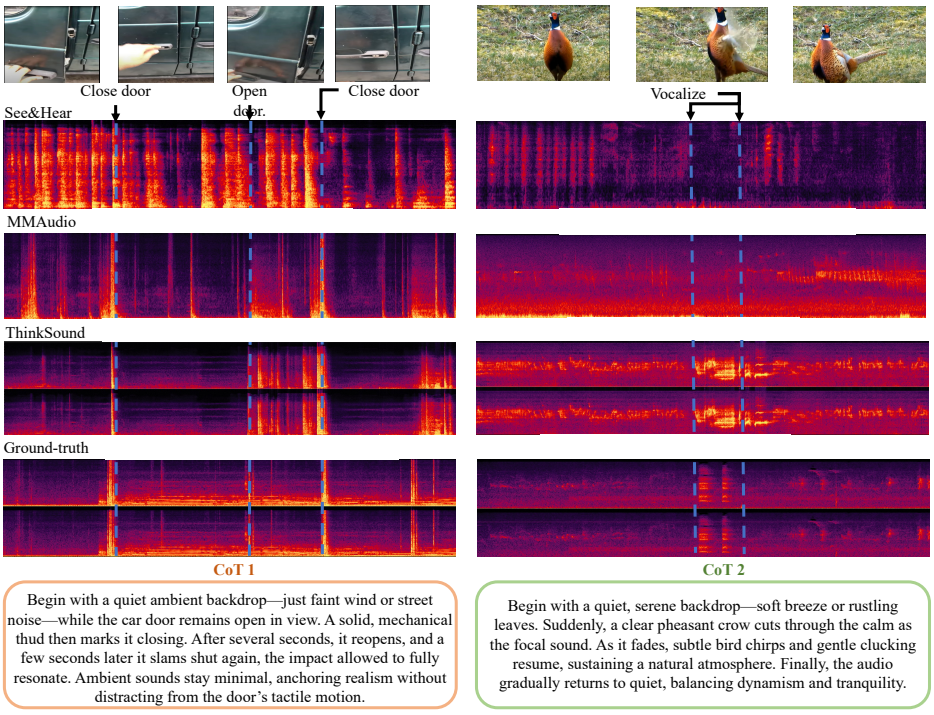

Figure 4: Qualitative Comparison: **Left:** Spectrograms for a car door movement sequence (closed → opened → closed), showing ThinkSound's precise alignment of each door sound versus the baseline's premature opening effect. **Right:** Spectrograms for a grassy-field pheasant scene (ambient bird calls → wing-flap chirp → ambient calls), illustrating ThinkSound's accurate detection and timing of the transient chirp compared to the baseline's omission or delay.

decreasing from 0.50 to 0.46. (2) The gated fusion mechanism outperforms simple element-wise addition and audio-only input across all metrics.

5.4 Case Study

In our qualitative analysis, we compare spectrograms of audio generated by ThinkSound and those produced by baselines, as illustrated in Figure 4. We make the following observations: (1) As demonstrated in case 1, a car door sequence—closed–opened–closed—is accurately reflected by ThinkSound, whereas the baseline models erroneously introduce an extra opening sound at the start. This highlights ThinkSound's ability, via CoT prompting, to track temporal and causal event order in structured scenes. (2) In case 2, a pheasant moving in a grassy field—accompanied by ambient bird calls, suddenly flapping its wings and emitting a sharp chirp before returning to background noise—is faithfully reproduced by ThinkSound. The baselines, however, often miss or delay this brief chirp. These cases underscore ThinkSound's enhanced temporal reasoning and sensitivity to subtle visual cues, resulting in more precise and context-aware audio synthesis.

6 Conclusion

This paper presented ThinkSound, a novel CoT reasoning framework for audio generation and editing that decomposed complex audio generation into three interpretable and user-centric stages. Our evaluations showed that ThinkSound outperformed state-of-the-art methods, producing contextually appropriate and temporally precise soundscapes. The framework's interactive nature enabled users to refine generated audio through intuitive pointing and natural language instructions, addressing the gap between creative intent and automated generation. While challenges remained in capturing nuanced acoustic properties, our AudioCoT dataset and training strategy established a foundation for future research in intelligent audio generation with applications across film, gaming, and social media. In future work, we will explore incorporating physical acoustic modeling and developing more sophisticated reasoning capabilities for complex multi-object sound interactions.

