Audio Generation with Multiple Conditional Diffusion Model

Zhifang Guo[1,2]**, Jianguo Mao**[1,2]

Rui Tao[2,3]**, Long Yan**[3]**, Kazushige Ouchi**[3]**, Hong Liu**[1]**, Xiangdong Wang**[1] *

[1]Bejing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Toshiba China R&D Center, Beijing, China
guozhifang21s@ict.ac.cn, mao@maojianguo.cn
{taorui, yanlong}@toshiba.com.cn, kazushige.ouchi@toshiba.co.jp, {hliu, xdwang}@ict.ac.cn

Abstract

Text-based audio generation models have limitations as they cannot encompass all the information in audio, leading to restricted controllability when relying solely on text. To address this issue, we propose a novel model that enhances the controllability of existing pre-trained text-to-audio models by incorporating additional conditions including content (timestamp) and style (pitch contour and energy contour) as supplements to the text. This approach achieves fine-grained control over the temporal order, pitch, and energy of generated audio. To preserve the diversity of generation, we employ a trainable control condition encoder that is enhanced by a large language model and a trainable Fusion-Net to encode and fuse the additional conditions while keeping the weights of the pre-trained text-to-audio model frozen. Due to the lack of suitable datasets and evaluation metrics, we consolidate existing datasets into a new dataset comprising the audio and corresponding conditions and use a series of evaluation metrics to evaluate the controllability performance. Experimental results demonstrate that our model successfully achieves fine-grained control to accomplish controllable audio generation.

Introduction

In recent years, significant progress has been made in generation models, with enhanced naturalness observed across diverse domains such as image (Ramesh et al. 2022; Chen et al. 2020; Nichol et al. 2022) and text (Radford et al. 2020; Schick et al. 2023; Kong et al. 2020). This is due to the more stable learning objectives of the models and the availability of extensive paired data. In addition to these fields, audio generation has also attracted significant attention, with strong potential for applications in video creation, virtual reality, and interactive systems.

Audio generation begins with label-to-audio (Pascual et al. 2023; Liu et al. 2021b), but the diversity of generated audio is limited by the classes of labels. Considering that text is a more natural way to express the need of audio, recent research on audio generation primarily focuses on using captions (referred to as *text condition*) to generate semantically coherent audio (Yang et al. 2022; Kreuk et al. 2022;
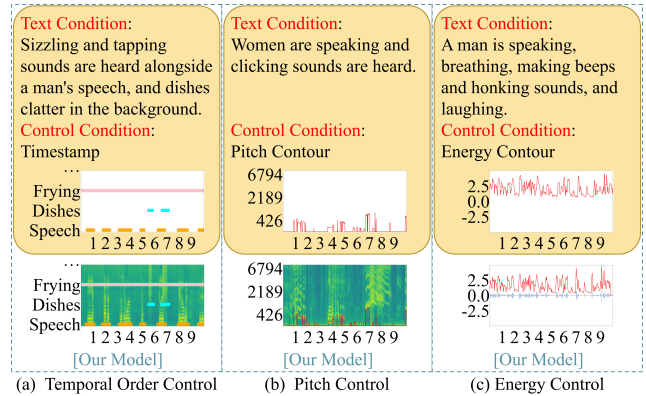

Figure 1: Our model supports fine-grained control by enabling the existing TTA models to accept control conditions. We show temporal order control in Subfigure (a), pitch control in Subfigure (b), and energy control in Subfigure (c).

Liu et al. 2023; Huang et al. 2023; Ghosal et al. 2023). However, in other domains, many generation models have been proposed to explore the use of other conditions (referred to as *control condition*), such as layout (Li et al. 2021a) and keypoint (He, Wandt, and Rhodin 2023), to control the generation process, even employing multiple conditions simultaneously (Li et al. 2023b). These control conditions serve as a guidance for the generation process to approximate specific distributions. However, within the audio domain, there is little research on audio generation with multiple conditions that are necessary for enriching the expression of the text condition. For instance, when generating audio for video or virtual reality, it is important to specify the temporal locations of sound events and precisely control the pitch and energy to enhance users' immersion in the scene. Obviously, it is difficult to precisely convey the intended fine-grained control over the audio through language alone. Therefore, it is reasonable to use other control conditions as a supplement to the text condition to overcome the limitations of language expression and facilitate the understanding of users' intention by the model. As shown in Figure 1, our focus is on three crucial control conditions: timestamp, pitch contour, and energy contour. These conditions are used to control the temporal order of sound events, as well as their pitch and

---

energy, respectively. Temporal order can control the content of the audio, while pitch and energy can control its style. All of these factors are important characteristics for audio.

Inspired by recent advances in the grounded text-to-image model GLIGEN (Li et al. 2023b), we present a new model that expands the capabilities of existing text-to-audio (TTA) models by incorporating additional control conditions while preserving the original text condition for fine-grained control. Our approach first employs a shared control condition encoder to perform encoding for all control conditions. We leverage the success of large language models (LLMs) across various domains (Chen et al. 2021b; Radford et al. 2021; Ruan et al. 2023) and utilize their semantic understanding capabilities to obtain distinguishable semantic representations of sound event classes. To ensure continuity of sound events, we adopt frame-level semantic representations. To preserve the diversity of generation while enabling the pre-trained model to support control conditions, we freeze the weights of the pre-trained model (Zhang and Agrawala 2023; Li et al. 2023b) and train a Fusion-Net to integrate control condition information into the audio generation process. Additionally, we enable one model to support multiple control conditions, allowing parameter sharing and enhancing feature representation, resulting in benefits for each type of control conditions and parameter efficiency.

