SEMANTICAUDIO: AUDIO GENERATION AND EDITING IN SEMANTIC SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, Text-to-Audio Generation has achieved remarkable progress, offering sound creators powerful tools to transform textual inspirations into vivid audio. However, existing models predominantly operate directly in the acoustic latent space of a Variational Autoencoder (VAE), often leading to suboptimal alignment between generated audio and textual descriptions. In this paper, we introduce **SemanticAudio**, a novel framework that decouples both audio generation and editing into semantic planning in a high-level semantic space followed by acoustic synthesis. We define this semantic space as a compact representation capturing the global identity and temporal sequence of sound events, distinct from fine-grained acoustic details. SemanticAudio employs a two-stage Flow Matching architecture: the **Semantic Planner** first generates these compact semantic features to sketch the global semantic layout, and the **Acoustic Synthesizer** subsequently produces high-fidelity acoustic latents conditioned on this semantic plan. Leveraging this decoupled design, we further introduce a training-free text-guided editing mechanism that enables precise attribute-level modifications on general audio without retraining. Specifically, this is achieved by steering the semantic generation trajectory via the difference of velocity fields derived from source and target text prompts. Extensive experiments demonstrate that SemanticAudio achieves competitive generation quality while providing consistently improved semantic alignment and enabling effective training-free audio editing.

Code and demo: `https://semanticaudio1.github.io/`

1 INTRODUCTION

Text-to-Audio (TTA) Generation (Liu et al., 2023; Evans et al., 2024; Liu et al., 2024) aims to synthesize diverse and high-fidelity auditory content directly from natural language textual prompts. This technology serves as a pivotal creative tool for applications including virtual reality, gaming, film post-production, and human-computer interaction. Recent years have witnessed a paradigm shift in this field, fueled by the scaling of data and model parameters alongside architectural innovations. In particular, the adoption of continuous generative objectives, exemplified by Diffusion Models and Flow Matching, has elevated the fidelity and controllability of synthesized audio.

Most mainstream TTA models perform modeling directly in the acoustic latent space, typically utilizing compressed representations from a Variational Autoencoder (VAE) (Liu et al., 2023; 2024). While this design excels at preserving low-level acoustic fidelity, it often falls short in high-level semantic understanding. These models frequently struggle to precisely capture the intent in textual prompts, resulting in insufficient *alignment*—defined here as the accurate correspondence between the presence and sequence of auditory events and the text description.

Addressing this limitation requires a clear distinction between the *semantic* and *acoustic* levels of audio. In this work, we define semantics as the high-level conceptual content—specifically the identity, occurrence, and temporal sequence of sound events—as distinct from fine-grained acoustic details. Audio signals exhibit significant semantic redundancy: high-level semantics are relatively compact and abstract compared to dense acoustic details. Drawing inspiration from two-stage semantic planning approaches in video generation, we hypothesize that directly modeling dense low-level representations in a high-dimensional acoustic latent space is suboptimal for achieving se-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mantic alignment. Instead, the generation process should be decomposed: first accomplishing global content planning in a compact high-level semantic space, followed by the progressive synthesis of acoustic details.

Motivated by this insight, we propose **SemanticAudio**, a novel two-stage Flow Matching-based framework. The core innovation lies in performing the audio generation process via a high-level semantic space. First, a **Semantic Planner** generates compact semantic features from text to sketch the global event layout. Second, conditioned on these features, an **Acoustic Synthesizer** produces high-fidelity VAE latent representations. This design effectively addresses the limitations in high-level semantic modeling inherent in conventional acoustic-space approaches.

Beyond generation, we apply the decoupled architecture to audio editing. In our experiments, training-free editing with acoustic-space models gave weaker results than editing in SemanticAudio's semantic space. We steer the generation trajectory using the difference of velocity fields from source and target prompts to modify audio attributes without additional training. This differs from supervised editing with Audit (Wang et al., 2023) and reference-guided latent optimization with AudioMorphix (Liang et al., 2025); both support multiple editing operations.

The main contributions of this work are summarized as follows:
SemanticAudio Framework: We propose a two-stage framework comprising a **Semantic Planner** and an **Acoustic Synthesizer**. This architecture performs audio generation directly in a high-level semantic space, effectively decoupling content planning from acoustic synthesis.

Superior Semantic Consistency: By first sketching the global event layout in the semantic space, our method improves substantially over our controlled acoustic-space baseline and is competitive with strong existing systems in high-level semantic alignment between generated audio and textual prompts.

Training-free Audio Editing: We introduce a training-free mechanism that enables flexible, text-driven manipulation of high-level semantics on general audio. By directly steering the semantic ODE trajectory, this approach achieves versatile attribute-level modifications without requiring additional training or complex inversion steps.

2 Related Work

Text-to-Audio Generation Many TTA models use an acoustic VAE latent space, including latent-diffusion systems such as AudioLDM (Liu et al., 2023), Make-An-Audio (Huang et al., 2023), and Tango 2 (Majumder et al., 2024). AudioGen (Kreuk et al., 2023) instead models discrete audio codes autoregressively. Flow Matching systems such as MeanAudio (Li et al., 2025) and LAFMA (Guan et al., 2024) also generate acoustic latents. Modeling event structure and acoustic details in a single space motivates our separation of semantic planning and synthesis.

Semantic Representations in Audio Prior work has already generated high-level audio representations: AudioLM (Borsos et al., 2023) models discrete semantic tokens, AudioLDM 2 (Liu et al., 2024) predicts AudioMAE-derived representations before acoustic synthesis, and SemanticVocoder (Xie et al., 2026) synthesizes audio from semantic latents. CLAP (Wu et al., 2024) learns global contrastive embeddings, whereas AudioMAE (Huang et al., 2022) uses masked spectrogram reconstruction. Our design uses continuous PE-A-Frame embeddings (Vyas et al., 2025) as a frame-level generation target, followed by a separately trained acoustic synthesizer.

Audio Editing Training-based approaches such as Audit (Wang et al., 2023) and RFM-Editing (Gao et al., 2025) learn instruction-following edits from paired data. Training-free methods use pre-trained generators; AudioMorphix (Liang et al., 2025) combines reference audio, spectrogram masks, and guided latent optimization for several editing operations. FlowEdit (Kulikov et al., 2025) introduced inversion-free vector-field composition for images. We apply this idea to audio in a continuous frame-level semantic space.