Acknowledgements

The research was supported by the NSFC (No. 62206234) of Mainland China.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. *arXiv preprint arXiv:2004.14368*, 2020.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.

Chen, Z., Seetharaman, P., Russell, B., Nieto, O., Bourgin, D., Owens, A., and Salamon, J. Video-guided foley sound generation with multimodal controls. *arXiv preprint arXiv:2411.17698*, 2024.

Cheng, H. K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., and Mitsufuji, Y. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*, 2024a.

Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024b.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Cramer, A. L., Wu, H.-H., Salamon, J., and Bello, J. P. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856. IEEE, 2019.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Evans, Z., Carr, C., Taylor, J., Hawley, S. H., and Pons, J. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

Floridi, L. and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Fonseca, E., Pons, J., Favory, X., Font, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. Freesound datasets: A platform for the creation of open audio datasets. In Cunningham, S. J., Duan, Z., Hu, X., and Turnbull, D. (eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pp. 486–493, 2017.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hershey, S., Ellis, D. P., Fonseca, E., Jansen, A., Liu, C., Moore, R. C., and Plakal, M. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370. IEEE, 2021.

Huberman-Spiegelglas, I., Kulikov, V., and Michaeli, T. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12469–12478, 2024.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Iashin, V., Xie, W., Rahtu, E., and Zisserman, A. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Koutini, K., Schlüter, J., Eghbal-Zadeh, H., and Widmer, G. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

Labs, B. F. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

Liang, Y., Chen, Z., Ding, C., and Di, X. Deepsound-v1: Start to think step-by-step in the audio generation from videos, 2025. URL `https://arxiv.org/abs/2503.22208`.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, H., Huang, R., Lin, X., Xu, W., Zheng, M., Chen, H., He, J., and Zhao, Z. Vit-tts: visual text-to-speech with scalable diffusion transformer. *arXiv preprint arXiv:2305.12708*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Liu, H., Huang, R., Liu, Y., Cao, H., Wang, J., Cheng, X., Zheng, S., and Zhao, Z. Audiolcm: Efficient and high-quality text-to-audio generation with minimal inference steps. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7008–7017, 2024b.

Liu, H., Wang, J., Huang, R., Liu, Y., Lu, H., Xue, W., and Zhao, Z. Flashaudio: Rectified flows for fast and high-fidelity text-to-audio generation. *arXiv preprint arXiv:2410.12266*, 2024c.

Liu, H., Wang, J., Huang, R., Liu, Y., Xu, J., and Zhao, Z. MEDIC: zero-shot music editing with disentangled inversion control. *CoRR*, abs/2407.13220, 2024d. doi: 10.48550/ARXIV.2407.13220. URL `https://doi.org/10.48550/arXiv.2407.13220`.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024e.

Liu, H., Luo, T., Jiang, Q., Luo, K., Sun, P., Wan, J., Huang, R., Chen, Q., Wang, W., Li, X., et al. Omniaudio: Generating spatial audio from 360-degree video. *arXiv preprint arXiv:2504.14906*, 2025.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Luo, S., Yan, C., Hu, C., and Zhao, H. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Mo, S., Shi, J., and Tian, Y. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*, 2024.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A., and Lima Netto, S. Variational autoencoder. In *Variational methods for machine learning with applications to deep networks*, pp. 111–149. Springer, 2021.

Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.-Y., Chuang, C.-Y., et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., Xiong, Y., Zhang, H., Li, F., Tang, P., Yu, K., and Zhang, L. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., Quitry, F. d. C., Chen, P., Badawy, D. E., Han, W., Kharitonov, E., et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Viertola, I., Iashin, V., and Rahtu, E. Temporally aligned audio for video with autoregression. *arXiv preprint arXiv:2409.13689*, 2024.

