AURELIUS: RELATION AWARE TEXT-TO-AUDIO GENERATION AT SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

We present *Aurelius*, a new framework that enables relation aware text-to-audio (TTA) generation research at scale. Given the lack of essential audio event and relation corpora, *Aurelius* contributes a large-scale audio event corpus *AudioEventSet* and another large-scale relation corpus *AudioRelSet*. Comprising of 110 event categories, *AudioEventSet* maximally covers all commonly heard audio events and each event is unique, realistic and of high-quality. *AudioRelSet* consists of 100 relations, comprehensively covering the relations that present in the physical world or can be neatly described by text. As the two corpora provide audio event and relation independently, they can be combined to create massive `<text,audio>` pairs with our pair generation strategy to support relation aware TTA investigation at scale. We comprehensively benchmark all existing TTA models from both general and relation aware evaluation perspective. We further provide in-depth investigation on scaling up existing TTA models' relation aware generation by either training from scratch or leveraging cross-domain general TTA knowledge. The introduced corpora and the findings through investigation in this work potentially facilitate future research on relation aware TTA generation.

1 INTRODUCTION

Text-to-audio (hereinafter TTA) generation task aims at generating acoustically high-fidelity audio with the content inferred by the input text. Owing to the success of generative modeling (*e.g.*, diffusion based (Ho et al., 2020; Xue et al., 2024), score based (Vahdat et al., 2021) and flow matching based (Lipman et al., 2023; Guan et al., 2024) methods) and the availability of large `<text,audio>` pair dataset (*e.g.*, AudioCaps (Kim et al., 2019), AudioSet (Gemmeke et al., 2017)), we have witnessed significant advancement in general TTA task in recent years (Ghosal et al., 2023; Hung et al., 2024; Liu et al., 2024). Despite these achievements, the relation aware TTA generation still remains as a challenging task as it jointly requires audio event generation and relation modeling. Audio events and their relation are two fundamental elements humans rely on for holistic acoustic scene understanding or engaging communication (Zacks et al., 2007; Hirsh et al., 1967; Lake et al., 2015). We humans can interpret the relation and audio events within the textual description at an ease to decide how the target audio looks like. Enabling TTA models with similar relational reasoning and event interpretation capability is therefore essential for bridging the gap between relation aware TTA model quality and human-level crossmodal reasoning.

The recent preliminary investigation by RiTTA (He et al., 2025) already shows the incapability of existing TTA models in relation aware generation, but the investigation runs on top of small relation and audio event corpora. The data corpora small scale issue naturally hinders further investigation. To enable relation aware TTA at scale, we introduce *Aurelius*, a novel framework contributes to relation aware TTA from both dataset benchmark and technical methodology aspects. From the dataset benchmark aspect, we meticulously curate two large-scale corpora: *AudioEventSet* and *AudioRelSet*. *AudioEventSet* is an audio event corpus that comprises 110 across fine-grained event classes across 7 main acoustic categories we commonly heard in our daily lives. In contrast to existing audio event datasets (Gemmeke et al., 2017; Kim et al., 2019; Fonseca et al., 2022) that are either noisy, polyphonic or label-missing, *AudioEventSet* provides a coarse-to-fine tree structured audio event corpus that is both internally distinctive and externally comprehensive. Each individual audio event in *AudioEventSet* is in high-quality, realistic and intra-class diverse. *AudioRelSet* is the large-scale relation corpus with up to 100 detailed relations completely covering the potential relations audio

events may present in the 3D physical world or text can describe succinctly. *AudioRelSet* is also tree structured and can be further scaled up to incorporate more relations. Each relation in *AudioRelSet* has an "arity" property that is further used to combine relation and audio events together to create `<text,audio>` pairs for relation aware TTA task. *AudioEventSet* and *AudioRelSet* are orders of magnitude larger than existing relevant dataset, enabling thorough and in-depth investigation for relation aware TTA task.

Based on the introduced audio event corpus *AudioEventSet* and relation corpus *AudioRelSet*, we further introduce a `<text, audio>` pair generation strategy that is capable of generating essential `<text,audio>` pairs highlighted by both audio event based and textual description diversity. As the audio event corpus is disentangled from relation corpus, our proposed strategy can generate nearly unlimited `<text,audio>` pairs tailored for various training requirements. In summary, as illustrated in Fig. 1, *Aurelius* advances relation-aware TTA research by contributing large-scale corpora of audio events and relations, together with a dedicated framework for relation-aware generation. The explicit disentanglement of audio events and relations, the hierarchical tree-structured design of each corpus, and the systematic `<text,audio>` creation strategy collectively provide a strong foundation for curating essential datasets in this domain. Building on this foundation, our proposed *AudioRelGen* framework tackles relation-aware TTA by decoupling audio event modeling from relation modeling, offering an essential first step toward structured audio generation. We believe this work will not only establish a new benchmark for relation-aware TTA but also inspire future research on modeling complex event–relation dynamics in sound.

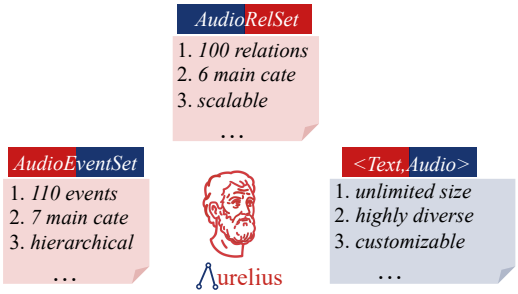

Figure 1: Aurelius contributes to relation aware TTA by introducing an audio event corpus *AudioEventSet*, a relation corpus *AudioRelSet* and `<text,audio>` pair generation strategy.

