BATON: Aligning Text-to-Audio Model Using Human Preference Feedback

Huan Liao[1] , **Haonan Han**[1] , **Kai Yang**[1] , **Tianjiao Du**[1] , **Rui Yang**[1] , **Qinmei Xu**[1]
Zunnan Xu[1] , **Jingquan Liu**[1] , **Jiasheng Lu**[2] and **Xiu Li**[1]

[1]Tsinghua University
[2]Huawei Technologies Co., Ltd.

{liaoh22, hhn22}@mails.tsinghua.edu.cn, li.xiu@sz.tsinghua.edu.cn

Abstract

With the development of AI-Generated Content (AIGC), text-to-audio models are gaining widespread attention. However, it is challenging for these models to generate audio aligned with human preference due to the inherent information density of natural language and limited model understanding ability. To alleviate this issue, we formulate the **BATON**, the first framework specifically designed to enhance the alignment between generated audio and text prompt using human preference feedback. Our BATON comprises three key stages: Firstly, we curated a dataset containing both prompts and the corresponding generated audio, which was then annotated based on human feedback. Secondly, we introduced a reward model using the constructed dataset, which can mimic human preference by assigning rewards to input text-audio pairs. Finally, we employed the reward model to fine-tune an off-the-shelf text-to-audio model. The experiment results demonstrate that our BATON can significantly improve the generation quality of the original text-to-audio models, concerning audio integrity, temporal relationship, and alignment with human preference. Project page is available at https://baton2024.github.io.

1 Introduction

Text-to-audio (TTA) generation is a promising application that concentrates on synthesizing diverse audio from text prompts. Recent advances in diffusion-based generative models, such as AudioLDM [Liu *et al.*, 2023a], Make-An-Audio [Huang *et al.*, 2023b], and TANGO [Ghosal *et al.*, 2023], have significantly facilitated audio generation. These models employ latent diffusion models [Rombach *et al.*, 2022] to create high-fidelity audio based on textual content, surpassing the performance of previous state-of-the-art TTA models. Nevertheless, there remains a bias between input text and generated audio, caused by the inherent information density of natural language and limited model ability. As shown in Figure 1, the upper example shows low integrity due to a missed audio event, and the lower example exhibits an incorrect temporal relationship because of a wrong time order.
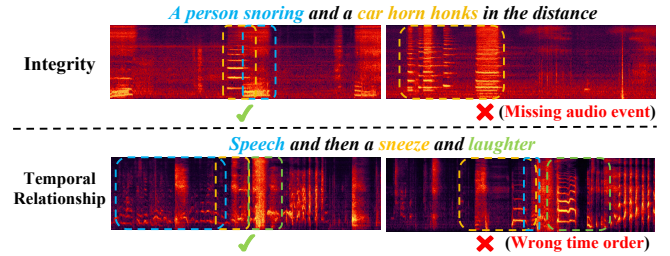

Figure 1: Showcases of two-label and three-label audio samples, with the left indicating alignment and the right indicating misalignment with prompts.

Recent studies have made efforts to mitigate this bias. For instance, Make-An-Audio2 [Huang *et al.*, 2023a] utilized large language model to generate structured captions, addressing issues of poor temporal consistency. Re-AudioLDM [Yuan *et al.*, 2024] tackled the missing generation of rare audio classes through retrieval augmentation. Auffusion [Xue *et al.*, 2024] explored the capabilities of different text encoders to capture fine-grained information such as temporal order and fine-grained events. However, none of these approaches leverage human preference feedback, extensively applied in language [Ouyang *et al.*, 2022; Achiam *et al.*, 2023] and image generation [Lee *et al.*, 2023], to address the aforementioned issue. Therefore, we aim to explore a new perspective: *Can effective alignment between text prompts and generated audio be achieved through human preference feedback in text-to-audio generation?*

To investigate this, we propose a new framework named BATON, designed for aligning TTA models using human feedback. To the best of our knowledge, BATON is the first work to improve the alignment of TTA models through human preference feedback. The BATON involves three crucial steps, as depicted in Figure 2: **(1)** We generate initial text prompts using GPT-4 [Achiam *et al.*, 2023], conditioned on audio labels obtained from AudioCaps [Kim *et al.*, 2019] and a self-build conjunction list, as it has proven to be a reliable annotator [Yang *et al.*, 2023c]. Next, audio is generated for each text prompt using an advanced TTA model, resulting in text-audio pairs. Notably, the generated multi-event audio emphasize the integrity and temporal relationships, aspects that pose challenges for existing TTA models [Xue *et al.*,

2024; Liu *et al.*, 2023a; Ghosal *et al.*, 2023]. Then, we collect binary human feedback for these pairs. **(2)** We introduce and train an audio reward model, building upon the human-annotated dataset. This model is tailored to predict human feedback based on the provided textual input and corresponding audio. **(3)** We fine-tune an off-the-shelf TTA model, TANGO [Ghosal *et al.*, 2023], through reward-weighted likelihood maximization to enhance the alignment of the TTA model with human preference, specifically focusing on the alignment between input text and generated audio. Here, the pre-trained loss of TANGO [Ghosal *et al.*, 2023] is also incorporated as a regularization mechanism, preventing the model from overfitting to the annotated dataset.