Wang, H., Ma, J., Pascual, S., Cartwright, R., and Cai, W. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 15492–15501, 2024a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wang, X., Wang, Y., Wu, Y., Song, R., Tan, X., Chen, Z., Xu, H., and Sui, G. Tiva: Time-aligned video-to-audio generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 573–582, 2024c.

Wang, Y., Ju, Z., Tan, X., He, L., Wu, Z., Bian, J., et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 71340–71357, 2023.

Wang, Y., Guo, W., Huang, R., Huang, J., Wang, Z., You, F., Li, R., and Zhao, Z. Frieren: Efficient video-to-audio generation network with rectified flow matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024d. URL https://openreview.net/forum?id=prXfM5X2Db.

Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., and Fei, H. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

Wu*, Y., Chen*, K., Zhang*, T., Hui*, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

Xie, Z., Yu, S., He, Q., and Li, M. Sonicvisionlm: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26866–26875, 2024.

Xing, Y., He, Y., Tian, Z., Wang, X., and Chen, Q. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 7151–7161. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00683. URL https://doi.org/10.1109/CVPR52733.2024.00683.

Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

Xu, M., Li, C., Tu, X., Ren, Y., Chen, R., Gu, Y., Liang, W., and Yu, D. Video-to-audio generation with hidden alignment. *arXiv preprint arXiv:2407.07464*, 2024.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., and Wang, L. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

You, Y., Wu, X., and Qu, T. Ta-v2a: Textually assisted video-to-audio generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zhang, Y., Gu, Y., Zeng, Y., Xing, Z., Wang, Y., Wu, Z., and Chen, K. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accurately reflect the paper's contributions and scope in the abstract and section 1 introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed limitations in section E Limitation and Future.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide theory assumptions and proofs in section 4

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we will open-source our codes(contains model or their access method), which is enough to reproduce our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: we specify all the training and test details

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We only use one fixed random seed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources in training details

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed Positive impacts in section 6 Conclusion and negative impacts in section F

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We take actions for safeguards , which is described in section G Safeguards

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Creators or original owners of assets, used in the paper, are properly credited and the license and terms of use explicitly mentioned and are properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our assets is well documented and the document is provided

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We carefully follow the LLM policy and fully describe the usage of LLMs

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A AudioCoT Dataset Details

A.1 Data Collection and Preprocessing

To ensure high data quality and consistency, we employ a comprehensive preprocessing pipeline. We begin by removing silent audio-video clips to retain only those with meaningful content. For the AudioSet subset (Gemmeke et al., 2017), we further exclude segments containing human voices based on tag information, as our focus is on non-speech audio. All audio-video clips are then segmented into fixed-length intervals of 9.1 seconds, with any shorter clips discarded to maintain uniformity. To achieve a balanced dataset, we maintain an approximately 1:1 ratio between music and sound effect samples, ensuring equal representation of both categories.

A.2 Quality Control for Automated Data Pipeline

To enhance the effectiveness of CoT reasoning in preserving audio characteristics while integrating visual context, we implemented a comprehensive multi-stage quality control pipeline:

Stage 1: Audio-Text Alignment Filtering We employ a systematic approach to ensure high-quality audio-CoT pairs. First, we calculate the CLAP score between each audio sample and its corresponding CoT description to quantify semantic alignment. For pairs exhibiting low CLAP scores (below 0.2), we regenerate the CoT using an enhanced prompt specifically designed to emphasize audio characteristics and features. After regeneration, we recalculate the CLAP score to assess improvement. Audio samples that continue to demonstrate poor alignment (persistent low CLAP scores) are excluded from the dataset to maintain quality standards.

Stage 2: Object Tracking Consistency To ensure reliable audio-visual correspondence, we retain only those video sequences containing at least one Region of Interest (ROI) that remains consistently visible throughout the entire duration. Videos with objects that disappear from view or exhibit inconsistent tracking are filtered out. This criterion ensures that our dataset maintains high-quality visual references for audio generation tasks, providing consistent visual anchors for the audio generation process.