Inspirations from Vision. Decoupling semantic planning from low-level synthesis has proven effective in image and video generation (Bai et al., 2025; Li et al., 2024; Ouyang et al., 2025). We adapt this paradigm to continuous frame-level semantic planning for audio generation and editing.

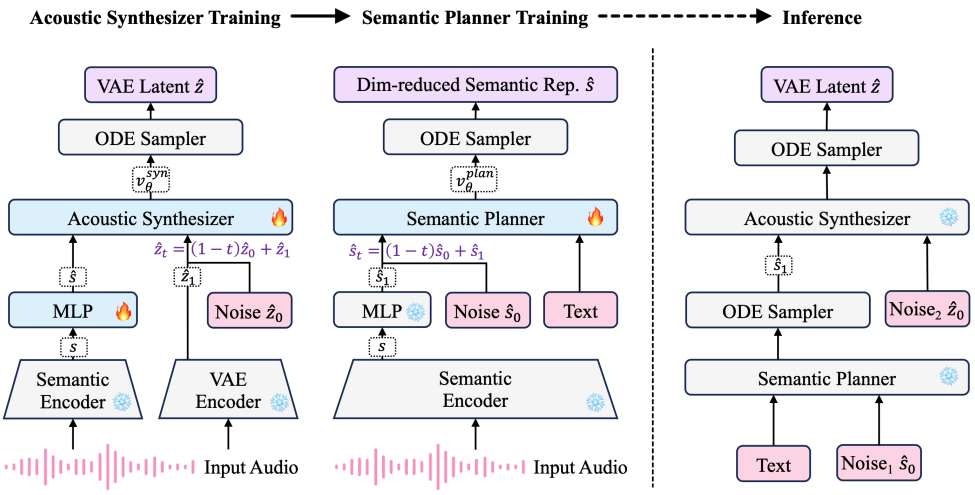

Figure 1: Overview of the SemanticAudio framework. The model employs a two-stage Flow Matching architecture: the Semantic Planner first generates low-dimensional semantic latents conditioned on text, followed by the Acoustic Synthesizer which produces high-fidelity acoustic latents for VAE decoding.

3 SemanticAudio Framework

In this section, we present the detailed architecture of the SemanticAudio framework. As illustrated in Figure 1, our framework effectively decouples text-to-audio generation into two distinct stages: (1) a **Semantic Planner** that sketches the global event layout in a compact semantic space, and (2) an **Acoustic Synthesizer** that produces high-fidelity acoustic details conditioned on the semantic plan. We first detail the representation spaces, followed by the design of the two generative stages.

3.1 Pre-trained VAE and Semantic Representation

SemanticAudio builds upon a pre-trained variational autoencoder (VAE) and a semantic encoder to bridge raw audio waveforms and high-level semantics.

Acoustic Representation. SemanticAudio leverages a variational autoencoder (VAE) to compress a raw audio waveform into a compact acoustic latent space $z \in \mathbb{R}^{T \times C}$. Formally, the encoder E_{VAE} maps the input waveform a to a latent representation $z = E_{\text{VAE}}(a)$, where T denotes the number of acoustic time steps and C represents the channel dimension. The decoder D_{VAE} reconstructs the audio from this latent, $\hat{a} = D_{\text{VAE}}(z)$, ensuring high perceptual fidelity. In this work, we adopt the continuous-latent DAC-VAE (Polyak et al., 2024), based on the Descript Audio Codec (Kumar et al., 2023), as our acoustic VAE.

Semantic Representation. To enable precise control over the temporal layout and content of sound events, we require a semantic encoder E_{sem} capable of extracting continuous, frame-level embeddings $s \in \mathbb{R}^{N \times D}$. Here, N corresponds to the number of semantic frames (determined by the frame rate of the encoder) and D is the embedding dimension. Unlike global descriptors that aggregate information into a single vector (e.g., CLAP (Elizalde et al., 2023)), frame-level representations are essential for preserving the fine-grained temporal structure required for event planning. Formally, we define the *semantic space* as the high-dimensional, understanding-oriented embedding space of a frame-level audio–text encoder, which encodes the identity, occurrence, and temporal position of sound events (*what* and *when*) rather than low-level acoustic detail. PE-A-Frame (D=1024, 25 fps) reports strong sound-event detection performance (Vyas et al., 2025), and a learned projection P_θ compresses it to a compact d-dimensional target that stays tractable for Flow Matching.

In this work, we adopt the pre-trained **Perception Encoder** (Vyas et al., 2025). This model is trained via fine-grained supervised contrastive learning on large-scale audio-text datasets. By explicitly aligning audio frames with their corresponding textual descriptions, it excels at capturing precise

semantic-temporal correspondences. This makes it uniquely capable of tasks requiring detailed event sequencing and distinguishing overlapping sound concepts, providing a robust foundation for our Semantic Planner. To enable tractable modeling in the generative process, we introduce a lightweight MLP projection head P_θ that reduces these high-dimensional embeddings ($D = 1024$) to a compact low-dimensional space:

$$\hat{s} = P_\theta(s) \in \mathbb{R}^{N\times d}, \quad d \ll D. \tag{1}$$

The projection head P_θ, which is randomly initialized, is trained jointly with the Acoustic Synthesizer and remains fixed during the subsequent training of the Semantic Planner. This design ensures that the reduced representations $\hat{s}$ preserve essential semantic content necessary for accurate acoustic synthesis.

3.2 Semantic Planner: Text-to-Semantic Generation

The **Semantic Planner** is responsible for high-level content planning. It learns to generate low-dimensional semantic representations directly conditioned on text prompts, effectively sketching the global event layout.

Given a text prompt y, we extract complementary semantic conditions using two distinct **pre-trained** encoders. We employ the text encoder from CLAP (Elizalde et al., 2023) to extract a global sentence embedding c_g, capturing the high-level atmosphere. Simultaneously, we use the Flan-T5 (Chung et al., 2022) encoder to extract a sequence of token-level embeddings c_d, preserving fine-grained syntactic structures and dynamic instructions. To simplify notation, we denote the full textual conditioning set as $C = \{c_g, c_d\}$. These representations serve as dual inputs to ensure both global coherence and local precision.

