SWIFT: SEQUENCE-AWARE WASSERSTEIN INFORMED FLOW TRANSPORT FOR CONDITIONAL AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing methods for coupling audio sequences with noise sequences in flow matching use a frame-level Euclidean ground metric that treats each sequence as a flat vector, entirely ignoring both the temporal structure of acoustic events and the semantic content of the conditioning signal. We propose **SWIFT** (**S**equence-aware **W**asserstein **I**nformed **F**low **T**ransport), a flow matching training framework for audio generation built on a novel ground metric called the Sliced Wasserstein Sequence Distance (SWSD). SWSD lifts each audio and noise sequence to a probability measure on its time-augmented path embedding, encoding temporal ordering explicitly into the geometry of the ground metric, and computes sliced Wasserstein distances using a semantic-aware slicing measure that concentrates projections along the acoustic subspace indicated by the text or visual conditioning vector. We validate SWIFT on two audio generation tasks: text-to-audio and video-to-audio. SWIFT outperforms baselines across evaluation metrics and inference budgets, with the largest improvements at low numbers of function evaluations, consistent with the design motivation that a semantically informed coupling produces straighter flow paths consistent with the test-time prior. We further demonstrate that a model pretrained with SWIFT provides a stronger initialisation for the reflow procedure, amplifying the benefit of trajectory straightening and achieving lower Fréchet Audio Distance at a fixed inference budget.

1 INTRODUCTION

Flow matching (Lipman et al., 2022) has rapidly emerged as the dominant paradigm for audio generation, underpinning state-of-the-art systems for text-to-speech (Mehta et al., 2024; Chen et al., 2025), text-to-audio (Hung et al., 2026; Li et al., 2025), video-to-audio (Wang et al., 2024; Cheng et al., 2025), and music generation (Prajwal et al., 2024). By learning a vector field that transports Gaussian noise to audio features, flow matching models offer both high generation quality and controllable inference cost through the number of function evaluations (NFE). Crucially, the *data–noise coupling* π determines the curvature of the learned probability paths (Liu et al., 2023b; Pooladian et al., 2023): under the independent coupling $\pi = q_1 \otimes q_0$, trajectories from different source–target pairs cross freely in the interior of the integration interval, forcing the vector field to average over conflicting directions and producing curved flows that require many NFE to integrate accurately. Replacing the independent coupling with an optimal transport (OT) plan significantly straightens the learned trajectories, reducing the required sampling steps by 30–60% on image benchmarks at fixed generation quality (Pooladian et al., 2023; Tong et al., 2024).

Motivated by these results, recent works have sought better couplings through improved minibatch OT solvers (Tong et al., 2024; Davtyan et al., 2026) and offline coupling methods (Liu et al., 2023b; 2024). However, all existing coupling methods compute the ground cost between source and target sequences with a *frame-level Euclidean distance* that treats each sequence as a flat vector, which is doubly unsuitable for conditional audio generation. First, audio feature sequences are temporally ordered: the same acoustic event at different time steps incurs a large Euclidean distance even when the sequences are perceptually near-identical, so couplings confuse temporal displacement with genuine acoustic dissimilarity. Second, the Euclidean cost entirely ignores the semantic conditioning vector z^i—the text or visual embedding that specifies what the audio should sound like. Indeed,

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.

Cheng & Schwing (2025) show that standard minibatch OT with a Euclidean cost can produce *worse* couplings than independent sampling on conditional generation tasks, yielding curved flows and degraded quality even with an OT solver in the loop. Euclidean OT coupling is therefore ill-suited to training flow matching models for text- or video-conditioned audio generation.

In this work, we address both limitations with the **S**equence-aware **W**asserstein **I**nformed **F**low **T**ransport (SWIFT) training method, built on our novel ground metric—the Sliced Wasserstein Sequence Distance (SWSD)—for audio–noise coupling in flow matching. SWSD has two complementary components. First, each audio feature sequence is lifted to a probability measure on its *time-augmented path embedding* (Liao et al., 2024; Chevyrev & Oberhauser, 2022), which prepends a normalised time coordinate to each frame and thereby encodes the temporal ordering of acoustic events directly into the geometry of the ground metric. Second, the sliced Wasserstein projections are steered toward the semantic content of z^i via a von Mises–Fisher slicing distribution, concentrating comparison in the acoustic subspace most relevant to the text or visual prompt. The result is a true metric on fixed-length sequences that is evaluated efficiently by Monte Carlo projection and 1-D quantile comparison, retaining the computational simplicity of minibatch OT while being sensitive to both temporal structure and semantic conditioning.

Our main contributions are as follows:

(i) We propose the SWIFT training method for flow matching audio generation, which uses the SWSD ground metric—a time-augmented path embedding that encodes temporal ordering combined with a condition-directed slicing distribution that incorporates semantic conditioning—to measure the discrepancy between sequences of audio latents and sequences of noise vectors. Experiments show that SWSD genuinely measures this discrepancy, leading to straighter flow paths consistent with the test-time prior.

