

# RiTTA: Modeling Event Relations in Text-to-Audio Generation

**Yuhang He**                      **Yash Jain**                      **Xubo Liu**  
Microsoft Research              Microsoft                      CVSSP, University of Surrey, UK  
yuhanghe@microsoft.com      yash.jain3599@gmail.com      xubo.liu@surrey.ac.uk

**Andrew Markham**  
CS Department, University of Oxford, UK  
andrew.markham@cs.ox.ac.uk

**Vibhav Vineet**  
Microsoft Research  
vibhav.vineet@microsoft.com

## Abstract

Existing text-to-audio (TTA) generation methods have neither systematically explored audio event relation modeling, nor proposed any new framework to enhance this capability. In this work, we systematically study audio event relation modeling in TTA generation models. We first establish a benchmark for this task by: (1) proposing a comprehensive relation corpus covering all potential relations in real-world scenarios; (2) introducing a new audio event corpus encompassing commonly heard audios; and (3) proposing new evaluation metrics to assess audio event relation modeling from various perspectives. Furthermore, we propose a gated prompt tuning strategy that improves existing TTA models’ relation modeling capability with negligible extra parameters. Specifically, we introduce learnable relation and event prompt that append to the text prompt before feeding to existing TTA models <sup>1</sup>.

## 1 Introduction

Text-based crossmodal content generation has gained significant attention in recent years as it opens up new possibilities for even amateur users to create professional content. Typical such methods include text-to-image (TTI) (Ho et al., 2020), text-to-music (TTM) (Copet et al., 2023), text-to-point (TTP) (Nichol et al., 2022), text-to-speech (TTS) (Ren et al., 2019) text-to-audio (TTA) (Liu et al., 2024; Huang et al., 2023b). Among all of them, text-to-audio (TTA) generation stands out as a particularly promising area, enabling the synthesis of complex acoustic environments or soundscapes directly from textual descriptions. Recent advances have demonstrated impressive progress in generating high-quality, detail-rich audio described in input text prompt (Liu et al., 2024, 2023; Huang et al., 2023b,a; Ghosal et al., 2023; Majumder et al., 2024; Kreuk et al., 2023).

<sup>1</sup>Code: <https://github.com/yuhanghe01/RiTTA>

*Text Prompt: generate dog barking audio,  
followed by cat meowing audio*

Method	Rel?	Remark
AudioLDM (2023)	✗	just cat meow
AudioLDM 2 (2024)	✗	output dog barking
MakeAnAudio (2023b)	✗	just cat meow
AudioGen (2023)	✗	output wrong audios
Tango (2024)	✗	two audios
Tango 2 (2024)	✗	can output two audios
TangoFlux (2024)	✗	can not satisfy relation

Table 1: A case study on existing TTA methods. “Rel” means “if the relation is correctly modeled?”.

When perceiving the physical world acoustically, whether through text or audio, the fundamental unit is the audio event, a distinct acoustic signal representing an independent source. The essence of perception lies in understanding the relationships emerging from events. Audio events are spatiotemporally distributed in the physical world. Together with relation, they contribute for holistic acoustic scene understanding (Qu et al., 2022). Studies in psychology (Zacks et al., 2007) and neuroscience (Lake et al., 2015; Hirsh et al., 1967) show that the human brain perceives the environment through discrete events and the relations between them. Humans are adept at using rich language to describe both audio events and their intricate relationships. While current TTA models can generate audios with high fidelity, their ability to generate audios that not only includes audio events but also preserves the text-informed relationships between them remains unexplored.

As a primary study, we prompt the latest six TTA models with an exemplar text with explicit audio events and their relation *generate dog barking audio, followed by cat meowing audio*. Next we check if the specified audio events are present and if so, their relations are correct in the generated audios. As is shown in Table 1, all existing TTA models fail to properly model temporal relationships in the generated audio, even when they succeed in generating the correct audio events. The generated audio waveform, spectrum and another case study with a

much complex text are shown in Fig. 1. The poor performance of current TTA models in modeling audio events relation, along with the lack of systematic discussion on this topic, motivates us to explore *Relation in TTA* (dubbed *RiTTA*) in depth in this work. We visualize the motivation in Fig. 1.

To systematically study *RiTTA*, we first benchmark it from four key perspectives: 1. we construct a comprehensive audio event relation corpus that captures common relationships found in the physical world. Unlike visual relations in cross-modal image tasks, which mainly focus on spatial aspects (*e.g.*, left, bottom) (Gokhale et al., 2022), audio events exhibit far more complex relationships spanning spatial, temporal, and compositional dimensions. Consequently, we define four primary relation categories: *Temporal Order*, *Spatial Distance*, *Count*, and *Compositionality*. 2. Accompanying the relation corpus, we build an audio event category corpus derived from five main sources, each of which is further linked to multiple seed audios. 3. devise a <text, audio> pair generation strategy emphasizing both text prompt and audio diversity. 4. propose a new relation aware evaluation protocol that assesses the relation in a multi-stage manner. The proposed benchmark will benefit the community to explore *RiTTA* in greater depth. Additionally, we introduce gated prompt tuning strategy to significantly improve existing TTA models’ relation modeling capability by simply introducing a negligible parameters.

1. We conduct extensive evaluation on existing TTA models’ inability in relation modeling.
2. We benchmark *RiTTA* by introducing three corpora: relation corpus, audio event corpus and seed audio corpus, as well as a new <text, audio> pair generation strategy.
3. We propose a new multi-stage relation aware evaluation framework.
4. We introduce gated prompt tuning to improve existing TTA models’ relation modeling capability by introducing tunable prompts.

## 2 Related Work

**Text-to-Audio (TTA) Generation** involves producing audio that faithfully reflects the acoustic content or behavior described by the input text. Recent advancements have significantly improved the quality and intelligibility of generated audio (Liu et al., 2024, 2023; Kreuk et al., 2023; Yang et al., 2022; Ghosal et al., 2023; Liao et al.,

2024). AudioLDM (Liu et al., 2023) builds on latent space (Rombach et al., 2022) to learn continuous representation. The most recent work TangoFlux (Hung et al., 2024) adopts flow matching to improve the performance. Despite the improvement, existing TTA methods still lag significantly in their ability to model relationships between audio events in the generated audio.

**Audio Events Relation Modeling** Based on how audio interact with the physical world in space, time and perceptual aspects, the resulting audio events exhibit complex relationships in spatial, temporal and compositional aspects. Prior work has partially addressed modeling certain temporal relations (*e.g.*, order) in TTA (Xie et al., 2024) and compositional reasoning (Ghosh et al., 2024) for discriminative tasks, such as audio classification and audio-text retrieval. While prior research has touched on modeling audio event relations, their potential in TTA remains largely underexplored. If we analogize an audio event to an object in image, the corresponding relationships exhibited in an image are mainly limited to 2D spatial relationship (*e.g.*, before, bottom, left). Despite object of interest spatial relationship learning and evaluation have received lots of attention in recent years (Krishna et al., 2016; Gokhale et al., 2022; Okawa et al., 2023), the research on audio event relation modeling has been almost ignored.

