Guided by the Plan: Enhancing Faithful Autoregressive Text-to-Audio Generation with Guided Decoding

Juncheng Wang[1]   **Zhe Hu**[1]   **Chao Xu**[2,3]   **Siyue Ren**[4]

Yuxiang Feng[5]   **Yang Liu**[2,3]   **Baigui Sun**[2,3*]   **Shujun Wang**[1*]

[1] The Hong Kong Polytechnic University   [2] IROOTECH TECHNOLOGY
[3] Wolf 1069 b Lab, Sany Group   [4] Shanghai Artificial Intelligence Laboratory
[5] Zhejiang University

Abstract

Autoregressive (AR) models excel at generating temporally coherent audio by producing tokens sequentially, yet they often falter in faithfully following complex textual prompts—especially those describing complex sound events. We uncover a surprising capability in AR audio generators: their early prefix tokens implicitly encode global semantic attributes of the final output, such as event count and sound-object category, revealing a form of *implicit planning*. Building on this insight, we propose **Plan-Critic**, a lightweight auxiliary model trained with a Generalized Advantage Estimation (GAE)-inspired objective to predict final instruction-following quality from partial generations. At inference time, Plan-Critic enables *guided exploration*: it evaluates candidate prefixes early, prunes low-fidelity trajectories, and reallocates computation to high-potential planning seeds. Our Plan-Critic-guided sampling achieves up to a **10 points improvement in CLAP score** over the AR baseline—establishing a new state of the art in AR text-to-audio generation—while maintaining computational parity with standard best-of-N decoding. This work bridges the gap between causal generation and global semantic alignment, demonstrating that even strictly autoregressive models can plan ahead. Codes will be available at https://github.com/wjc2830/Siren.git.

1 Introduction

Text-to-audio generation aims to synthesize audio signals given textual descriptions (Liu et al., 2023; Majumder et al., 2024; Xing et al., 2024; Tian et al., 2025). Most existing approaches adopt non-autoregressive (NAR) frameworks, such as diffusion models (Cheng et al., 2025; Wang et al., 2025b), that operate over continuous latent representations. While highly effective, they often neglect the inherently sequential and causal nature
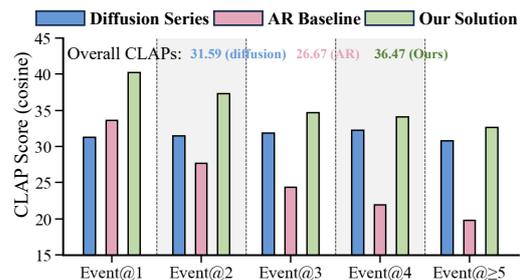

Figure 1: Instruction-following performance (y-axis) of different methods, measured by CLAP scores. The x-axis labels, Event@N, indicate prompts with N distinct audio events, where a higher number of events corresponds to increased complexity.

of audio, which unfolds temporally in a strictly forward direction. To better capture such dynamics, recent works such as Siren (Wang et al., 2025a) has revisited AR paradigm (Kreuk et al., 2023; Copet et al., 2024), generating audio token-by-token. It naturally mirrors the temporal progression of sound and yields superior temporal coherence.

Despite their success in producing high-fidelity audio, AR models suffer from a persistent limitation: *they struggle to faithfully follow complex textual prompts*. When prompts describe multiple sequential or complex audio events, AR-generated outputs frequently omit or misalign key semantic elements, resulting in degraded instruction-following. As shown in Figure 1, while diffusion models maintain stable performance as prompt complexity increases, AR baselines exhibit a consistent drop in CLAP score (Wu et al., 2023)—highlighting a fundamental gap between strong temporal modeling and robust semantic alignment.

The main reason for such disparity lies in architectural differences. Diffusion models, often built on UNet (Rombach et al., 2022) or bidirectional transformer (Peebles and Xie, 2023), allow every token to attend to both past and future states, enabling global coordination and coarse-to-fine planning across the entire timeline. AR models, by

contrast, generate outputs strictly left-to-right in token level, without explicit global control (Li et al., 2022; Hua and Wang, 2020; Shen et al., 2019). As a result, they excel at local temporal coherence but often lack global semantic alignment, a weakness that becomes more pronounced as prompt complexity grows. This raises a central challenge: *how can AR models strengthen instruction following while preserving their causal generation advantages?*

We draw inspiration from human sequential production, such as speech or writing, where people plan ahead and rely on this plan as a global control during output generation (Hu et al., 2022; Hovy, 1990; Pan and McKeown, 1998; McKeown, 1985). Motivated by this analogy, we first conduct a preliminary analysis with global attribute probing. Our results reveal that AR models exhibit similar signs of implicit planning: the prefix tokens (e.g., the first 32 of a 288-token sequence) already encode predictive information about high-level audio attributes, including the *event count* and the *sound object category*. This finding suggests that AR models, though restricted to causal decoding, implicitly embed global structural cues early in the generation process as *implicit planning*.

Building on this observation, we propose a guided decoding method that leverages implicit planning as a control signal to steer AR model generation <u>at inference time</u>. Specifically, we introduce **Plan-Critic**, an auxiliary model trained to predict instruction-following quality (measured by CLAP score) from partial audio sequences. In this way, we transfer the implicit planning into a rectified score that enables early selection and guidance of the subsequent full sequence generation. A key challenge in training such critic is credit assignment: the model must infer the overall sequence quality based on partial generations. To address this, we design a **Generalized Advantage Estimation (GAE)-inspired training framework**: The critic produces value estimates at each generation step, with direct supervision applied only to the final-step prediction using the ground-truth CLAP score. To bridge the gap, intermediate predictions are regularized through temporal consistency losses, propagating supervisory signals backward along the sequence. This strategy mitigates the credit assignment problem and allows reliable assessment during early generation satge.

Equipped with Plan-Critic, we reframe inference as *guided exploration* rather than blind left-to-right rollout. Our **Plan-Critic-guided sampling** strategy evaluates candidate prefixes early, prunes low-scoring trajectories, and reallocates computation to promising planning seeds. Unlike full-sequence methods like best-of-N sampling—which generate complete outputs before reranking (Stiennon et al., 2020)—our approach expands the effective search space over high-quality global structures while avoiding redundant computation. This yields more efficient inference-time scaling and significantly stronger alignment with complex prompts.