The Semantic Planner is a Flow Matching model $\mathcal{F}_{\text{plan}}$ that learns a velocity field $v_\theta^{\text{plan}}(t, \hat{s}_t, C)$ to transport noise $\hat{s}_0 \sim \mathcal{N}(0, I)$ to the target semantic latent $\hat{s}_1$. The training objective follows the Flow Matching (Lipman et al., 2023) loss:

$$\mathcal{L}_{\text{FM}}^{\text{plan}} = \mathbb{E}_{t,\hat{s}_t,C} \left\| v_\theta^{\text{plan}}(t, \hat{s}_t, C) - (\hat{s}_1 - \hat{s}_0) \right\|^2, \tag{2}$$

where $\hat{s}_t = (1-t)\hat{s}_0 + t\hat{s}_1$, and the target $\hat{s}_1 = P_\theta(E_{\text{sem}}(a_{\text{gt}}))$ is computed using the frozen projection head P_θ learned during Acoustic Synthesizer training. Here, a_{gt} denotes the target ground-truth audio waveform from the training pair, and E_{sem} is the fixed semantic encoder. The Semantic Planner thus requires no manual annotation: its supervision is self-supervised regression onto these fixed PE-derived targets $\hat{s}_1$.

During inference, we sample $\hat{s}_0 \sim \mathcal{N}(0, I)$ and integrate the learned velocity field from t=0 to t=1 using the Euler method with N discrete steps:

$$\hat{s}_{t+\Delta t} = \hat{s}_t + v_\theta^{\text{plan}}(t, \hat{s}_t, C) \cdot \Delta t, \quad \Delta t = \tfrac{1}{N}, \tag{3}$$

yielding the planned semantic features $\hat{s}_1$.

3.3 Acoustic Synthesizer: Semantic-to-Acoustic Synthesis

The **Acoustic Synthesizer** bridges abstract semantic plans and concrete auditory signals. Conditioned on the semantic features $\hat{s}_1$, it learns to synthesize high-fidelity acoustic latents $z_1 \in \mathbb{R}^{T\times C}$.

Training Strategy. A critical aspect of our framework is that **the Acoustic Synthesizer is trained prior to the Semantic Planner**. We jointly optimize the synthesizer and the projection head P_θ. This ensures that the projected semantic features $\hat{s} = P_\theta(E_{\text{sem}}(a_{\text{gt}}))$ retain sufficient information for reconstruction while discarding redundant noise. Once trained, P_θ is frozen to provide target labels for the Semantic Planner.

Modeling. The synthesizer adopts the same Flow Matching formulation as the Semantic Planner (Eq. 2). It learns a velocity field $v_\theta^{\text{syn}}(t, z_t, \hat{s}_1)$ to map noise z_0 to the ground-truth acoustic latents $z_1 = E_{\text{VAE}}(a_{\text{gt}})$, conditioned on $\hat{s}_1$.

Inference. The full generation pipeline is executed sequentially: we first generate the semantic plan $\hat{s}_1$ using the Semantic Planner, which then serves as the condition for the Acoustic Synthesizer to generate z_1. Finally, the waveform is reconstructed via the VAE decoder $\hat{a} = D_{\text{VAE}}(z_1)$.

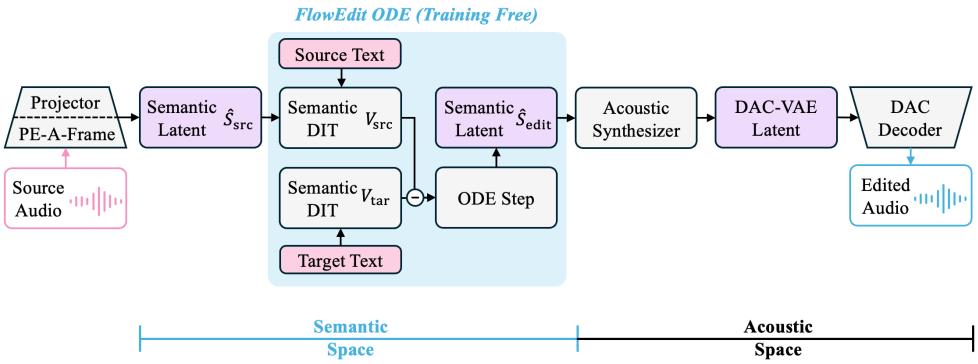

Figure 2: Overview of our training-free text-guided audio editing method. The process leverages the pre-trained velocity fields of the Semantic Planner to perform semantic-level editing in the low-dimensional latent space via difference velocity integration, followed by high-fidelity reconstruction using the Acoustic Synthesizer. The method requires no additional training, inversion, or optimization.

3.4 Training-Free Text-Guided Audio Editing

A key advantage of our decoupled SemanticAudio framework is its inherent capability for training-free audio editing. Unlike pixel- or acoustic-space editing methods that often struggle to disentangle semantic content from background noise, our approach operates directly on the high-level semantic layout. This allows users to modify specific auditory events while preserving the underlying temporal structure, all without requiring model fine-tuning.

Building upon this insight, we introduce a mechanism inspired by FlowEdit (Kulikov et al., 2025), as shown in Figure 2. It directly leverages the velocity fields learned by the **Semantic Planner** to perform precise semantic-level modifications, while the **Acoustic Synthesizer** ensures high-fidelity acoustic reconstruction.

Given a source audio a_{src} and its semantic latent $\hat{s}_{\text{src}}$, we define the editing trajectory using a **Delta Velocity Field** v_{Δ}^{t}. This field represents the directional difference between the Semantic Planner's velocity fields conditioned on the target (C_{tgt}) and source (C_{src}) prompts:

$$v_{\Delta}^{t}(\hat{s}^{t}, t) = v_{\theta}^{\text{plan}}(\hat{s}^{t}, t, C_{\text{tgt}}) - v_{\theta}^{\text{plan}}(\hat{s}^{t}, t, C_{\text{src}}). \tag{4}$$

where C_{src} can be the source text embedding or **null conditioning** if the source text is unavailable.

In practice, to ensure stability against stochastic variations, we approximate v_{Δ}^{t} by averaging over N_{avg} noisy realizations at each timestep:

$$v_{\Delta}^{t} \approx \frac{1}{N_{\text{avg}}} \sum_{i=1}^{N_{\text{avg}}} \left[v_{\theta}^{\text{plan}}(\hat{s}_{\text{tgt},i}^{t}, t, C_{\text{tgt}}) - v_{\theta}^{\text{plan}}(\hat{s}_{\text{src},i}^{t}, t, C_{\text{src}}) \right]. \tag{5}$$

Starting from the source semantic latent $\hat{s}^{1} = \hat{s}_{\text{src}}$, we integrate this delta field backward to $t = 0$ using standard discrete steps (e.g., Euler method) to obtain the edited semantic latent $\hat{s}_{\text{edit}}$. This construction does not assume the flow is globally linear: following FlowEdit, the difference of conditional velocity fields yields a first-order editing direction that is integrated stepwise along the trajectory, and the N_{avg} averaging above controls stochastic variation at each step. Finally, $\hat{s}_{\text{edit}}$ is decoded by the **Acoustic Synthesizer** into the final audio.