**Prompt Tuning** (Jia et al., 2022; Liang et al., 2025; Lester et al., 2021) is originally proposed in Natural Language Processing (NLP) (Lester et al., 2021) as an efficient alternative to full fine-tuning for large pre-trained models. Prompt tuning method proposes learnable prompts as task-specific continuous vectors that are directly optimized via gradients during fine-tuning. In recent years, prompt tuning has been successfully adopted in computer vision as visual prompt tuning (VPT) (Jia et al., 2022; Sohn et al., 2023) and audio as audio prompt tuning (APT) (Liang et al., 2025; Oiso et al., 2024). Inspired by the prompt tuning, we introduce gated prompt tuning strategy that significantly improves existing TTA models performance on relation-aware generation in a parameter-efficient way.

## 3 Benchmark Relation-Aware TTA

### 3.1 Audio Event Relation Corpus

An audio event refers to a distinct acoustic signal occurrence with specific frequency, duration

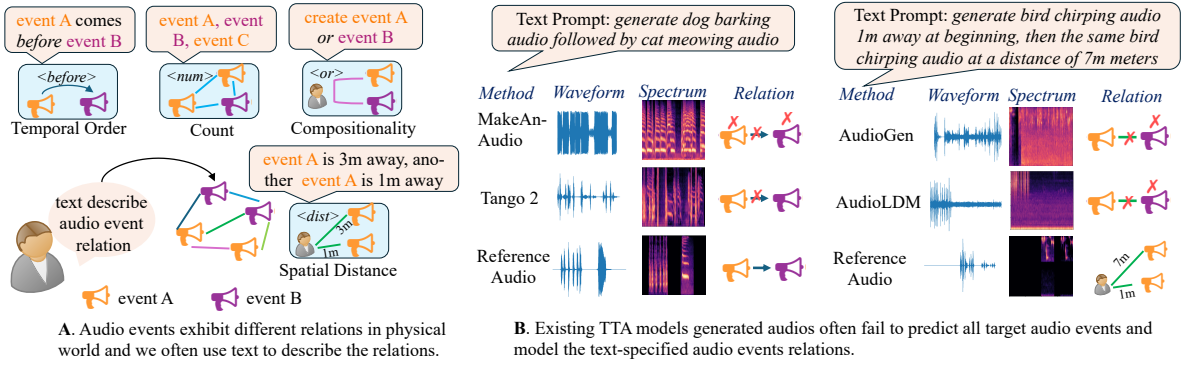


Figure 1: *RiTTA* Motivation: The acoustic world is rich with diverse audio events that exhibit various relationships. While text can precisely describe these relationships (Fig. A), current TTA models struggle to capture both the audio events and the relations conveyed by the text (Fig. B). This challenge motivates us to systematically study *RiTTA*.

Main Relation	Sub-Relation	Sample Text Prompt
Temporal Order	before; after; simultaneity	generate dog barking audio, followed by cat meowing;
Spatial Distance	close first; far first; equal dist.	generate dog barking audio that is 1 meter away, followed by another 5 meters away.
Count	count	produce 3 audios: dog barking, cat meowing and talking.
Compositionality	and; or; not; if-then-else	create dog barking audio or cat meowing audio.

Table 2: Audio Events Relation Corpus.

Main Category	Sub-Category
Human Audio	baby crying; talking; laughing; coughing; whistling
Animal Audio	cat meowing; bird chirping; dog barking; rooster crowing; sheep bleating
Machinery	boat horn; car horn; door bell; paper shredder; telephone ring
Human-Object Interaction	vegetable chopping; door slam; footstep; keyboard typing; toilet flush
Object-Object Interaction	emergent brake; glass drop; hammer nailing; key jingling; wood sawing

Table 3: Audio Events Category Corpus.

and context characteristics that can be attributed to distinguish an independent sound source (He et al., 2021) in an environment. Audio event is ubiquitous in the physical world and serves as the fundamental entity to analyze and interpret the acoustic scene. We embrace the audio event as the fundamental element to construct the relation corpus.

We construct the audio events relation corpus based on two key aspects. First, we consider relations commonly found in the physical world, such as those arising from spatial and temporal variations, which test TTA models’ ability to replicate audio events’ interactions in real-world scenarios. Second, we focus on relations that challenge TTA models’ logical reasoning, evaluating their ability to determine both which audio events to generate and how to generate them. These two aspects partially overlap. Specifically, we define five main audio event categories, each associated with five subcategories of audio events. The detailed relation corpus is provided in Table 2, including,

1. **Number Count:** The number of audio events included in audio, testing TTA models’ ability to address acoustic polyphony challenge.

2. **Temporal Order:** Temporal order refers to the sequence of audio events in the generated audio. We include three basic temporal relations for two audio events: before, after, and simultaneity, testing the TTA models’ ability to distinguish and generate the correct event order as specified in the input text prompt.
3. **Spatial Distance:** Spatial distance refers to the variation in relative spatial distances inferred from the generated audio. It evaluates the TTA models’ ability to capture the spatial distance differences specified in the text prompt. Since we focus on mono-channel audio, obtaining the absolute distance for each audio event is nearly impossible (He et al., 2021). Therefore, we rely on loudness differences within intra-class audio events to verify their spatial distance variations.
4. **Compositionality:** Compositionality relation describes how multiple individual audio events are integrated together to form a complex auditory structure that specified in the input text prompt. It tests TTA models’ logical reasoning capability in determining which audio events to generate and how to structure them, by fol-

lowing the guidance illustrated in the input text prompt. Specifically, we incorporate four main compositionality relations: Conjunction (And, *e.g.*, generate audio A and audio B together); Disjunction (Or, *e.g.*, generate audio A or Audio B, not both); Negation (Not, exclude one particular audio event, *e.g.*, do not generate dog barking audio); Condition (if-then-else, either generate two audio events if the condition is met, otherwise generate the third audio if the condition is not met).

Most of the relations relate to two audio events (see Table 3 for more detail). Expanding to include more complex relations with a greater number of audio events is left for future work.

### 3.2 Audio Event Category Corpus

Alongside the relation corpus presented in Sec. 3.1, we further construct a comprehensive audio event category corpus. The two corpora serve as fundamental dataset for constructing text prompts for TTA models. Since different audio event signals are generated from various sources or through different interactions, we first establish four main audio source categories, further detailing each category with five sub-categories. These constructed audio categories encompass the majority of ubiquitous audio events encountered in our daily lives. Specifically, the audio event category corpus contain,

1. **Human Audio:** the audio generated by human beings in our daily life, including *baby crying, coughing, laughing, whistling, female speech and male speech*.
2. **Animal Audio:** the audio generated by animals, including *cat meowing, dog barking, bird chirping, horse neighing, rooster crowing, sheep bleating and pig oinking*.
3. **Machinery Audio:** audio generated by various machinery devices while they are working, including *car horn, doorbell, telephone ring, paper shredder and boat horn*.
4. **Human-Object Interaction Audio:** human-object interaction audios include *vegetable chopping, keyboard typing, toilet flushing, door slamming and foot step*.
5. **Object-Object Interaction Audio:** we further incorporate object-object interaction audios, including *glass dropping, car emergency brake, hammering nail, wood sawing and keys jingling*.