Stage 3: Semantic Pairing of Audio Components For tasks requiring paired audio components, we utilize GPT-4.1-nano to analyze tag categories from VGGSound (Chen et al., 2020) based on two critical criteria. First, we ensure semantic distinctiveness, where tags must be sufficiently distinct to avoid confusion during audio extraction and removal tasks. Second, we verify contextual plausibility, ensuring that the co-occurrence of paired sounds is contextually reasonable within the same acoustic scene. This balanced approach ensures that our audio pairs are both semantically meaningful and practically useful for audio generation tasks.

Human Verification Protocol To validate our automated filtering processes, we implement a rigorous manual review at each pipeline stage. No less than 5% of the total data volume undergoes human inspection to ensure quality. This verification step helps validate the effectiveness of our automated filtering criteria and ensures the overall reliability and quality of our dataset. The human reviewers assess both the technical aspects of alignment and the perceptual quality of the audio-visual correspondence. When samples fail human verification, they are immediately removed from the dataset. Additionally, if the human rejection rate for any filtering criterion exceeds 5%, we recalibrate the corresponding automated filtering parameters and reprocess the entire batch to maintain dataset integrity. This feedback loop between automated filtering and human verification ensures continuous improvement of our quality control pipeline.

A.3 Benchmark Construction

We evaluate the performance of ThinkSound on three different tasks: video-to-audio generation, object-focused audio generation, and audio editing. For the video-to-audio generation task, we use the VGGSound test set as the in-distribution evaluation set while the MovieGen Audio Bench is the out-of-distribution evaluation set. For the VGGSound test set, we use the same quality filtering protocol as our training data preparation. Given that our primary focus is on video-to-sound/music generation, we construct three different difficulty levels based on the complexity of the audio-visual

Table 7: Overview of datasets used in our work.

Dataset	Modality	Text Format	Hours
VGGSound (Chen et al., 2020)	Audio-Video	Caption	**453.6**
AudioSet (Gemmeke et al., 2017)	Audio-Video	Caption	287.5
AudioSet-SL (Hershey et al., 2021)	Audio-Text	Caption	262.6
Freesound (Fonseca et al., 2017)	Audio-Text	Caption	**1286.6**
AudioCaps (Kim et al., 2019)	Audio-Text	Caption	112.6
BBC Sound Effects [5]	Audio-Text	Tags	128.9
Total Hours	–	–	**2531.8**

relationships. Specifically, we distinguish the difficulty levels based on a multi-dimensional scoring system that evaluates:

- **Semantic Consistency**: The alignment between the audio and the visual content is evaluated by the ImageBind score (0.3+ for easy, 0.25-0.3 for medium, 0.2-0.25 for hard), and the CLAP score between audio and CoT (0.4+ for easy, 0.3-0.4 for medium, 0.2-0.3 for hard).
- **Temporal Synchronization**: The degree of synchronization between visual events and corresponding sounds evaluated by DeSync score (0-0.3 for easy, 0.3-0.6 for medium, 0.6+ for hard).
- **Acoustic Scene Complexity**: The audio events' numbers (one dominant sound for easy, 2-3 distinct sounds for medium, multiple overlapping sounds for hard)

According to the above evaluation criteria, we input the scores and criteria into GPT-4.1-nano to generate the difficulty level for each sample. The final difficulty assignment follows a tertile distribution: the lowest-scoring third is classified as "easy," the middle third as "medium," and the highest-scoring third as "hard." This stratified approach ensures balanced representation across difficulty levels while maintaining meaningful distinctions in task complexity. For each difficulty level, we construct a benchmark subset containing around 2000 samples.

For stages 2 and 3, we maintain methodological consistency with our training protocols while adapting the evaluation criteria to each task's specific requirements. For stage 2, we select samples with clearly identifiable visual objects that produce distinct sounds, while for stage 3, we focus on samples with different audio categories suitable for manipulation tasks. Each evaluation subset contains approximately 2,000 samples.

B Model Configurations and Architecture

B.1 Model Configurations

ThinkSound consists of two primary components: a hierarchical variational autoencoder (VAE) for audio compression and reconstruction, and a flow-matching multimodal transformer.