In summary, our main contributions are:

• We provide empirical evidence that early tokens in AR audio generation encode predictive information about global semantic attributes, revealing an intrinsic capacity for latent planning.

• We propose Plan-Critic—a novel auxiliary model—and a GAE-inspired training paradigm that quantifies implicit planning, enabling accurate prediction of final instruction-following quality from partial sequences.

• We introduce a Plan-Critic-guided sampling algorithm that improves both efficiency and semantic fidelity, achieving up to a **10 points gain in CLAP score** over the AR baseline and establishing a new state-of-the-art in AR text-to-audio generation.

2 On Prefix Tokens and Implicit Planning

2.1 Preliminaries of AR Audio Generation

Consider an audio waveform $x \in \mathbb{R}^{T_{\text{wav}} \times C_{\text{wav}}}$, where C_{wav} denotes the number of channels and T_{wav} the temporal duration. An AR generator first encodes x into discrete tokens using a residual vector quantization (RVQ) tokenizer (Kumar et al., 2023), yielding a token sequence $\mathbf{q} \in [V]^{R \times T}$, where V is the vocabulary size, R the number of RVQ layers, and T the sequence length in tokens.

Following Kreuk et al. 2023; Wang et al. 2025a, a causal transformer is employed to model the joint distribution over tokens in a left-to-right manner, predicting the next R codes at each time step:

$$p(\mathbf{q}) = p(q_1^1, \ldots, q_1^R, \ldots, q_T^1, \ldots, q_T^R) \qquad (1)$$

$$= \prod_{t=1}^{T} p(q_t^1, \ldots, q_t^R \mid q_1^1, \ldots, q_1^R, \ldots, q_{t-1}^R, \Phi(c)),$$

where $\Phi(c)$ is the encoded textual prompt. After T AR steps, the generated token sequence $\mathbf{q}$ is decoded back into a waveform via a detokenizer.
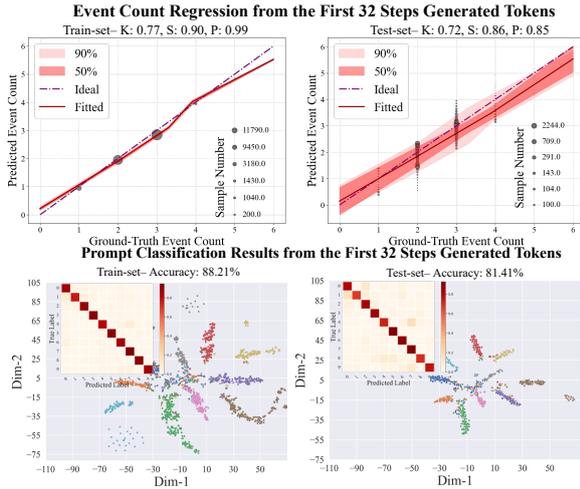
Figure 2: **Upper row**: Event counting regression results, with Kendall (K), Spearman (S), and Pearson (P) correlation coefficients reported between predicted and ground-truth counts. **Lower row**: t-SNE visualization of prefix representations, colored by the object category associated with each prefix. The confusion matrix for the classification results is also included. **Takeaways**: the non-trivial generalization performance of event counting and sounding object classification demonstrate the high correlation between the prefix tokens with posterior global attributes. *Best viewed in color.*

2.2 Implicit Planning for Future Generation in Prefix Audio Tokens

Unlike diffusion-based models (Haji-Ali et al., 2024; Tian et al., 2025; Cheng et al., 2025), which leverage bidirectional context to globally coordinate generation across time, the strictly causal nature of AR models limits their ability to align with complex textual instructions (Hua and Wang, 2020; Shen et al., 2019). This often results in omissions or misalignments of key semantic elements.

Inspired by human cognitive processes in speech and writing, where individuals formulate high-level plans before articulating details (Hu et al., 2022; Hovy, 1990), we hypothesize that AR audio generators similarly engage in *implicit planning*: early generated prefix tokens encode predictive signals about global properties of the full audio sequence.

To test this, we conduct correlation probing experiments: we train lightweight MLPs to predict global attributes of the complete generated audio from representations of its prefix tokens. Successful generalization would imply that such attributes are implicitly encoded during early generation.

Correlation Probing Framework With prompt $\Phi(c)$, we use pretrained generator $\pi_\theta(\mathbf{q} \mid \Phi(c))$ to synthesize full audio sequences $\mathbf{q} \in [V]^{R \times T}$. In the first T_{prefix} steps, we extract the penultimate-layer (for its richest semantics of model) hidden states corresponding to the prefix tokens, $\mathbf{h}_{\text{prefix}} \in \mathbb{R}^{R \times T_{\text{prefix}} \times C}$, with C as hidden dimension.

We define an *attribute function* $f(\mathbf{q})$ that maps the full token sequence to a global semantic attribute (e.g., number of sound events). If $\mathbf{h}_{\text{prefix}}$ contains predictive information about $f(\mathbf{q})$, then a mapping $g(\mathbf{h}_{\text{prefix}}) \mapsto f(\mathbf{q})$—realized by an MLP g—should generalize beyond training data.

Attribute Design The chosen attributes must be (i) global in scope, (ii) not inferable from local prefix content alone, and (iii) distributed across the full temporal span. We select two such attributes that directly influence instruction-following fidelity:

1. **Event count**: the number of distinct audio events in the generated output. To extract this, we use a multimodal large language model (MLLM) (Comanici et al., 2025) to caption the audio and segment the caption into atomic sound events. For example, `Motor noise` is followed by a `horn honking` and a `siren wailing` is parsed into three events, yielding a label of 3.

2. **Sound-producing object category**: the dominant object responsible for sustained sound production. We prompt a LLM (Comanici et al., 2025) to generate diverse captions for each of 10 predefined objects (e.g., `rain`), such as `Heavy rain is drumming on the window` or `The rain is tapping quietly on the window during the storm`. All captions derived from the same object share its category label.

For attribute (1), the MLP is trained as a regressor; for attribute (2), as a classifier.