As the first exploration of audio generation with multiple conditions, there currently lacks datasets and evaluation metrics designed for this task. Therefore, we design a dataset and evaluation metrics as a benchmark for future research: (1) Dataset: We integrate existing datasets to create a new dataset for this task, which contains audio, corresponding text, and control conditions. (2) Evaluation metrics: We employ sound event detection (SED) system detection results to measure the performance of temporal order control. Moments are used to assess the pitch distribution of generated audio, and mean absolute error (MAE) is employed to calculate the similarity between the energy of generated audio and the corresponding reference value. Additionally, we ask experts to listen to the location of sound events as a subjective assessment to complement the results of SED system. Audio samples and our dataset are publicly available[1].

The contributions are summarized as follows:

- We introduce a new task that generates audio guided by both text and control conditions, enabling fine-grained customization of the audio with timestamp, pitch contour, and energy contour.
- We integrate the existing datasets to create a new dataset comprising the audio and corresponding conditions and use a series of evaluation metrics to assess performance, which can serve as a benchmark for future work.
- We propose an audio generation model based on existing pre-trained TTA models, which accepts not only text as a condition but also incorporates other control conditions to achieve finer-grained and more precise control on audio generation. Experimental results demonstrate the effectiveness of our model in generating audio with greater control.

[1] https://conditionaudiogen.github.io/conditionaudiogen/

Related Work

Text-Guided Audio Generation

Diffusion models (Ho, Jain, and Abbeel 2020) have demonstrated remarkable success in generation models across various fields, including image (Chen et al. 2021a), speech (Kong et al. 2021b), and video (Singer et al. 2022). However, the iterative processing of high-dimensional features in diffusion models can lead to slow inference speed. To address this challenge, researchers reduce the dimensions of features during the iteration process and most audio generation models are built on this foundation: first, an encoder is employed to convert audio features into latent representation. Subsequently, a diffusion model operates within this latent space, and the predicted outcomes are rebuilt back to audio features with a decoder. Notably, various improvements have been built upon this foundation: DiffSound (Yang et al. 2022) introduces VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) in a discrete diffusion model to process mel-spectrograms. Instead of training over mel-spectrograms, AudioGen (Kreuk et al. 2022) learns directly from the raw waveform and utilizes classifier free guidance (Jonathan and Tim 2022) under an autoregressive structure for audio generation. More recently, Make-an-Audio (Huang et al. 2023) and AudioLDM (Liu et al. 2023) have conducted an exploration of audio generation within a continuous space. In addition, to enhance training efficiency, Tango (Ghosal et al. 2023) utilizes the language under ing capabilities from a frozen pre-trained LLM to train a diffusion model on a limited training dataset. However, the current conditions mainly involve text, which poses challenges to conveying additional information by natural language. Therefore, we enhance the control capabilities of the text-guided audio generation model by enabling it to be guided by additional conditions including content (timestamp) and style (pitch contour and energy contour), which is simple for users to create the audio that meets their requirements.