We conducted extensive experiments to illustrate that BATON is capable of achieving significant improvements in text-audio alignment. Specifically, BATON demonstrates gains of $+2.3\%$ and $+6.0\%$ in CLAP scores for integrity and temporal relationship tasks, respectively. When evaluated by human annotators, BATON achieves a MOS-Q of 4.55 for the integrity task, surpassing the original model's score of 4.05, and attains a MOS-F of 4.41 for the temporal relationship task, outperforming the original model by 0.58 MOS-F.

In summary, our main contributions are as follows:

- We generate a dataset with 4.8K text-audio pairs across 200 audio event categories, in which 2.7K samples are annotated by human annotators.

- Using the constructed dataset, we train an audio reward model to predict the TTA alignment score reflecting human preference. Our experiments demonstrate that scores predicted by the audio reward model closely correlate with human preference.

- The audio reward models are utilized to enhance an off-the-shelf TTA model in terms of the integrity and temporal relationship of audio events. Detailed experiments provide strong support for its effectiveness.

2 Related Works

2.1 Text-to-Audio Generative Models

The TTA generative models are designed to generate audio that is semantically consistent with text prompt. Currently, their predominant architecture is founded upon the latent diffusion model [Rombach *et al.*, 2022; Ho *et al.*, 2020]. These models engage in a process of noise injection and denoising within the latent space, which is obtained by encoding audio features. And the generated audio representation undergoes transformation into a waveform through a vocoder [Kong *et al.*, 2020a].

Diffsound [Yang *et al.*, 2023a] employs a discrete diffusion model for audio generation from text, operating within the latent space obtained by quantized VAE(VQ-VAE) [Van Den Oord *et al.*, 2017] trained on mel-spectrograms. On the other hand, Audiogen [Kreuk *et al.*, 2022] is built upon a VQ-VAE trained on raw waveform and utilizes an autoregressive language model for the generation of audio guided by text. For the continuous latent space, AudioLDM [Liu *et al.*, 2023a] utilizes audio features obtained from contrastive

language-audio pretraining (CLAP) [Wu *et al.*, 2023] as training conditions and guides the generation of audio through textual features encoded by CLAP, which effectively compensates for the scarcity of paired data. TANGO [Ghosal *et al.*, 2023] harnesses the remarkable text representation capability of the pretrained Large Language Model (LLM), achieving superior performance within the same latent space as AudioLDM, even with limited paired data.

However, challenges remain for generating audio that matches the details of prompt and human perception. Some recent studies have tried to handle this problem in terms of prompt processing and model components design. Make-An-Audio2 [Huang *et al.*, 2023a] employs LLM to generate structured captions to tackle poor temporal consistency. Yuan [Yuan *et al.*, 2024] solves missing generation of rare audio classes with retrieval augmentation. Guo [Guo *et al.*, 2024] uses timestamps condition for controllable audio generation. Auffusion [Xue *et al.*, 2024] explores the ability of different text encoders for capturing fine-grained information like temporal order and fine-grained events.

2.2 Fine-Tuning with Human Preference Feedback

Fine-tuning guided by human preference feedback, stands as an indispensable component within the Reinforcement Learning from Human Feedback (RLHF). This approach proves particularly vital in machine learning, especially when dealing with intricate or ambiguous objectives. Its efficacy spans diverse applications, ranging from gaming, as demonstrated in Atari [Christiano *et al.*, 2017], to more intricate tasks in robotics [Ziegler *et al.*, 2019; Casper *et al.*, 2023], where it significantly enhances the agent's success rate in completing tasks.

The integration of RLHF into the development of large language models (LLMs) signifies a noteworthy milestone in the field. Notable models such as OpenAI's GPT-4 [Achiam *et al.*, 2023], Anthropic's Claude [Anthropic, 2023], Google's Bard [Google, 2023], and Meta's Llama 2-Chat [Touvron *et al.*, 2023] leverage this approach to improve their performance and relevance. Collecting human judgments on response quality is often more feasible than obtaining expert demonstrations. Subsequent works have fine-tuned LLMs using datasets reflecting human preferences, resulting in enhanced proficiency in translation [Kreutzer *et al.*, 2018], summarization [Stiennon *et al.*, 2020], story-telling [Ziegler *et al.*, 2019], and instruction-following [Ouyang *et al.*, 2022]. Additionally, RL and RLHF has been applied to enhance spoken language understanding and train language models for various objectives [Ranzato *et al.*, 2015; Wu and Hu, 2018; Cheng *et al.*, 2023]. Presently, RLHF has been employed to fine-tune diffusion models, contributing to improved image equality, text-image alignment, and image aesthetic scores [Black *et al.*, 2023; Dong *et al.*, 2023; Yang *et al.*, 2023b].