Empirical Results For event count prediction, we generate 5,000 diverse prompts and synthesize multiple audio samples per prompt using π_θ. An MLLM labels each sample with its event count. The dataset is split into train, validation, and test sets. As shown in the upper panel of Figure 2, the MLP regressor achieves strong generalization on the test set. Correlation metrics—Kendall, Spearman, and Pearson—all indicate significant alignment between predicted and ground-truth counts.

For object category classification, we generate 2,000 prompts per object across 10 categories. The MLP classifier, trained on prefix representations, achieves high accuracy in distinguishing object categories, as visualized in the lower panel of Figure 2.
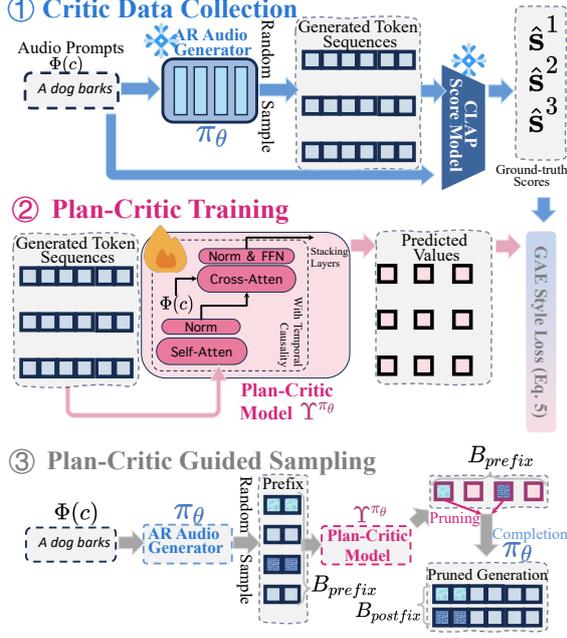
Figure 3: Overall pipeline of the proposed plan-critic guided framework. ① illustrates the process of collecting training data for the plan-critic model. ② shows the training of the plan-critic model, where we conduct sparse GAE supervision with certain step t is involved in loss (elaborate in Section 4.1). ③ depicts how the trained critic guides the generator's inference sampling.

In summary, these results demonstrate that prefix tokens in AR audio generation encode rich predictive signals about global semantic attributes of the full sequence. This provides empirical evidence for *implicit planning*—a latent capacity to anticipate future content during early decoding stages.

3 Plan-Critic Guided Generation

Our above insight suggests a practical strategy: *rather than retraining the large AR generator, we can enhance its alignment with complex prompts by leveraging early-generation signals as a form of implicit plan*. To this end, we introduce **Plan-Critic**, a lightweight auxiliary model that maps prefix tokens to a scalar score estimating the eventual instruction-following quality of the full audio. This score enables early evaluation and selection of high-potential generation trajectories, guiding inference toward semantically faithful outputs, with Figure 3 depicting pipeline.

3.1 Plan-Critic Model

The Plan-Critic model, denoted Υ^{π_θ}, is a transformer-based critic initialized from a subset of parameters of the pre-trained AR generator π_θ. Its

purpose is to evaluate, at intermediate generation step t, how well the generated audios will perform on instruction from current partial sequence.

Architecture and Initialization   Given a prefix token sequence $\mathbf{q}_{\text{prefix}} \in [V]^{R \times T_{\text{prefix}}}$, we first embed it into $\mathbf{e}_{\text{prefix}} \in \mathbb{R}^{R \times T_{\text{prefix}} \times C}$ using the embedding layer. Following standard practice (Copet et al., 2024; Wang et al., 2025a), we collapse the R quantization layers via summation to obtain a single temporal sequence $\mathbf{e}'_{\text{prefix}} \in \mathbb{R}^{T_{\text{prefix}} \times C}$.

The prompt $\Phi(c)$ is encoded as a sequence of contextualized text embeddings. The Plan-Critic then processes $\mathbf{e}'_{\text{prefix}}$ and $\Phi(c)$ through interleaved causal self-attention and cross-attention layers. The self-attention captures temporal dynamics within the audio prefix, while cross-attention models alignment between audio and text.

The resulting contextualized representation is:

$$\mathbf{h}^\Upsilon = \Upsilon_h(\mathbf{e}'_{\text{prefix}}, \Phi(c)), \quad \text{where } \mathbf{h}^\Upsilon \in \mathbb{R}^{T_{\text{prefix}} \times C}. \tag{2}$$

To ensure compatibility with the implicit planning signals identified in Section 2.2—which were from the penultimate layer of π_θ—we initialize Υ^{π_θ} by corresponding layers of the generator. This alignment ensures the critic operates in the same representational manifold as the planning cues.

Finally, a MLP regression head Υ_s maps the hidden state at step t, $\mathbf{h}^\Upsilon_{\leq t}$, to a scalar value:

$$s_t = \Upsilon_s(\mathbf{h}^\Upsilon_{\leq t}). \tag{3}$$

Generalized Advantage Estimation Training The goal of Plan-Critic is to predict, from a partial sequence, the eventual instruction-following of the completed audio—quantified by CLAP score. A naive approach would assign the final CLAP score uniformly to all intermediate steps. However, it ignores the *temporal credit assignment problem*: each step tokens actually contribute differently to the final performance.

To address this, we adopt a training objective inspired by Generalized Advantage Estimation (GAE) from reinforcement learning (Schulman et al., 2017). For a generated sequence $\mathbf{q}$ of length T, the Plan-Critic produces a sequence of value estimates $\{s_t\}_{t=1}^{T}$. Let $\hat{s}$ denote the ground-truth CLAP score of the full sequence. We define a sparse reward signal $\hat{\mathbf{s}} \in \mathbb{R}^T$ where $\hat{\mathbf{s}}_T = \hat{s}$ and $\hat{\mathbf{s}}_t = 0$ for $t < T$.