Variational Autoencoder The encoder consists of five convolutional blocks with channel multipliers [1, 2, 4, 8, 16] and strides [2, 4, 4, 8, 8], projecting the stereo waveform into a 128-dimensional latent space. The decoder mirrors this architecture with transposed convolutions to reconstruct 64-dimensional latent representations back into the time-domain waveform.

Multimodal Diffusion Transformer ThinkSound employs an enhanced Multi-modal Diffusion Transformer (MM-DiT) with a hidden size dimension of 1024. It comprises 14 multi-stream transformer layers and 7 single-stream transformer layers, with 16 attention heads. We further attach our different model scale parameters for reference. We use the large model by default.

B.2 Model Architecture

The architecture of multistream transformers is depicted in Figure 5.

Table 8: Diffusion Transformer Configurations at Different Model Sizes

Model Scale	Hidden Size	Depth	Attention Heads	Multistream Layers	Singlestream Layers	Total Parameters
Large	1024	21	16	14	7	1.3B
Medium	768	21	12	14	7	724M
Small	512	18	8	12	6	533M

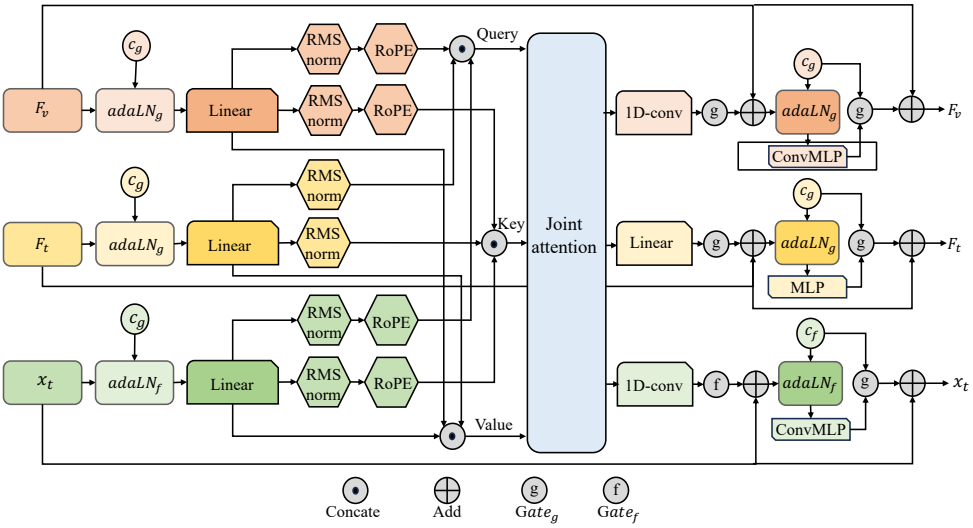

Figure 5: Multi-stream blocks: F_v is the video features, F_t is the text features, x_t is the audio latents, and c_g denotes the global condition.

C Evaluation

C.1 Objective Metrics

To comprehensively evaluate the generated audio, we adopt a set of objective metrics targeting different aspects: perceptual quality, semantic consistency, temporal alignment, and cross-modal correspondence.

Feature Distribution Alignment: We project both generated and reference audio into the OpenL3 embedding space Cramer et al. (2019); Evans et al. (2024) and compute the **Fréchet Distance (FD)** Kilgour et al. (2018); Copet et al. (2024) to assess the similarity between their distributional statistics. We chose OpenL3 because it accepts signals of up to 48kHz while VGGish operates at 16kHz, which is more suitable for our 44kHz audio. Following the previous work Evans et al. (2024), we extend the FD to evaluate stereo signals by projecting left and right channels separately and then averaging the results. Moreover, to evaluate whether the generated audio matches the reference in terms of its distribution, we compute the **Kullback-Leibler (KL) Divergence**Copet et al. (2024) between class probability distributions predicted by the PaSST modelKoutini et al. (2021) and PaNNs model (Kong et al., 2020) as classifiers.