Similar to the RLHF method mentioned above, we also trained a reward model for evaluating the quality of audio and applied it to fine-tuning the pre-trained model. Differently, our approach does not use reinforcement learning algorithms to update the original model. Instead, we used the reward model as a weighting factor to increase the probability of audio with higher rewards.
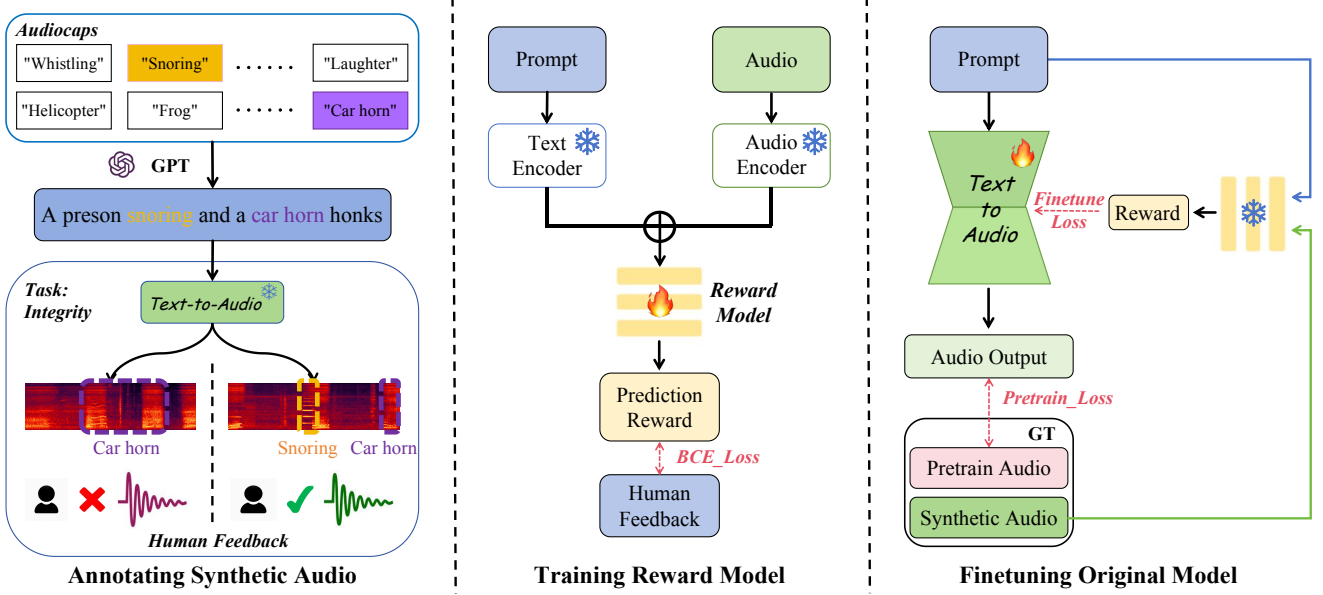
Figure 2: **The framework of BATON.** BATON integrates three modules: (1) An audio generation unit using LLM-augmented prompts, with human-scored annotations; (2) A reward model trained on synthetic data to emulate human alignment preference; (3) A fine-tuning mechanism that enhance the original generative model using reward model combined human-labeled and pre-training datasets.

3 Method

The framework of BATON, as shown in Fig 2, consists of three components. Firstly, we generated a synthetic dataset of audio-text pairs using selected prompts and then annotated it with task-specific human-derived annotations in Section 3.1. Furthermore, we employed the generated data, augmented with human feedback, to train a reward model designed to predict human preferences in audio content, which is illustrated detailedly in Section 3.2. Finally, as elaborated in Section 3.3, we utilized a reward model reflecting human preferences to fine-tune the text-to-audio model, thereby enhancing its task-specific alignment performance.

3.1 Dataset Construction

Recent studies [Huang *et al.*, 2023a; Yuan *et al.*, 2024; Xue *et al.*, 2024] have found that text-to-audio models face challenges in producing audio aligning with human preferences, primarily manifest in low integrity, incorrect temporal relationships, and the like. As shown in Figure 1, the upper example exhibits poor integrity because of a missed audio event and the lower example has an incorrect temporal relationship due to a wrong time order. For simplicity, we here focus on evaluating the integrity and temporal relationships of generated audio to explore the effectiveness of human preference feedback in text-to-audio models. To start, we generate text-audio samples related to these two attributes, and subsequently, human evaluators rate these generated samples.

Data Collection. The left part of Figure 2 illustrates the process of generating audio data concerning integrity and temporal relationships. We specifically select audio event categories that ranked among the top 200 in occurrence within AudioCaps [Kim *et al.*, 2019], constituting our meta labels

Given a label group, X_{label}, the two labels in it will be described into an audio event group: $\{Event_1, Event_2\}$. Then, please join the audio event group with conjunction to form an audio caption, defined as "$Event_1, Conjunction, Event_2$".

Conjunction list: X_{conj}

Please ensure to randomly select conjunctions from the list and avoid relying on a single conjunction.

Please try to generate captions that match human language expressions as much as possible.

Table 1: GPT instruction for integrity task.

denoted as $X_{meta} = \{l_1, ..., l_{200}\}$. Based on that, we randomly select 2 or 3 labels to create a label group, denoted as X_{label}. Conditioned upon the composed X_{label} and a predefined conjunction list X_{conj}, we instruct the GPT-4, named M_T, with prompt P whereby attaining a complete sentence that matches human language:

$$\mathbb{D}_{text} \sim M_T(P|X_{label}, X_{conj}). \tag{1}$$

The prompt P includes the prefix system message and the suffix prompt, guiding M_T to produce the desired audio captions. The prompt for the group with 2 labels is detailed in Table 1. The conjunction list X_{conj} encompasses terms such as *"and"*, *"with"*, *"followed by"*, and so forth (refer to the Appendix for the complete list). Following the generation process, we manually refine and rephrase certain captions to align them more closely with human oral language. As a re-

Category	Labels	Generated Captions
Integrity	"Baby cry, infant cry", "Waves, surf"	A baby cries while waves crash onto the shore.
	"Canidae, dogs, wolves", "Rustle"	Dogs and wolves howl and bark, rustling leaves in the background.
Temproal Relationships	"Crying, sobbing", "Toilet flush", "Female singing"	Young child crying at first, then a toilet flushes, as a woman's singing begins.
	"Rain on surface", "Helicopter", "Engine starting"	Rain on a surface, followed by a helicopter, and then an engine starts.