The detailed audio event corpus is given in Table 3. With the constructed relation and audio event corpus, we can create relation aware text prompts for TTA models.

## 4 Seed Audio and Text-Audio Pair Creation

1. generate audio A succeeded by B;
2. start with A, followed by B;
3. play A initially, B afterwards;
4. generate A preceded by B;
5. A in the beginning, B coming next;

Figure 2: GPT-4 augmented prompts (before relation).

In order to create the corresponding audio for any constructed text prompt, we instantiate each audio event presented in Sec. 3.2 in the main paper with five exemplar seed audios collected from [freesound.org](https://freesound.org)<sup>2</sup>. Since most audio files on [freesound.org](https://freesound.org) are uploaded by volunteers who recorded them in their daily lives, incorporating five exemplar audios for each individual audio event category enhances both the diversity and realism of the seed audio. For instance, in the case of the dog barking audio event, the five selected audios vary in terms of dog breeds and barking styles. To further enhance an audio event’s temporal length diversity, we randomly slice each seed audio into non-overlapping clips ranging from 1 sec to 5 secs. In summary, we have constructed 11 relations (see Table 2 Sub-Relation column), and 25 audio events across five main audio events categories. Each audio event has been associated with 5 diverse audio clips ranging from 1 sec to 5 secs collected from [freesound.org](https://freesound.org) dataset.

**Text Prompt Generation:** a proper audio events relation aware text prompt comprises of two parts: a relation (*e.g.*, <before>) and audio events categories. The audio event categories can be either intra-class or inter-class, and the audio event number depends on the relation. We first instantiate an initial text prompt describing this relation. For example, for the temporal order before relation, the initial text prompt can be like: *generate audio A, followed by audio B*. To enrich the text prompts, we further use the initial text prompt to query LLM (in our case GPT-4) to provide more text prompts with diverse descriptive language for the same relation. One such GPT-4 augmented text prompts is shown

<sup>2</sup>since [freesound.org](https://freesound.org) does not contain meaningful people talking audio, we collect people talking audio from VCTK (Yamagishi et al., 2019)



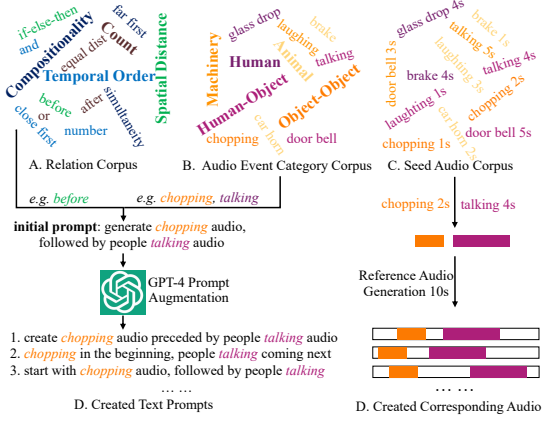


Figure 3: Relation aware <textprompt, audio> pair creation pipeline. It introduces large diversity in both text prompt and audio.

in Fig. 2, which illustrates that the same relation can be exactly expressed by multiple different text prompts. By incorporating GPT-4, we create 5 text prompts for each individual relation.

**Audio Generation:** Given the aforementioned audio events categories and relation, we randomly select an exemplar seed audio for each audio event and further linearly blend them together by satisfying the specified relation. For example, the relation <before> requires two audio events, the two selected audios can be blended together to form the final audio as long as the two seed audios satisfy the <before> relation (Fig. 3, D). Notably, unlike blending two objects in an image that requires careful consideration of factors like occlusion and viewing angle, combining two audio signals simply involves linearly adding them together (Pierce, 2019). This offers an advantage for audio generation, as it eliminates the need for additional operations beyond the specified relation.

The generation of the <text, audio> pair is further illustrated in Fig. 3. With the proposed <text, audio> pair generation strategy, we can create massive diverse pairs even for the same audio events and the same relation, significantly enhancing the diversity and generalization capability of our generated dataset.

#### 4.1 Relation-Aware Evaluation Protocol

Existing TTA methods adopt general evaluation metrics to assess the similarity between generated audio and reference audio, including Fréchet Audio Distance (FAD), Fréchet Distance (FD) (Heusel et al., 2017), Kullback–Leibler (KL) divergence, Fréchet Inception Distance (FID) etc., among oth-

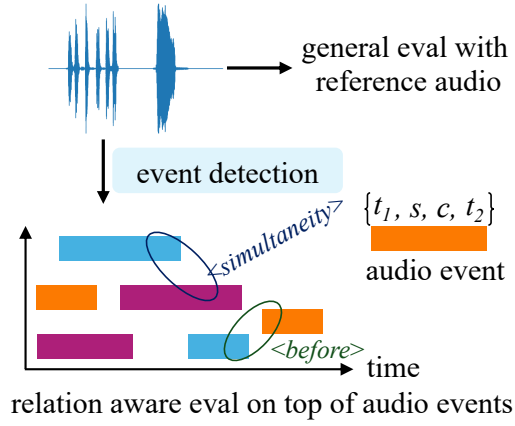


Figure 4: Relation aware evaluation. Audio event detection model is applied to get audio events. The meta data of each event contains start time  $t_1$ , end time  $t_2$ , confidence score  $s$  and class label  $c$ . Various relations can be discovered from these audio events.

ers. While those general evaluation metrics give an overall estimation of the similarity between the two comparing audios, they do not offer direct relation-aware evaluations. In addition to incorporating general evaluation metrics, we further propose multi-stage relation-aware evaluation metrics, with which we can gain insight on how the method performs w.r.t. difference relations.

**General Evaluation Metric:** We incorporate three widely used general evaluation metrics: the objective evaluation FAD, FD and KL divergence scores. FAD and FD measure the distribution similarity with feature embedding extracted from pre-trained on VGGish model (Hershey et al., 2017).

**Relation aware Evaluation Metric:** To directly measure how accurately the text-indicated relation is reflected in the generated audio, we incorporate relation aware metrics for each specified relation.

In relation aware evaluation, we base on the individual audio event to compute the metrics, which allows us to measure the relation between audio events. Let’s denote  $(\mathcal{A}_g, \mathcal{T}, \mathcal{R}, \mathcal{A}_p)$  by ground truth audios, text prompts, relations and generated audios, respectively. We first extract audio events  $\mathcal{E}$  from generated audios  $\mathcal{A}_p$ . For example, for the  $i$ -th generated audio  $a_i^p$ , we apply pre-trained audio event detection model (we use fine-tuned PANNS (Kong et al., 2020), see Sec. A in Appendix) to extract all potential audio events involved in the audio  $E_{a_i^p} = \{(e_j, m_j) | s\}_{j=1}^k$  by a given event confidence threshold  $s \in \mathcal{S}$ , where  $e_j$  is the  $j$ -th audio event and  $m_j$  is the corresponding meta data (e.g., class label, confidence score, temporal start and end time, see Fig. 4). To ob-

tain audio events data for ground truth audios, we can either apply the same pre-trained model or directly extract from text prompts. Finally, we can get  $(\mathcal{A}_g, \mathcal{T}, \mathcal{R}, \mathcal{A}_p, \mathcal{E}_p, \mathcal{E}_g)$ , the relation aware evaluation function  $f(\cdot)$  depends on the audio events  $\mathcal{E}_p, \mathcal{E}_g$  and relations  $\mathcal{R}$ ,  $f(\mathcal{E}_p, \mathcal{E}_g | \mathcal{R}, s)$ . We adopt a multi-stage relation aware evaluation strategy.