2 Related Work

Text-to-Audio Generation aims at generating the audio waveform that semantically aligns well with the input text. The fast development of generative modeling techniques (Ho et al., 2020; Vahdat et al., 2021; Lipman et al., 2023) in recent years has largely advanced the TTA generation in terms of high-fidelity and high-intelligibility (Liu et al., 2024; 2023; Kreuk et al., 2023; Yang et al., 2022; Ghosal et al., 2023; Liao et al., 2024), alongside with other crossmodal generation tasks including but not limited to text-to-music (TTM, *e.g.*, MusicGen (Copet et al., 2023) and MusicLM (Agostinelli et al., 2023)), image-to-audio (I2A, *e.g.*, RegNet (Chen et al., 2020), Img2Wav (Sheffer & Adi, 2023) and SpecVQGAN (Iashin & Rahtu, 2021) and text-to-image (T2I). Although the promising achievement in generating realistic and semantically text-aligned audio, existing TTA methods still perform poorly in relation aware TTA generation. Prior work like RiTTA (He et al., 2025) and CompA (Ghosh et al., 2024) have preliminarily explored relation aware TTA and shown the incapability of existing TTA methods through limited audio event and relation corpora, which inevitably hinders future investigation at scale. Moreover, publicly available audio event corpora (audioset) are directly collected from either online video data or audio sharing platform without proper quality check, resulting in the audio events label-missing, noisy and ambiguous. Our circumvents these barriers by introducing a meticulously curated audio event corpus *AudioEventSet* that is of high-quality, distinctive and realistic, potentially covering all commonly heard audio events.

Relation Modeling has been widely discussed within modalities, including image (Liu et al., 2022; Zerroug et al., 2022), natural language processing (Wadhwa et al., 2023) and acoustics (Xie et al., 2025a; Ghosh et al., 2024; He et al., 2025). In the context of 2D image, the objects of interest can exhibit compositional and spatial relation (Liu et al., 2022; Zerroug et al., 2022). In the context of 3D physical world, audio event is the most fundamental acoustic signal and multiple audio events join together to represent the 3D physical world via more sophisticated relations than image-based relations, ranging from basic spatial, temporal, perceptual relation to their nested combination. Prior works (Xie et al., 2025a; Ghosh et al., 2024; He et al., 2025; Xie et al., 2025b) have discussed

audio event relations in small-scale and with minimal complexity, making them hard to scale up to accommodate the potential relation complexity that present in either 3D physical environment or textual description. To fill in this gap, we curate *AudioRelSet*, a large-scale relation corpus that reflect the relation potentially present in the physical world and can be neatly describe by text.

Text-to-Audio Generation Techniques. Existing TTA methods can be technically divided into two main categories: while the early methods are diffusion based (Liu et al., 2024; 2023; Kreuk et al., 2023; Yang et al., 2022; Ghosal et al., 2023; Liao et al., 2024; Xue et al., 2024), the latest methods are flow-matching based (He et al., 2025; Hung et al., 2024; Guan et al., 2024). The flow-matching based methods are usually faster during both training and inference, and can give better performance than diffusion based methods. We completely benchmark all these methods on our introduced corpora, and further provide in-depth investigation to reveal potential ways scale up existing TTA methods' relation aware TTA capability.

3 AURELIUS BENCHMARK: AUDIOEVENTSET AND AUDIORELSET

3.1 AUDIO EVENT CORPUS: *AudioEventSet*

An audio event refers to an auditory signal occurring over a specific period of time, typically representing an independent, human-recognizable sound. To support the relation-aware TTA research, the desired audio event corpus should be: 1. diverse enough so as to maximally accommodate the wide variety of audio events potentially present in the 3D physical world; 2. clean and of high-fidelity so as to enable reliable in-depth technical investigation; 3. distinctive so that they can be easily distinguished without any ambiguity; 4. hierarchically organized w.r.t. their genre so as to enable investigation at different granularity. After thorough investigation on existing audio event related dataset, however, we find all existing datasets fall short in exhibiting the four properties. As is shown in Table 1, existing audio event dataset (*e.g.*, AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019), AudioTime (Xie et al., 2025a) and FSD50K (Fonseca et al., 2022)) are either noisy, label-missing, polyphonic (multiple events temporally overlap) or semantically ambiguous (where multiple event classes correspond to the same audio). To address this dilemma, we introduce *AudioEventSet*, a meticulously curated audio event corpus that is intrinsically clean, diverse, distinctive and hierarchically organized.

Table 1: Audio Event Dataset Comparison.

Dataset	Characteristic
AudioSet (2017) FSD50K (2022) AudioCaps (2019) AudioTime (2025a)	*polyphonic*, *ambiguous*, *noisy, label-missing*
AudioEventSet	*distinctive*, *high-quality*, *clean, hierarchical* *coarse-to-fine* *intra-class diversity* *inter-class discriminative*

AudioEventSet ontology is tree-structured and the tree depth is three. From the root node to the leaf node, each audio event is organized in coarse-to-fine granularity. As is shown in Fig. 2 and Table I in Appendix, we base on RiTTA (He et al., 2025) to categorize *AudioEventSet* into seven main categories: five singular-source categories *Animal*, *Human*, *Machinery*, *Music* and *Nature*, two interaction-based categories *Human-Object* and *Object-Object* interactions. The seven categories maximally cover the commonly heard audio events in the 3D physical world. Each main category associates with multiple subcategories, each of which is further associated with multiple fine-grained event classes. For example, the *Human* main category contains *human voice*, *human speech*, *hands action*, *group action* and *locomotion* subcategories, comprehensively categorizing the human centered audio event from various aspects.