Temporal Alignment: To evaluate the synchronization between generated audio and its corresponding video, we adopt the **DeSync** score predicted by the Synchformer model Iashin et al. (2024), following the practice of Cheng et al. (2024a). For each sample, we truncate the video to match the duration of the generated audio and compute the DeSync score using Synchformer, which operates with a 4.8-second context window. Specifically, we extract both the first and last 4.8-second segments from each video-audio pair, calculate the DeSync score for each segment, and report the average as the final temporal alignment metric. **Text-Audio Correspondence:** To assess the semantic alignment between generated audio and textual prompts, we utilize the **CLAP score** Wu* et al. (2023); Chen et al. (2022), which measures similarity in a shared audio-text embedding space. Specifically, we report CLAP_{cap} for evaluating alignment with the original video captions and CLAP_{CoT} for alignment

with our constructed CoT descriptions. As discussed in Section 5.2, the original VGGSound captions are often low quality and yield lower CLAP scores, whereas our CoT annotations provide richer semantic detail and achieve higher alignment. **Consequently, we primarily use the CoT-audio alignment metric in our evaluations, except in Table 1 where caption-based alignment is also reported.**

C.2 Subjective Metrics

Our subjective evaluation framework employs the Mean Opinion Score (MOS) methodology along two critical dimensions to comprehensively assess the generated audio:

Audio Quality Assessment (MOS-Q) We evaluate the intrinsic perceptual quality of generated audio through a rigorous assessment protocol where participants are instructed to focus on four specific aspects:

- *Clarity*: The absence of unwanted artifacts, distortions, or noise
- *Naturalness*: How realistic and non-synthetic the audio sounds
- *Fidelity*: The richness and accuracy of acoustic characteristics
- *Overall impression*: The holistic listening experience

Each listener rates these qualities using a standard 5-point Likert scale (1: Poor, 2: Fair, 3: Good, 4: Very Good, 5: Excellent). The final MOS-Q score for each audio sample represents the average rating across all evaluators, providing a robust measure of perceived audio quality.

Semantic and Temporal Alignment Assessment (MOS-A) To evaluate the cross-modal coherence between generated audio and visual content, we assess both semantic relevance and temporal synchronization (we also provide CoT text as the auxiliary information for semantic alignment):

- *Semantic alignment*: How well the audio content matches the objects, actions, and environment depicted in the video
- *Temporal alignment*: How accurately sound events correspond to visual events in time

Participants judge the alignment according to three categories on the same 5-point scale:

- *Fully aligned* (4-5 points): Complete semantic correspondence with precise temporal synchronization
- *Mostly aligned* (2.5-3.9 points): Good semantic match with occasional minor temporal misalignments
- *Partially aligned* (1-2.4 points): Noticeable discrepancies in either semantic content or temporal synchronization

Evaluation Protocol To ensure evaluation reliability and consistency, we implemented the following protocol:

- All assessments were conducted in controlled in-person sessions with standardized audio equipment
- 15 raters with normal hearing ability were recruited and briefed on the evaluation criteria
- Each rater evaluated a random subset of 50 video-audio pairs from our test collection
- Samples were presented in randomized order to prevent ordering bias
- Reference examples of each quality level were provided before the evaluation sessions
- Raters were given sufficient time to carefully evaluate each sample

D Additional Quantitative Results

D.1 Details on Video-to-Audio Comparison

For the results in Table 1, we reproduce the results of Seeing and Hearing (Xing et al., 2024), V-AURA (Viertola et al., 2024), FoleyCrafter (Zhang et al., 2024), and MMAudio (Cheng et al., 2024a) using the official code and pre-trained models. For the other baselines, we use the generated samples provided by the authors, i.e., Frieren, V2A-Mapper, and Movie Gen (Polyak et al., 2024). Furthermore, the CLAP_{cap} scores of MMAudio, MovieGen, and ThinkSound are 0.43, 0.44, and 0.49, respectively.