(ii) We validate SWIFT on text-to-audio (AudioCaps (Kim et al., 2019), MeanAudio (Li et al., 2025) backbone) and video-to-audio (VGGSound (Chen et al., 2020), Frieren (Wang et al., 2024) backbone), where it consistently outperforms MinibatchOT (Tong et al., 2024) and LOOM-CFM (Davtyan et al., 2026) across metrics and inference budgets.

(iii) We show that a model pretrained with SWIFT is a stronger initialisation for reflow (Liu et al., 2023b), amplifying the benefit of trajectory straightening and achieving lower FAD at a fixed inference budget.

2 Background of Flow Matching

Flow matching generative models. Given a source distribution $q_0(X_0)$ and a target data distribution $q_1(X_1)$, a flow matching model learns a time-dependent vector field $v_t(x;\theta)$ by regressing onto a target field $u_t(x)$ that generates a probability path $p_t(x)$ interpolating between q_0 and q_1:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x\sim p_t(x)} \big\| v_t(x;\theta) - u_t(x) \big\|^2, \tag{1}$$

subject to $p_0 = q_0$ and $p_1 = q_1$. This objective is intractable because p_t and u_t are unknown in closed form. Lipman et al. (2022) showed that $\mathcal{L}_{\mathrm{FM}}$ has the same gradient as the tractable *conditional flow matching* (CFM) objective, which conditions the probability path on individual data samples $x_1 \sim q_1$:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\, x_1\sim q_1,\, x\sim p_t(x|x_1)} \big\| v_t(x;\theta) - u_t(x \mid x_1) \big\|^2, \tag{2}$$

where $p_t(x \mid x_1)$ satisfies $p_0(x \mid x_1) = q_0(x)$ and $p_1(x \mid x_1) = q_1(x)$. The standard choice is the *conditional optimal transport (OT) path* (Lipman et al., 2022), a linear displacement interpolant between q_0 and q_1 whose conditional vector field and sample are

$$u_t(x \mid x_1) = x_1 - x_0, \qquad x_t = (1-t)\, x_0 + t\, x_1, \tag{3}$$

with $x_0 \sim q_0$. Substituting Eq. equation 3 into Eq. equation 2 and making the dependence on the data–noise pair explicit yields

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\, x_1\sim q_1,\, (x_1,x_0)\sim\pi} \Big\| v_t\big((1-t)\, x_0 + t\, x_1;\, \theta\big) - (x_1 - x_0) \Big\|^2, \tag{4}$$

where $\pi \in \Pi(q_0, q_1)$ is a coupling whose marginals are q_0 and q_1.

Data–noise coupling. The coupling π in Eq. equation 4 critically affects training efficiency. Under the *independent coupling* $\pi = q_1 \otimes q_0$, trajectories from different source–target pairs cross freely in the interior of the time interval, so $p_t(x)$ is a broad mixture and the learned vector field must average over conflicting directions (Liu et al., 2023b; Pooladian et al., 2023); the resulting flows are curved and require many function evaluations at inference. A principled remedy is to replace the independent coupling with the *optimal transport plan* between q_0 and q_1: if π^* minimises the expected transport cost

$$\pi^* = \underset{\pi \in \Pi(q_0, q_1)}{\arg\min} \; \mathbb{E}_{(x_1, x_0)\sim\pi}\big[c(x_1, x_0)\big], \tag{5}$$

then by Theorem D.5 of Liu et al. (2023b), training under π^* yields a lower expected transport cost than the independent coupling for any convex cost c, straightening probability paths and enabling faster inference. Since the exact OT plan between high-dimensional continuous distributions is intractable, a *minibatch* approximation is used: for B data and B noise samples, the batch-level plan is computed with an off-the-shelf solver such as the Sinkhorn algorithm (Cuturi, 2013), with the ℓ_2 distance as the ground cost c.

3 Proposed Method

3.1 Problem Setting

We consider audio generation via flow matching. Let $X_1 = (x_1^1, \ldots, x_1^T) \in \mathbb{R}^{d\times T}$ denote a sequence of d-dimensional audio feature vectors from a fixed audio encoder, and $X_0 = (x_0^1, \ldots, x_0^T) \in \mathbb{R}^{d\times T}$ a sequence of noise vectors with frames $x_0^k \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$. Both sequences are temporally ordered, with marginals $q_1(X_1)$ and $q_0(X_0)$. Each audio sequence X_1^i is associated with a semantic conditioning vector $z^i \in \mathbb{R}^{d_z}$, a text embedding from a fixed language model encoder or a visual embedding from a vision encoder. The goal is to learn a vector field $v_t(X_t; \theta)$ that defines a conditional probability path $p_t(X_t \mid X_1)$ transporting q_0 to q_1, as in Eq. equation 4.