The GAE-based target for step t is:

$$r_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \left(\hat{\mathbf{s}}_{t+l} + \gamma s^{\text{old}}_{t+l+1} - s^{\text{old}}_{t+l} \right) + s^{\text{old}}_t, \tag{4}$$

where s^{old} denotes the value estimates from a detached (non-gradient) copy of the critic, and $\gamma, \lambda \in [0, 1]$ are discount and smoothing hyperparameters. The critic is trained to minimize:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_t \left[\frac{1}{2}(s_t - r_t)^2 \right]. \tag{5}$$

Remark (Intuition behind Eq. 5). The Eq. 5 target r_t propagates credit backward from the final reward while discounting distant contributions (via γ) and smoothing temporal differences (via λ). This enables the critic to assign higher value to prefixes that genuinely support semantic alignment.

3.2 Plan-Critic Guided Sampling

Equipped with a trained Plan-Critic, we reformulate inference as a guided search over generation trajectories. Standard AR sampling performs a blind left-to-right rollout, allocating equal computational resources to all time steps. However, our analysis in Section 2.2 shows that the prefix tokens encode a high-level plan that largely determines semantic fidelity. Moreover, empirical results (Section 4.3) confirm that once the prefix is fixed, further exploration in the postfix yields diminishing returns in instruction-following performance.

This motivates a *prefix-first* inference strategy: allocate more sampling budget to exploring diverse prefix candidates, evaluate them early using the Plan-Critic, and prune low-scoring trajectories before committing to full-sequence generation.

Our guided sampling proceeds as follows:

1. Given a prompt $\Phi(c)$, sample a large batch of prefixes $\{\mathbf{q}^i_{\text{prefix}}\}_{i=1}^{B_{\text{prefix}}}$ from $\pi_\theta(\cdot \mid \Phi(c))$.

2. At step T_{prefix}, critic each prefix' s planning to obtain scores $s^i_{T_{\text{prefix}}} = \Upsilon_s(\mathbf{q}^i_{\text{prefix}}, \Phi(c))$.

3. Select the top-B_{postfix} prefixes ($B_{\text{postfix}} < B_{\text{prefix}}$) with the highest scores.

4. Continue AR generation from these selected prefixes to produce full audio sequences.

This approach dynamically reallocates computation toward high-potential planning seeds, improving both efficiency and semantic alignment.

In our experiments, we fix the total token budget to match that of the baseline AR model (i.e., $B_{\text{prefix}} \cdot T_{\text{prefix}} + B_{\text{postfix}} \cdot (T - T_{\text{prefix}}) = B_{\text{baseline}} \cdot T$), ensuring a fair comparison. Details on hyperparameter selection are provided in Section 4.4.

4 Experiment

4.1 Settings

Datasets To train Plan-Critic model on base generator unseen prompts, we use a large language model (LLM) to synthesize 5,000 pseudo audio captions and prompt our baseline generator, Siren (Wang et al., 2025a), to produce 32 audio samples per caption.

To mitigate potential domain shift from LLM-generated prompts, we additionally curate prompts from the official test sets of AudioCaps (Kim et al., 2019) and VGGSound (Chen et al., 2020a)—ensuring they were not seen during Siren's original training. From each dataset, we reserve 1,000 such prompts as held-out evaluation environments for assessing final performance.

Training Details To enable effective credit assignment across long sequences, we adopt a sparse supervision strategy for the Generalized Advantage Estimation (GAE) objective: the critic outputs scores only at every 32nd time step, i.e., for all t such that $mod(t, 32) = 0$. This ensures gradient signals from the final reward can propagate meaningfully back to the prefix tokens. Moreover, to avoid overfitting, we randomly add a Gaussian noise upon the computed CLAP scores.

The plan-critic is trained using the AdamW optimizer with a learning rate of $1e-4$ for $40k$ steps.

Inference Details During inference, we employ a prefix-first search strategy: we generate $B_{prefix} = 128$ candidate prefixes in parallel and retain only the top $B_{postfix} = 2$ based on plan-critic scores for full-sequence completion. The prefix length is fixed at $T_{prefix} = 32$ tokens.

For a full audio sequence of 288 tokens per RVQ layer (with 12 RVQ layers total), our method consumes $128 \times 32 + 2 \times (288 - 32) = 4,608$ tokens per RVQ layer, matching the token budget of Siren' s official best-of-N (BoN-16) baseline ($16 \times 288 = 4,608$). For notational simplicity, we omit the factor of 12 (RVQ layers) in subsequent discussions, reporting token counts per layer.

All experiments are conducted on a single-node AMD MI300X server.

Method	Instruction Following ↑						Audio Quality			
	Overall CLAP	CLAP@1	CLAP@2	CLAP@3	CLAP@4	CLAP@≥5	FAD↓	FD↓	IS↑	KL↓
Solutions with Bidirectional Generators										
AudioLDM2	24.46	30.89	24.75	22.12	21.93	20.30	2.04	37.76	8.32	1.72
MagNet	22.04	19.68	22.51	22.22	24.51	19.26	4.64	33.24	8.20	1.99
AudioX	27.12	27.14	27.19	27.11	28.49	25.36	3.02	28.68	11.28	1.54
MMAudio	31.21	32.94	31.13	31.13	30.78	28.17	5.51	19.06	13.78	1.37
TangoFlux	33.43	31.36	32.94	34.68	34.28	**35.05**	2.89	25.54	<u>13.75</u>	**1.16**
GenAU	<u>34.59</u>	<u>33.68</u>	<u>34.57</u>	<u>34.54</u>	**35.36**	<u>34.59</u>	**1.73**	23.01	13.63	1.28
Solutions with Unidirectional Generators										
DelayPattern+BoN-16	20.01	21.26	20.27	19.44	19.76	17.76	3.58	17.82	10.67	2.08
Siren+BoN-16	26.67	33.63	27.68	24.38	21.95	19.82	<u>1.95</u>	<u>17.63</u>	10.41	1.91
Siren+Ours	**36.47**	**40.20**	**37.35**	**34.72**	34.09	32.59	1.96	**15.70**	**13.82**	<u>1.27</u>

Table 1: Main results on AudioCaps. The best performed metric is in **bold**, and the second best is <u>underlined</u>. Among solutions with bidirectional solutions, MagNet also utilizes a RVQ discrete tokenizer, but is equipped with a BERT-like transformer in its generator. For unidirectional solutions, DelayPattern and Siren are all equipped with inference decoding method of best-of-16. For Siren and ours solution, we share the same generated token budget, where for the first policy transformer, the generated token numbers are all 4,608.