D.2 Impact of Model Size

We compare three model size of ThinkSound: **Large (1.3B)**, **Medium (724M)**, and **Small (533M)**. The Large model achieves the best performance across all metrics as shown in Table 9. These results demonstrate that model capacity significantly enhances audio quality and improves alignment with ground truth distribution. As model size decreases, performance degrades substantially, highlighting the necessity of adequate model capacity for effective audio generation.

Table 9: Impact of model size results.

Size	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	CLAP_{CoT} ↑
Small	43.26	1.64	1.39	0.50	0.43
Medium	37.62	1.56	1.34	0.47	0.44
Large	**34.56**	**1.52**	**1.32**	**0.46**	**0.46**

D.3 Performance across different difficulty levels

To better validate the performance of our CoT-Guided generation, we also report the results in the video-to-audio generation of different difficulty levels. We illustrate the construction of different difficulty levels in Section A A.3. The results are shown in Table 10, and we can conclude that (1) As expected, the performance of all models decreases as the difficulty level increases, and (2) Our CoT-Guided generation outperforms other baselines across all difficulty levels.

Table 10: Performance across different difficulty levels.

Difficulty	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	CLAP_{CoT} ↑
Easy	**31.32**	**1.35**	**1.16**	**0.42**	**0.52**
Medium	35.45	1.46	1.31	0.46	0.49
Hard	48.78	1.63	1.40	0.57	0.41

D.4 Performance Comparisons between coarse-grained and fine-grained CoT

To further validate the effectiveness of our fine-grained CoT, we compare the performance of our model with the coarse-grained CoT. The results are shown in Table 11. We can conclude that our fine-grained CoT outperforms the coarse-grained CoT across all metrics.

Table 11: Performance comparisons between coarse-grained and fine-grained CoT.

Granularity	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	CLAP_{CoT} ↑
Coarse	42.72	1.58	1.41	0.52	0.34
Fine	**34.56**	**1.52**	**1.32**	**0.46**	**0.46**

D.5 Effectiveness of T5 Encoder for CoT Structure

A key assumption of our work is that the T5 encoder can effectively capture the logical structure within the CoT reasoning steps. We provide both a theoretical rationale and empirical validation for this.

Theoretical Rationale While T5 is pre-trained on general web text (C4), its Transformer architecture is fundamentally designed to capture long-range dependencies and compositional structure. The reasoning steps in our CoT (e.g., "first do A, then B happens, which causes C") are expressed through natural language syntax and logical connectives. T5's contextual embeddings are well-suited to capture these structural and causal cues. The goal is not for T5 to "understand" reasoning in a human sense, but to produce a rich, structured conditioning signal that the downstream diffusion transformer can effectively leverage.

Empirical Validation To prove that the T5 encoder effectively utilizes the structure of the CoT, we conducted an ablation study comparing our full model against two degraded variants: (1) **Tags Only**, which provides the model with only extracted keywords (e.g., "car," "rain," "dog bark"), removing all reasoning structure; and (2) **Randomized CoT**, which randomly shuffles the sentences within the CoT, preserving all keywords but destroying the logical flow. As shown in Table 12, both ablations lead to a significant performance drop. This strongly indicates that the performance gain is attributable to the stepwise logical structure of the CoT—as encoded by T5—and not merely the presence of more keywords.

Table 12: Ablation study on the importance of CoT structure. We compare our full model against variants with only keywords (Tags Only) and shuffled reasoning sentences (Randomized CoT). Results show that the logical structure is critical for performance.

Setting	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	$CLAP_{CoT}$ ↑
Randomized CoT	40.52	1.56	1.35	0.51	0.43
Tags Only	39.35	1.60	1.38	0.50	0.42
Ours (Ordered CoT)	**34.56**	**1.52**	**1.32**	**0.46**	**0.46**

D.6 Impact of CoT Verbosity on Generation Quality

To mitigate the risk of verbose or hallucinated CoT text, we employ a two-pronged strategy: (1) rigorous quality control during the creation of our AudioCoT training data (see Appendix A.2), and (2) direct constraints on CoT generation during inference (e.g., ≤ 3 sentences, ≤ 77 tokens). To empirically validate this approach, we conducted an ablation study quantifying the impact of verbosity. We removed our length constraints to generate "Overly Detailed" CoT prompts and compared them against our proposed "Concise CoT." The results in Table 13 demonstrate a clear performance degradation with verbose reasoning, confirming that our dual strategy is effective and that concise reasoning is critical for achieving high-quality results.