4 Experiments

In this section, we empirically evaluate SemanticAudio on three primary tasks: text-to-audio generation, text-to-audio benchmark evaluation, and training-free semantic editing. We aim to verify our core hypothesis: decoupling global semantic planning from acoustic synthesis leads to improved semantic alignment without compromising audio fidelity.

4.1 Datasets and Evaluation Protocols

Training Data. We train all models on a combined corpus of three public datasets: AudioCaps (Kim et al., 2019) ($\sim$46k clips, $\sim$128 hours), AudioSet (Gemmeke et al., 2017) (a large-scale audio event dataset with $\sim$2 million 10-second clips), and WavCaps (Mei et al., 2024) ($\sim$400k audio-caption pairs harvested from diverse web sources). This combination provides both high-quality human annotations (AudioCaps) and broad acoustic coverage (AudioSet, WavCaps), enabling the model to learn from diverse sound categories and complex scene compositions. All audio clips are standardized to a 10-second duration at 48 kHz via silence padding or truncation.

Test Set for Generation. For standard text-to-audio generation, we utilize the official AudioCaps (Kim et al., 2019) test split (957 clips). Following standard protocols (Li et al., 2025), we randomly select one caption per clip as the generation prompt.

TTA-Bench. We additionally evaluate on the 1,500-prompt accuracy subset of TTA-Bench (Wang et al., 2026), which spans sound categories and temporal relations. The full benchmark contains 2,999 prompts; the accuracy subset includes 400 AudioCaps prompts.

Protocol for Training-Free Editing. We construct a 200-pair editing benchmark derived from the AudioCaps test split: **Source Audio:** We use 50 representative clips as source audio. **Instruction Generation:** Using GPT-5, we generate diverse editing instructions per clip (e.g., timbre modification, event replacement) based on the original caption. **Benchmark Construction:** We retain pairs with low source-to-target CLAP and include 200 pairs across edit types. Details are in Appendix B.

4.2 Implementation Details

Implementation. SemanticAudio comprises two decoupled DiT Flow-Matching modules (28 layers, 16 heads, hidden dim 1152, $\sim$610M parameters each); the **Base Model** uses one DiT backbone without a Semantic Planner. Text conditioning uses CLAP and FLAN-T5-large, while PE-A-Frame-large supplies frame-level audio semantics; the acoustic space is DAC-VAE ($C = 128$). The primary semantic dimension is $d = 128$ (Section 4.5). The Planner and Base Model have matched DiT backbones, but the two-stage system has an additional $\sim$610M-parameter synthesizer, so total capacity and compute are not controlled by this comparison. We explore $d \in \{32, 64, 128, 1024\}$; at d=1024 we compare *rawPE* (raw PE, no projection) and *DDT* (a learned linear projection). Generation runs use 500k-step checkpoints (AdamW, LR 10^{-4}, batch 32 per GPU, effective batch 256 on 8 H20 GPUs) trained on AudioCaps+AudioSet+WavCaps. At inference, the Planner uses CFG 4.5 (50 steps) and the Synthesizer unguided sampling (25 steps); editing uses FlowEdit (N_{avg}=1, target CFG 5.0, $n_{\min}$=0). Further details are in Appendix A.

4.3 Evaluation Metrics

We adopt a multi-faceted evaluation protocol to assess the model across four distinct dimensions.

Reconstruction Quality. To verify the **Acoustic Synthesizer's** ability to decode semantic plans into high-fidelity waveforms, we report the **Mel-spectrogram loss** and **Multi-Scale STFT loss**.

Text-to-Audio Generation. We employ standard objective metrics on the AudioCaps test set: **CLAP$_\text{L}$** (LAION-CLAP (Wu et al., 2024)) for semantic alignment, **Fréchet Distance (FD)** computed on PANN features (Kong et al., 2020) for distributional fidelity, and **Inception Score (IS)** for sample quality and diversity. We additionally conduct a human listening study: 20 listeners rate 30 randomly selected samples per model on a 5-point Mean Opinion Score (MOS) scale for overall audio quality.

TTA-Bench Metrics. On the 1,500-prompt accuracy subset, we report **CLAP$_\text{M}$** (MS-CLAP (Elizalde et al., 2023)) for semantic alignment along with **Aesthetics (AES)** scores from Audiobox Aesthetics (Tjandra et al., 2025): AES-CU (Content Usefulness, higher is better) estimates how suitable the audio is as source material for content creation, and AES-PQ (Production Quality, higher is better) evaluates technical production quality.

Editing Metrics. For the editing benchmark (200 pairs), we report CLAP$_\text{L}$ (LAION-CLAP) similarity between the edited audio and the target caption, the improvement over the original source

audio (Δ CLAP_L), and MOS from human listeners (10 listeners, 20 samples per model) evaluating overall edit quality and naturalness.

4.4 Results and Analysis

We evaluate SemanticAudio on text-to-audio generation, TTA-Bench, and training-free editing. We compare against state-of-the-art baselines Resonate (Li et al., 2026) and TangoFlux (Hung et al., 2025), as well as our controlled **Base Model**. As an additional structure-sensitive check that goes beyond global CLAP/FD, we further validate event ordering and editing preservation with an independent audio-LLM judge (Appendix C).

4.4.1 Text-to-Audio Generation

Table 1: Text-to-audio generation on AudioCaps test set (957 samples). Lower is better (↓) for FD; higher is better (↑) for IS, CLAP, and MOS.