Metrics Following prior work (Cheng et al., 2025; Tian et al., 2025; Wang et al., 2025a), we evaluate audio quality using Fréchet Audio Distance (FAD), Fréchet Distance (FD), Kullback–Leibler (KL) divergence, and Inception Score (IS). And CLAP for overall instruction following. To be fine grained, we assess model behavior across complexity, where we use LLM to classify test prompts by the number of audio events and report performance across different event counts.

4.2 Main Results

Comparison with Other Audio Generators In Table 1 and 4, we compare our method with some bidirectional solutions, including AudioLDM2 (Liu et al., 2024a), MagNet (Ziv et al., 2024), AudioX (Tian et al., 2025), MMAudio (Cheng et al., 2025), TangoFlux (Hung et al., 2024), and GenAU (Haji-Ali et al., 2024). As for unidirectional solutions, we pick DelayPattern (Copet et al., 2024) and our baseline Siren (Wang et al., 2025a). As the tables shown, our method (Siren+Ours) achieves consistently nontrivial improvement over baseline Siren. This demonstrates that critic-guided prefix search enables superior instruction alignment.

Comparison with Other Advanced Decoding Algorithms With plan-critic model trained as an instant reward model for ARs, it can be incorporated with existing guided decoding methods, including BoN, importance sampling (Owen and Zhou, 2000) and TreeBoN (Qiu et al., 2024). Results in Table 2 and 3 show our presented sampling best cooperates

Method	CLAP↑	FAD↓	FD↓	IS↑	KL↓
BoN	26.67	**1.95**	<u>17.63</u>	<u>10.41</u>	1.91
Importance Sampling	29.72	2.52	27.58	9.91	2.90
TreeBoN	<u>34.04</u>	1.97	18.26	10.10	<u>1.38</u>
Ours	**36.47**	<u>1.96</u>	**15.70**	**13.82**	**1.27**

Table 2: Comparison with other advanced decoding algorithms on AudioCaps, where these methods share the same our trained plan-critic model.

Method	CLAP↑	FAD↓	FD↓	IS↑	KL↓
BoN	25.62	4.63	32.23	9.56	3.02
Importance Sampling	27.37	4.40	30.84	9.19	<u>3.01</u>
TreeBoN	30.90	**3.64**	<u>29.41</u>	10.01	3.33
Ours	**35.88**	<u>3.70</u>	**28.02**	**10.65**	**2.87**

Table 3: Comparison with other advanced decoding algorithms on VGGSound.

with our plan-critic model as the best.

4.3 Finding Studies

In this section, we conduct findings experiments to support our major insights: the prefix tokens indeed contain predictive information to the global attributes of posteriorly generated audio.

Correlation with attributes is not because prefix tokens decoded audio content To support that the correlation obtained in Section 2.2 is not because the prefix tokens decoded audios have contained corresponding attributes, we conduct a comparison experiment that we utilize the detokenizer to decode the prefix tokens into audio, and replace the MLP network with a VGGish audio feature extraction model, and try to regress the event count.

Method	Instruction Following ↑						Audio Quality			
	Overall CLAP	CLAP@1	CLAP@2	CLAP@3	CLAP@4	CLAP@≥5	FAD↓	FD↓	IS↑	KL↓
AudioX	<u>33.93</u>	<u>37.71</u>	<u>33.60</u>	32.32	<u>29.61</u>	<u>28.52</u>	4.44	35.86	9.39	2.94
MMAudio	31.69	36.81	31.39	29.45	26.02	22.81	<u>3.57</u>	<u>29.10</u>	<u>10.64</u>	2.90
GenAU	33.03	36.76	32.90	<u>33.60</u>	27.04	<u>21.10</u>	**3.44**	34.55	10.10	**2.52**
Siren+BoN-16	25.62	28.60	26.51	22.51	21.53	16.37	4.63	32.23	9.56	3.02
Siren+Ours	**35.88**	**38.83**	**35.56**	**35.53**	**32.86**	**28.86**	3.70	**28.02**	**10.65**	<u>2.87</u>

Table 4: Main results on VGGSound, where the textual prompts are annotated by Gemini-2.5 pro, and further polished with a CLAP score restriction with original audio.
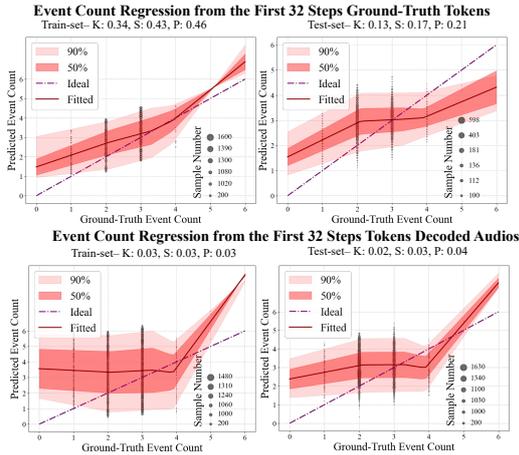

Figure 4: **Upper row**: Using prefix tokens from real audio to regress the total event count of the full audio. **Low row**: Event count regression using prefix tokens from decoded (reconstructed) audio.

As shown in the lower row of Figure 4, we fail to build this correlation from audio, which reflects the correlation is not because audio content.

Moreover, we conduct a comparison, that given some ground-truth audios collected from real world, we utilize Siren's tokenizer (Kumar et al., 2023) to transfer them into token sequences, where we find that the first 32-step tokens fail to regress event counts, which further demonstrates **implicit-planning** is the property of AR generators.