During *AudioEventSet* ontology construction, we guarantee each curated audio event emits distinctive, unique and human-distinguishable audio. Audio event emitting ambiguous or nondistinctive audio is discarded. For example, *engine idling* in AudioSet (Gemmeke et al., 2017) audio differs significantly by various engines, and it easily confuses with another audio event such as *working fan* and *hairdryer*. We thus exclude the all of them from the corpus. Moreover, we account for both audio event source origin, event category and the audio generation physical mechanism for *AudioEventSet* ontology construction. For example, in the Object-Object main category, we exhaustively consider the impact, friction, dropping and explosion audio generation mechanism. In summary, we have curated 110

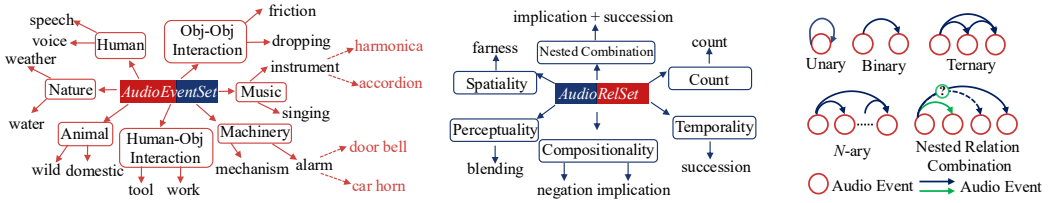

Figure 2: *AudioEventSet* and *AudioRelSet* corpora illustration: we visualize the *AudioEventSet* ontology in sub-figure A. It is tree-structured with depth 3 and contains 7 main categories and 110 event categories (leaf node) in total. We just show part of the leaf nodes (with red dotted arrow) for the seek of clear visualization. The detailed event ontology in given in Table I in Appendix. The *AudioRelSet* ontology in sub-figure B, it is tree-structured with depth 2. It contains 6 main categories and 100 categories in total. The detailed relation ontology is given in Table II in Appendix. In sub-figure C, we conceptually illustrate the relation "arity", which is used to connects relation and audio event to generate audios.

audio events, which is four times larger than audio event corpus proposed in RiTTA (He et al., 2025), each leaf audio event is associated with around 75 realistic audio snippets ranging from 1 s to 5 s

For each leaf node audio event, we collect exemplar audios from either copyright-free `freesound.org` platform or FSD50K (Fonseca et al., 2022). As most audios from `freesound.org` and FSD50K [1] real audios shared by volunteers across the globe, the collected audios for each audio event are diverse and realistic enough to reflect the audio event we can hear in the physical world. Manually verification is adopted to ensure the collected exemplar audios content correctness, label consistency. We argue that the curated *AudioEventSet* can be potentially applied to other tasks other than TTA, we anticipate much wider usage of the curated dataset.

3.2 Relation Corpus: *AudioRelSet* AudioRelSet

Prior works (Xie et al., 2025a; He et al., 2025; Ghosh et al., 2024) have explored audio events relation from various perspectives, but only on a small scale. For example, AudioTime (Xie et al., 2025a) and CompA (Ghosh et al., 2024) have discussed temporal relations. RiTTA (He et al., 2025) has additionally introduced spatial, compositional and count relations, resulting in a total of 11 relations. In this section, we introduce *AudioRelSet*, a meticulously curated large-scale relation corpus with up to 100 distinct relations. To ensure *AudioRelSet* to exhibit both real scenario practicability, text-manageable complexity and relation scalability, we follow 3 guidelines to curate *AudioRelSet*: 1. maximally cover the potential relations audio events can present in the 3D physical world; 2. enough relation complexity but can still be efficiently and neatly described by text; 3. the relation corpus can be scaled up to accommodate more sophisticated relations. To this end, we construct 6 main fundamental relations, in which 4 main relations describe the relations present in the 3D physical world, one main relation focuses on TTA model's logical reasoning capability and last one relation derives from the nested combination of the five main relations.

AudioRelSet ontology is tree-structured and the tree depth is 2, the root node connects 6 main relations, each of which further associates with multiple sub-relations. Let $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ denote the audio events in *AudioEventSet* introduced in Sec. 3.1, $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$ denote the relations to be constructed. *AudioRelSet* is represented as follows,

1. **Temporality** describes the sequence or overlap of audio events in time domain, it contains 4 sub-relations: *Precedence*: $E_1 \prec E_2$ (event E_1 occurs before E_2); *Succession*: $E_1 \succ E_2$ (event E_1 occurs before E_2); *Simultaneity*: $E_1 \parallel E_2$ (E_1 and E_2 occur concurrently); *Repetitiveness*: $\sim E_1$ (event E_1 occur repetitively in the time domain).

2. **Spatiality** defines the relative spatial positions or motion status between or within audio events, it contains 5 sub-relations: *Proximity*: $d(E_1, E_2) \leqslant \tau$ (E_1 E_2 are within distance τ); *Closeness*:

[1] FSD50K (Fonseca et al., 2022) data is also sourced from `freesound.org`

$d(E_1) < d(E_2)$ (E_1 is closer than E_2); *Farness*: $d(E_1) > d(E_2)$ (E_1 is further than E_2); *Approaching*: $\frac{d}{dt}d_{E_1}(t) < 0$ (E_1 is moving close); *Departuring*: $\frac{d}{dt}d_{E_1}(t) > 0$ (E_1 is moving away).

3. **Count** focuses on the number of audio events take place within a period of time: *Count*: $|\mathcal{E}| = N, N \in \mathbb{Z}^+$. (cardinality $\mathcal{E}$ is the number).

4. **Perceptuality** introduces 6 acoustic effects to an audio event,

- *Balancing*: $\mathcal{R}_{\text{balance}}(E_1, E_2, \sigma)$ (level balance between E_1 and E_2 by balancing factor σ, so that one event dominates and the other serves as the background audio).
- *Blending*: $\mathcal{R}_{\text{blend}}(E_1, E_2, \theta)$ (mix E_1 and E_2 together by factor θ so as to be indistinguishable).
- *Reverberation*: $\mathcal{R}_{\text{reverb}}(E_1)$ applies reverberation effect to E_1, as if it is heard in the canyon.
- *Time-stretching*: $\mathcal{R}_{\text{stretch}}(E_1, \alpha)$, where α is the time-stretching factor and E_1 listens slowly.
- *Amplification*: $\mathcal{R}_{\text{amp}}(E_1, \beta)$, where β is the amplification factor and E_1 listens to be louder.
- *Attenuation*: $\mathcal{R}_{\text{att}}(E_1, \gamma)$, where γ is the attenuation factor and E_1 listens to be quieter.