Model	FD↓	IS↑	CLAP_L↑	MOS↑
Resonate	**16.0**	11.04	**0.403**	3.78
TangoFlux	22.6	**12.05**	0.361	**3.85**
Base Model	25.0	7.11	0.318	3.42
Ours (d=32)	22.5	7.35	0.348	3.48
Ours (d=64)	22.2	7.40	0.356	3.52
Ours (d=128)	**19.1**	**9.13**	**0.381**	**3.72**
Ours (d=128, REPA)	21.7	8.24	0.358	3.66
Ours (d=1024)	20.9	8.76	0.373	3.65
Ours (d=1024, rawPE)	33.8	6.65	0.320	3.18

As shown in Table 1, SemanticAudio ($d = 128$) attains a LAION-CLAP score of 0.381, surpassing TangoFlux (0.361) and our Base Model (0.318) and approaching the Resonate baseline, which is highest at 0.403. In terms of distributional fidelity (FD), our model achieves 19.1, approaching Resonate (16.0) while substantially outperforming TangoFlux (22.6) and the Base Model (25.0). Human evaluation (MOS) shows TangoFlux (3.85) slightly edges Resonate (3.78) in perceptual quality, with our $d = 128$ model (3.72) close behind—indicating competitive perceptual quality while improving text-audio alignment over TangoFlux and the Base Model on this test set.

The consistent improvement from $d = 32$ to $d = 128$ demonstrates that richer semantic representations enable better alignment. The $d = 1024$ variant achieves competitive performance (CLAP 0.373, FD 20.9). Notably, the rawPE variant ($d = 1024$, without learned projection) shows significantly degraded results (CLAP 0.320, FD 33.8), demonstrating that the learned projection P_θ is critical—raw PE embeddings are not directly suitable as generation targets.

The separate $d = 128$ REPA run reports FD-PANN 21.7, ISC-PANN 8.24, LAION-CLAP 0.358, and MOS 3.66 on the same 957-clip AudioCaps test split. Its additional scores are FD-VGG 3.1309, FD-PASST 231.2200, and ISC-PASST 5.8299 ± 0.4477.

4.4.2 TTA-Bench Evaluation

Table 2 presents results on the 1,500-prompt TTA-Bench accuracy subset. SemanticAudio ($d = 128$) achieves CLAP of 0.459, matching Resonate and approaching TangoFlux (0.464). In audio quality (AES-PQ 5.936), our model nearly matches Resonate (5.942) while substantially outperforming TangoFlux (5.701). For content usefulness (AES-CU 5.187), our model outperforms TangoFlux (5.006) and approaches Resonate (5.258). The improvement over the Base Model (+0.033 CLAP, +0.068 AES-CU) extends beyond evaluation on the AudioCaps test prompts alone. The new $d = 128$ REPA run reports CLAP 0.460, AES-CU 4.822, and AES-PQ 5.603. Its CLAP is similar to the original $d = 128$ run, but its aesthetics scores are lower; since it uses 50 Synthesizer steps instead of 25, this comparison does not isolate REPA's effect.

Table 2: Evaluation on the TTA-Bench accuracy subset (1,500 prompts). Higher is better (↑) for all metrics. AES-CU and AES-PQ are from Audiobox Aesthetics (Tjandra et al., 2025).

Model	CLAP_M↑	AES-CU↑	AES-PQ↑
Resonate	0.459	**5.258**	**5.942**
TangoFlux	**0.464**	5.006	5.701
Base Model	0.426	5.119	5.897
Ours (d=32)	0.422	4.896	5.648
Ours (d=64)	0.425	4.911	5.652
Ours (d=128)	0.459	**5.187**	**5.936**
Ours (d=128, REPA)	**0.460**	4.822	5.603
Ours (d=1024)	0.458	5.045	5.735

4.4.3 Acoustic Reconstruction Quality

Table 7 in Appendix D evaluates the Acoustic Synthesizer with ground-truth semantic features. Our $d = 128$ model obtains Mel and STFT losses of 1.413 and 1.828; the $d = 32$ variant has higher losses of 1.817 and 2.270, while the 1024-dimensional rawPE variant has the lowest losses (1.202 and 1.670). The target for our rows is the waveform reconstructed by the frozen DAC-VAE, so these values isolate synthesizer error beyond DAC reconstruction rather than measuring the complete text-to-audio pipeline. The SemanticVocoder reference uses a different target and native audio resolution; its losses in Appendix D are contextual, not directly comparable.

4.4.4 Training-Free Semantic Editing

Table 3: Editing evaluation on 200 AudioCaps-derived pairs across 14 edit types. All methods are reported with FlowEdit hyperparameters cfg=5.0, $n_{\min}$=0, and 50 steps in different latent spaces.

Method	CLAP_L↑	$\Delta\text{CLAP}_\text{L}$↑	MOS↑
Ours (Semantic)	**0.351**	**+0.094**	**3.71**
TangoFlux	0.339	+0.081	3.54
Resonate	0.322	+0.065	3.48
Base (DAC)	0.290	+0.033	3.32
Original Source	0.257	—	3.89

We report editing on 200 pairs spanning 14 categories (Table 3). They cover subject and event replacement, sound addition/removal, environment and weather changes, and other attribute edits. Most instructions change a single semantic element. Construction and category results are in Appendix B.

Editing Results. SemanticAudio's average CLAP rises from 0.257 for the source to 0.351 for the edit (ΔCLAP $+0.094$). The reported ΔCLAP values for TangoFlux, Resonate, and the DAC baseline are +0.081, +0.065, and +0.033, respectively. Mean MOS is 3.71 for our edits versus 3.54 for TangoFlux and 3.48 for Resonate (source: 3.89).

Interpreting Semantic FlowEdit. We hypothesize that applying the delta velocity $v_\Delta = v_{\text{tar}} - v_{\text{uncond}}$ in the compact semantic space makes target attributes easier to steer than in acoustic latents. These results are consistent with this explanation, but the comparison does not isolate the mechanism or control total model capacity.

Editing Configuration. We use null source conditioning, target-text CFG 5.0, $n_{\min} = 0$, and 50 Euler steps for the semantic edit. The Acoustic Synthesizer decodes the edited latent with 25 steps. Editing requires no additional model training or per-example optimization. Appendix A.3 gives the full settings.

4.5 Ablation and Analysis

Semantic Encoder Capabilities. PE-A-Frame-Large (Vyas et al., 2025) reports frame-level sound-event detection AUROC of 0.96 on AudioSet-Strong and 0.97 on DESED. These results motivate its use for modeling *what* events occur and *when*; our $d = 128$ projection is evaluated in Appendix D (Table 7).