Table 2: Examples of generated audio captions.

Category	Amount of audio	Human feedback (%)		
		Aligned	Not	Skip
Integrity	1431	56.1	28.4	15.5
Temporal	1332	37.8	59.4	2.8
Total	2763	47.3	43.4	9.3

Table 3: Details of amounts of human feedback dataset.

sult, $\mathbb{D}_{\text{text}}$ encompasses a total of 962 audio captions, with samples presented in Table 2. Subsequently, we employ the TANGO [Ghosal *et al.*, 2023], named M_{A}, to synthesize 5 different audio for each audio caption:

$$\mathbb{D}_{\text{data}} \sim \text{M}_{\text{A}}(\mathbb{D}_{\text{text}}). \tag{2}$$

Thus, we obtained $4,810$ text-audio pairs in total.

Human Annotation. Upon the collected dataset $\mathbb{D}_{\text{data}}$, we selected 2.7 K samples, denoted as $\mathcal{D}_{\text{human}}$, and subjected them to evaluation by human annotators. As presented in Table 3, where $1,431$ samples are related to integrity, and $1,332$ samples are associated with temporal relationships. Annotators employed a binary rating system, assigning either 0 or 1 to each text-audio pair. This is because the nature of integrity or temporal relationship often yields a binary outcome (true or false). Annotators only need to make binary feedback instead of rating a continuous quality score. Admittedly, the more informative human feedback should be adopted for more complex or subjective aspects, such as audio quality and emotional expression. Nevertheless, binary scoring helps to minimize the effects of subjectivity and individual differences for the integrity and temporal relationship tasks, which improves the consistency of the assessment. In practice, we implement a task-specific human preference scoring system, where each audio sample receives an evaluation and is assigned a binary score reflecting its alignment with the given text. A score of **1** indicates adherence to the desired preference (**aligned**), whereas a score of **0** denotes non-conformance (**not aligned**). We have also set the option "skip" to help the annotators resolve ambiguities of the synthetic audio. As recorded in Table 3, 28.4% text-audio samples in the integrity task are incorrect and 59.4% text-audio samples in the temporal relationships task have wrong order, highlighting a substantial gap between generated audio and human preference.

3.2 Audio Reward Model

As shown in Figure 2, upon the established dataset $\mathcal{D}_{\text{human}}$, we developed an audio reward model $r_\phi(c, x)$ (parameterized by ϕ) to predict a scalar reward, which reflects the human preference for the alignment between the provided text c and audio x. Such a reward model comprises a text encoder $\text{E}_{\text{C}}(\cdot)$, an audio encoder $\text{E}_{\text{X}}(\cdot)$, and MLP layers. When given the input text c and generated audio x, the text encoder and audio encoder extract the respective text embedding e_c and audio embedding e_x. Then, e_c and e_x are concatenated along the channel dimension and forwarded to the subsequent MLP layers. The last MLP layer yields a reward that signifies human preference. This process can be summarized as:

$$
\begin{aligned}
e(c, x) &= \text{Cat}(\text{E}_{\text{C}}(c), \text{E}_{\text{X}}(x)), \\
r_\phi(c, x) &= \text{MLP}(e(c, x)).
\end{aligned} \tag{3}
$$

In practice, we exploit the audio encoder $\text{E}_{\text{X}}(\cdot)$ and text encoder $\text{E}_{\text{C}}(\cdot)$ from the CLAP model [Wu *et al.*, 2023] that pre-trained on various text-audio samples. Since the human feedback in $\mathcal{D}_{\text{human}}$ is binarized, i.e., $y \in \{0, 1\}$, we consider the prediction of human preference as a classification problem. Accordingly, the audio reward model is trained by minimizing the binary cross-entropy loss:

$$\mathcal{L}(\phi) = \mathop{\mathbb{E}}_{(c,x,y)\sim\mathcal{D}_{\text{human}}} [-y \log(r_\phi(c, x)) - (1-y) \log(1-r_\phi(c, x))]. \tag{4}$$

The resulting audio reward model is capable of producing an alignment reward that emulates human preference.