5. **Compositionality** indicates the logical operation within audio events TTA models need to reason before deciding what audio events to generate. It contains 5 sub-relations.

- *Conjunction*: $E_1 \wedge E_2$ (both events occur).
- *Disjunction*: $E_1 \vee E_2$ (at least one event occurs, or both occur).
- *Negation*: $\neg E_1$ (the absence of the event E_1 in the generated audio).
- *Exclusive Or*: $(E_1 \vee E_2) \wedge \neg(E_1 \wedge E_2)$ (either E_1 or E_2 occur, but not both).
- *Implication*: $E_1 \Rightarrow E_2, \neg E_1 \Rightarrow E_3$ (if E_1 occur, then E_2 occur, else E_3 occur).

6. **Nested Combination** is a hierarchical structuring of multiple basic relations (*e.g.*, the aforementioned *Temporality*, *Spatiality*), such that the output of one relation serves as the input or context for another, forming a directed acyclic relation structure. Nested combination allows for capturing complex relation interactions among audio events. For example, by nesting *Implication*, *Approaching* and *Conjunction*, we can generate a more complex text prompt showing below,

Nest Combination Example: *Implication*, *Approaching* and *Conjunction*

If generated both {A} event and {B} event, → *Conjunction*
then continue to generate {C} audio event,
else just generate {D} audio event that is gradually approaching close. →*Approaching*

Mathematically, the relation $R_{\text{nested}}(E)$ resulting from nested combination can be represented as,

$$R_{\text{nested}}(E) = R_n(R_{n-1}(\dots R_2(R_1(E))\dots)) \tag{1}$$

where $E = \{e_1, e_2, \dots, e_m\}$ represents a finite set of audio events. We combine relations arising from the introduced 5 basic relations to construct nested combination relations and have created 79 nested combination relations.

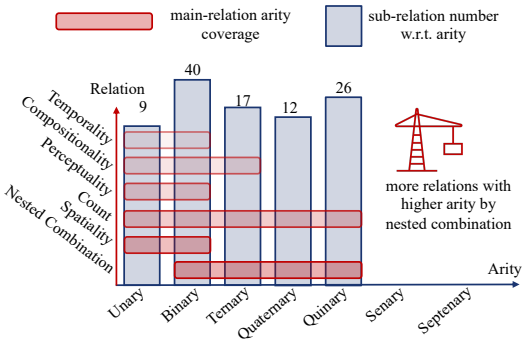

Figure 3: Arity coverage in *AudioRelSet*.

It is worth noting that the nested combination is scalable and we can theoretically construct more complex nested relations (even infinite relations) by simply involving more basic relations into the nested combination process. In this work, we constrain the nested combination up to maximumly involving 5 audio events (*Quinary*), it remains as future research topic to explore more complex nested combination, and the key challenge remains on how to construct the corresponding concise and precise textual description for

the given highly complex nested relation. Moreover, during the nested combination process, we explicitly run internal nested relations logic correctness and feasibility check before accepting the nested relations as a new relation, any nested relation violating the correctness and feasibility rule is abandoned. For example, the combination of *Count* and *Conjugation* internally equals to *Count*.

Relation Arity. each relation in *AudioRelSet* is associated with an "arity" property, which indicates the audio event number it requires to represent the relation. The visual illustration of arity is shown in Fig. 2 C. The arity coverage across *AudioRelSet* main relation categories is given in Fig, 3, from which we can see that the arity ranges from 1 to 5 (unary to quinary) and most main relation cuts across multiple arities. Moreover, the construction of more complex relations introduces higher arity. We use "arity" to create `<text,audio>` pairs (see Sec. 3.3) and experiment evaluation (see Sec. 4).

3.3 TEXT-AUDIO PAIR CREATION: <Text, Audio>

With the constructed audio event corpus in Sec. 3.1 and relation corpus in Sec. 3.2, we can further construct relation aware `<text,audio>` pairs. Specifically, as is shown in Fig. 4, we first associate each of the 100 relations in the relation corpus with meticulously curated 5 text description templates. We either manually write or query GPT-4o to generate 5 text prompt templates precisely describing the relation and accommodating the large language usage variation (see Fig. 4 line 4-8). Each template contains audio events name placeholder, we instantiate the template by replacing the placeholder with real audio event name to obtain the text prompt. To accommodate the synonymy of audio event name, we maintain a synonym list for each audio event name, and randomly select one each time when instantiating the template. For example, the audio event name "hammer nailing" can be synonymously replaced by one of [*hitting*, *slapping*, *smacking*, *punching*].

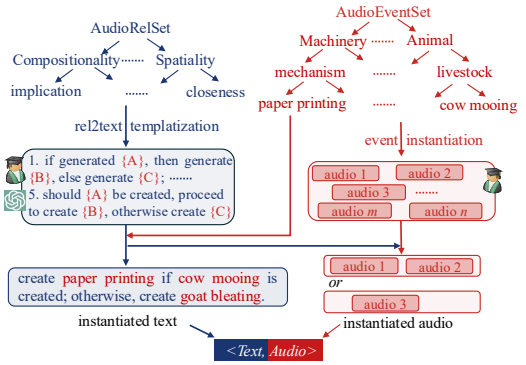

Figure 4: `<text,audio>` pair generation illustration, which can generate nearly unlimited pairs with high diversity.