**Stage 1: Target Audio Events Presence (Pre).** The paramount requirement for a successful audio generation is the presence of text-specified audio events in the generated audio. In this evaluation, the ground truth audio events and generated audio events are treated as *set*. For a given ground truth and generated audio events pair  $(E_g, E_p)$ , we iterate over each audio event  $e_g$  in the ground truth  $E_g$  to check if it exists in the generated audio events  $E_p$ , regardless of its number and temporal position.

$$f_p(E_p, E_g) = \frac{1}{k} \sum_{e_g \in E_g} \mathbb{1}(e_g, E_p); \quad (1)$$

$$\mathbb{1}(e_g, E_p) = \begin{cases} 1, & \text{if } e_g \in E_p, \\ 0, & \text{otherwise.} \end{cases}$$

where  $k$  is event number in the ground truth.  $s_l(e_g)$  is a potential event meeting the confidence threshold in the generated audio. We select the event with the highest confidence score as the target.

**Stage 2: Relation Correctness (Rel).** Once confirming the aforementioned target audio presence, we further investigate if these audio events obey text-specified relation. The relation is correctly modeled if at least a subset of generated audio events meet the relation. We give score 1 if relation is correctly modeled, otherwise score 0.

$$f_r(E_p | R) = \max_{E_t \in E_p \cap E_g} \{\mathbb{1}(E_t, R)\};$$

$$\mathbb{1}(E_t, R) = \begin{cases} 1, & \text{if } E_t \text{ meets } R, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

**Stage 3: Audio Parsimony (Par).** Apart from requiring to generate all target audios, we should discourage the model from generating excessive intra-class audio events or irrelevant inter-class audio events. We call this property *Audio Parsimony*. Once it is violated, we introduce extra penalty,

$$f_s(E_p, E_g) = \exp(-w_s \cdot |n(E_p) - n(E_g)|) \quad (3)$$

where  $n(\cdot)$  indicates event number.  $w_s$  is the weight adjusting the penalty (in our case,  $w_s = 0.1$ ). The higher audio event number incurs lower parsimony score, the resulting parsimony score lies

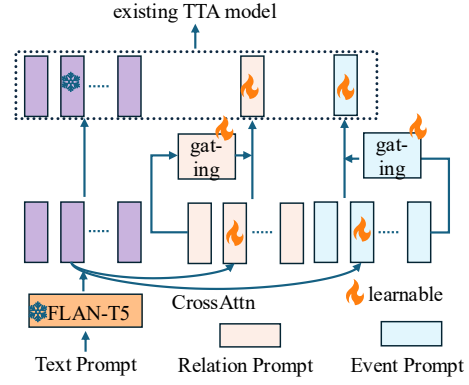


Figure 5: Gated Prompt Tuning Illustration.

within  $(0, 1)$ . The final relation aware score based on event confidence threshold  $s$  equals to the multiplication of the three stage scores,

$$f(\mathcal{E}_p, \mathcal{E}_g | \mathcal{R}, s) = \frac{1}{N} \sum_{(E_p, E_g, R) \in (\mathcal{E}_p, \mathcal{E}_g, \mathcal{R})} f_p(E_p, E_g) \cdot f_r(E_p | R) \cdot f_s(E_p, E_g) \quad (4)$$

where  $N$  is data number. The final average MSR (AMSR) score  $f(\mathcal{E}_p, \mathcal{E}_g | \mathcal{R}, s)$  lies within  $[0, 1)$  (the higher of the score, the better of the model's performance). Following prior COCO object detection evaluation strategy (Lin et al., 2014), we further average across multiple discrete audio event confidence thresholds to get the mean average MSR score (mAMSR),  $f(\mathcal{E}_p, \mathcal{E}_g | \mathcal{R})$ ,

$$f(\mathcal{E}_p, \mathcal{E}_g | \mathcal{R}) = \frac{1}{K} \sum_{s \in \mathcal{S}} f(\mathcal{E}_p, \mathcal{E}_g | \mathcal{R}, s) \quad (5)$$

where  $K$  is the discrete audio event confidence thresholds number. In our case we use uniformly sample four confidence thresholds in range  $[0.5, 0.8]$  with step size 0.1.

## 5 Gated Prompt Tuning

We introduce gated prompt tuning, a new strategy that enables parameter-efficient and task-adaptive relation-modeling without explicitly intervening existing TTA models' design. Specifically, drawing inspiration from recent advancement in prompt tuning (Liang et al., 2025; Jia et al., 2022), we introduce learnable relation prompt for each relation and event prompt for each audio event class, and further feed these prompts alongside text prompts to the TTA neural network for optimization.

Formally, we construct a learnable one-dimensional prompt for each relation and each audio event  $[P_r, P_e], (P_r \in \mathbb{R}^{T_r \times d}, P_e \in \mathbb{R}^{T_e \times d}, T_r$  is the relation number,  $T_e$  is the audio event class

number), the prompt size equals to text prompt token embedding size (in our case, 1024). Instead of directly concatenating all prompts with text prompt tokens, we first condition the learnable prompts on the input text tokens via cross attention (Vaswani et al., 2017). Afterwards, we compute a gated combination for  $P_r$  and  $P_e$  separately, resulting in one integrated relation prompt and another integrated event prompt. By appending these two integrated prompts to text tokens and feeding them to an existing TTA model, we jointly tuning the prompts and the TTA model to instill relation modeling capability into the TTA model.

As each input text just relies on sparse audio events and relations, we adopt the  $\text{entmax}_{1.5}$  (Peters et al., 2019) gating mechanism to encourage the model to focus on a small subset of prompts. Unlike softmax, the  $\text{entmax}_{1.5}$  transformation yields sparse probability distributions, enabling some prompts to be assigned zero weights. To compute this gating, we first extract a summary representation of the prompts using mean average pooling, which is further fed to fully-connect layers to learn the gating logits (the value before  $\text{entmax}_{1.5}$ ). With the  $\text{entmax}_{1.5}$  computed weight, we weight-sum the relation prompts (or event prompts) to get one relation prompt (or one event prompt, accordingly). The whole pipeline is illustrated in Fig. 5.

Since each added prompt is associated with a specific relation or event class, we add a classification loss to each learnable prompt during training phase to encourage each prompt to learn its designated class. It is worth noting that the introduced prompts’ parameter size (5 M) is negligible with comparing with the existing TTA model parameter size (e.g., the Tango parameters is 866 M), and it does not intervene existing TTA model architecture. We experimentally show the effectiveness of this design in distilling relation modeling capability into existing TTA models.