3.3 Fine-Tuning the Text-to-Audio Model with Audio Reward

After obtaining the audio reward model $r_\phi(c, x)$, we combine it with the original conditional audio distribution $p(x|c)$, resulting in a new probability distribution $\tilde{p}(x|c) = f(r_\phi(c, x))p(x|c)$. Here, $f(\cdot)$ is a monotonically increasing function and we choose $f(\cdot) = exp(\cdot)$ by default. As the original generative text-to-audio model [Ghosal *et al.*, 2023], our objective is to update the model p with parameters θ by maximizing the probability, i.e., minimizing the following loss function:

$$\mathcal{L}_1(\theta) = \mathbb{E}[-\log \tilde{p}_\theta(x|c)] = \mathbb{E}[-r_\phi(c, x) \log p_\theta(x|c)]. \tag{5}$$

Since diffusion models [Ho *et al.*, 2020; Rombach *et al.*, 2022] are in principle able to model conditional distributions with a conditional denoising autoencoder $\epsilon_\theta(x_t, t, c)$,

Task	Model	FD↓	FAD↓	IS↑	KL↓	S_{CLAP}↑	MOS-Q↑	MOS-F↑
Integrity	TANGO[†][Ghosal *et al.*, 2023]	**36.79**	**2.81**	4.78	**1.03**	33.0%	4.05±0.07	3.70±0.07
	AudioLDM-L[Liu *et al.*, 2023a]	53.63	4.18	4.53	1.76	27.5%	3.70±0.08	2.95±0.08
	AudioLDM2-L[Liu *et al.*, 2023b]	44.01	3.13	4.84	1.44	28.8%	3.93±0.08	3.34±0.08
	BATON (ours)	38.54	3.02	**5.07**	1.11	**35.3%**	**4.55±0.05**	**4.42±0.05**
Temporal	TANGO[†][Ghosal *et al.*, 2023]	36.86	4.02	4.46	1.16	35.9%	4.15±0.06	3.83±0.06
	AudioLDM-L[Liu *et al.*, 2023a]	56.48	4.74	4.66	1.87	27.8%	3.39±0.07	2.79±0.08
	AudioLDM2-L[Liu *et al.*, 2023b]	45.34	**3.05**	4.69	1.50	31.9%	3.98±0.07	3.51±0.08
	BATON (ours)	**36.19**	4.31	**4.88**	**1.13**	**41.9%**	**4.68±0.05**	**4.41±0.05**

Table 4: **Comparison of different text-to-audio models.** TANGO[†] is the baseline model which is ultilized by BATON to fine-tune with human preference feedback.

the Eq. 5 can be simplified as:

$$\mathcal{L}_2(\theta) = \mathbb{E}_{t,(x_t,c)\sim\mathcal{D}_{\text{data}}}[r_\phi(c,x_t)\|\epsilon - \epsilon_\theta(x_t,c,t)\|^2], \quad (6)$$

where x_t is the noise version of the input x. Compared to the original denoising loss, the reward term behaves as a modulating factor: when an input sample aligns with human preference, the modulating factor is large, thus the text-to-audio model pays more attention to learning this good sample, and vice versa. In addition, since our constructed dataset is relatively small compared to the pre-trained dataset, we introduce the denoising loss from the original text-to-audio model [Ghosal *et al.*, 2023]. Consequently, the final loss is:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,(x_t,c)\sim\mathcal{D}_{\text{data}}}[r_\phi(c,x_t)\|\epsilon - \epsilon_\theta(x_t,c,t)\|^2]$$
$$+ \beta\mathbb{E}_{t,(x_t,c)\sim\mathcal{D}_{\text{pretrain}}}[\|\epsilon - \epsilon_\theta(x_t,c,t)\|^2], \quad (7)$$

in which $\mathcal{D}_{\text{pretrain}}$ refers to the sampled data from the pre-trained dataset [Kim *et al.*, 2019] and β is a hyper-parameter. This regularized loss prevents the model from overfitting to our self-build dataset, ensuring it does not deviate heavily from the original model.

4 Experiments

4.1 Implementation Details

Employing TANGO (Full-FT-Audiocaps) [Ghosal *et al.*, 2023] as our baseline, we integrated a human feedback-driven fine-tuning framework. From the Audiocaps dataset, top occurrences 200 labels were chosen, and GPT-4 [Achiam *et al.*, 2023] was utilized to expand these into prompts for text-to-audio synthesis, resulting in a dataset of 962 audio captions and their 4810 resultant audio outputs of which default length is set to 10 seconds. Human annotators then binary score the text-audio pairs, assessing their alignment with preference criteria. We trained the audio reward model on synthetic dataset over 50 epochs, with a batch size 64 and learning rate 0.01 using the Adam [Kingma and Ba, 2014]. During the fine-tuning of the original model, we assigned a weight parameter β of 0.5 to the pretrain loss. The ratio of human-labeled data to reward model-scored data was maintained at 1:1 for both sub-tasks. $\mathcal{D}_{\text{data}}$ and $\mathcal{D}_{\text{pretrain}}$ involve 4.8 K and 2.5K samples, respectively. The fine-tuning process, conducted over 10 epochs with settings including a learning rate of 1×10^{-5}, batch size of 6, and the default optimizer AdamW [Loshchilov and Hutter, 2017], utilized a total of 48GB*2 GPU memory.