Conditional Generation Model

With the increasing availability of diverse paired data, there has been rapid progress in the development of generation models conditioned by various types of information, such as text (Ramesh et al. 2022; Liu et al. 2021a), image (Huang et al. 2023; Li et al. 2023a), audio (Zelaszczyk and Mandziuk 2021; Wang et al. 2023), and layout (Jahn, Rombach, and Ommer 2021; Li et al. 2021b), but they rely solely on one condition, limiting the information transmitted. Therefore, it is crucial to explore how to leverage the existing knowledge within these generation models to extend to new conditions. GLIGEN (Li et al. 2023b) revolutionize text-to-image models by incorporating grounding conditions to control the generation of image and preserving the original text condition. We want to explore the possibilities of extending conditions in the audio domain. In addition, we also want to excavate the relationships between these different conditions and their effects on the audio generation.
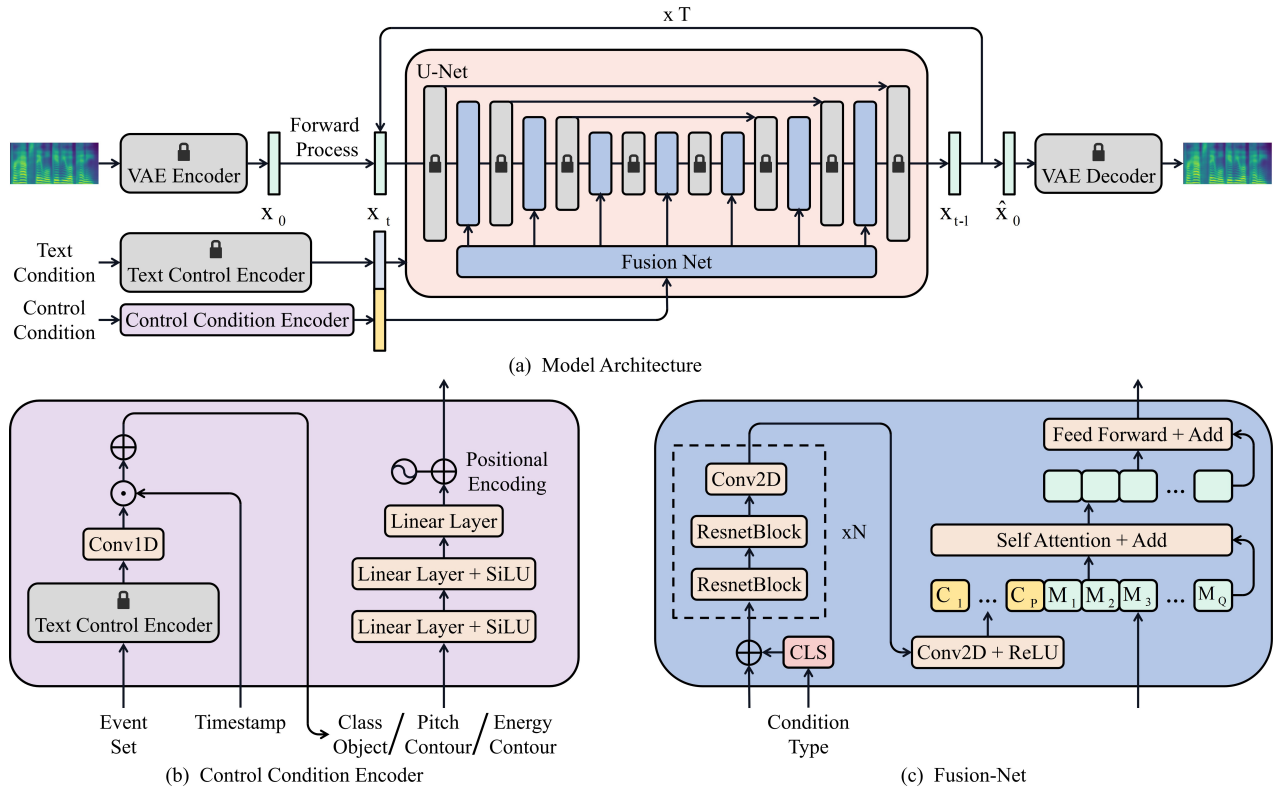
Figure 2: The overview of our model. Subfigure (a) is the architecture of our model that synthesizes audio guided by the text and control conditions. Subfigure (b) shows the control condition encoder that takes the control conditions including timestamp, pitch contour, and energy contour as input and extract control embedding to guide the audio generation. Subfigure (c) demonstrates the Fusion-Net between each layer of the U-Net, enabling the fusion of control embedding.

Method

In this section, we begin with presenting the architecture of our model. Subsequently, we introduce the method and specific details for handling control conditions. Finally, we illustrate the method employed to augment the existing TTA model with control conditions.

Overview

Our model consists of a text condition encoder, a control condition encoder, a conditional latent diffusion model (LDM), a variational auto-encoder (VAE) (Kingma and Welling 2013), and a Fusion-Net. The text and control condition encoders extract the corresponding embeddings, respectively. Under the guidance of the text and control embeddings, the noised mel embedding (for training) from the VAE encoder or standard Gaussian noise (for testing) is used to construct latent representation , which is rebuilt back to the mel-spectrogram by the VAE decoder. And we use a vocoder to generate the final audio. In order to preserve the generation ability, we freeze the VAE, text condition encoder, and conditional latent diffusion model and use the weights from a pre-trained TTA model, Tango (Ghosal et al. 2023). The overview of our model is shown in Figure 2 (a).

Text Condition Encoder As the instruction-tuned LLM makes a significant improvement in the performance of the TTA model (Ghosal et al. 2023), we employ the pre-trained LLM, *FLAN-T5-LARGE* (Chung et al. 2022), as the text condition encoder to generate text embedding with rich semantic information. And the parameters of the text condition encoder are frozen during the training stage to keep the original semantic representation ability.

Control Condition Encoder As shown in Figure 2 (b), the control condition encoder takes the control condition as input, including timestamp, pitch contour, and energy contour. For the temporal order control, we leverage the language knowledge derived from *FLAN-T5-LARGE* to generate the semantic representation of the event classes and convert it with timestamp into the class object. Then the class object, pitch contour, and energy contour are processed using the same multi-layer perceptron and position embedding to generate the control embeddings of the same dimensions, which will be discussed in *Subsection Control Condition Preprocessing*.

Conditional Latent Diffusion Model Our model can be seen as a LDM that relies on the conditions including the noised mel embedding $\mathbf{x}_t$, the text embedding $\mathbf{c}_{\text{text}}$, and the control embedding $\mathbf{c}_{\text{control}}$. The mel embedding $\mathbf{x}_t$ is ob-

tained through the diffusion (forward) processing of the initial mel embedding $\mathbf{x}_0$ with variance preserving:

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + (1 - \overline{\alpha}_t)\epsilon, \tag{1}$$

where t is uniformly sampled from time steps $\{1, \cdots, T\}$, ϵ is standard Gaussian noise, and $\overline{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$ is based on the predefined noise schedule β_s. Our aim is to produce denoising (reverse) process on the mel embedding $\mathbf{x}_t$ or the standard Gaussian noise ϵ back to the mel embedding $\hat{\mathbf{x}}_0$. For each step t, the LDM training objective solves the denoising problem:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{x,\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})}\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{control}}) - \epsilon\|_2^2, \tag{2}$$

where ϵ_θ is the $(t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{control}})$-conditioned noise estimation. The LDM architecture primarily relies on a U-Net (Yang et al. 2022) with a cross attention mechanism (Vaswani et al. 2017) comprising a series of ResNet (He et al. 2016) and Transformer blocks (Vaswani et al. 2017). Similarly to the standard diffusion model, it takes the noised mel embedding $\mathbf{x}_t$ and time step t as input conditioned by the text embedding $\mathbf{c}_{\text{text}}$ to predict the noise estimation ϵ_θ. Within each Transformer block, a cross attention layer is incorporated to fuse the time step t and text embedding $\mathbf{c}_{\text{text}}$. In order to preserve the generation capability of the original TTA model, we freeze the parameters of each layer of the U-Net.

Fusion-Net In order to incorporate the control embedding $\mathbf{c}_{\text{control}}$ into the LDM, we introduce a trainable Fusion-Net between each layer of the frozen U-Net, which will be discussed in *Subsection Control Condition Fusion*.

Variational Auto-Encoder The VAE (Kingma and Welling 2013) consists of an encoder and a decoder that compresses the mel-spectogram into the mel embedding $\mathbf{x}_0$ and reconstructs the mel-spectogram from the mel embedding $\hat{x}_0$. The VAE is constructed using ResUNet blocks (Kong et al. 2021a) and are trained using a combination of ELBO (Kingma and Welling 2013) and adversarial loss (Isola et al. 2017). We use the VAE checkpoint from AudioLDM (Liu et al. 2023) and freeze its parameters.

Control Condition Preprocessing

In audio generation, besides the text condition, there are various other conditions that can control the generation process. Our control conditions include timestamp, pitch contour, and energy contour and we demonstrate how to extract and process these control conditions in this section.

Timestamp Previous layout-to-image models (Li et al. 2023b; Jahn, Rombach, and Ommer 2021; Li et al. 2021b) commonly employ the bounding box labeled with classes to control the positions of objects in generated image, which map these inputs into a series of fixed-length tokens and every token contains not only category information but also top left and bottom right coordinates. However, when dealing with audio, more accurate representation of timestamp is necessary, as audio generation requires more continuous sound events.

As shown in Figure 2 (b), the event set is a collection of sound event classes with the number of D and the frame-level timestamp $\mathbf{i}_{\text{timestamp}} \in \mathbb{R}^{D \times L}$ is a matrix and L represents the number of frames, within which 1 indicates the presence of a sound event and 0 indicates the absence of a sound event. To distinguish sound events from a semantic perspective, we firstly employ a frozen *FLAN-T5-LARGE* (Ghosal et al. 2023; Chung et al. 2022), which is identical to the text condition encoder, to convert the sound event classes into the semantic representation. To maintain the independence of each sound event, we utilize a trainable 1×1 convolutional layer to convert the output of *FLAN-T5-LARGE* into a label embedding $\mathbf{i}_{\text{label}} \in \mathbb{R}^{D \times H}$ and H refers to the hidden size of the label embedding. Finally, we multiple the label embedding $\mathbf{i}_{\text{label}}$ with with the corresponding position of the frame-level timestamp and sum the dimensions of sound event classes to obtain the frame-level semantic representation class object $\mathbf{i}_{\text{object}} \in \mathbb{R}^{L \times H}$:

$$\mathbf{i}_{\text{object}} = \sum_D \mathbf{i}_{\text{label}} \odot \mathbf{i}_{\text{timestamp}} \tag{3}$$

Pitch Contour Pitch is a crucial characteristic of sound effects, music, and speech, but current TTA models primarily rely on the text for coarse-grained control (Liu et al. 2023) and we aim to achieve fine-grained control by incorporating the pitch contour $\mathbf{i}_{\text{pitch}} \in \mathbb{R}^{L \times H}$. In the domain of text-to-speech, pitch contour exhibits considerable variations and the distribution of pitch poses challenges for models to comprehend (Ren et al. 2021). Therefore, we employ the continuous wavelet transform to decompose the continuous pitch contour and quantize each frame of it into 256 possible values on a logarithmic scale (Suni et al. 2013; Ren et al. 2021), which serves as the control condition and is converted into the control embedding by the control condition encoder.

Energy Contour Energy control is equally important in the field of text-to-speech and we incorporate energy contour $\mathbf{i}_{\text{energy}} \in \mathbb{R}^{L \times H}$ as a control condition during audio generation. Drawing inspiration from the method outlined in Fastspeech 2 (Ren et al. 2021), we calculate the energy of each short-time Fourier transform frame by computing the L2-norm of its amplitude and quantizing on a logarithmic scale, which is encoded using the control condition encoder.

In order to achieve one model supporting multiple control conditions (timestamp, pitch contour, and energy contour), we also apply the operations above to standardize the control conditions to the same dimensions. Subsequently, we utilize a shared multi-layer perceptron to encode them, employing the condition type as a prompt to distinguish different control conditions within the Fusion-Net.

Control Condition Fusion

The existing TTA models have been pre-trained on large scale datasets comprising text-audio pairs, enabling them to generate audio based on diverse and complex text. Our objective is to maintain the high-quality audio generation capability of these large models, which have been trained on billions of audio samples, while also enhancing their capability to support additional conditions. Similarly to the prac-

tice in the image field (Zhang and Agrawala 2023; Li et al. 2023b), we freeze the original weights, and gradually adapt the model by tuning the Fusion-Net.