As discussed in Section 2, an informative data–noise coupling π equation 5 straightens the probability paths learned by v_t, reducing trajectory crossings and enabling faster inference. The natural ground cost is the *frame-level Euclidean cost* $c(X_1^i, X_0^j) = \sum_{k=1}^T \big\|x_1^{i,k} - x_0^{j,k}\big\|^2$, or its mean-frame variant $c(X_1^i, X_0^j) = \big\|\frac{1}{T}\sum_{k=1}^T x_1^{i,k} - \frac{1}{T}\sum_{k=1}^T x_0^{j,k}\big\|^2$. Both are inadequate for audio–noise coupling. First, both are insensitive to the temporal ordering of acoustic events: the frame-level cost compares frames at fixed time indices, so two sequences identical in content but shifted by one time step incur a large distance despite being perceptually near-identical, while the mean-frame cost collapses each sequence to a single mean vector, discarding temporal structure entirely and treating two time-shifted sequences as equidistant from a noise sequence as two acoustically unrelated ones. Second, neither cost incorporates z^i. Since z^i encodes the linguistic or acoustic content that X_1^i is supposed to realise, ignoring it may couple a noise sequence to an audio sequence that is close in raw feature space but semantically misaligned with the intended target, producing curved probability paths and degraded generation quality.

We therefore replace the Euclidean cost with the proposed **Sliced Wasserstein Sequence Distance** (**SWSD**), which addresses both limitations. (1) **Temporal robustness:** rather than comparing frames at fixed time indices, SWSD lifts each sequence to a probability measure on its time-augmented path embedding (Liao et al., 2024; Chevyrev & Oberhauser, 2022) and computes a sliced Wasserstein distance between the resulting measures, which optimally transports mass between them and is thus more robust to temporal displacement than the frame-level Euclidean distance (Thorpe et al., 2017; Janati et al., 2020). (2) **Semantic-aware projection:** SWSD incorporates z^i directly into the projection geometry by concentrating slicing directions, via a slicing measure, along the acoustic subspace indicated by the conditioning vector.

3.2 Sliced Wasserstein Sequence Distance

Time-Augmented Embedding. An audio encoder $\mathcal{E} : \mathcal{A} \to \mathbb{R}^d$ maps raw audio frames to d-dimensional latent vectors, so an audio signal of fixed duration yields a temporally ordered sequence of T latent frames $\mathbf{X} = (x^1, \ldots, x^T) \in \mathbb{R}^{d\times T}$ with $x^k = \mathcal{E}(\mathcal{A}_k)$. To encode temporal position into the representation, we apply the *time augmentation* transformation (Liao et al., 2024; Chevyrev &

Oberhauser, 2022), which prepends a normalised time coordinate to each frame, lifting the sequence from $\mathbb{R}^d$ to a path in $\mathbb{R}^{d+1}$. The **time-augmented embedding** of $\mathbf{X}$ is the map $\hat{\mathbf{X}} : \{1, \dots, T\} \to \mathbb{R}^{d+1}$ defined by

$$\hat{x}^k = \left(\frac{k-1}{T-1},\ x^k \right) \in \mathbb{R}^{d+1}, \qquad 1 \le k \le T, \tag{6}$$

with associated empirical measure $\mu_{\hat{\mathbf{X}}} = \frac{1}{T} \sum_{k=1}^{T} \delta_{\hat{x}^k} \in \mathcal{P}(\mathbb{R}^{d+1})$. The first coordinate $t_k = (k-1)/(T-1) \in [0,1]$ encodes normalised temporal position; the remaining d coordinates encode the audio latent content at frame k. For sequences of fixed length $T \ge 2$, $\mu_{\hat{\mathbf{X}}_1} = \mu_{\hat{\mathbf{X}}'_1}$ if and only if $\hat{X}_1 = \hat{X}'_1$: the atoms $\{\hat{x}_1^k\}_{k=1}^T$ have strictly distinct time coordinates t_k, so equality of two T-atom empirical measures forces a bijection between atom sets that respects the t_k ordering, and since both sequences share the same time coordinates this implies $x_1^k = x_1'^k$ for all k.

Condition-Directed Projection and Slicing Distribution. The semantic conditioning vector $z^i \in \mathbb{R}^{d_z}$ must be mapped into the time-augmented embedding space $\mathbb{R}^{d+1}$ before it can steer projection directions. Given a mapping $\phi : \mathbb{R}^{d_z} \to \mathbb{R}^d$ (specified below), the condition direction is

$$\tilde{z}^i = \frac{(0,\ \phi(z^i))}{\|(0,\ \phi(z^i))\|_2} \in \mathbb{S}^d. \tag{7}$$

The zero time-component confines the condition direction to the d amplitude dimensions, so steering projections toward $\tilde{z}^i$ sharpens comparison along the latent axes relevant to the condition without distorting the temporal structure of the embedding.