Impact of Semantic Dimension. Our ablation across semantic dimensions ($d \in \{32, 64, 128, 1024\}$) reveals a clear trend: $d = 128$ achieves the best balance between semantic alignment and acoustic fidelity across all benchmarks. Lower dimensions ($d = 32, 64$) create information bottlenecks that limit both FD and CLAP scores in generation. On AudioCaps, CLAP improves from 0.348 ($d = 32$) to 0.381 ($d = 128$), a 9.5% relative gain. Similarly on TTA-Bench, CLAP improves from 0.422 to 0.459. The 1024-dimensional DDT variant achieves competitive generation (CLAP 0.373 on AudioCaps, 0.458 on TTA-Bench) but does not surpass the compressed $d = 128$ representation. This suggests that moderate compression acts as a beneficial regularizer for the Semantic Planner, forcing it to focus on the most discriminative semantic features while filtering out redundant variation.

Comparison with Baselines. Resonate (Li et al., 2026) and TangoFlux (Hung et al., 2025) generate in acoustic VAE latent spaces without a separate frame-level semantic planner, although their VAE spaces and optimization methods differ. The Base Model shares our Planner's DiT backbone and training recipe but omits semantic planning. The full two-stage system has greater total capacity than that single-module baseline; these results therefore do not isolate the effect of semantic planning from model size.

Why Semantic Space Helps. The fundamental insight is that text-to-audio alignment benefits from an intermediate semantic representation that bridges the modality gap. When a Flow Matching model operates directly in the 128-dimensional DAC acoustic space, it must simultaneously learn (1) what sounds to generate (semantic content) and (2) how to render them (acoustic details). By factoring these into separate stages, the Semantic Planner can focus exclusively on "what," leveraging the pre-trained PE-Audio encoder's strong text-audio alignment. The Acoustic Synthesizer then handles "how," conditioned on an explicit semantic plan. This decomposition is particularly beneficial for complex prompts involving multiple sequential events, where the temporal ordering must be correctly resolved before acoustic synthesis.

TTA-Bench Accuracy. SemanticAudio ($d = 128$) matches Resonate and approaches TangoFlux on the diverse TTA-Bench accuracy subset despite differences in training strategies. Its improvement over the Base Model (+0.033 CLAP) extends beyond the AudioCaps test prompts alone. Because this subset includes AudioCaps prompts, we do not interpret the aggregate score as a separate test of out-of-distribution generalization.

5 Conclusion

We presented SemanticAudio, a two-stage Flow Matching framework that decouples text-to-audio generation into semantic planning and acoustic synthesis. The semantic stage models event identity and timing in compact frame-level features, and the acoustic stage converts the resulting plan into DAC latents for waveform decoding. The Semantic Planner attains strong CLAP on AudioCaps (0.381, surpassing TangoFlux at 0.361 and our Base Model at 0.318) and competitive TTA-Bench performance, while the Acoustic Synthesizer renders high-fidelity audio. The same design enables training-free editing via differential velocity fields (FlowEdit), with ΔCLAP $+0.094$ on the 200-pair editing benchmark versus a reported $+0.081$ for the strongest acoustic-space baseline. Together, the generation and editing results suggest that semantic planning can bridge text intent and acoustic realization. Future work will extend the method to variable-length audio, multimodal conditioning, and compositional multi-edit instructions.

Ethical Statement

SemanticAudio synthesizes and edits audio and could be misused to fabricate or tamper with soundscapes. We train on AudioCaps, AudioSet, and WavCaps, which draw on audio from multiple sources with differing use conditions. The method can alter perceived vocal attributes, a capability that carries impersonation risk. We encourage watermarking and provenance tracking for AI-generated audio. At our institution, the human listening studies did not require ethics review or an exemption.

AI Use Statement

GPT-5 was used to generate editing instructions from source captions for the audio-editing benchmark (Appendix B). Gemini-2.5 was used as an audio-understanding judge for the supplementary evaluation (Appendix C). GPT-5.6, via OpenAI Codex, assisted with manuscript polishing and demo-page construction.

Reproducibility Statement

Architecture and training details are given in Appendix A; the edit-set construction and evaluation protocol are described in Appendices B and C. An anonymized code and audio-demo release is linked from this manuscript.

References

Jianhong Bai, Xiaoshi Wu, Xintao Wang, Xiao Fu, Yuanxing Zhang, Qinghe Wang, Xiaoyu Shi, Menghan Xia, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Kun Gai. Semanticgen: Video generation in semantic space, 2025. URL https://arxiv.org/abs/2512.20619.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation, 2023. URL https://arxiv.org/abs/2209.03143.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. In *Proc. ICASSP*, pp. 1–5, 2023.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024. URL https://arxiv.org/abs/2407.14358.

Liting Gao, Yi Yuan, Yaru Chen, Yuelan Cheng, Zhenbo Li, Juan Wen, Shubin Zhang, and Wenwu Wang. Rfm-editing: Rectified flow matching for text-guided audio editing, 2025. URL https://arxiv.org/abs/2509.14003.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingyang Hong, and Yong Qin. Lafma: A latent flow matching model for text-to-audio generation. In *Proc. Interspeech*, pp. 4813–4817, 2024.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2207.06405.

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, 2023. URL https://arxiv.org/abs/2301.12661.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization, 2025. URL https://arxiv.org/abs/2412.21037.

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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

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, 2020. URL https://arxiv.org/abs/1912.10211.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation, 2023. URL https://arxiv.org/abs/2209.15352.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models, 2025. URL https://arxiv.org/abs/2412.08629.

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

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method, 2024. URL https://arxiv.org/abs/2312.03701.

Xiquan Li, Junxi Liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen, and Xie Chen. Meanaudio: Fast and faithful text-to-audio generation with mean flows, 2025. URL https://arxiv.org/abs/2508.06098.

Xiquan Li, Junxi Liu, Wenxi Chen, Haina Zhu, Ziyang Ma, and Xie Chen. Resonate: Reinforcing text-to-audio generation via online feedback from large audio language models, 2026. URL https://arxiv.org/abs/2603.11661.

Jinhua Liang, Yuanzhe Chen, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhuo Chen, Yuping Wang, and Yuxuan Wang. Audiomorphix: Training-free audio editing with diffusion probabilistic models, 2025. URL https://arxiv.org/abs/2505.16076.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

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. In *Proceedings of the International Conference on Machine Learning*, pp. 21450–21474, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. doi: 10.1109/TASLP.2024.3399607.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization, 2024. URL https://arxiv.org/abs/2404.09956.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

Wenqi Ouyang, Zeqi Xiao, Danni Yang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Tokensgen: Harnessing condensed tokens for long video generation, 2025. URL https://arxiv.org/abs/2507.15728.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models, 2024. URL https://arxiv.org/abs/2410.13720.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning, 2025. URL https://arxiv.org/abs/2512.19687.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. TTA-Bench: A comprehensive benchmark for evaluating text-to-audio models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33512–33520, 2026. doi: 10.1609/aaai.v40i39.40639.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. Audit: Audio editing by following instructions with latent diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, 2024. URL https://arxiv.org/abs/2211.06687.