As shown in Figure 2 (c), the Fusion-Net takes the control embedding as input added with the condition type embedding CLS. We adjust the dimension of the control embedding to match that of the mel-spectogram, after a series of structures resembling the VAE encoder, which includes ResnetBlocks and convolutional layers (Kingma and Welling 2013). The 2-D convlutional layers with different kernel sizes and strides are applied to generate control condition tokens $(C_1, ..., C_P)$ of different dimensions. Between each layer of the fixed U-Net, we concatenate the control condition tokens $(C_1, ..., C_P)$ with the mel tokens $(M_Q, ..., M_Q)$ obtained from the previous layer in the U-Net. Subsequently, we employ a self-attention and only select the output tokens corresponding to the mel tokens for the feed forward block (Li et al. 2023b).

Classifier-Free Guidance

During the process of denoising (reverse) process, we incorporate the classifier-free guidance (Jonathan and Tim 2022) of text embedding $\mathbf{c}_{\text{text}}$ and control embedding $\mathbf{c}_{\text{control}}$ to control the degree of guidance:

$$\hat{\epsilon}_\theta = \omega \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{control}}) + (1 - \omega)\epsilon_\theta(\mathbf{x}_t, t), \qquad (4)$$

where ω refers to the guidance scale. To reduce the model's reliance on guidance, we randomly exclude the guidance for 10% of the training samples.

Dataset

Given that there are no TTA datasets with both text and control conditions, we integrate the existing datasets to a new dataset called *AudioCondition* that consists of the audio along with corresponding conditions. We use audio in AudiosetStrong (Hershey et al. 2021) which contains audio from Audioset (Gemmeke et al. 2017). As for the text condition, we obtain from WavCaps (Mei et al. 2023), a caption dataset based on ChatGPT and processed through a three-stage pipeline to filter noisy data and produce high-quality text, which has designed the text condition for the audio in AudiosetStrong. For the timestamp, we obtain it from AudiosetStrong that annotates audio with frame-level timestamp for 456 sound events. For the pitch contour and energy contour, we extract the values of them from audio with signal processing tools (Morise, Yokomori, and Ozawa 2016; McFee et al. 2015). We employ sound event detection (SED) systems to provide evaluation metrics, but the number of sound events supported by the current SED (Janek and Reinhold 2021b,a, 2022) systems is limited. Therefore, we only select audio that includes sound events [2] supported by SED systems for test set, but our model supports for controlling all sound event classes. We randomly split Audio-Condition into three sets: 89557 samples for training, 1398 samples for validation, and 1110 samples for testing, which are publicly available[1].

[2] Alarm bell ringing, blender, cat, dishes, dog, electric shaver toothbrush, frying, running water, speech, and vacuum cleaner.

Evaluation Metrics

In this section, we introduce evaluation metrics for temporal order, pitch, and energy control: (1) Temporal order control metrics: inspired by the practice in the image field (Li et al. 2023b; Jahn, Rombach, and Ommer 2021) which evaluates the location of generated object using a object detection model, we employ a SED system to provide event-based measures (Eb) and clip-level macro F1 score (At) to assess the temporal order control capability (Mesaros, Heittola, and Virtanen 2016). These metrics evaluate the presence of sound events in the generated audio, as well as the onsets and offsets using the first-place SED system in DCASE 2022 Task 4, PB-SED (Janek and Reinhold 2022, 2021a,b)[3], on AudioCondition test set. (2) Pitch control metrics: to compare the distribution of audio pitch, we computed several moments used in the speech field (Ren et al. 2021) including standard deviation (σ), skewness (γ), and kurtosis (κ) (Bistra et al. 2014; Niebuhr and Skarnitzl 2019) as well as the average dynamic time warping distance (DTW) (Müller 2007) of the pitch distribution between the ground-truth audio and synthesized audio. (3) Energy control metrics: we compute the MAE between the frame-wise energy extracted from the generated audio and the energy in the ground-truth audio (Ren et al. 2021).

Experiment Setup

In this section, we present the experiment setup, including the model configuration and baseline models.

Model Configuration

Our model mainly consists of the frozen module, trainable module, and vocoder. In the frozen module, the structure primarily involves the pre-trained TTA model, Tango (Ghosal et al. 2023),which does not require training. The trainable module comprises the control condition encoder and Fusion-Net. We utilize HiFi-GAN (Kong, Kim, and Bae 2020) as the vocoder for fairness and consistency.

Baseline Models

Control Performance Given the absence of previous work in generating audio guided by both text and control condition or event controlling audio generation solely based on the control condition, we compared our model with Tango (Ghosal et al. 2023) and AudioLDM (Liu et al. 2023) that are able to control temporal order, pitch, and other attributes at a coarse-grained level with the text condition to demonstrate that our model has fine-grained control capability. In order to demonstrate the superior performance of our model, drawing inspiration from the single-condition generation models in other domains (Jahn, Rombach, and Ommer 2021; Li et al. 2023a; Radford et al. 2020), we also design a baseline model that only accepts the control condition without the text condition called control-condition-to-audio (referred to **CCTA**) as baseline.