Random projection. A suitable choice for ϕ is the fixed random projection $\phi(z^i) = Pz^i$, where $P \in \mathbb{R}^{d \times d_z}$ has near-orthonormal rows from the QR decomposition of a random Gaussian matrix, spanning a uniformly random d-dimensional subspace of $\mathbb{R}^{d_z}$; P is sampled once before training and frozen, so $\tilde{z}^i$ is a deterministic function of z^i across minibatches. This map assumes nothing about how semantic information is distributed across the coordinates of the conditioning space; its guarantee is distributional: random orthogonal projections preserve inner products and angles up to a distortion decaying as $\mathcal{O}(1/\sqrt{d})$, so conditioning vectors that are well separated in $\mathbb{R}^{d_z}$ remain well separated after projection with high probability. Since Eq. equation 7 retains only the direction of $\phi(z^i)$, this preservation of angular structure is exactly what σ_κ^i requires to concentrate slicing directions along a subspace faithful to the conditioning signal.

From the normalised condition direction $\tilde{z}^i$, we define the *semantic-aware slicing distribution*

$$\sigma_\kappa^i(\cdot \mid \tilde{z}^i) = \mathrm{vMF}(\tilde{z}^i,\ \kappa), \tag{8}$$

the von Mises–Fisher distribution on $\mathbb{S}^d$ with mean direction $\tilde{z}^i$ and concentration $0 < \kappa < \infty$. As $\kappa \to \infty$ it collapses to a Dirac at $\tilde{z}^i$, reducing the distance to a single-direction projection that is blind to acoustic variation orthogonal to the condition axis; a finite positive κ balances the smoothness essential for distance comparison against the information from $\tilde{z}^i$.

Definition 1 (Sliced Wasserstein Sequence Distance) *For concentration $\kappa > 0$ and order $p \ge 1$, the Sliced Wasserstein Sequence Distance (SWSD) between audio embedding sequences $\hat{X}_1$ and $\hat{X}_0$ is*

$$\mathrm{SWSD}_p^{\tilde{z}^i}(\hat{X}_1, \hat{X}_0; \kappa) = \left(\mathbb{E}_{\psi \sim \sigma_\kappa^i(\psi \mid \tilde{z}^i)} \left[W_p^p \left(\psi_\sharp\, \mu_{\hat{X}_1},\ \psi_\sharp\, \mu_{\hat{X}_0} \right) \right] \right)^{1/p}, \tag{9}$$

where $\psi_\sharp\, \mu$ denotes the pushforward of measure μ under the scalar projection $\langle \psi, \cdot \rangle$, and the one-dimensional p-Wasserstein distance admits the closed form

$$W_p^p(\psi_\sharp\, \mu_{\hat{X}_1},\ \psi_\sharp\, \mu_{\hat{X}_0}) = \int_0^1 \left| F^{-1}_{\psi_\sharp \mu_{\hat{X}_1}}(u) - F^{-1}_{\psi_\sharp \mu_{\hat{X}_0}}(u) \right|^p du, \tag{10}$$

where F^{-1} denotes the quantile function of the respective projected measure.

SWSD is essentially a distributional sliced Wasserstein distance between $\mu_{\hat{X}_1}$ and $\mu_{\hat{X}_0}$ (Nguyen et al., 2021; Nguyen & Ho, 2024). In practice, Eq. equation 9 is approximated by drawing L directions $\psi_1, \dots, \psi_L \overset{\text{iid}}{\sim} \sigma_\kappa^i$, projecting and sorting both sequences, and averaging the one-dimensional Wasserstein distances.

Remark 1 (Metricity) *For any conditioning vector $z^i \in \mathbb{R}^{d_z}$, finite κ, and $p \geq 1$, $\mathrm{SWSD}_p^{\tilde{z}^i}$ is a metric on the space of fixed-length audio latent sequences: it is non-negative, symmetric, satisfies the triangle inequality, and $\mathrm{SWSD}_p^{\tilde{z}^i}(\hat{X}_1, \hat{X}_0; \kappa) = 0 \iff \hat{X}_1 = \hat{X}_0$. Full support of σ_κ^i and injectivity of the time-augmented embedding imply that distinct sequences have distinct projected measures on some direction in the support of σ_κ^i, so $\mathrm{SWSD}_p^{\tilde{z}^i}(\hat{X}_1, \hat{X}_0; \kappa) > 0$ whenever $\hat{X}_1 \neq \hat{X}_0$. Symmetry follows from the symmetry of W_p, and the triangle inequality from Minkowski's integral inequality; see Nguyen (2025) for a detailed discussion.*