Table 13: Impact of CoT verbosity. We compare our standard concise CoT against an overly detailed version generated without length constraints. Verbose reasoning leads to a clear performance degradation.

Setting	FD↓	KL_{PaSST} ↓	KL_{PaNNs} ↓	DeSync↓	$CLAP_{CoT}$ ↑
Overly-detailed CoT	43.56	1.61	1.55	0.54	0.35
Concise CoT (Ours)	**34.56**	**1.52**	**1.32**	**0.46**	**0.46**

D.7 Dedicated MLLM Reasoning Evaluation

The quality of the CoT reasoning is the linchpin of our framework. To substantiate our claims, we conducted a comprehensive evaluation of the generated CoT quality using both expert human annotators and automated LLM-based metrics.

Human Evaluation We randomly sampled 100 VGGSound test cases. Two expert annotators rated each generated CoT on a 0-1 scale across five dimensions: (1) multimodal integration, (2) specificity of audio details, (3) feasibility for audio generation, (4) logical consistency, and (5) brevity/format compliance. The final score is the sum (max 5).

LLM-Based Similarity Evaluation For large-scale analysis, we compared the generated CoTs to ground-truth CoTs using GPT-4-nano. The model was prompted to score similarity on a 0–5 scale, focusing on reasoning structure, causal/temporal relationships, and object-sound associations.

Results The results, presented in Table 14, provide strong empirical evidence that our fine-tuned MLLM (ThinkSound) greatly outperforms other powerful MLLMs in generating high-quality, structured CoT for video-to-audio reasoning.

Table 14: Evaluation of MLLM-generated CoT quality. Our model is compared against baselines using human evaluation and LLM-based similarity scoring, demonstrating superior performance in generating high-quality CoT.

Model	Human Score (0-5)	LLM CoT Similarity (0-5)
Qwen2.5-VL-7B	3.78	3.95
Qwen2-Audio-7B	3.82	4.09
ThinkSound (VideoLLaMA2)	**4.13**	**4.31**

E Limitation and Future

While the current MLLMs are capable of a strong understanding and reasoning of semantic information, they still have limitations in understanding the precise temporal and spatial information of video. For example, in the case of locating the exact timestamp of the sound event, MLLMs often fail to provide accurate results or provide wrong results. Moreover, the current open-source video-audio datasets for audio generation are limited in diversity and coverage, which may lack rare or culturally specific sound events. In the future, we will continue to explore more diverse and comprehensive datasets to improve the performance of our model. Furthermore, we will explore more effective methods to improve the temporal and spatial alignment of generated audio.

F Potential Negative Societal Impacts

ThinkSound carries potential risks if misused. Malicious actors could exploit the system to generate fake audio for synthetic media, thereby contributing to the spread of misinformation. Moreover, if the training data underrepresents certain cultures or environments, the model may unintentionally amplify biases—for instance, by reinforcing stereotypes or misassociating sounds with particular demographic groups.

F.1 Ethical Considerations

The dataset used in this research is strictly for academic and non-commercial purposes. We implemented several measures to ensure compliance with ethical standards, as follows.

- **Data Transparency and Anonymization.** We only provide ASR transcripts after rigorous text anonymization processes, visual features of video clips, our annotations, and links to the original videos, to ensure transparency regarding the data sources and their usage while maintaining anonymity.
- **Authorization.** Any personal data should be used only with express authorization, ensuring lawful and fair processing in accordance with applicable laws.

G Safeguards

We used a diverse training dataset covering a wide range of acoustic scenes to minimize reinforcing stereotypes or incorrect associations between sounds and specific demographic groups. The model will be released in stages to better assess its impact and improve safeguards. However, once the model is openly released, we cannot control how others use it. Therefore, we provide clear usage guidelines to encourage responsible use and help mitigate potential misuse.