To accurately describe the audio event with text, we adopt the "Head-Modifier Structure with Progressive Verb Form" approach. In this approach, the description begins with the subject or entity producing the audio (*e.g.*, "food") as the head, emphasizing the primary source of the sound. The action is then specified using its present participle form (*e.g.*, "frying") as the modifier to convey a sense of immediacy and highlight that the audio event is ongoing. For instance, instead of describing a sound as "frying food" or "fry food" it is labeled as "food frying audio," where the subject ("food") is foregrounded, and the action ("frying") contextualizes the nature of the audio. This approach ensures clarity, aligns with the temporal context of the audio, and effectively captures the dynamic nature of the event. With the same audio events name, we can retrieve its relevant audio waveform data and generate the audio by following the relation (He et al., 2025).

4 EXPERIMENT

4.1 DATASET CONSTRUCTION

Following the common setup in existing TTA model, the created audio is 10 second long with sampling rate 16 kHz. Based on the data creation method introduced in Sec. 3.3, in training phase we randomly construct 360 `<text,audio>` pairs for each relation, and in total we have created 36,000 pairs. In testing phase, we randomly construct 100 `<text, audio>` pairs for each relation, ensuring no constructed appear in the training dataset. As we follow the prior TTA models setting to create the audio to be 10 second long with sampling rate 16 kHz, the training audio dataset is 100 hours, and testing audio dataset is 28 hours. Since we decouple relation from audio events during dataset construction and the texts in training dataset are different from the texts in the testing dataset,

the constructed training and testing <text,audio> pairs have no overlap and differ from each other significantly.

4.2 Evaluation Metric

We accommodate both classic general evaluation metrics and relation-aware evaluation metrics. For general evaluation, we follow traditional TTA works (Liu et al., 2024; 2023; Ghosal et al., 2023; Majumder et al., 2024) and adopt three metics: Fréchet Audio Distance (FAD), Fréchet Distance (FD) (Heusel et al., 2017), Kullback–Leibler (KL) divergence. These three metrics measure the overall similarity in embedding space between reference audio and generated target audio without explicitly taking relation into account. Specifically, following the practice in prior TTA works, we extract the embeddings from VGGish (Hershey et al., 2017) model for FAD and KL metrics, embeddings from PANNs (Kong et al., 2020) model for FD metric.

For relation-ware evaluation, we adopt the multi-stage relation-aware (*MSR*) evaluation protocol introduced in RiTTA (He et al., 2025). In *MSR* protocol, we first explicitly extract out audio events and relations (E', R') from generated audio, the further compare them with reference audio events and relations (E, R). To reflect if the model has generated but just generated the designated audio events and relations, *MSR* adopts *Pre*sence, *Rel*ation correctness and *Par*simony score to gauge the quality of generated audio from different perspectives. Specifically, we report `mAPre`, `mARel` and `mAPar` scores for either separate relations or across all relations. More detailed information about *MSR* metric refers to RiTTA (He et al., 2025). To extract out audio event from generated audio, we finetune an audio event detection and tagging model on top of the pre-trained PANNs (Kong et al., 2020) model with 1 million training dataset. The mAP on 100,000 testing dataset achieves 0.91 for audio event detection, ensuring the finetuned model can extract out all potential audio events with high precision. To classify acoustic effect, we train another 7 acoustic effects classification model on top of the pre-trained PANNs model with 1 million training dataset. The accuracy rate on 100 k testing dataset achieves 95%.

4.3 Benchmarking Methods

We exhaustively benchmark 9 most recent general TTA models: AudioLDM (Liu et al., 2023), AudioLDM 2 (Liu et al., 2024), MakeAnAudio (Huang et al., 2023), AudioGen (Kreuk et al., 2023), Tango (Ghosal et al., 2023), Tango 2 Majumder et al. (2024), LAFMA (Guan et al., 2024), Affusion (Xue et al., 2024) and TangoFlux (Hung et al., 2024). They are pretrained on general TTA dataset (Gemmeke et al., 2017; Kim et al., 2019). For benchmarking, we choose their released model to generate a 10 second audio from the text prompt, detailed configuration is in Table III in Appendix.

We further benchmark two agentic workflow based methods, in which we leverage open-sourced Qwen family LLM (**?**) acting as an agent to analyze the input text and output the separate audio events an TTA model needs to generate. At the same time, the same LLM works as the third agent to output the python code that merges the audios generated by the TTA model. The reason of experimenting agentic flow is to see if we can decompose the relation-aware generation task into simple single audio event generation task.

4.4 Benchmarking Result on Existing TTA Models

The benchmarking result is given in Table 2, from which we can observe that that all existing TTA models perform poorly on relation aware TTA generation. Similar to RiTTA (He et al., 2025), we also find the contradictory evaluation result between general evaluation and relation aware evaluation, which shows the speciality of relation aware TTA task. Among all the benchmarking methods, AudioGen (Kreuk et al., 2023) and TangoFlux (Hung et al., 2024) perform the best. While AudioGen (Kreuk et al., 2023) achieving the best in mAPar (relation parsimony) and mAMSR, TangoFlux (Hung et al., 2024) stays the best-performing in mAPre and mARel which mean it excels at accurately generating the target audio events and corresponding relation. However, almost all benchmarking methods achieves less than 10% percent accuracy rate across all relation aware evaluation metrics, which in turn verify the necessity to introduce new large-scale benchmark tailored for relation aware TTA research.