3.3 The Sequence-aware Wasserstein Informed Flow Transport method

General formulation. Let $q_1(X_1, z)$ denote the joint distribution of audio embedding sequences and their conditioning vectors, and $q_0(X_0)$ the distribution of Gaussian noise sequences. The optimal coupling between q_1 and q_0 is defined as the solution to the regularised optimal transport problem

$$\pi^* = \underset{\pi \in \Pi(q_1, q_0)}{\arg\min} \; \mathbb{E}_{(X_1, X_0) \sim \pi} \left[\mathrm{SWSD}_p^{\tilde{z}^i}(\hat{X}_1,\, \hat{X}_0; \kappa) \right], \tag{11}$$

where $\Pi(q_1, q_0)$ is the set of all joint distributions with marginals q_1 and q_0, and the ground cost is the semantic-aware SWSD (Definition 1) between the time-augmented embeddings of the audio and noise sequences. It measures the distributional discrepancy of each audio–noise pair in the joint time-feature space, with the projection geometry of each audio sequence steered by its conditioning direction $\tilde{z}^i$, so that sequences semantically aligned with z^i are recognised as proximate; under π^*, each noise sequence is thus paired with the most distributionally and semantically compatible audio sequence, reducing trajectory crossings and producing straighter interpolation paths.

Minibatch approximation. Computing the coupling over the full data distribution is intractable, so we approximate π^* within each minibatch of B audio–condition pairs $\{(X_1^i, z^i)\}_{i=1}^B$ and B noise sequences $\{X_0^j\}_{j=1}^B$. We first construct the $B \times B$ ground cost matrix $\mathbf{C}$ of SWSDs between audio and noise sequences,

$$\mathbf{C}_{ij} = \mathrm{SWSD}_p^{\tilde{z}^i}(\hat{X}_1^i,\, \hat{X}_0^j), \tag{12}$$

where each row i uses L projection directions drawn from the semantic-aware slicing measure σ_κ^i (Eq. equation 9). The minibatch coupling is then obtained by solving the entropy-regularised OT problem over $\mathbf{C}$,

$$\pi^* = \underset{\pi \in \Pi(\mathbf{1}/B,\, \mathbf{1}/B)}{\arg\min} \; \langle \pi,\, \mathbf{C} \rangle_F + \lambda \, \mathrm{KL}\Big(\pi \,\Big\|\, \tfrac{1}{B^2} \mathbf{1}\mathbf{1}^\top \Big), \tag{13}$$

where $\Pi(\mathbf{1}/B, \mathbf{1}/B)$ is the set of doubly stochastic matrices with uniform marginals and $\lambda > 0$ controls the entropic regularisation strength; Eq. equation 13 is solved efficiently via the Sinkhorn algorithm (Cuturi, 2013). The full coupling procedure is given in Algorithm 1 in the appendix.

Flow matching objective. Each coupled pair (X_1^i, X_0^j) sampled from π^* defines the straight-line interpolant and regression target for the flow matching objective in Eq. equation 4:

$$X_t = (1 - t)\, X_0^j + t\, X_1^i, \qquad u_t(X_t \mid X_1^i) = X_1^i - X_0^j. \tag{14}$$

Because noise sequences are paired with audio sequences that are distributionally proximate in the joint time-feature space and semantically aligned via $\tilde{z}^i$, the resulting interpolation paths are straighter than those induced by an isotropic Euclidean ground cost, reducing the curvature of the learned vector field $v_t(X_t; \theta)$ and enabling faster inference with fewer function evaluations.

4 Experiments

4.1 Experimental Setup

Datasets. For *text-to-audio generation*, we use AudioCaps (Kim et al., 2019), approximately $50{,}000$ ten-second audio clips sourced from AudioSet (Gemmeke et al., 2017), and evaluate on the standard test split of 957 clips; since each clip has five captions, one is selected uniformly at random as the text condition at inference time. For *video-to-audio generation*, we use VGGSound (Chen et al., 2020),

approximately 200,000 ten-second YouTube clips; following the train/test splits of Wang et al. (2024), which account for clips no longer publicly available, the training and test sets contain approximately 182,600 and 15,300 clips.

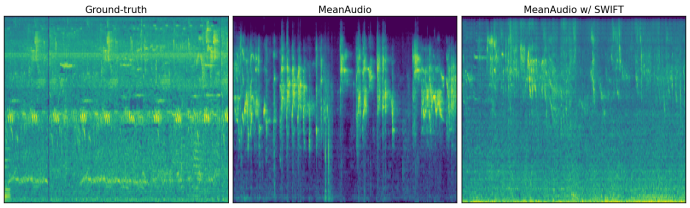
Evaluation Metrics. We report Fréchet Audio Distance (FAD), Kullback–Leibler Divergence (KL), and Inception Score (IS) for both tasks. For text-to-audio generation, we additionally report the CLAP score (Wu et al., 2023), the cosine similarity between the text condition and the generated audio in a shared audio–language embedding space, which is only meaningful when captions are available. We also evaluate the path straightness of pretrained models as defined by Pooladian et al. (2023). All results are computed five times and averaged.