## 6 Experiment

We run two experiments: benchmarking existing TTA methods on our curated 22 hrs dataset (aka testing dataset); comparing gated prompt tuning strategy with other baselines to show its efficiency.

### 6.1 More Discussion on Data Creation

We follow the strategy presented in Sec. 4 to create the dataset. Specifically, for each of the 11 sub-relations in Table 2, we create 720 (2 hrs)

<text, audio> pairs for testing (aka benchmark dataset) and 1440 pairs (4 hrs audio) for training (aka finetuning dataset). The highlight of training/testing dataset is in Table III Appendix.

To ensure all relations can be effectively evaluated, we applied two key constraints during the data creation process. First, to make the audio events countable without ambiguity, we selected inter-category audio events to form the <text, audio> pairs. This avoids the ambiguity that arises when using intra-category events, especially for those with repetitive, similar local occurrences (e.g., multiple instances of dog barking). Second, for the *Spatial Distance* relation, we introduced a temporal order constraint to ensure that the two audio events do not overlap in time. Temporal overlap would require complex source separation models (Petermann et al., 2023) to distinguish individual events. By enforcing this non-overlapping constraint, the evaluation of *Spatial Distance* becomes manageable using an audio event detection model (see Sec. A in Appendix). To make all the proposed relation measurable, we approximate spatial distance by loudness distance. More evaluation setup is given in Sec. B in Appendix.

### 6.2 Relation-Aware Benchmarking Result

We benchmark our curated test dataset on 8 most recent TTA models: AudioLDM (Liu et al., 2023) (two versions), AudioLDM 2 (Liu et al., 2024), MakeAnAudio (Huang et al., 2023b), AudioGen (Kreuk et al., 2023), Tango (Ghosal et al., 2023), Tango 2 (Majumder et al., 2024) and TangoFlux (Hung et al., 2024). We directly depend on their released models to generate a 10 second audio from each text prompt, detailed configuration is given in Table I in Appendix. We further benchmark agentic workflow based methods, in which we leverage LLM (GPT-4) 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.

The quantitative evaluation results across all relations are shown in Table 4. From this table we can observe that the general evaluation results are inconsistent with our proposed relation aware evaluation metrics. The best performing methods un-

Benchmark Method	Model	#param	General Evaluation			Relation Aware Evaluation ( $\uparrow$ )			
			FAD $\downarrow$	KL $\downarrow$	FD $\downarrow$	mAPre	mARel	mAPar	mAMSR
Zero-Shot	AudioLDM (S-Full) (2023)	185 M	5.65	38.95	37.30	2.76	0.50	2.52	0.04
	AudioLDM (L-Full) (2023)	739 M	5.47	38.42	37.96	3.09	0.77	2.56	0.08
	AudioLDM 2 (L-Full) (2024)	844 M	6.68	29.07	35.85	12.26	2.41	10.01	3.39
	MakeAnAudio (2023b)	452 M	9.46	82.72	45.98	8.14	1.68	6.47	1.02
	AudioGen (2023)	1.5 B	6.43	28.01	32.04	9.61	2.12	8.60	2.27
	Tango (2023)	866 M	10.79	90.26	39.46	11.13	2.27	9.88	3.10
	Tango 2 (2024)	866 M	13.84	89.66	44.03	16.63	4.40	12.53	11.55
	TangoFlux (2024)	515 M	8.07	32.80	47.92	33.83	7.02	29.01	76.57
LLM+Agentic+TTA Model	Tango (2023)	866 M	11.88	92.19	41.44	12.33	9.21	11.28	19.17
	Tango 2 (2023)	866 M	14.76	93.10	44.89	19.33	9.37	14.13	20.31
	TangoFlux (2024)	515 M	8.93	32.99	49.00	35.19	9.69	28.22	79.43

Table 4: Benchmark quantitative result across all relations. mAPre, mARel and mAPar are in  $10^{-2}$ . mAPre and mARel can be treated as *presence*, *relation correctness* percentage ratio, in range  $[0, 100]$ . mAPar score also lies within  $[0, 100]$ . mAMSR ( $10^{-4}$ ) lies in range  $[0, 1]$ . The LLM is delimited by GPT-4. The top-, second- and third- performing methods are labeled in different colors.

Model	General Evaluation			Relation Aware Evaluation ( $\uparrow$ )				mAMSR Across Four Main Relations			
	FAD $\downarrow$	KL $\downarrow$	FD $\downarrow$	mAPre	mARel	mAPar	mAMSR	Count	TempOrder	SpatDist	Compos
Tango (finetune)	4.60	23.92	27.03	21.23	10.78	20.35	48.67	8.04	324.10	1.88	44.42
Tango (ours GPT)	3.12	20.21	25.11	25.77	15.38	27.19	59.13	10.11	378.90	3.12	54.87
TangoFlux (finetune)	2.94	20.10	21.09	37.12	11.11	33.99	83.44	13.56	368.77	5.10	58.88
TangoFlux (ours GPT)	1.60	17.98	18.20	43.11	15.33	39.10	127.98	13.04	425.98	3.10	69.56

Table 5: The quantitative comparison across general and relation-aware evaluation between finetuning strategy and our proposed gated prompt tuning strategy. The gated prompt tuning (GPT) just introduces extra 5 M parameters.

der generational evaluations (the two AudioLDM versions) perform the worst under relation aware evaluations, and vice versa. These discrepancies highlight the necessity of proposing evaluation metrics specifically tailored for audio events relations. Additionally, while the performance differences among the seven benchmarking methods under general evaluation are relatively minor, the corresponding differences under relation aware evaluation are significantly more pronounced (e.g., Tango 2 outperforms AudioLDM (S-Full) by about 200 times). However, even the top-performing method, TangoFlux (Hung et al., 2024), still struggles to model audio events relations, as both its presence accuracy and relation accuracy rate are small (less than 30% accuracy rate on average). Moreover, agentic workflow performs slightly better under relation-aware evaluation but worse under general evaluation, which shows LLM involved agentic workflow cannot address relation modeling sufficiently. All of these observations demonstrate the limitations of existing TTA methods in modeling audio events relation and the necessity to systematically study audio events relation in TTA, highlighting the importance of our proposed work. More experimental result is given in Sec. F in Appendix.

In summary, we conclude that, 1. existing TTA models lack the ability to model audio events relation described by the text prompt in the generated

audio, emphasizing the importance of our work in systematically study audio events relation in TTA. 2. Existing TTA evaluation metrics fall short in accurately measuring audio events relations from the generated audio. Our proposed multi-stage relation evaluation framework suffices to measure the relation accuracy from various aspects. 3. LLM involved agentic workflow does not suffice to address relation modeling.