Table 2: Quantitative benchmarking result on our introduced benchmark. mAPre, mARel and mAPar are in 10^{-2}. mAPre and mARel can be treated as *presence*, *relation correctness* percentage ratio, they lie in range $[0, 100]$. mAPar score also lies within $[0, 100]$. mAMSR (%) lies in range $[0, 1]$

Eval Way	Model	#Param	General Evaluation			Relation Aware Evaluation %(↑)			
			FAD ↓	KL ↓	FD ↓	mAPre	mARel	mAPar	mAMSR
Zero-Shot	AudioLDM (s-full) 2023	185 M	4.02	21.23	22.36	3.47	0.91	2.95	0.73
	AudioLDM (l-full) 2023	739 M	4.13	22.05	23.03	3.10	0.79	2.63	0.63
	AudioLDM 2 (l-full) 2024	844 M	4.54	22.90	30.53	0.35	0.04	0.31	0.03
	MakeAnAudio 2023	452 M	5.10	50.97	30.49	4.75	0.88	4.05	0.73
	AudioGen 2023	1.5 B	7.97	25.19	32.29	11.3	2.84	9.13	2.22
	LAFMA 2024	272 M	25.85	269.54	65.27	0.96	0.15	0.45	0.07
	Affusion 2024	1.1 B	4.13	42.59	31.17	6.71	1.41	4.07	0.79
	Tango 2023	866 M	7.47	64.10	28.28	4.46	0.98	3.67	0.79
	Tango 2 2024	866 M	9.59	65.24	35.50	9.68	2.48	5.49	1.29
	TangoFlux 2024	576 M	6.01	26.73	30.00	12.38	3.34	7.28	1.77
Agentic	Qwen2 7B+TangoFlux	-	9.98	142.87	39.20	3.53	0.77	2.25	0.04
	Qwen2.5 32B+TangoFlux	-	9.70	140.56	38.65	3.79	0.96	2.41	0.60

Furthermore, both the two comparing agentic flow baselines perform poorly, they perform substantially worse than most existing existing TTA approaches. This poor performance highlights a critical limitation: simply scaling up current TTA methods without fundamentally enhancing their relation-aware modeling capability is unlikely to succeed. In this light, the benchmark introduced in this paper is not merely a comparison tool but a catalyst—providing the necessary structure, evaluation, and motivation to drive genuine advances in relation-aware TTA research.

4.5 Two Intuitive Ways to Improve Relation Aware Modeling

Table 3: Quantitative result comparison on testset between finetuning (ft) and training from scratch (scratch) on curated 100 hours dataset.

Train Way	Model	#Param	General Evaluation			Relation Aware Evaluation %(↑)			
			FAD ↓	KL ↓	FD ↓	mAPre	mARel	mAPar	mAMSR
ft	Tango 2023	866 M	3.88	33.26	21.30	14.58	4.18	10.16	2.73
	Tango 2 2024	866 M	4.06	22.39	20.32	15.53	4.63	10.21	2.86
	TangoFlux 2024	576 M	1.29	9.68	16.44	28.57	8.02	20.84	5.58
scratch	Tango 2023	866 M	3.63	22.34	20.16	14.89	3.69	10.98	2.64
	TangoFlux 2024	576 M	1.64	17.82	11.72	16.68	3.82	12.01	2.58

Two intuitive strategies to enhance relation aware modeling in existing TTA methods are (i) finetuning on our curated dataset and (ii) training from scratch. This dual perspective not only tests the feasibility of our benchmark but also evaluates the potential of transferring general TTA domain knowledge into relation-aware settings. To this end, we apply both training strategies to three representative baselines: Tango (Ghosal et al., 2023), Tango 2 (Majumder et al., 2024), and TangoFlux (Hung et al., 2024). The results in Table 3 reveal a clear trend: both finetuning and training from scratch substantially improve relation aware performance, validating the effectiveness of our benchmark as a testing ground for relation-aware TTA. Notably, TangoFlux benefits the most from finetuning, indicating that cross-domain TTA knowledge can be effectively transferred to relation aware tasks. In contrast, Tango shows little difference between the two strategies, suggesting that model architecture and inductive bias may affect the extent to which general TTA knowledge can be leveraged. These findings highlight our benchmark's unique role in uncovering such model-specific behaviors and point to an open research direction: how to best exploit general TTA knowledge to scale up relation aware TTA, and conversely, how relation aware training can reciprocate general TTA advances. We visualize the generated audio comparison

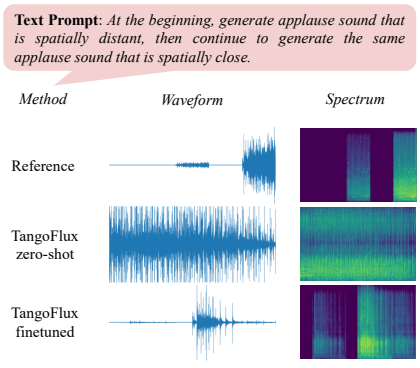

Figure 5: Qualitative comparison between zero-shot and finetune based TangoFlux inference on one text prompt.

between TangoFlux in zero-shot and finetuned base inference mode in Fig. 5, from this figure we can clearly see that finetuning on our curated dataset benefits relation aware modeling.

To further investigate the role of training datasize, we extend both finetuning and training from scratch experiments to larger datasets of 200 hours and 300 hours. As is shown in Fig. 6, the $mAMSR$ trend reveals two distinct behaviors: finetuning yields strong early gains but quickly saturates when the datasize approaches 300 hours, whereas training from scratch continues to improve substantially with increasing data. This divergence underscores an important insights: scaling relation aware TTA models ultimately requires massive datasets, and reliance on finetuning alone may be insufficient for long-term progress. Our benchmark is therefore essential: it not only provides the controlled scaling environment needed to expose these trends, but also offers the first practical platform to systematically study how training strategy and datasize interact in advancing relation aware TTA.

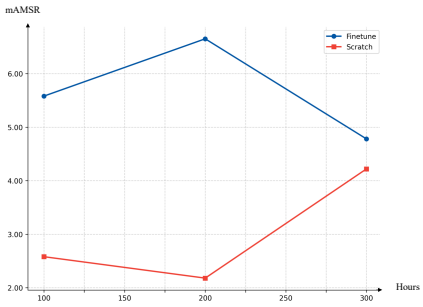

Figure 6: `<text,audio>` pair generation illustration, which can generate nearly unlimited pairs with high diversity.