Baselines. We compare against four coupling methods applied to the same backbone models. MinibatchOT (Tong et al., 2024) and LOOM-CFM (Davtyan et al., 2026) both compute an optimal transport plan within each training batch using the Euclidean ground cost. C2OT (Cheng & Schwing, 2025) accounts for the semantic condition when coupling noise and data by adding a condition-dependent weighting term to the cost matrix. SWIFT operates in the same online minibatch setting as MinibatchOT and LOOM-CFM, but replaces the Euclidean ground cost with the semantic-aware SWSD of Section 3.2. Implementation details for both the flow and reflow training stages are given in Appendix A.

4.2 Main results

Table 1: Data–noise coupling methods on text-to-audio and video-to-audio generation at two inference budgets (NFE). **Bold**: best; underline: second best; ↓/↑: lower/higher is better.

Task	NFE	Method	FAD ↓	KL ↓	IS ↑	CLAP ↑
Text-to-Audio	5	MeanAudio w/ independent coupling	19.25	1.48	6.85	0.283
		w/ MinibatchOT coupling	18.94	1.57	6.32	0.259
		w/ LOOM-CFM coupling	18.60	1.52	6.66	0.263
		w/ C2OT coupling	19.05	1.46	6.65	0.274
		w/ SWIFT (Ours)	**17.71**	**1.41**	**7.22**	**0.286**
	25	MeanAudio w/ independent coupling	18.26	1.46	6.98	0.299
		w/ MinibatchOT coupling	18.11	1.55	6.76	0.271
		w/ LOOM-CFM coupling	17.86	1.51	6.80	0.278
		w/ C2OT coupling	18.59	1.47	6.73	0.288
		w/ SWIFT (Ours)	**17.29**	**1.45**	**7.34**	**0.301**
Video-to-Audio	5	Frieren w/ independent coupling	24.16	3.14	8.03	N/A
		w/ MinibatchOT coupling	24.26	3.47	6.29	N/A
		w/ LOOM-CFM coupling	23.84	2.96	7.83	N/A
		w/ C2OT coupling	24.07	3.12	7.92	N/A
		w/ SWIFT (Ours)	**23.39**	**2.93**	**8.90**	N/A
	25	Frieren w/ independent coupling	14.21	2.89	11.51	N/A
		w/ MinibatchOT coupling	17.19	3.74	8.54	N/A
		w/ LOOM-CFM coupling	**13.88**	2.83	9.29	N/A
		w/ C2OT coupling	14.18	2.88	11.54	N/A
		w/ SWIFT (Ours)	13.97	**2.82**	**12.08**	N/A



Prompt: "A group of birds call and chirp for a while."

Figure 1: Mel-spectrogram comparison between the ground truth, MeanAudio, and MeanAudio trained with SWIFT at NFE $= 5$.

Table 2: Straightness and FAD at very low inference budgets on text-to-audio generation. Straightness is reported as mean (std) over the AudioCaps test set. **Bold**: best result. ↓ lower is better.

Method	NFE = 4		NFE = 2	
	Straightness ↓	FAD ↓	Straightness ↓	FAD ↓
MeanAudio	8.60(±3.34)	<u>22.52</u>	9.87(±3.90)	76.87
MinibatchOT	**6.06**(±2.67)	24.56	<u>6.34</u>(±3.11)	74.56
LOOM-CFM	<u>6.13</u>(±2.42)	22.77	**6.10**(±2.13)	<u>73.34</u>
C2OT	7.05(±3.11)	22.38	10.40(±5.04)	74.08
Ours	6.48(±2.23)	**19.75**	8.23(±3.33)	**68.71**