### 6.3 Gated Prompt Tuning Result

We run gated prompt tuning on two most advanced TTA models, Tango and TangoFlux, by initializing their model weights from pretrained model. All the learnable prompts are randomly initialized, each prompt is of size 1024. For Tango, we tune learnable prompts (5 M) and latent diffusion model (UNet, 866 M). For TangoFlux, we tune learnable prompts and Transformer blocks (515 M). We use Adam optimizer with the learning rate  $3 \times 10^{-5}$ , batch size of 16, SNR gamma value 5. We finetune 40 epochs on 4 A100 GPUs. The results, shown in Table 5, show that 1. finetuning either Tango or TangoFlux results in significant performance improvement in both relation-aware and general evaluation, and TangoFlux achieves better performance than Tango. 2. The introduced gated prompt tuning further improves the performance drastically. Considering the fact that gated prompt



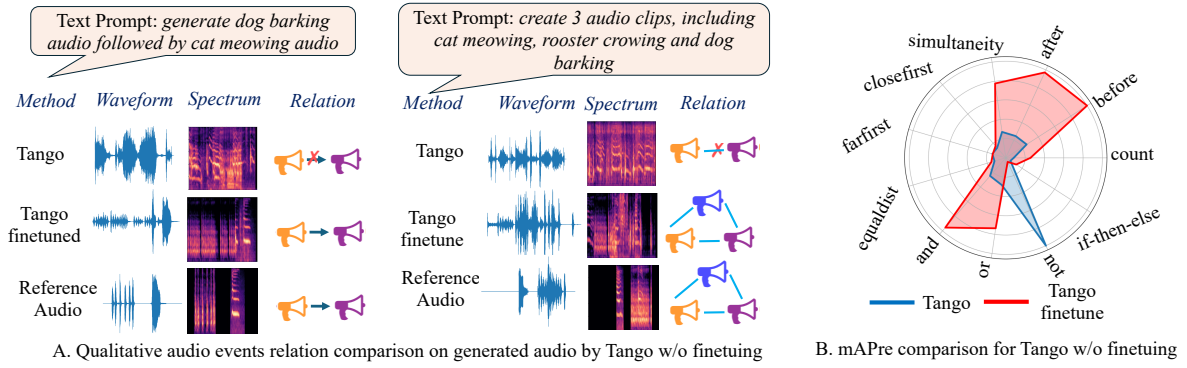


Figure 6: Qualitative visualization comparison of Tango w/o finetuning (A) and mAPre w.r.t. 11 sub-relations.

tuning just introduced negligible parameters, the performance gain directly shows the effectiveness of our proposed gated prompt tuning design.

Two qualitative examples of w/o finetuning Tango are in Fig. 6 A. It is evident that the finetuned Tango successfully models the <before> relation (Table 1 and Fig. 1 show all existing TTA models fail on this case), and <count> relation. The mAPre score w.r.t. the 11 600 sub-relations is shown in Fig. 6 B (the mARel, mA- 601 Par, mAMSR are in Fig. I in Appendix).

#### 6.4 Ablations on Gated Prompt Tuning

GPT-only	GPT-Rel	GPT-Event	PT-only	GPT (ours)
40.12	91.98	87.65	102.33	127.98

Table 6: Ablation Study Result on TangoFlux Model. We report mAMSR ( $\uparrow$ ) score ( $10^{-4}$ ).

We run 4 main ablations on TangoFlux (Hung et al., 2024) based gated prompt tuning (GPT) to validate the effectiveness of our gated prompt tuning design. 1. **GPT-only**: just train gated prompts without tuning the existing TTA model, which helps to test if tuning existing TTA model is necessary; 2. **GPT-Rel**, just introduce learnable relation prompts, which helps to test if jointly tuning event prompts is necessary; 3. **GPT-Event**, just introduce learnable event prompts, which in turn tests if jointly tuning relation prompts is necessary; 4. **PT-only**, prompt tuning without gating mechanism (just mean-average all prompt to obtain the final single prompt).

The mAMSR metric is given in Table 6. From this table we can observe that just tuning gated prompts leads to obvious performance drop (it performs even worse than zero-shot based benchmark, see Table 4). We assume that such large difference results from the domain gap between conventional TTA and our introduced relation-aware TTA task,

which naturally requires to tune large number of parameters to fill in the gap; Discarding either relation prompts (GPT-Event) or events prompts (GPT-Rel) also leads to obvious performance drop; Removing gating mechanism also sees performance drop (but still performs better than GPT-Rel and GPT-Event); All of these ablation study results shows the importance of each of our introduced gated prompt tuning component.

## 7 Conclusion

We demonstrate existing TTA models struggle with relation modeling. Despite its importance, relation modeling has received rare attention in previous research. To address this gap, we introduce a new benchmark, a relation-aware evaluation metric, and a gated prompt tuning strategy. More discussion is in Sec. G in Appendix.

## 8 Limitations

There are two main limitations in this work.

First, in this work, we incorporate 11 relation and 25 audio event in the relation corpus and event corpus, respectively. They are not sufficient enough to reflect the potential relations and audio events existing in the real scenarios. It is desirable to scale up the benchmark by introducing more audio event categories and accommodating more complex relations (e.g., the nested combination of the 25 relations to generate more complex relation).

Second, a powerful relation-aware TTA model should be scalable to extend to incorporate new relations or new audio events automatically (opened setting). Our current setting is close-ended, disallowing novel audio events or relation handling. This opened relation-aware TTA model also remains as a future research direction.

## References

- Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-Audio Generation using Instruction Tuned LLM and Latent Diffusion Model. In *ACM International Conference on Multimedia (ACMMM)*.
- Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models. In *International Conference on Learning Representations (ICLR)*.
- Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2022. Benchmarking Spatial Relationships in Text-to-Image Generation. *arXiv preprint arXiv:2212.10015*.
- Francois Grondin, James Glass, Iwona Sobieraj, and Plumbley Mark D. 2019. A study of the complexity and accuracy of direction of arrival estimation methods based on gcc-phat for a pair of close microphones. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*.
- Yuhang He and Andrew Markham. 2023. SoundSynp: Sound Source Detection from Raw Waveforms with Multi-Scale Synperiodic Filterbanks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Yuhang He, Niki Trigoni, and Andrew Markham. 2021. SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform. In *International Conference on Machine Learning (ICML)*.
- 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. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- IJ Hirsh, C Milliman, and F Darley. 1967. Brain mechanisms underlying speech and language.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Neural Information Processing Systems (NeurIPS)*.
- Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023a. [Make-An-Audio 2: Temporal-Enhanced Text-to-Audio Generation](#). Preprint, arXiv:2305.18474.
- Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: Text-to-audio Generation with Prompt-enhanced Diffusion Models. *International Conference on Machine Learning (ICML)*.
- Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. 2024. [TangoFlux: Super Fast and Faithful Text to Audio Generation with Flow Matching and Clap-Ranked Preference Optimization](#). Preprint, arXiv:2412.21037.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representation (ICLR)*.
- Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark Plumbley. 2020. PANNS: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.
- Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually Guided Audio Generation. *International Conference on Learning Representations (ICLR)*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*.
- Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. Human-level Concept Learning through Probabilistic Program Induction. *Science*, 350(6266):1332–1338.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Jinhua Liang, Xubo Liu, Wenwu Wang, Mark D. Plumbley, Huy Phan, and Emmanouil Benetos. 2025. Acoustic Prompt Tuning: Empowering Large Language Models With Audition Capabilities. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*.
- Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jisheng Lu, and Xiu Li. 2024. BATON: Aligning Text-to-Audio Model with Human Preference Feedback. *International Joint Conference on Artificial Intelligence (IJCAI)*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 41–45.
- Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *International Conference on Machine Learning (ICML)*.
- Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.
- Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning Diffusion-based Text-to-Audio Generations Through Direct Preference Optimization. In *ACM International Conference on Multimedia (ACMMM)*.
- Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. In *arXiv preprint arXiv:2212.08751*.
- Hideyuki Oiso, Yuto Matsunaga, Kazuya Kakizaki, and Taiki Miyagawa. 2024. Prompt Tuning for Audio Deepfake Detection: Computationally Efficient Test-time Domain Adaptation with Limited Target Dataset. In *Interspeech*.
- Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hideori Tanaka. 2023. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Darius Petermann, Gordon Wichern, Aswin Subramanian, and Jonathan Le Roux. 2023. Hyperbolic Audio Source Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse Sequence-to-Sequence Models. In *Proc. ACL*.
- Allan D. Pierce. 2019. Acoustics: An Introduction to Its Physical Principles and Applications. In *Springer International Publishing*.
- Yuanyuan Qu, Xuesheng Li, Zhiliang Qin, and Qidong Lu. 2022. Acoustic Scene Classification based on Three-dimensional Multi-channel Feature-correlated Deep Learning Networks. In *Scientific Reports*.
- Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kihyuk Sohn, Yuan Hao, José Lezama, Luisa Polania, Huiwen Chang, Han Zhang, Irfan Essa, and Lu Jiang. 2023. Visual prompt tuning for generative transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jacob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyu Wu. 2024. AudioTime: A Temporally-aligned Audio-Text Benchmark Dataset. *arXiv preprint arXiv:2407.02857*.
- Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).
- Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diff-sound: Discrete Diffusion Model for Text-to-sound Generation. *IEEE Transactions on Audio, Speech and Language Processing*.
- Jeffrey M Zacks, Nicole K Speer, Khen M Swallow, Todd S Braver, and Jeremy R Reynolds. 2007. Event Perception: A Mind-Brain Perspective.