4.6 More Investigation on Existing TTA Model Analysis

Relation aware TTA demands not only the correct presence of target audio events but also the faithful preservation of their underlying relations. However, current TTA methods (Hung et al., 2024; Ghosal et al., 2023; Xue et al., 2024) remain narrowly focused on single-event generation, leaving them ill-equipped to handle multi-event, relation-aware prompts. Table 4 makes this gap explicit: while TangoFlux (Hung et al., 2024), the state-of-the-art general TTA model, achieves 75% accuracy on single-event prompts, its performance collapses to just 12% for multi-event correctness and a mere 3% for relation fidelity. This dramatic degradation exposes a fundamental blind spot in existing approaches—relation-aware modeling is virtually unaddressed. Our benchmark directly targets this deficiency, offering the first systematic platform to quantify and dissect these failures. By doing so, it not only diagnoses the shortcomings of current TTA methods but also establishes the essential foundation for driving genuine advances in relation-aware TTA.

Table 4: Audio event and relation accuracy of TangoFlux generation under different setting.

Description	Accu.
Event (single event, no relation)	75%
Event (multi-event, relation-aware)	12%
Relation (multi-event, relation-aware)	3%

5 Conclusion

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating Music from Text. In *arXiv preprint arXiv:2301.11325*, 2023. 2

Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating Visually Aligned Sound from Videos. *IEEE Transactions on Image Processing (TIP)*, 2020. 2

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an Open Dataset of Human-labeled Sound Events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 30:829–852, 2022. 1, 3, 4

J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 1, 3, 7

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-Audio Generation using Instruction Tuned LLM and Latent Diffusion Model. In *ACM International Conference on Multimedia (ACMMM)*, 2023. 1, 2, 3, 7, 8, 9, 15

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 4

Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingyang Hong, and Yong Qin. LAFMA: A Latent Flow Matching Model for Text-to-Audio Generation. In *Interspeech*, 2024. 1, 3, 7, 8, 15

Yuhang He, Yash Jain, Xubo Liu, Andrew Markham, and Vibhav Vineet. RiTTA: Modeling Event Relations in Text-to-Audio Generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. 1, 2, 3, 4, 6, 7, 14

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 7

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 7

IJ Hirsh, C Milliman, and F Darley. Brain Mechanisms Underlying Speech and Language, 1967. 1

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio Generation with Prompt-enhanced Diffusion Models. *International Conference on Machine Learning (ICML)*, 2023. 7, 8, 15

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. TangoFlux: Super Fast and Faithful Text to Audio Generation with Flow Matching and Clap-Ranked Preference Optimization, 2024. 1, 3, 7, 8, 9, 15

Vladimir Iashin and Esa Rahtu. Taming Visually Guided Sound Generation. In *British Machine Vision Conference (BMVC)*, 2021. 2

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 1, 3, 7

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark Plumbley. PANNS: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2020. 7

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually Guided Audio Generation. *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 7, 8, 15

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level Concept Learning hrough Probabilistic Program Induction. *Science*, 350(6266):1332–1338, 2015. 1

Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. BATON: Aligning Text-to-Audio Model with Human Preference Feedback. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 2, 3

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *International Conference on Machine Learning (ICML)*, 2023. 2, 3, 7, 8, 15

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2024. 1, 2, 3, 7, 8

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional Visual Generation with Composable Diffusion Models. In *European Conference on Computer Vision (ECCV)*, 2022. 2

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning Diffusion-based Text-to-Audio Generations Through Direct Preference Optimization. In *ACM International Conference on Multimedia (ACMMM)*, 2024. 7, 8, 15

Roy Sheffer and Yossi Adi. I Hear Your True Colors: Image Guided Audio Generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 2

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based Generative Modeling in Latent Space. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Revisiting relation extraction in the era of large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 2

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyu Wu. AudioTime: A Temporally-aligned Audio-Text Benchmark Dataset. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025a. 2, 3, 4

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. PicoAudio: Enabling Precise Temporal Controllability in Text-to-Audio Generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025b. 2

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the Power of Diffusion and Large Language Models for Text-to-Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2024. 1, 3, 7, 8, 9, 15

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete Diffusion Model for Text-to-sound Generation. *IEEE Transactions on Audio, Speech and Language Processing*, 2022. 2, 3

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event Perception: A Mind-Brain Perspective, 2007. 1

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A Benchmark for Compositional Visual Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

A APPENDIX

.1 AGENTIC AUDIO SCENE GENERATION WORKFLOW

We design an agentic workflow that integrates large language model (LLM) reasoning with state-of-the-art text-to-audio (TTA) generation. The workflow operates in three stages:
1. **Scene Planning**
2. **Segment Synthesis**
3. **Timeline Stitching**

.2 AUDIO EVENT CATEGORY CURATION DETAIL

A AUDIO EVENTS RELATION CORPUS

A.1 BENCHMARKING MODEL INFERENCE SETTING

Table I: *AudioEventSet* corpus detail. We list all 110 event classes, which are deriving from 7 main categories and 23 sub-categories.