Postfix tokens yields diminishing influence in the instruction following We investigate how much instruction-following quality is determined by prefix tokens. For 10,000 fixed prefixes, we generate 1,000 full audio completions each and compute the mean and standard deviation of the CLAP scores over 1,000 post completions. In Figure 5, the means $[-10, 20]$, which means significant prefix influences. While standard deviations are tightly clustered in $[2, 3]$. This narrow variance indicates that postfix variations have minimal impact.
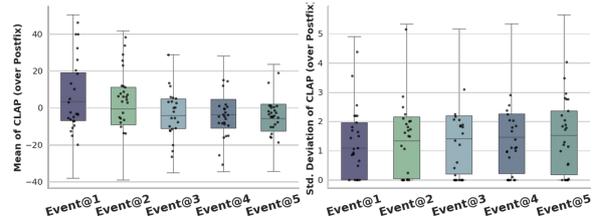

Figure 5: Given a set of prefix tokens, each is expanded into 1,000 postfix sequences. The two plots show the distributions of the **mean** and **standard deviation** of CLAP scores computed across the 1,000 generations per prefix, aggregated over 10,000 distinct prefixes. To simplify the figure, we randomly select and visualize 25 points for each event count.

Components	CLAP↑	FAD↓	FD↓	IS↑	KL↓
Baseline	26.67	1.95	17.63	10.41	1.91
Sparse Supervision	10.67	3.01	37.66	8.16	3.52
Random Guess	12.86	2.47	23.19	8.50	3.09
Ours	36.47	1.96	15.70	13.82	1.27

Table 5: Ablating plan-critic model training's supervision signal.

4.4 Ablation Studies

In this section, we conduct extensive ablation study over the choices of hyper-parameters involved in our method. In the following tables, we highlight the rows with our final option in gray.

Ablating plan-critic model training In Table 5, we compare our plan-critic model training's signal with a naive sparse signal (tokens within a sequence shares the same CLAP score). As a comparison, we compare it with a random guess method (using the same inference sampling strategy expect critic-then-prune). In Table 6, we ablate different choices of GAE supervision interval steps.

Sampling Strategy	Cost Tokens↓	CLAP↑	FAD↓	FD↓	IS↑	KL↓
Baseline	4608	26.67	1.95	17.63	10.41	1.91
Interval-16	2592	31.27	1.90	17.02	12.57	1.64
Interval-32	4608	36.47	1.96	15.70	13.82	1.27
Interval-64	8640	37.34	1.76	15.21	13.48	1.21

Table 6: Ablating the interval step of GAE training.

Critic Time	CLAP↑	FAD↓	FD↓	IS↑	KL↓
Prefix-32	27.52	1.86	18.07	11.12	1.85
Prefix-64	18.81	2.40	21.84	8.88	2.64
Prefix-96	18.56	2.43	22.19	8.88	2.68
Postfix-64	18.40	2.55	22.31	8.86	2.69
Postfix-32	18.01	2.50	21.91	9.01	2.71

Table 7: Ablating the critic temporal step.

Method	Instruction Following		Audio Fidelity	
	General	Complex	General	Complex
vs. Siren	78.36	93.20	53.53	55.20
vs. AudioX	70.49	67.89	60.81	65.37
vs. GenAU	65.01	63.37	45.61	50.19

Table 8: User study results. We report the win-rate (%) of our solution when comparing with different methods.

Ablating sampling strategies In Table 7, we ablate different choices for the timing of Plan-Critic evaluation, specifically the prefix length T_{prefix} at which the critic is applied. To clarify it: *Prefix-x* denotes applying critic after the first x temporal steps, while *Postfix-x* is applying it x steps before the end of sequence (i.e., at step $288 - x$).

To ensure that the *Postfix-32* configuration can be executed within the memory constraints of a single GPU, we use a reduced prefix batch size of $B_{\text{prefix}} = 8$ uniformly across all entries in Table 7.

Additional ablation studies exploring alternative sampling strategies are provided in the Appendix.

4.5 User Study

In Table 8, we conducted a human evaluation study with 30 domain experts. Each annotator assessed approximately 100 prompts per semantic category using pairwise A/B tests. For each pair, they judged which audio better followed the prompt instructions and which sounded more natural. This yielded over 18,000 comparisons in total, providing human-validated support for our method's gains in both semantic fidelity and audio quality.

5 Related Work

Audio Generation. Denoising diffusion models (Ho et al., 2020; Song and Ermon, 2019; Song et al., 2020; Dhariwal and Nichol, 2021; Lu et al., 2022; Rombach et al., 2022; Peebles and Xie, 2023) excel at audio synthesis. They are typically applied in latent space: 1D tokenizers operate on waveform (Evans et al., 2025; Lee et al., 2024), while 2D architectures process mel-spectrograms using U-Nets (Liu et al., 2022, 2023, 2024a; Xue et al., 2024; Evans et al., 2024; Xing et al., 2024; Du et al.,

2023; Liu et al., 2024b; Agostinelli et al., 2023; Majumder et al., 2024; Wang et al., 2025b) or diffusion transformers (DiT) (Lee et al., 2024; Valle et al., 2025). Recent work integrates flow matching (Lipman et al., 2023; Liu et al., 2025) to accelerate sampling and achieve sota results (Cheng et al., 2025; Valle et al., 2025). However, the bidirectional feature interaction breaks the nature of causal flow of audio data.

In parallel, ARs adopt discrete representations via vector quantization (VQ) (Chen et al., 2020b; Chang et al., 2022; Li et al., 2023; Yu et al., 2023a; Zheng et al., 2022; Chang et al., 2023; Tian et al., 2024). Early hybrids suffered from fidelity loss due to VQ compression (Yu et al., 2023b; Mentzer et al., 2023), prompting the adoption of residual VQ (RVQ) (Wu and Yu, 2019; Défossez et al., 2022; Kumar et al., 2023). A key advance is Siren (Wang et al., 2025a), which employs a parallel transformer and achieves sota audio quality. Yet, it exhibits weak instruction-following capability.

Test-Time Compute. Our work is also related to recent advances in test-time compute, which aim to improve model performance by generating multiple candidate outputs and then selecting the best one according to specific criteria (Snell et al., 2024; Ji et al., 2025; Muennighoff et al., 2025). Common strategies include majority voting (Wang et al., 2022; Toh et al., 2024) and Best-of-N selection based on predefined rewards (Cobbe et al., 2021; Levi, 2024). A central component of this paradigm is the use of reward models to evaluate either the final outcomes (Liu et al., 2024c; Xin et al., 2024) or the intermediate reasoning steps (Ma et al., 2023; Khalifa et al., 2025). In this work, we extend test-time compute to text-to-audio generation by introducing an implicit planning–based partial reward model, enabling effective and efficient scaling of AR models for high-quality audio synthesis.