[3] https://github.com/fgnt/pb_sed/

Settings	Temporal Order		Pitch				Energy
	Eb ↑	At ↑	σ	γ	κ	DTW ↓	MAE ↓
GT	43.37	67.53	54.21	2.02	7.31	−	−
AudioLDM	9.75	46.55	50.96	3.17	14.31	40.87	0.669
Tango	5.20	45.11	40.38	1.23	3.24	36.11	0.553
CCTA	14.57	18.27	50.85	1.80	5.74	11.24	0.230
Ours	**29.07**	**47.11**	**55.68**	1.93	**6.77**	**10.79**	**0.200**

Table 1: The control performance between Ours and baseline models. "GT" stands for the ground-truth recordings.

Module Effectiveness Since our primary focus on the control condition encoder lies in timestamp, we design two baseline models using the weights from Tango (Ghosal et al. 2023) to verify the effectiveness of the control condition encoder: **Tango-Timestamp** uses the architecture of Tango and is fine-tuned on AudiosetStrong with the text condition like *"Music from 4.11 to 10.00 and Squeak from 3.17 to 3.50."* to control the location of sound events with only the text condition. **Ours-Box** replaces our model's control condition encoder with GLIGEN's bounding box process (Li et al. 2023b), which is fine-tuned on AudioCondition with the text and bounding box conditions. We also design a setting **Ours-Add** to verify the effectiveness of the Fusion-Net, where we eliminate the Fusion-Net and integrate the control embedding into the mel embedding $\mathbf{x}_t$.

Result

In this section, we evaluate the performance of our model (referred to **Ours**) and the baseline models in temporal order, pitch, and energy control. Then we compare our model trained under multiple control conditions (that **Ours** employ this setting by default) and single control conditions (referred to **Ours-Sin**). We also perform an ablation study on the design of the control condition encoder and Fusion-Net. Furthermore, we investigate the advantages of classifier-free guidance scales and inference steps. Finally, we ask experts to provide a subjective evaluation of the temporal order control as a supplement to the SED systems' detection.

Control Performance

Temporal Order Control Performance According to an Eb and At in Table 1, the following observations can be made: (1) The excellent performance achieved by PB-SED on GT demonstrates that PB-SED can effectively serve as an evaluation metric. (2) The low performance of CCTA indicates that the text condition is crucial for controlling audio generation, and the more input conditions there are, the more helpful it is for the model to understand users' intention. (3) Ours achieves an Eb of 29.07 % and At of 47.11 %, indicating its capability to generate temporal-order-controlled audio. (4) Ours surpasses the baseline models in terms of both Eb and At, primarily due to its acceptance of control conditions, enabling a fine-grained control to generate audio in a more consistent style and content.

Pitch Control Performance We refer to Table 1 and observe that the σ, γ, and κ of the audio generated by Ours

Settings	Temporal Order		Pitch				Energy
	Eb ↑	At ↑	σ	γ	κ	DTW ↓	MAE ↓
GT	43.37	67.53	54.21	2.02	7.31	−	−
Ours-Sin	29.06	46.91	51.32	**1.95**	6.76	11.23	0.267
Ours	**29.07**	**47.11**	**55.68**	1.93	**6.77**	**10.79**	**0.200**

Table 2: The control performance between multiple and single control condition.

Settings	Temporal Order	
	Eb ↑	At ↑
GT	43.37	67.53
Ours-Box	4.79	40.59
Tango-Timestamp	9.45	46.33
Ours-Sin	**29.11**	**46.91**

Table 3: The control performance of temporal order with different designs for the control condition encoder.

are closer to that of GT. Additionally, the DTW of Ours is significantly smaller compared to the other settings. These results indicate that Ours exhibits superior performance in terms of pitch control.

Energy Control Performance According to Table 1, it is evident that Ours achieves the lowest MAE value, indicating that the energy of Ours is closer to the reference value, which has the best energy control capabilities.

Performance Between Multiple and Single Control Conditions

We conducted a comparative analysis on AudioCondition test set between Ours-Sin that is trained and inferred under a single control condition and Ours that is trained under multiple control conditions and inferred under a single control condition. The results depicted in Table 2 demonstrate that the performance of Ours slightly outperforms Ours-Sin in most evaluation metrics, indicating that training multiple tasks simultaneously enables information sharing and mutual complementation, leading to performance enhancements. Additionally, Ours-Sin consists of three separate models requiring 3177M parameters, while Ours only needs 1076M parameters, which can reduce redundant parameters and enhance the efficiency.

Ablation Study

Control Condition Encoder Design The design of the control condition encoder for the timestamp is a significant contribution of our paper and the encoding process for different control conditions exhibits similarities. Consequently, this section is to show the effectiveness of the control condition encoder by comparing the performance of Tango-Timestamp, Ours-Box, and Ours-Sin to control the location of sound events. The results, as depicted in Table 3, indicate that the baseline models perform similarly to Ours-Sin in terms of At, thereby demonstrating their ability to generate audio with the desired sound events. However, Ours-Sin outperforms the baseline models significantly in terms of Eb,

Guidance	Step	Settings	Temporal Order		Pitch				Energy
			Eb ↑	At ↑	σ	γ	κ	DTW ↓	MAE ↓
–	–	GT	43.37	67.53	54.21	2.02	7.31	–	–
1	200	Ours	11.40	30.68	46.65	1.78	5.11	14.85	0.201
3	200		27.90	46.72	56.73	1.84	5.85	11,08	0.208
5	200		**29.07**	**47.11**	**55.68**	**1.93**	**6.77**	**10.79**	0.200
10	200		25.76	45.78	51.15	1.79	5.82	11.67	**0.176**
5	10	Ours	21.54	39.32	57.39	1.84	5.88	**8.90**	0.242
5	50		30.02	50.14	55.74	1.81	5.55	10.34	0.209
5	100		**31.26**	**50.27**	**55.06**	1.89	6.41	10.77	0.205
5	200		29.07	47.11	55.68	**1.93**	**6.77**	10.79	**0.200**

Table 4: The control performance with different classifier-free guidance scales and inference steps.