Zeyu Xie, Chenxing Li, Qiao Jin, Xuenan Xu, Guanrou Yang, Wenfu Wang, Mengyue Wu, Dong Yu, and Yuexian Zou. Semanticvocoder: Bridging audio generation and audio understanding via semantic latents, 2026. URL https://arxiv.org/abs/2602.23333.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9e42b4d7163931f3689d6d6fbaa11d0-Abstract-Conference.html.

A Model and Training Details

A.1 Architecture

Generative Backbone. Both the Semantic Planner and Acoustic Synthesizer employ a Diffusion Transformer (DiT) architecture consisting of 28 transformer layers with 16 attention heads, a hidden dimension of 1152, and a total of ∼610M parameters each. We use rotary positional embeddings and adaptive layer normalization conditioned on the diffusion timestep. The Base Model (controlled ablation) shares this identical architecture for fair comparison.

Conditioning Setup. Text prompts are encoded via two complementary encoders:

- **Global condition** (c_g): CLAP text encoder (Elizalde et al., 2023) produces a 1024-dimensional global sentence embedding, injected via adaptive layer norm.
- **Cross-attention condition** (c_d): Flan-T5-Large (Chung et al., 2022) encoder produces token-level embeddings (1024-dim, variable length), attended to via cross-attention layers.

Acoustic VAE. We adopt the pre-trained DAC-VAE (Polyak et al., 2024) operating at 48 kHz with a 128-dimensional continuous latent space. For 10-second audio, the VAE produces 250 latent frames (T=250, C=128). The VAE is frozen throughout all training.

Semantic Encoder. PE-A-Frame-Large (Vyas et al., 2025) extracts 1024-dimensional frame-level embeddings at 25 fps, yielding 250 frames for 10-second audio (N=250, D=1024). The encoder is frozen; only the downstream MLP projection P_θ is trained.

MLP Projection Head (P_θ). A 2-layer MLP ($1024 \to 512 \to d$) with GELU activation. Trained jointly with the Acoustic Synthesizer. For d=1024 (DDT variant), the projection is a single linear layer.

A.2 Training Configuration

- **Data:** AudioCaps (Kim et al., 2019) (∼46k clips) + AudioSet (Gemmeke et al., 2017) (∼2M clips) + WavCaps (Mei et al., 2024) (∼400k clips). All clips resampled to 48 kHz mono and padded/trimmed to 10 seconds.
- **Optimization:** AdamW, learning rate 10^{-4}, β_1=0.9, β_2=0.999, weight decay 10^{-2}. Linear warm-up over 1k steps followed by step decay (0.5× at 200k and 400k steps).
- **Batch size:** 32 per GPU × 8 H20 GPUs = 256 effective batch size.
- **Training duration:** Generation comparisons use 500k-step checkpoints. Reconstruction checkpoints vary by semantic dimension, as noted in Appendix D.
- **Time step sampling:** Logit-normal distribution with μ=0.4, σ=1.0.
- **CFG dropout:** 10% unconditional dropout during training for both global and cross-attention conditions.
- **Training order:** Acoustic Synthesizer + P_θ trained first; P_θ is then frozen and used to provide targets for Semantic Planner training.

A.3 Inference Configuration

- **Semantic Planner:** Euler ODE solver, 50 steps, CFG scale 4.5.
- **Acoustic Synthesizer:** Euler ODE solver, 25 steps, no CFG (scale 1.0).
- **Editing:** FlowEdit with null source conditioning (`force_uncond`=True for v_{src}), target CFG scale 5.0, 50 Euler steps, $n_{\min}$=0, VAE decode 25 steps.

We also evaluate a 128-dimensional REPA variant (Yu et al., 2025) using PE-A-Frame-Large and FLAN-T5-Large encoders. Its Semantic DiT generates semantic latents, which the Acoustic DiT maps to DAC latents for 48 kHz audio decoding. For this experiment, both stages use 50 sampling steps, with CFG 4.5 and seed 42. Its additional TTA-Bench metrics are AES-CE 3.0625 and AES-PC 3.5422.

Table 4: Editing performance by category on the 200-pair benchmark (cfg=5.0, $n_{\min}$=0). Means are recomputed from the released pair-level benchmark records.

Edit Type	N	CLAP↑	ΔCLAP
Sound Removal	22	0.312	+0.149
Subject Replace	31	0.384	+0.146
Other Replace	18	0.294	+0.110
Action Replace	7	0.369	+0.099
Weather Change	22	0.333	+0.100
Event Replace	16	0.390	+0.094
Sound Addition	18	0.350	+0.080
Intensity Change	6	0.419	+0.048
Material Change	6	0.389	+0.067
Spatial Change	6	0.332	+0.069
Vehicle Replace	12	0.342	+0.056
Atmosphere Change	12	0.350	+0.046
Environment Add	14	0.338	+0.041
Temporal Change	10	0.389	+0.037

A.4 Baseline Models

Resonate (Li et al., 2026): A flow-matching Transformer operating on mel-spectrogram VAE latents; the published method also includes Flow-GRPO alignment. Its published pre-training corpus combines AudioCaps, AudioSet, Clotho, VGGSound, WavCaps, and AudioStock. Our released evaluation script loads the locally named pre-training checkpoint `Resonate_PT.pth` with CFG 4.5, 25 Euler steps, seed 42, and 44.1 kHz sampling.

TangoFlux (Hung et al., 2025): A rectified-flow model using the 44.1 kHz stereo VAE from Stable Audio Open (Evans et al., 2024). Its CLAP-ranked preference optimization (CRPO) iteratively constructs preference pairs and optimizes the model. We use the `declare-lab/TangoFlux` checkpoint with 50 inference steps and CFG scale 4.5.

Base Model: Our controlled ablation sharing the identical DiT backbone (28 layers, 1152 dim, 610M params), same conditioning encoders (CLAP + Flan-T5), same training data and schedule, but operating directly in the DAC acoustic space without the Semantic Planner stage.