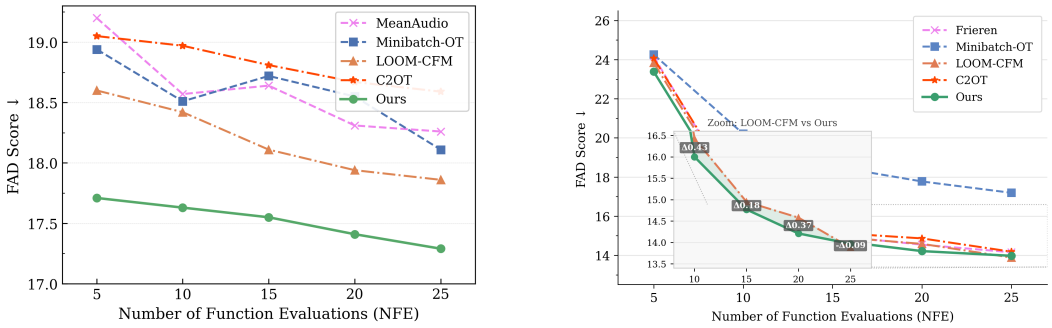
Overall performance for audio generation tasks. Table 1 summarises results across both tasks and inference budgets. Our method is best on the majority of metrics in every setting, the single exception being FAD on video-to-audio at NFE = 25, where it ranks second behind LOOM-CFM by 0.09. On text-to-audio it outperforms all baselines on every metric at both budgets; at NFE = 5 it reduces FAD from 19.25 to 17.71, improves IS from 6.85 to 7.22, and raises CLAP from 0.283 to 0.286. The more pronounced gains at low NFE align with the design motivation of SWSD (see the NFE analysis below). On video-to-audio, our method achieves the best IS at both budgets (8.03 → 8.90 at NFE = 5 and 11.51 → 12.08 at NFE = 25), indicating that the condition-directed projection geometry measures the data–noise distance more precisely in the latent space and thereby straightens the probability path between the two distributions. Two comparisons isolate the contribution of the SWSD ground metric. Against LOOM-CFM, which uses the same minibatch OT solver, SWIFT is better on 12 of 14 metric–setting combinations, attributing the gain to the ground cost rather than the coupling procedure; against C2OT, which also injects conditioning into the OT problem, it is better on all 14, showing that the improvement comes from the specific geometry of condition-directed slicing rather than from conditioning the transport problem. By jointly encoding temporal ordering, distributional structure, and semantic conditioning via $\tilde{z}^i$, SWIFT yields acoustically and semantically compatible couplings that straighten the learned flow paths and improve quality across both tasks and budgets. Figure 1 shows a qualitative example at NFE = 5; more are in Appendix F.

Path straightness evaluation. Table 2 reports trajectory straightness and FAD at NFE = 4 and NFE = 2. Our method achieves the best FAD in both settings (22.52 → 19.75 at NFE = 4 and 76.87 → 68.71 at NFE = 2) while ranking third on mean straightness. Straightness and quality are in fact anti-correlated here: MinibatchOT is the straightest at NFE = 4 (6.06) yet has the worst FAD (24.56), while C2OT is the least straight at NFE = 2 (10.40) yet beats it on FAD. This is the train–test gap of Cheng & Schwing (2025): a coupling that ignores the condition skews the training prior $P(x_0 \mid z)$, producing paths that are geometrically short but misdirected once noise is drawn independently at test time. Encoding $\tilde{z}^i$ in the ground cost keeps the training and test-time priors aligned, so SWIFT still straightens paths relative to independent coupling, dominates C2OT on both axes, and attains the lowest straightness variance at NFE = 4—uniformly well-behaved trajectories rather than a mix of very straight and misaligned ones. The operative quantity is thus straightness consistent with the test-time prior, not straightness alone.

NFE analysis. Figure 2 plots FAD against NFE for both tasks. On text-to-audio, our method achieves the lowest FAD at every NFE from 5 to 25, with a clear margin over MinibatchOT, LOOM-CFM, and C2OT throughout. On video-to-audio, it also outperforms LOOM-CFM over most of the range, improving FAD by 0.44, 0.18, and 0.37 at NFE = 10, 15, and 20 (see the zoomed inset); at NFE = 25, LOOM-CFM is better by 0.09, the only exception to this trend. The gap is largest at low NFE and narrows as NFE increases, consistent with the design rationale of SWSD: a semantically informed coupling straightens the flow trajectories, an advantage most pronounced when a coarse solver's few integration steps would otherwise accumulate substantial error.

4.3 SWIFT WITH REFLOW

The main results show that SWIFT improves models trained from scratch by straightening the data–noise paths within each minibatch. We further hypothesise that a model pretrained with SWIFT generates high-quality data–noise pairs that serve as a superior initialisation for reflow, yielding additional improvements. To validate this, we pretrain Frieren on the VGGSound training set



(a) Text-to-audio generation.

(b) Video-to-audio generation.

Figure 2: FAD versus NFE across coupling methods on both tasks (lower is better).

Table 3: Combining SWIFT with reflow on video-to-audio generation (VGGSound test set). Δ / Δ: gain / degradation relative to the row above.

	Method	FAD↓	KL↓	IS↑	CLAP↑
NFE=5	Frieren	24.16	3.14	8.03	N/A
	+ reflow	17.42$_{\Delta 6.74}$	2.76$_{\Delta 0.38}$	6.73$_{\Delta 1.3}$	N/A
	Frieren w/ SWIFT	23.39	3.01	8.90	N/A
	+ reflow	15.68$_{\Delta 7.71}$	2.67$_{\Delta 0.34}$	6.86$_{\Delta 2.04}$	N/A
NFE=10	Frieren	16.72	2.93	10.61	N/A
	+ reflow	16.29$_{\Delta 0.43}$	2.77$_{\Delta 0.16}$	7.42$_{\Delta 3.19}$	N/A
	Frieren w/ SWIFT	16	2.85	10.98	N/A
	+ reflow	14.87$_{\Delta 1.13}$	2.68$_{\Delta 0.17}$	7.63$_{\Delta 3.35}$	N/A

with SWIFT and use it to generate the data–noise pairs on which a reflowed model is trained following Liu et al. (2023b).

