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

Anonymous Author(s)
Affiliation
Address
email

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.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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.

Table 1: 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 [1]	Audio-Text	Tags	128.9
Total Hours	–	–	**2531.8**

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 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 evaluated by Imagebind score (0.3+ for easy, 0.25-0.3 for medium, 0.2-0.25 for hard), and 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 transformers.

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.

Table 2: 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

B.2 Model Architecture

The architecture of multistream transformers is depicted in Figure 1.

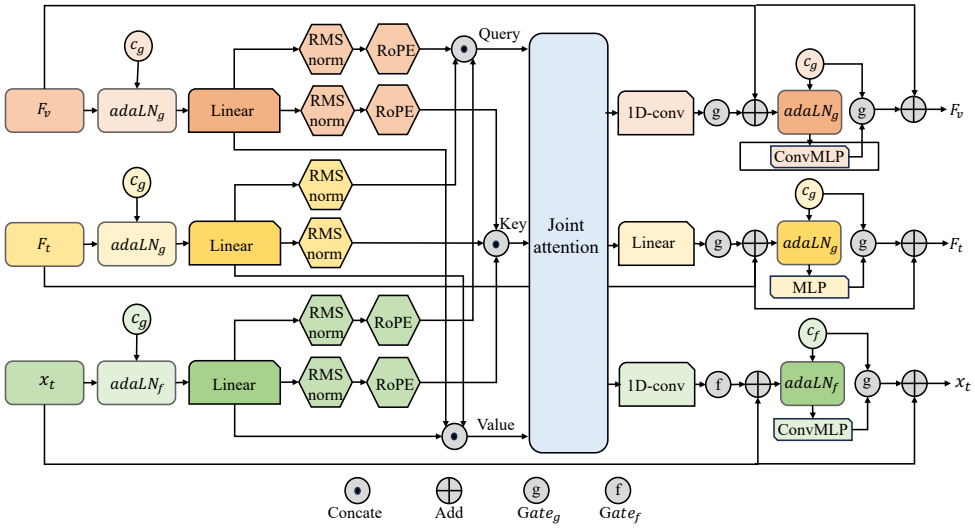

Figure 1: Multi-stream blocks: F_v is the video features, F_t is the text features, x_t is the audio latents, and c_g denotes the global condition.

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. (2024). 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., 2024) 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 3. 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 3: 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 4, 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 4: 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 5. We can conclude that our fine-grained CoT outperforms the coarse-grained CoT across all metrics.

Table 5: 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**

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. What's more, 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.

References

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.

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*, 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.

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

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.

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.

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.

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.

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

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.

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.