4.2 Main Result

Evaluation method. For evaluation, we filtered the Audiocaps test set using words such as *and, as, then, while, before, after, followed* to obtain a multi-labeled test set. Subsequently, we curated subsets comprising 148 two-label and 165 three-label prompts from the filtered data. These two subsets are used to evaluate audio integrity and temporal consistency respectively. Consistent with prior research [Huang *et al.*, 2023a; Liu *et al.*, 2023a; Liu *et al.*, 2023b; Xue *et al.*, 2024], our objective evaluation of audio quality and fidelity employs metrics such as Fréchet Distance (FD), Fréchet Audio Distance (FAD) [Kilgour *et al.*, 2019], Kullback-Leibler divergence (KL) [Yang *et al.*, 2023a]. We adapt Inception Score (IS) to assess the quality and variety of samples. Additionally, we utilized the Cross-Modal Language-Audio Perceptual (CLAP) metric S_{CLAP} to objectively assess the alignment between audio and text. For subjective analysis, 40 participants were recruited to rate the perceived quality and text alignment of the audio separately on a 5-point scale, with 5 being the highest possible score. The Mean Opinion Score (MOS) method, with a 96% confidence interval, is employed to evaluate two aspects: audio quality (MOS-Q) and the faithfulness of text-to-audio alignment (MOS-F).

Quantitative experiment. BATON showcases competitive performance across multiple metrics in text-to-audio tasks, as detailed in Table 4. In evaluating the similarity between generated and real samples, the FAD metric exhibits a slight reduction compared to the original model in both tasks. Considering that FD relies on PANNs [Kong *et al.*, 2020b] classifiers and the labels in our fine-tuned dataset constitute a subset of the PANNs sound class, using FD as the primary metric for assessing audio quality appears to be more reliable. In temporal task, BATON surpasses the original model in terms of FD, IS and KL, indicative of its robustness in maintaining semantic cohesion while minimizing divergence from the target distribution. Additionally, BATON achieves the highest IS in both subtasks, showcasing excellent audio variety. Notably, the superior performance of BATON in S_{CLAP}, particularly a 2.3% improvement over the baseline TANGO in the integrity task and a significant 6.0% enhancement in the temporal task, underscores its effectiveness in producing audio that aligns closely with human preferences. This enhancement can be attributed to our fine-tuning approach that leverages human

Original model   **Finetuned model**

(a) *A **motor hums** softly followed by **spraying***

(b) ***Running** then **spraying***

(c) *A **man talking** then **laughing** as a **duck quack***

(d) *A **man talking** as **music is playing** followed by a **frog croaking***
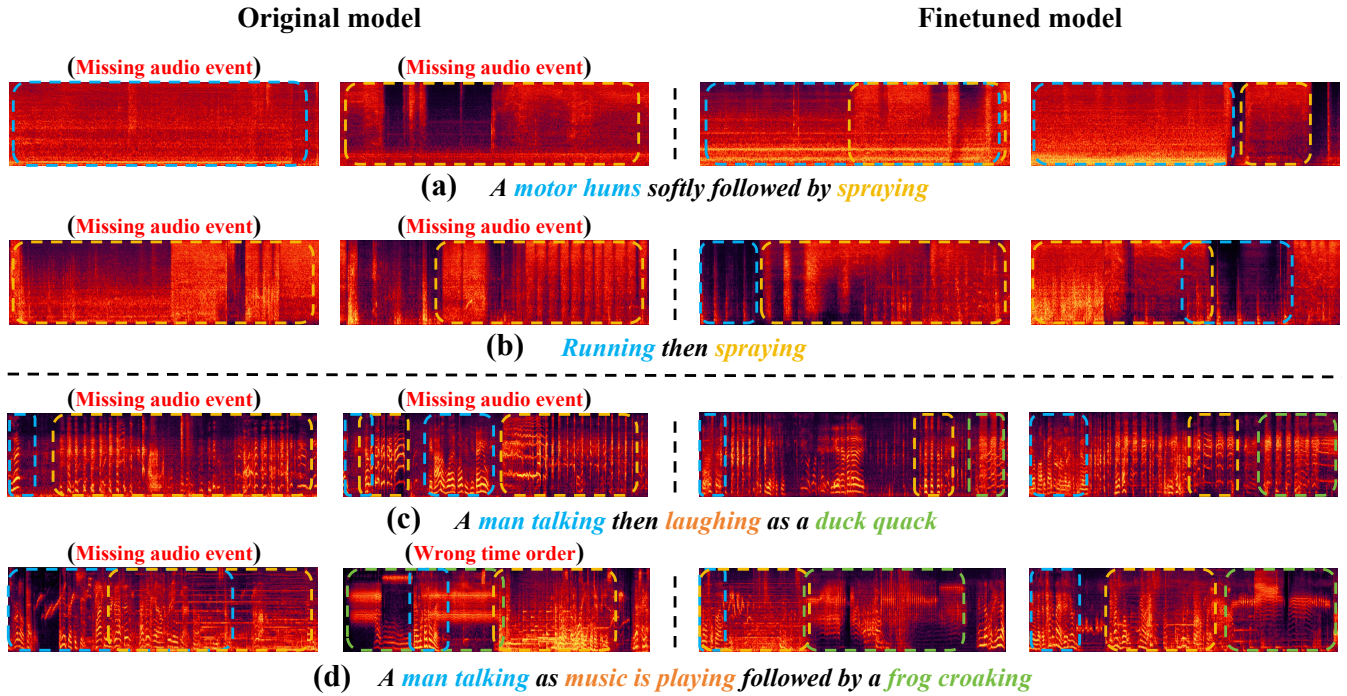
Figure 3: Generated samples comparison of TANGO (Original model-Left two columns) and BATON (Finetuned model-Right two columns). Sample (a), (b) and (c) show that the finetuned model produces more complete audio events in both tasks. Sample (d) demonstrate that the original model generates audio with a confused sequence, whereas the finetuned model adheres to the sequence of prompt in temporal task.

feedback, allowing BATON to better understand and distinguish the nuances of audio content that are valued by human.