B Editing Benchmark Construction

We construct a comprehensive editing benchmark from the AudioCaps test split through the following procedure:

1. **Edit Generation:** For each source clip, we create minimal textual edits by replacing a single semantic element in the caption (e.g., "a *man* speaking" → "a *woman* speaking"). This yields over 500 candidate edit pairs spanning 14 categories.
2. **Semantic Difficulty:** We retain pairs where the source audio has low similarity to the target ($\text{CLAP}_{\text{orig}\to\text{tar}} < 0.42$), ensuring the editing task is non-trivial.
3. **Category Coverage:** The final benchmark contains 200 pairs across 14 edit categories.

Edit Type Distribution. Table 4 provides the per-category breakdown and performance.

Across the 200 pairs, the largest mean improvements are for *sound removal* (+0.149) and *subject replacement* (+0.146).

C INDEPENDENT AUDIO-LLM EVALUATION

Global metrics such as CLAP, FD, and IS summarize semantic similarity but cannot reliably distinguish event *ordering* (e.g., "A followed by B" vs. "B followed by A") or verify that an edit preserves non-target content. To provide a structure-sensitive check, we additionally evaluate with an independent audio-understanding model, Gemini-2.5, used as a judge.

Protocol. The judge is a *different model family* from the PE-A-Frame encoder used inside SemanticAudio. It receives only the clip audio (and, for editing, the source and edited clips) together with the text prompt; clips are unlabeled by method. To provide an audit trail, the judge first transcribes the sound events it hears with onset/offset timestamps, and only then scores. Each clip is judged three times and averaged (temperature 0–0.2, strict-JSON output). Full prompts and all per-clip outputs are released with our code.

Metrics. For generation we report event recall/precision, *Temporal-Order Match* (0–1: does the audio realize the event order stated in the prompt), *Temporal-Order Accuracy* (TOA: fraction of clips with fully correct order), and a holistic *Alignment* score (1–5). For editing we report *edit-success*, *non-target preservation*, *temporal-structure preservation*, and *naturalness* (each 1–5). We evaluate 30 prompts per model (15 AudioCaps + 15 TTA-Bench) for generation and 19 editing pairs.

Table 5: Independent judge: event ordering and alignment (30 prompts/model). Ground truth is scored on a subset containing reference audio and shown only as a sanity-check reference.

Model	TOM	TOA	Rec.	Prec.	Align
SemanticAudio	**0.86**	**0.50**	0.83	0.95	**4.09**
TangoFlux	0.76	0.40	0.87	0.97	4.04
Resonate	0.74	0.37	0.72	0.91	3.44
Base	0.64	0.23	0.70	0.88	3.41
GT (ref.)	0.97	0.63	0.98	0.99	4.83

Table 6: Independent judge: editing success vs. preservation for SemanticAudio (19 pairs, 1–5).

Axis	Score
Edit success	4.37
Non-target preservation	4.70
Temporal-structure preservation	4.84
Naturalness	4.40

Findings. SemanticAudio attains the highest temporal ordering (TOM 0.86, TOA 0.50) and alignment (4.09) among the four generative systems in this 30-prompt evaluation, compared with the Base Model's TOM 0.64 and TOA 0.23 (Table 5). TangoFlux has higher event recall (0.87 versus 0.83) but lower ordering (0.76 versus 0.86) on these prompts. Ground-truth audio scores highest under the same judge (TOM 0.97), a limited sanity check on the rubric rather than proof of judge validity. For editing (Table 6), 19 edits receive mean success 4.37/5, non-target preservation 4.70/5, and temporal-structure preservation 4.84/5. These small-sample judge scores are descriptive.

Failure cases. The evaluation also surfaces where the model struggles: (i) dense multi-event AudioCaps prompts, where recall and ordering both drop; (ii) explicit order-reversal prompts, the hardest ordering case; (iii) overlapping/simultaneous events, which the judge often cannot temporally separate; and (iv) large semantic edits, where non-target preservation decreases as the requested change grows. These indicate that improving multi-event and long-range temporal consistency is the most promising direction for future work.

D ACOUSTIC RECONSTRUCTION QUALITY

Table 7: AudioCaps reconstruction losses evaluated at 48 kHz (lower is better). Our rows compare with the input after DAC encode–decode (957 clips); SemanticVocoder (Xie et al., 2026) compares with the original waveform (847 valid clips). The reference targets differ.

Config	Mel Loss ↓	STFT Loss ↓
Ours d=1024 (rawPE)	**1.202**	**1.670**
Ours d=1024 (DDT)	1.256	1.695
Ours (d=128)	1.413	1.828
Ours (d=128, REPA)	1.413	1.820
Ours (d=64)	1.595	1.992
Ours (d=32)	1.817	2.270
SemanticVocoder (reference)	1.927	2.886

Table 7 evaluates the **Acoustic Synthesizer** conditioned on ground-truth semantic features. For our rows, the target is the input clip reconstructed by the frozen 48 kHz DAC-VAE; the loss therefore isolates error beyond the DAC encode–decode stage. The $d = 1024$ rawPE variant has the lowest losses among our variants (Mel 1.202, STFT 1.670); DDT has Mel 1.256 and STFT 1.695. The primary reconstruction checkpoints are at 500k steps for rawPE, DDT, and $d = 32$; 290k for $d = 128$; and 415k for $d = 64$. Losses rise as the dimension decreases from $d = 128$ to $d = 32$, showing a reconstruction cost for stronger compression. SemanticVocoder uses a native 24 kHz waveform generator (Xie et al., 2026); in our independent evaluation its output is resampled to 48 kHz for loss computation, and its target is the original audio, not a DAC reconstruction. Its 847 valid clips also differ from our 957. The SemanticVocoder row (Mel 1.927, STFT 2.886) is provided only as context and should not be used for a direct ranking of the acoustic synthesizers. These numbers are not taken from SemanticVocoder's published 24 kHz table.

1024-d Variants: rawPE vs. DDT. We compare unprojected PE features (*rawPE*) with a learned linear projection (*DDT*). rawPE gives lower reconstruction Mel loss (1.202 versus 1.256; Table 7) but worse end-to-end AudioCaps generation CLAP (0.320 versus 0.373; Table 1). Both variants have trainable Semantic Planners; the results show a reconstruction–generation tradeoff, not that raw PE lacks a Flow Matching target.