Settings	Temporal Order		Pitch			DTW ↓	Energy
	Eb ↑	At ↑	σ	γ	κ		MAE ↓
GT	43.37	67.53	54.21	2.02	7.31	–	–
Ours-Add	12.40	43.28	69.13	1.87	6.43	24.72	0.606
Ours	**29.07**	**47.11**	**55.68**	**1.93**	**6.77**	**10.79**	**0.200**

Table 5: The control performance with different designs for the Fusion-Net.

Settings	Temporal Order	
	Eb ↑	At ↑
GT	100.00	100.00
AudioLDM	25.45	90.00
Tango	15.45	90.00
CCTA	25.75	43.33
Ours-Box	12.71	90.00
Tango-Timestamp	22.45	90.00
Ours-Add	32.71	90.00
Ours-Sin	76.90	96.67
Ours	**84.28**	**96.67**

Table 6: The control performance of temporal order with different designs for subjective controllability evaluation.

implying that the baseline models struggle to precisely control the fine-grained temporal order of these sound events.

Fusion-Net Design To assess the efficacy of Fusion-Net, we substitute it with a summation operation and conduct a comparative analysis of three control conditions. As depicted in Table 5, the findings demonstrate that Ours surpasses the Ours-Add across all metrics, indicating that Fusion-Net plays a crucial role in integrating control conditions into the existing TTA model.

Effectiveness of Classifier-Free Guidance Scales and Inference Steps

Ours is based on the weights of Tango whose audio quality is related to the classifier-free guidance scale and inference steps (Ghosal et al. 2023). Similarly on AudioCondition test set in Table 4, we have observed that these two parameters also play a significant role in control: (1) Classifier-free guidance scales: in the absence of classifier-free guidance whose scale is set to 1, the performance is poor only when the control condition is the timestamp. By increasing the scale to 5, improved performance is achieved across most evaluation metrics. However, further increment in scale leads to a decline in performance, which may be the diversity brought by larger scales hinders the controllability of the model. (2) Inference steps: temporal order control reaches its optimum at step 100, energy control reaches its optimum at step 200, but pitch control does not have a clearly suitable step that allows all indicators to reach their best values, since pitch could be more difficult to model.

Subjective Controllability Evaluation

Considering that the SED system is employed for this task for the first time, and the number of the sound events sup-

ported by the PB-SED system is also limited, we randomly selected 20 audio clips from the test set containing all sound events that are not limited to the 10 sound events detected by PB-SED. And the corresponding generated audio clips are listened by experts to determine the presence of sound events and their onsets and offsets. The results in Table 6 show a similar distribution to the results obtained from the SED system above, indicating that the SED system can serve as an evaluation metric for this task. Furthermore, the performance of Ours on other sound events achieves an Eb of 84.28 % and At of 96.67 %, demonstrating that Ours can also perform well on all sound events.

Conclusion

In this work, we aim to investigate the potential for enhancing the current text-to-audio model by improving its control capability. To achieve fine-grained controllable audio generation, we propose a novel model that incorporates not only the text condition but also the control conditions as supplementary information. To facilitate this research, we integrate existing datasets containing audio and corresponding conditions and introduce a series of evaluation metrics serving as a benchmark for future studies in the audio domain. Experimental results demonstrate that our model significantly improves the precision of audio generation control. In the future, we plan to expand our model's control capability to accept a wider range of control conditions, enabling it to process and utilize more diverse supplementary information.

Acknowledgements

References

Bistra, A.; Grazyna, D.; Bernd, M.; Frank, Z.; Jeanin, J.; and Magdalena, O.-P. 2014. Differences of pitch profiles in Germanic and slavic languages. *International Speech Communication Association (Interspeech)*.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative Pretraining From Pixels. *International Conference on Machine Learning (ICML)*.

Chen, N.; Zhang, Y.; Zen, H.; Weiss, R. J.; Norouzi, M.; and Chan, W. 2021a. WaveGrad: Estimating Gradients for Waveform Generation. *International Conference on Learning Representations (ICLR)*.

Chen, T.; Saxena, S.; Li, L.; Fleet, D. J.; and Hinton, G. 2021b. Pix2seq: A language modeling framework for object detection. *International Conference on Learning Representations (ICLR)*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ghosal, D.; Majumder, N.; Mehrish, A.; and Poria, S. 2023. Text-to-Audio Generation using Instruction Tuned LLM and Latent Diffusion Model. *arXiv preprint arXiv:2304.13731*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, X.; Wandt, B.; and Rhodin, H. 2023. LatentKeypoint-GAN: Controlling GANs via Latent Keypoints. *Conference on Computer and Robot Vision (CRV)*.