## Appendix

### A Audio Event Detection Model Finetune

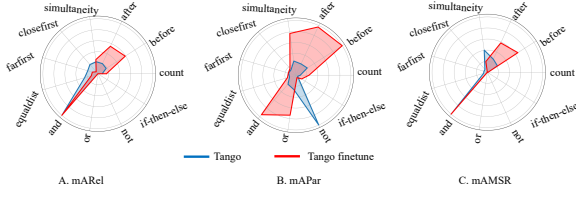


Figure I: The comparison of mARel, mAPar, mAMSR on Tango w/o finetuning.

To detect the audio events from generated audio, we employ a pre-trained audio event detection model (in our case, we adopt PANNS (Kong et al., 2020)) to detect all audio events, each detected event has class label with a confidence score, start time and end time. Analyzing these detected audio events can uncover various audio events relations (see Fig. 4 in the main paper).

The PANNS model (Kong et al., 2020) is pre-trained on the large-scale 527 class AudioSet dataset (Gemmeke et al., 2017). It contains an audio tagging model and an audio event detection model. Directly applying the pre-trained detection model to detect audio events from our generated audios inevitably results in false positive and ambiguous detections. For instance, a *door slam* sound may be incorrectly detected as speech or music with high confidence scores. To mitigate the ambiguity and inaccuracies, we finetune the detection model (“Cnn14\_DecisionLevelMax” variant) on our specially curated 100 k dataset by just tuning the last classification layer. Finally the finetuned model achieves mAP 0.57 on our curated 10k test sets, far outperforming the original model with mAP 0.43.

We based on the pretrained PANNS (Kong et al., 2020) audio event detection model to finetune it on our curated 100 k audio training dataset. Each audio is 10 s long with sampling rate 16 kHz. Moreover, each audio randomly contains one to five audio events, each event has a random start time position in the 10 s long audio. The input is 10 s long audio waveform. The output is a confidence map of shape [20, 25], where 20 is the time steps with the temporal resolution 0.5 s and 25 is the audio event class number. Potential audio events are extracted from the confidence map by thresholding the confidence map, audio events with too short time duration (in our case, less than 0.5 s) are discarded.

The training and testing datasets size are 100 k and 10 k respectively. We adopt Adam (Kingma and Ba, 2015) to train the model with initial learning rate 0.0001 but decays every 200 epochs with decaying rate 0.5. Finally, we train 350 epochs. The loss function is binary cross-entropy loss (BCE). On the testing dataset, the finetuned model achieves mAP 0.57. We use the finetuned audio event detection model to detection audio events from the generated audios.

### B More Discussion on RiTTA Evaluation

We specifically adjust the audio generation process for relations under *Compositionality* and *Spatial Distance* to so as to ensure these relations can be accurately evaluated under our proposed framework.

First, we skip general evaluation for <Not> as it lacks a corresponding ground truth reference audio. During finetuning, we generate silent audio for <Not> for create finetuning pairs. Second, for the <if-then-else> and <Or> sub-relations, which correspond to two possible ground truth audios, we handle evaluation by computing the L2 distance (in the time domain) between the generated audio and the two reference audios. For example, for the prompt *if event A then event B, else event C*, the first reference is the combination of events A and B, while the second contains only event C. We use the reference audio with smaller L2 distance to the generated audio for general evaluation.

Third, precise evaluation of the three sub-relations (<closefirst>, <farfirst>, and <equaldist>) under *Spatial Distance* from unconstrained audio requires sound event detection and localization (SELD (He and Markham, 2023; Grondin et al., 2019)) techniques to spatially localize each audio event, which is impossible with mono-channel audio. To address this, we approximate spatial distance by calculating the loudness, which can be estimated using the L2 norm of the audio waveform. The rationale behind this approach is that greater distances result in a dampening of waveform amplitude (and consequently reduced loudness) due to energy decay along the audio propagation path. When the loudness difference exceeds a predefined threshold (for <closefirst>, <farfirst>) or is within that threshold (for <equaldist>), we consider the evaluation accurate. Specifically, we use a loudness reduction ratio  $\sigma_1$  (with  $\sigma_1 = 0.2$  in our case). For <closefirst>, if the closer



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) (2023b)	guidance_scale=3.5, random_seed=45, n_candidates=3
MakeAnAudio (2023b)	ddim_steps = 100, scale = 3.0
AudioGen (2023)	model name: audiogen-medium
Tango (2023)	num_steps = 200, guidance=3, num_samples=1
Tango 2 (2024)	num_steps = 200, guidance=3, num_samples=1

Table I: Detail setting for each TTA method.

event’s loudness is at least  $\sigma$  times greater than the further event’s loudness, the relation is considered correct. Similarly, for `<equaldist>`, the loudness difference between the two events should be within  $\sigma_2$  (with  $\sigma_2 = 0.4$  in our case) of the louder event’s loudness. This estimation is also reflected in the data generation process (see Sec 6.1).