When subjectively evaluated on the integrity task, BATON demonstrates a superior MOS-Q of 4.55±0.05, outperforming TANGO (4.05±0.07). Concurrently, BATON's MOS-F is 4.42±0.05, indicating a more faithful text-to-audio alignment compared to TANGO's 3.70±0.07. Significantly, BATON continues to demonstrate superiority in subjective metrics for the temporal task, with its MOS-F surpassing the original model by 0.58 points. This observation indicates its adept handling of temporal information. These findings affirm that BATON not only generates high-quality audio that maintains or even exceeds the original model, but also sustains an audio reconstruction of the textual details.

Qualitative experiment. Figure 3 compares the original and fine-tuned models in terms of generation faithfulness for two-label integrity and three-label temporal samples. In our qualitative analysis, discernible discrepancies are noted in the audio generation by the original model. For integrity task, in response to the prompt "*A **motor hums** softly followed by **spraying***", the generated audio included only one of the elements, but omitting specific audio event. Similarly, for temporal task, the original model misrepresented the sequence of events. An example is the prompt "*A **man talking** as **music is playing** followed by a **frog croaking***", where the generated audio incorrectly prioritized the "*frog croaking*" event over the initial "*man talking*". In contrast, BATON demonstrates enhanced performance, achieving more faithful generation on diverse audio events and temporal order shown in prompts.

4.3 Ablation Study

Considering the increased intricacy posed by the three-label temporal task, which involves handling more audio events and challenges in temporal alignment, our ablation examination focuses on temporal task.

Human feedback. As detailed in Table 5, to dissect whether the enhancement in model performance is attributed to human feedback or increased data volume, we conducted a comparative analysis of the impact of data with and without feedback annotations on the model's performance. In the case without feedback annotations, 'PD', 'HD‡' and 'RD‡' denote fine-tuning with text-audio pairs data based on pretrain loss only. In the case with feedback annotations, we opted for BATON (HF). The baseline model TANGO achieved CLAP score of 35.9%. Incorporating 'PD' yielded slight gains in KL and CLAP accuracy but FD and FAD decreased, while continually integrating synthetic dataset 'HD‡' and 'RD‡' slightly improved CLAP score to 36.2% and 37.5%. Fine-tuning with human feedback achieves the highest CLAP accuracy at 41.9%, demonstrating this uptick is primarily attributed to the feedback of human preference instead of the boost of training data volume.

Preference data. The ablation study detailed in Table 6 investigates the influence of different preference data sources on BATON. Initially, employing only human-annotated data provides a CLAP score of 38.8%, indicating a moderate enhancement over the baseline. The exclusive use of reward model labeled data achieves an improved CLAP accuracy of 38.8%, suggesting that reward model alone is beneficial

PD	HD[‡]	RD[‡]	HF	FD↓	FAD↓	IS↑	KL↓	S_{CLAP}↑
				36.86	**4.02**	4.46	1.17	35.9%
✓				37.06	4.08	4.45	1.11	36.1%
✓	✓			37.78	4.49	4.70	**1.09**	36.2%
✓	✓	✓		37.23	4.09	4.80	1.15	37.5%
			✓	**36.19**	4.31	**4.88**	1.13	**41.9%**

Table 5: Ablation study for efficacy of human feedback on the test set. 'PD', 'HD[‡]' and 'RD[‡]' separately denote fine-tuning with pretrain data, data for human annotating and data for reward model labeling (without feedback value) only using pretrain loss.

as well. The most substantial improvements were observed when both human-annotated and reward model labeled data are utilized together, optimizing the performance with an FD score of 36.19 and a peak CLAP score of 41.9%. This configuration demonstrates the synergistic effect of combining human annotation with reward score, significantly improving the generation with human preference.

HA	RA	FD↓	FAD↓	IS↑	KL↓	S_{CLAP}↑
		36.86	**4.02**	4.46	1.17	35.9%
✓		37.95	4.80	4.87	**1.07**	38.8%
	✓	37.50	4.52	**4.88**	1.14	38.8%
✓	✓	**36.19**	4.31	**4.88**	1.13	**41.9%**

Table 6: Ablation study for preference data source on the test set. 'HA' and 'RA' denote utilizing human-annotated data and rewardmodel-labeled data (with feedback value).

Pretrain loss. The ablation study, outlined in Table 7 evaluates the efficacy of pretrain loss in fine-tuning. Incorporating pretrain loss improved the FD and FAD score to 36.2 and 4.31 while IS and KL score slightly decrease, and CLAP score rises to 41.9%, underscoring its significance in enhancing alignment and preventing overfitting on a relatively small amount of preference data.

PL	FD↓	FAD↓	IS↑	KL↓	S_{CLAP}↑
w/o	36.93	4.48	**4.99**	**1.07**	39.7%
w	**36.19**	**4.31**	4.88	1.13	**41.9%**

Table 7: Ablation study for pretrain loss on the test set. 'PL' represents for fine-tuning with pretrain loss.