Main Category	Sub-Category	Names	Description
Animal (22)	wild ground animal	lion roaring, wolf howling, donkey braying, cricket chirping, frog croaking, horse neighing	live in the wild
	domestic animal	dog barking, cat meowing, dog growling, cat spurring	live in domestic setting
	livestock	pig oinking, sheep bleating, cow mooing, rooster crowing, duck quacking	domesticated livestock
	wild animal	cuckoo calling, birds chorus, seagull cawing, peacock rattling blue jay whistling, nightingale singing, fly buzzing	animals in the wild
Human (21)	human voice	baby crying, laughing, shouting, whistling, coughing, snoring, sneezing, chewing, burping, farting	human use vocal tract
	human speech	male speech, female speech, child speech, group talk	speech audio
	hands action	finger snapping, clapping	audio by action
	group action	group clapping, cheering, group talking	audio by a group
	locomotion	running, footsteps	audio by movement
Machinery (13)	alarm	siren, door bell, car horn, bicycle bell, telephone ringing, telephone dialing, boat horn	machinery alarming
	mechanism	ratchet and pawl clicking, camera shuttering, printer printing, engine revving, clock ticking, paper shredding	mechanism audio
Human-Obj Interaction (18)	tools	hammer nailing, wood sawing, pen writing, wood chopping, rasping	human use tools
	culinary	dish audio, silverware audio, food frying, vegetable chopping	in kitchen setting
	work	toilet flushing, pouring water, keyboard typing, door slamming, cupboard open or close, drawer open or close, packing tape, dentist drilling, door knocking	audio during work
Obj-Obj Interaction (15)	impact audio	key jingling, ball bouncing, pen clicking, wind chime	impact effect
	friction audio	car emergency braking knife sharpening, sandpaper scraping, plastic scratching, string rubbing	friction effect
	dropping audio	coin dropping, glass clinking, metal dropping	dropping effect
	explosion	gunshot, firework, artillery fire	explosion effect
Music (11)	music instrument	plucked string, piano keyboard, bowed string, wind string, brass, harmonica, accordion	musical instruments
	singing	female singing, male singing, child singing, group singing	singing audio
Nature (10)	water	water bubbling, ocean wave, water dripping, water flowing, water boiling	water movement
	weather	thunder, wind, rain	nature weather
	nature change	wood cracking, rustling leaves	natural change

Table II: *AudioRelSet* corpus detail. We introduce 21 basic relations, and advanced 79 nested combination relations, resulting in a total of 100 relations – 9 times larger than the relation corpus proposed in RiTTA (He et al., 2025). *AudioRelSet* maximumly covers all potential relations that audio events may exhibit in either the physical world or linguistic description. It is worth noting that *AudioRelSet* is open-ended. By nesting existing relations, we can potentially construct massive new relations.

Category	Relation Name	Explanation	Event Arity	Sample prompt
Temporality (4)	precedence	before	binary	audio {A} followed by {B}
	succession	after	binary	create audio {A} after {B}
	simultaneity	same time	binary	{A} and {B} simultaneously
	periodicity	cyclic	unary	create audio {A} periodically
Spatiality (5)	closeness	spatial close	binary	{A} is closer than {B}
	farness	spatial far	binary	{A} is farther than audio {B}
	proximity	equal-dist	binary	{A} and {B} the same dist
	approaching	moving close	unary	{A} is moving closer
	departuring	moving away	unary	{A} is moving further away.
Count (1)	count	number	n-ary	3 audios: {A}, {B} and {C}
Perceptuality (6)	balancing	level balance	binary	{A} dominates, {B} fades
	blending	mix audios	binary	{A} and {B} are mixed
	reverberation	reverberant	unary	generate audio {A} in canyon
	time-stretching	speed manipulate	unary	stretch audio {A} in time scale
	amplification	become louder	unary	amplify {A} to be louder
	attenuation	less loudly	unary	attenuate {A} to be quieter
Composition-ality (5)	conjunction	logical AND	binary	create both {A} and {B}
	disjunction	logical OR	binary	create {A} or {B}, or both
	negation	logical NOT	unary	do not generate audio {A}
	exclusive-or	logical XOR	binary	generate {A} or {B}, not both
	implication	if-then-else	ternary	if {A}, then {B}, else just {C}
Nested Combination (79)	Temp + Spat (4)	Temp + Spat	binary	{A} before approaching {B}
	Temp + Percep (8)	Temp + Percep		reverb. {A}, succeeded by {B}
	Percep + Comp (12)	Percep + Comp		stretched {A} or {B}, not both
	Spat + Comp (4)	Spat + Comp		approaching {A} or {B}, not both
	Temp + Comp (6)	Temp + Comp	ternary	{A} first, then {B} or {C}
	Percep + Comp (1)	Percep + Comp		mix {A} with {B}, or {C}
	Comp + Comp (1)	Comp + Comp		{A} and {B}, or {A} and {C}
	Spat + Comp (5)	Comp + Comp		{A} and {B}, or {A} and {C}
	Spat + Comp + Percep (2)	Comp + Comp		{A} and {B}, or {A} and {C}
	Temp + Comp (4)	Temp + Comp	quaternary	audio {A} or {B} first, followed by {C} or {D}
	Comp + Comp (7)	Comp + Comp		{A} or {B} first, then {C} or {D}
	Temp + Comp (3)	Temp + Comp	quinary	{A} before {B} first, then {C} before {D} or {E}
	Spat + Comp (9)	Spat + Comp		if {A} closer than {B}, then {C} closer than {D}, else {E}
	Comp + Comp (9)	Comp + Comp		if {A} and {B}, then {C} and {D} else {E}
	Count + Comp (4)	Count + Comp		if {A}, {B}, {C}, then {D}, else {E}

Methods	Setting
AudioLDM (S-Full) (2023)	guidance_scale=5, random_seed=42, n_candidates=3
AudioLDM (L-Full) (2023)	guidance_scale=5, random_seed=42, n_candidates=3
AudioLDM 2 (L-Full) (2023)	guidance_scale=3.5, random_seed=45, n_candidates=3
MakeAnAudio (2023)	ddim_steps = 100, scale = 3.0
AudioGen (2023)	model name: audiogen-medium
Auffusion (2024)	num_steps = 100, guidance=7.5, num_samples=1
LAFMA (2024)	num_steps = 200, guidance=3, num_samples=1
Tango (2023)	num_steps = 200, guidance=3, num_samples=1
Tango 2 (2024)	num_steps = 200, guidance=3, num_samples=1
TangoFlux (2024)	num_steps = 50, guidance=3, num_samples=1

Table III: Detail setting for each TTA method.