### C Existing TTA model Setting

We test 8 most recent TTA models: AudioLDM (Liu et al., 2023) (two versions), AudioLDM 2 (Liu et al., 2024), MakeAnAudio (Huang et al., 2023b), AudioGen (Kreuk et al., 2023), Tango (Ghosal et al., 2023), Tango 2 (Majumder et al., 2024) and TangoFlux (Hung et al., 2024). We depend on their released pre-trained model and use their recommended hyperparameter setting for benchmarking (from their Github page). The detailed setting for each TTA method is given in Table I.

### D More Result on Tango Finetuning

The mARel, mAPar and mAMSR score w.r.t. 11 sub-relations is given in Fig. I.

### E <Text, Audio> pair generation and RiTTA benchmark highlight

The RiTTA <Text, Audio> pair generation pipeline is illustrated in Fig. 3 and RiTTA benchmark summary is highlighted in Table III.

### F Key Findings of TTA models on RiTTA Benchmark

The quantitative evaluation results (mAMSR score) w.r.t the four main relation categories are presented in Table II. We observe that both general and relation-aware evaluations show better performance on *Temporal Order* and *Compositionality* compared to *Count* and *Spatial Distance*. This suggests that the *Count* and *Spatial Distance* relations

pose significant challenges for TTA tasks. Additionally, we visualize the detailed relation aware evaluation results for the 11 sub-relations, highlighting the top three performing methods AudioLDM 2 (Liu et al., 2024), Tango (Ghosal et al., 2023), and Tango 2 (Majumder et al., 2024), in Fig. II. We can observe that all the three methods 1. achieve exceedingly high presence score on Not relation, which is expected since a high **Presence** score (Subfig. A) can be easily obtained by simply not generating the specified audio event. 2. perform well in modeling And relation (Subfig. B) (then `<equaldist>` and the three relations in *Temporal Order*); 3. exhibit strength in generating concise audios particularly for Not relation (Subfig. C). Overall, all the three methods excel in modeling And relation and then the three sub-relations in *Temporal Order*, which is also reflected by the result in Table II. The key findings from the relation-aware benchmarking are summarized in the Table IV.

The key findings of TTA models is summarized in Table IV.

### G Conclusion and Future Works

Complex relationships within audio bring the world to life. While text-to-audio (TTA) generation models have made remarkable progress in generating high-fidelity audio with fine-grained context understanding, they often fall short in capturing the relational aspect of audio events in real-world. The world around us is composed of interconnected audio events, where audio event rarely occurs in isolation. Simply generating single sound sources is insufficient for producing realistic audio that reflects the richness of the world.

To analyze the capabilities of current state-of-the-art TTA generative models, we first conduct a systematic study of these models in audio event relation modeling. We introduce a benchmark for this task by creating a comprehensive relational corpus

Table II: Benchmark quantitative result w.r.t. the four main relations. We report FAD score and mAMSR score for general evaluation and relation aware evaluation, respectively.

Model	General Evaluation (FAD ↓)				Relation Aware Eval. (mAMSR ↑)			
	Count	TempOrder	SpatDist	Compos	Count	TempOrder	SpatDist	Compos
AudioLDM (S-Full) (Liu et al., 2023)	3.85	6.86	4.56	9.36	0.00	0.05	0.00	0.18
AudioLDM (L-Full) (Liu et al., 2023)	3.68	6.45	4.10	8.98	0.00	0.05	0.06	0.17
AudioLDM 2 (L-Full) (Huang et al., 2023b)	5.03	8.94	4.72	9.41	0.14	1.87	1.46	9.89
MakeAnAudio (Huang et al., 2023b)	6.02	10.21	8.18	12.78	0.12	0.66	0.44	2.40
AudioGen (Kreuk et al., 2023)	6.14	8.39	3.38	9.98	0.32	3.83	0.48	4.18
Tango (Ghosal et al., 2023)	8.54	10.25	10.11	13.97	0.16	3.44	0.82	8.10
Tango 2 (Majumder et al., 2024)	10.01	13.91	13.23	17.04	0.96	20.92	1.92	23.25
TangoFlux (Hung et al., 2024)	6.02	8.00	52.20	60.22	1.72	111.51	0.00	47.71

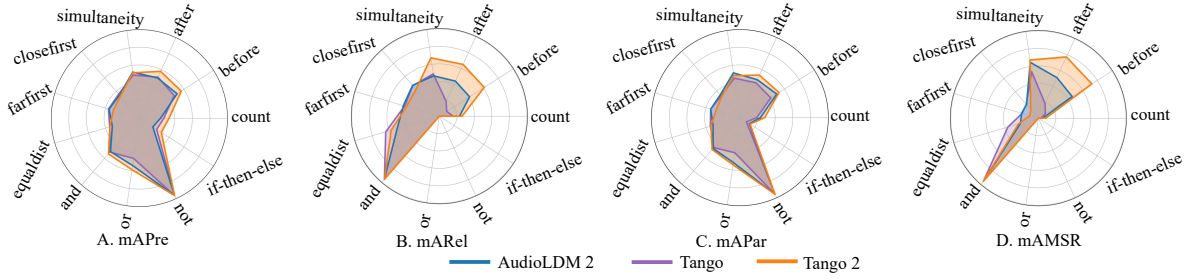


Figure II: Top 3 performing in audio events relation modelling TTA methods' performance w.r.t. the 11 sub-relations. We report mAPre, mARel, mAPar and mAMSR scores separately.

covering all potential relations in the real-world scenarios. Further, we propose new evaluation metric framework to assess audio event relation modeling from various perspectives. Additionally, we propose a finetuning strategy to boost existing models' ability in modelling audio events relation, and we show improvement across all relation metrics. Finally, we will release both the dataset and the code for the evaluation metrics, which will be useful for future research in this domain.

Going forward, our work provides a unique research opportunity to bring the world to life by exploring ways to generate long-term audio events to acoustically understand the physical world. Further, understanding the successes and failures of these models in generating such complex audio events is another promising research direction. This analysis could lead to further improvements in TTA models and their applications in areas such as virtual reality, cinema and immersive media.

Entry	Highlight
seed audio	one event has 5 audios each has 1 s-5 s audio clips
audio category corpus	5 main categories 25 sub-categories
relation corpus	4 main 11 sub relations
relation - event number	count: 2-5 events; Not: 1 event; if-then-else: 3 events others: 2 events.
train and test data info	each audio is 10 s long sampling rate 16 kHz train: 44 hrs, 1.6 k pairs test: 22 hrs, 0.8 k pairs
data creation constraint	count inter-category audio <i>SpatialDist</i> intra-category and require temporder
audio diversity	one event → multi-audios; seed audio → multi time len; seed audios various start time
text prompts diversity	GPT-4 augmented prompts; one template → multi-events.

Table III: RiTTA benchmark highlights.

- 
1. generation eval. contradicts with RiTTA eval.
  2. *TemOrder/Compos* better than *Count/SpatDist*
  3. event presence in Not is the highest;
  4. relation correctness in And is the highest;
  5. parsimony score in Not is the highest;
  6. event presence accuracy rate is below 1%;
  7. relation correctness accuracy rate is below 1%;
  8. An average of 2 redundant audio events;
- 

Table IV: Key findings from experiments of TTA models on our RiTTA benchmark.