6 Conclusion

In this work, we show that AR text-to-audio models inherently perform implicit planning: early tokens encode global semantic attributes of the full output. Based on this, we propose **Plan-Critic**, a lightweight model trained with a GAE-inspired objective to estimate instruction-following quality from partial sequences. Leveraging Plan-Critic, we further introduce a guided sampling strategy that reallocates computation to high-potential prefixes, yielding up to a **10% CLAP score gain** over

Siren and setting a new state of the art. Our results demonstrate that even strictly causal models can achieve global alignment—when guided wisely.

Limitations

While our Plan-Critic framework significantly improves instruction-following fidelity in autoregressive (AR) text-to-audio generation, several limitations remain:

Dependency on CLAP as a Proxy Metric: Our Plan-Critic is trained to predict CLAP scores, which serve as a proxy for semantic alignment. However, CLAP—like any embedding-based metric—may not fully capture nuanced aspects of human perception, such as temporal plausibility, emotional tone, or fine-grained event ordering. To remedy this, our future work may expand from designing a human preference aligned reward model.

Fixed Prefix Length Assumption: Our method assumes that global planning is encoded within a fixed-length prefix (e.g., the first 32 tokens). While empirical results support this for the evaluated model and audio duration (288 tokens), this may not generalize to longer or more complex audio sequences (e.g., multi-minute soundscapes), where planning might unfold over multiple stages or require hierarchical structure. To remedy this, the potential future work may explore the prefix length with a fixed ratio with the whole length.

Training Data Bias: The Plan-Critic is trained on audio generated by a single AR model (Siren) using LLM-synthesized and curated prompts. This introduces potential biases: the critic may overfit to Siren's specific failure modes or stylistic tendencies and may not transfer well to other AR generators or domains with different acoustic characteristics (e.g., music vs. environmental sounds). However, the scaling capacity of transformer architecture along with training data has been well validated. The following works may address this from scaling up the critic model size and training data.

Acknowledgment

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, and 1 others. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020a. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020b. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. 2023. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, Sergey Tulyakov, and Vicente Ordonez. 2024. Taming data and transformers for audio generation. *arXiv preprint arXiv:2406.19388*.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and 1 others. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

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

Eduard H Hovy. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2):153–197.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li,

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. *arXiv preprint arXiv:2412.21037*.

Yixin Ji, Juntao Li, Yang Xiang, Hai Ye, Kaixin Wu, Kai Yao, Jia Xu, Linjian Mo, and Min Zhang. 2025. A survey of test-time compute: From intuitive inference to deliberate reasoning. *arXiv preprint arXiv:2501.02497*.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. Process reward models that think. *arXiv preprint arXiv:2504.16828*.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\`echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

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. In *International Conference on Learning Representation*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. 2024. Etta: Elucidating the design space of text-to-audio models. *arXiv preprint arXiv:2412.19351*.

Noam Levi. 2024. A simple model of inference scaling laws. *arXiv preprint arXiv:2410.16377*.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. 2023. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2301.12503*.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024a. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028.

Peng Liu, Dongyang Dai, and Zhiyong Wu. 2025. Rfwave: Multi-band rectified flow for audio waveform reconstruction. In *International Conference on Learning Representations*.

Xiulong Liu, Kun Su, and Eli Shlizerman. 2024b. Tell what you hear from what you see–video to audio generation through text. *arXiv preprint arXiv:2411.05679*.

Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024c. Acemath: Advancing frontier math reasoning with post-training and reward modeling. *arXiv preprint arXiv:2412.15084*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning. *arXiv preprint arXiv:2310.10080*.

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 *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572.

Kathleen R McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial intelligence*, 27(1):1–41.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Art Owen and Yi Zhou. 2000. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Shimei Pan and Kathleen McKeown. 1998. Learning intonation rules for concept to speech generation. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1003–1009, Montreal, Quebec, Canada. Association for Computational Linguistics.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Chenhao Zhu, Xinzhe Juan, Ling Yang, Huazheng Wang, Kaixuan Huang, and 1 others. 2024. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. *arXiv preprint arXiv:2410.16033*.

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*, pages 10684–10695.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. 2019. Towards generating long and coherent text with multi-level latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2079–2089, Florence, Italy. Association for Computational Linguistics.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*.

Vernon YH Toh, Deepanway Ghosal, and Soujanya Poria. 2024. Not all votes count! programs as verifiers improve self-consistency of language models for math reasoning. *arXiv preprint arXiv:2410.12608*.

Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, Joao Felipe Santos, Shuqi Dai, Siddharth Gururani, Aya Aljafari, and 1 others. 2025. Fugatto 1: Foundational generative audio transformer opus 1. In *International Conference on Learning Representations*.

Juncheng Wang, Chao Xu, Cheng Yu, Zhe Hu, Haoyu Xie, Guoqi Yu, Lei Shang, and Shujun Wang. 2025a. Language model based text-to-audio generation: Anti-causally aligned collaborative residual transformers. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Juncheng Wang, Chao Xu, Cheng Yu, Lei Shang, Zhe Hu, Shujun Wang, and Liefeng Bo. 2025b. Synchronized video-to-audio generation via mel quantization-continuum decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. 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.

Ze-bin Wu and Jun-qing Yu. 2019. Vector quantization: a review. *Frontiers of Information Technology & Electronic Engineering*, 20(4):507–524.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161.

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

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and 1 others. 2023a. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, and 1 others. 2023b. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. 2022. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*.

A Data Engine

We detail our data pipeline for curating rollout data generated by a baseline audio synthesis model, using either synthesized or collected textual prompts.

A.1 Synthesized Prompts

To enrich the diversity and coverage of our training data, we employ a large language model (LLM) (Comanici et al., 2025) to automatically generate a wide range of textual prompts that describe various auditory scenes. These prompts are carefully designed to span diverse acoustic environments, sound sources, spatial configurations, and contextual scenarios. The LLM is guided by a set of curated seed phrases and constraints to ensure semantic coherence, acoustic plausibility, and alignment with the capabilities of the audio generation model. This synthetic prompt generation process significantly expands the breadth of our dataset while maintaining control over the thematic and structural properties of the resulting audio scenes.