Reward model. As shown in Table 8, comparison of encoders and training losses for the audio reward model reveals that the CLAP text encoder, coupled with BCE loss, significantly outperforms the Flan-T5 [Shen *et al.*, 2023] encoder in test set performance, yielding superior alignment outcomes. This efficacy likely stems from ability of CLAP, through large-scale pretraining based on contrastive learning, to extract more aligned text feature and audio feature, and from aptness of BCE loss for binary classification (0 and 1) tasks within reward model training.

To assess the accuracy of reward model in predicting human preferences, we applied it to two subtasks test sets, each

CLAP	T5	MSE	BCE	S_{CLAP}↑
	✓	✓		40.9%
	✓		✓	40.9%
✓	✓	✓		41.0%
✓			✓	**41.9%**

Table 8: Ablation study for feature extractors and loss adapted in reward model on the test set.

comprising 120 text-audio pairs, and annotated by two separate individuals. Figure 4 presents the findings. The predictions prominently feature two high-frequency peaks at 0 ("Not Aligned") and 1 ("Aligned"), suggesting a strong alignment of the model with human ratings. The concentration of predictions at these extremes, with few near the mid-point of 0.5, indicates the model's effectiveness in differentiating text-audio pair quality, particularly in tasks of integrity and temporal relationships.
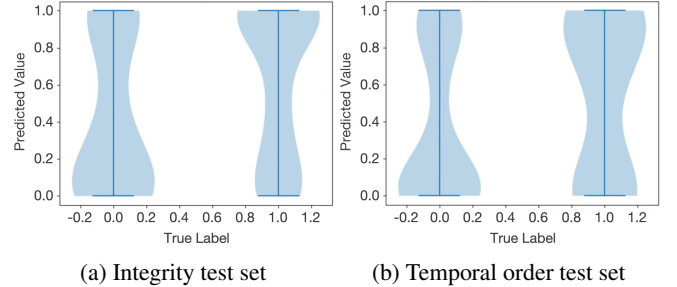

(a) Integrity test set (b) Temporal order test set

Figure 4: Prediction distribution of audio reward models.

5 Conclusion and Discussion

Conclusion. In this paper, we propose a novel framework (BATON), using human feedback to enhance the alignment between text prompts and generated audio in TTA models. Our BATON involves three steps: collecting human feedback for the constructed text-audio pairs, training an audio reward model on human-annotated data, and fine-tuning an off-the-shelf TTA model using the audio reward model. Extensive experiments manifest that BATON effectively leverages human feedback to mitigate biases and improve alignment in TTA models, making a valuable contribution to the ongoing advancements in audio synthesis from textual inputs. We anticipate that BATON can pave the way for aligning TTA models through human feedback.

Discussion. As a plug-and-play fine-tuning framework for various TTA models, BATON has some limitations and areas for future exploration. Firstly, the data-driven characteristic of BATON means the performance of alignment depends on the quality and quantity of human feedback. Secondly, the reliance of our approach on a two-stage training process, rather than using reinforcement learning strategies for online fine-tuning, restricting BATON to offline updates. Future work should further explore reinforcement learning with human feedback in text-audio alignment.

Contribution Statement

Huan Liao and Haonan Han contributed equally to the design and implementation of the project, Jiasheng Lu and Xiu Li served equally as corresponding authors, providing supervision and guidance throughout the research. All authors have contributed to the design of the methodology, implementation of the experiments, evaluation of the model performance, and writing of the paper.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Anthropic, 2023] Anthropic. Introducing claude, 2023.

[Black *et al.*, 2023] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[Casper *et al.*, 2023] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[Cheng *et al.*, 2023] Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10495–10505, 2023.

[Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[Dong *et al.*, 2023] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[Ghosal *et al.*, 2023] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[Google, 2023] Google. Bard, 2023.

[Guo *et al.*, 2024] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18153–18161, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2023a] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023.

[Huang *et al.*, 2023b] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *ICML*, 2023.

[Kilgour *et al.*, 2019] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech 2019*, Sep 2019.

[Kim *et al.*, 2019] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[Kingma and Ba, 2014] DiederikP. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv: Learning,arXiv: Learning*, Dec 2014.

[Kong *et al.*, 2020a] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[Kong *et al.*, 2020b] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, page 2880–2894, Jan 2020.

[Kreuk *et al.*, 2022] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[Kreutzer *et al.*, 2018] Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Lee *et al.*, 2023] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[Liu *et al.*, 2023a] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation

with latent diffusion models. *Proceedings of the International Conference on Machine Learning*, 2023.

[Liu *et al.*, 2023b] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Ranzato *et al.*, 2015] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *Computer Science*, 2015.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Shen *et al.*, 2023] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705*, 2023.

[Stiennon *et al.*, 2020] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[Wu and Hu, 2018] Yuxiang Wu and Baotian Hu. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Wu *et al.*, 2023] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation.

In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Xue *et al.*, 2024] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.

[Yang *et al.*, 2023a] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[Yang *et al.*, 2023b] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023.

[Yang *et al.*, 2023c] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023.

[Yuan *et al.*, 2024] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Retrieval-augmented text-to-audio generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585. IEEE, 2024.

[Ziegler *et al.*, 2019] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.