Hershey, S.; Ellis, D. P. W.; Fonseca, E.; Jansen, A.; Liu, C.; Channing Moore, R.; and Plakal, M. 2021. The Benefit of Temporally-Strong Labels in Audio Event Classification. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. *arXiv preprint arXiv:2301.12661*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jahn, M.; Rombach, R.; and Ommer, B. 2021. High-Resolution Complex Scene Synthesis with Transformers. *arXiv preprint arXiv:2105.06458*.

Janek, E.; and Reinhold, H.-U. 2021a. Forward-Backward Convolutional Recurrent Neural Networks and Tag-Conditioned Convolutional Neural Networks for Weakly Labeled Semi-Supervised Sound Event Detection. *arXiv preprint arXiv:2103.06581*.

Janek, E.; and Reinhold, H.-U. 2021b. Self-Trained Audio Tagging and Sound Event Detection in Domestic Environments. *Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Janek, E.; and Reinhold, H.-U. 2022. Pre-Training and Self-Training for Sound Event Detection in Domestic Environments. *Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Jonathan, H.; and Tim, S. 2022. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *Transactions on Audio, Speech, and Language Processing*, 2880–2894.

Kong, Q.; Cao, Y.; Liu, H.; Choi, K.; and Wang, Y. 2021a. Decoupling Magnitude and Phase Estimation with Deep ResUNet for Music Source Separation. *International Society for Music Information Retrieval Conference(ISMIR)*.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021b. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *International Conference on Learning Representations (ICLR)*.

Kreuk, F.; Synnaeve, A., Gabrieland Polyak; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2022. AudioGen: Textually Guided Audio Generation. *arXiv preprint arXiv:2209.15352*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Li, Y.; Lu, J.; Shechtman, E.; Lee, Y. J.; and Singh, K. K. 2021a. Collaging Class-specific GANs for Semantic Image Synthesis. *International Conference on Computer Vision (ICCV)*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023b. GLIGEN: Open-Set Grounded Text-to-Image Generation. *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Z.; Wu, J.; Koh, I.; Tang, Y.; and Sun, L. 2021b. Image Synthesis from Layout with Locality-Aware Mask Adaption. *International Conference on Computer Vision (ICCV)*.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *arXiv preprint arXiv:2301.12503*.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; Liu, P.; and Zhao, Z. 2021a. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *AAAI Conference on Artificial Intelligence (AAAI)*.

Liu, X.; Iqbal, T.; Zhao, J.; Huang, Q.; Plumbley, M. D.; and Wang, W. 2021b. Conditional Sound Generation Using Neural Discrete Time-Frequency Representation Learning.

McFee; Brian; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *python in science conference*.

Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2023. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *arXiv preprint arXiv:2303.17395*.

Mesaros, A.; Heittola, T.; and Virtanen, T. 2016. Metrics for Polyphonic Sound Event Detection. *Applied Sciences*.

Morise, M.; Yokomori, F.; and Ozawa, K. 2016. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*.

Müller, M. 2007. Dynamic Time Warping. *Information Retrieval for Music and Motion*, 69–84.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *Proceedings of Machine Learning Research (PMLR)*.

Niebuhr, O.; and Skarnitzl, R. 2019. Measuring a speaker's acoustic correlates of pitch-but which? A contrastive analysis for perceived speaker charisma. *International Congress of Phonetic Sciences (ICPhS)*.

Pascual, S.; Bhattacharya, G.; Yeh, C.; Pons, J.; and Serrà, J. 2023. Full-Band General Audio Synthesis with Score-Based Diffusion. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2020. Language Models are Few-Shot Learners. *International Conference and Workshop on Neural Information Processing Systems (NeurIPS)*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; et al. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *International Conference on Learning Representations (ICLR)*.

Ruan, L.; Hu, A.; Song, Y.; Zhang, L.; Zheng, S.; and Jin, Q. 2023. Accommodating Audio Modality in CLIP for Multimodal Processing. *AAAI Conference on Artificial Intelligence (AAAI)*.

Schick, T.; Yu, J. A.; Jiang, Z.; Petroni, F.; Lewis, P.; Izacard, G.; You, Q.; Nalmpantis, C.; Grave, E.; and Riedel, S. 2023. PEER: A Collaborative Language Model. *International Conference on Learning Representations (ICLR)*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; and Taigman, Y. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv preprint arXiv:2209.14792*.

Suni, A.; Aalto, D.; Raitio, T.; Alku, P.; and Vainio, M. 2013. Wavelets for intonation modeling in HMM speech synthesis. *ISCA Workshop on Speech Synthesis, Proceedings, Barcelona*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. *International Conference and Workshop on Neural Information Processing Systems (NeurIPS)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*.

Wang, P.; Wang, S.; Lin, J.; Bai, S.; Zhou, X.; Zhou, J.; Wang, X.; and Zhou, C. 2023. ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. *arXiv preprint arXiv:2305.11172*.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2022. Diffsound: Discrete Diffusion Model for Text-to-sound Generation. *arXiv preprint arXiv:2207.09983*.

Zelaszczyk, M.; and Mandziuk, J. 2021. Audio-to-Image Cross-Modal Generation. *arXiv preprint arXiv:2109.13354*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.