A.2 Realistic Prompts Collection

To mitigate the domain gap between synthetically generated prompts and naturally occurring audio descriptions, we curate a set of realistic prompts from the official test sets of two established audio-captioning benchmarks: AudioCaps (Kim et al., 2019) and VGGSound (Chen et al., 2020a). These datasets provide realistic audio collected from real world, that reflect authentic listening experiences and real-world acoustic contexts.

To ensure a fair and standardized evaluation, we reserve 1,000 prompts from each dataset—drawn exclusively from their official test splits—as our held-out evaluation set. The remaining prompts from the test sets (where permitted by the dataset licenses and evaluation protocols) are repurposed to augment our training data, thereby enriching the realism and linguistic diversity of our prompt distribution without compromising the integrity of the final benchmark results. This strategy allows us to bridge the gap between synthetic and real-world prompts while maintaining compatibility with established evaluation frameworks.

A.3 Rollout Data Collection

Using the combined set of synthesized and realistic prompts described above, we generate rollout data with the pre-trained autoregressive audio generator *Siren* (Wang et al., 2025a). For each prompt, we perform random sampling to produce 32 distinct audio rollouts, resulting in a diverse set of acoustic realizations conditioned on the same textual instruction.

For every generated audio–prompt pair, we compute an instruction-following score using the CLAP model (Wu et al., 2023). Specifically, CLAP provides a similarity-based alignment score between the generated audio and its corresponding prompt, which we interpret as a proxy for how well the audio adheres to the semantic and contextual intent of the instruction. These scores are assigned to each rollout trajectory, enabling us to assess and later rank or filter generations based on their fidelity to the given prompt. This annotated dataset of (prompt, audio, CLAP score) triples forms the foundation for subsequent critic optimization stages in our pipeline.

B Metrics Computation

The primary objective evaluation metrics we employ are Fréchet Distance (FD), Inception Score (IS), and Kullback–Leibler (KL) divergence, all computed using features extracted from PANNs (Pretrained Audio Neural Networks) (Kong et al., 2020), a state-of-the-art audio classification model.

- **Fréchet Distance** (FD), analogous to the Fréchet Inception Distance (FID) in image generation, quantifies the similarity between the distributions of generated and reference (real) audio samples in the PANNs embedding space. A lower FD indicates better alignment between the generated and target audio distributions.

- **Inception Score** (IS) assesses both the quality and diversity of generated samples by measuring the entropy of conditional label distributions predicted by PANNs: high-quality samples yield confident predictions (low conditional entropy), while diverse samples produce a broad range of labels (high marginal entropy).

- **KL divergence** is computed at the individual (paired) sample level—comparing the PANNs-predicted class distribution of a generated audio clip against that of its corresponding ground-truth or reference audio—and then averaged across the dataset to yield a final score. Lower KL divergence reflects closer semantic and acoustic alignment.
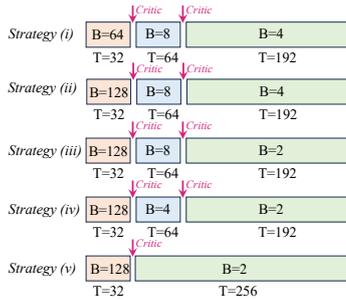
Figure 6: Schematic illustration of different sampling strategies, where B is batch size, T the temporal length.

Sampling Strategy	Cost Tokens↓	CLAP↑	FAD↓	FD↓	IS↑	KL↓
Strategy (i)	3456	32.46	1.87	16.30	12.00	1.50
Strategy (ii)	5504	36.98	1.96	15.79	13.51	1.30
Strategy (iii)	5056	36.75	1.94	15.75	13.56	1.29
Strategy (iv)	4800	36.74	1.93	15.74	13.57	1.30
Strategy (v)	4608	36.47	1.96	15.70	13.82	1.27

Table 9: Ablating different sampling strategies.

In addition to PANNs-based metrics, we also report the **Fréchet Audio Distance (FAD)** (Kilgour et al., 2018), which follows a similar distributional comparison principle but uses embeddings from VGGish (Hershey et al., 2017). While VGGish is a widely used audio representation model, it is generally considered less powerful than PANNs for modern audio understanding tasks; thus, FAD serves as a complementary, albeit potentially less discriminative, benchmark.

Finally, to evaluate semantic alignment between audio and text, we adopt the CLAP score (Wu et al., 2023), which leverages the contrastively trained CLAP model to compute the cosine similarity between the embedded representations of generated audio and their associated prompts. Higher CLAP scores indicate stronger cross-modal coherence and better instruction following.

C Additional Ablate Study on Difference Sampling Strategies

In Table 9, we compare different kinds of sampling strategies with a fixed critic step of $T_{\text{prefix}} = 32$, where the schematic illustration of each strategy can be found in Figure 6.

Strategy (i) generates 64 prefixes and completes all of them, resulting in modest instruction-following performance (CLAP: 32.46). Strategy (ii) explores a larger pool of 128 prefixes and prunes to the top 8 for completion, yielding a significant CLAP improvement (36.98) at a higher token cost (5,504), which exceeds our budget constraint.

Strategies (iii)–(v) operate within the 4,608-token budget. Strategy (iii) mirrors our main method: 128 prefixes are generated, scored by Plan-Critic at step 32, and only the top 2 are completed, achieving a CLAP score of 36.75. Strategy (iv) retains 4 completions instead of 2, achieving comparable performance (CLAP: 36.74) with slightly higher cost (4,800 tokens). Strategy (v) serves as a strong baseline: it generates 128 prefixes and directly completes the top 2 *without* an intermediate critique step, attaining CLAP 36.47—matching the result reported in Table 1 for our full method.

These results highlight two key insights: (1) early evaluation via Plan-Critic enables more effective pruning than selecting prefixes based on generation likelihood alone, and (2) allocating more computation to diverse prefix exploration—followed by aggressive pruning—yields substantial gains in semantic fidelity without increasing total inference cost. Our chosen configuration (Strategy iii) thus offers the best trade-off between performance and efficiency.