Table 3 examines the interaction between SWIFT and reflow on video-to-audio at NFE = 5 and 10. Before reflow, our method improves the base model (FAD 24.16 → 23.39 and IS 8.03 → 8.90 at NFE = 5; same pattern at NFE = 10), confirming that a semantically informed coupling improves single-pass quality independently of any post-hoc straightening. More importantly, SWIFT provides a stronger initialisation for reflow: at NFE = 5, SWIFT with reflow reaches FAD = 15.68 (Δ7.71 from its base) versus 17.42 (Δ6.74) for Frieren with reflow, with the same advantage at NFE = 10 (14.87 vs. 16.29). This suggests that the straighter, test-time-prior-consistent paths produced by our method give reflow a better trajectory prior, amplifying the benefit of straightening. Although IS declines after reflow in all configurations—a known diversity–fidelity trade-off (Liu et al., 2023b)—SWIFT with reflow retains higher IS than Frieren with reflow (6.86 vs. 6.73 at NFE = 5; 7.63 vs. 7.42 at NFE = 10), indicating that the condition-directed coupling partially mitigates this diversity loss.

4.4 Ablation Studies

Table 4 ablates the three core design choices of SWIFT on text-to-audio. Removing the time-augmented path embedding and computing SWSD on raw latent frames degrades FAD from 17.71 to 18.53 at NFE = 5 and from 17.29 to 17.62 at NFE = 25, confirming that explicitly encoding temporal ordering into the ground metric is essential for well-matched audio–noise couplings. Setting κ=0 reduces the slicing measure to a fully random projection that discards all conditioning information from the cost matrix; this yields the worst FAD (19.25 and 17.93 at NFE = 5 and 25) and the lowest CLAP (0.265 and 0.284), demonstrating that the semantic-aware component is critical for alignment between the generated audio and the text prompt. Setting κ=∞ collapses the slicing measure to the single direction $\tilde{z}^i$, which partially recovers FAD over κ=0 but still underperforms the full model, consistent with the argument that a single-direction projection loses sensitivity to acoustic variation orthogonal to the condition axis. The full model, combining time-augmented embeddings with the slicing measure at intermediate κ, achieves the best FAD, KL, and CLAP in both settings, so all

Table 4: Ablation study comparing our method against variants.

	Method	FAD↓	KL↓	IS↑	CLAP↑
NFE=5	Ours (full)	**17.71**	**1.41**	7.22	**0.286**
	w/o time-augmented	18.53	1.44	7.30	0.279
	κ=0, random projection	19.25	1.47	**7.53**	0.265
	κ=∞, single-direction projection	18.29	1.46	7.40	0.271
NFE=25	Ours (full)	**17.29**	**1.45**	7.34	**0.301**
	w/o time-augmented	17.62	1.46	7.15	0.291
	κ=0, random projection	17.93	1.48	**7.36**	0.284
	κ=∞, single-direction projection	17.85	1.47	7.28	0.288

three components contribute complementary, non-redundant improvements. The only exception is IS, where κ=0 is slightly higher than the full model (7.53 vs. 7.22 at NFE = 5; 7.36 vs. 7.34 at NFE = 25); we attribute this to the greater diversity of random projections encouraging broader mode coverage at the cost of reduced semantic fidelity. Ablations on batch size and κ are reported in Appendix D.

Temporal displacement evaluation. We conduct an ablation study on temporal displacement to justify the effectiveness of the time-augmented module. We use the HTSAT audio encoder (Chen et al., 2022) as the feature extractor to obtain frame-level features from the ground-truth and generated audio. We then apply Dynamic Time Warping (Sakoe & Chiba, 1978) to measure the temporal displacement between the two feature sequences, using the distance $\mathrm{d}(s_i, s_j) = 1 - \mathrm{cosine}(s_i, s_j)$. Table 5 reports the temporal displacement of our method with and without the time-augmented component across two audio generation tasks. The consistently lower temporal displacement with the time-augmented component indicates that the temporal ordering of acoustic events in the generated audio aligns more closely with the ground truth.

Task	NFE	Method	TD (HTSAT) ↓
TTA	5	w/o time-aug.	0.248
		SWIFT	**0.240**
	25	w/o time-aug.	0.242
		SWIFT	**0.233**
V2A	5	w/o time-aug.	0.162
		SWIFT	**0.154**
	25	w/o time-aug.	0.159
		SWIFT	**0.150**

Table 5: Ablation on time augmentation. Temporal displacement (TD) measured with HTSAT features (lower is better).

5 Conclusion

We presented SWIFT, a flow matching training method for conditional audio generation that couples data and noise using the Sliced Wasserstein Sequence Distance (SWSD). By combining time-augmented path embeddings with condition-directed slicing, SWSD captures both temporal structure and semantic content at minibatch-OT cost. On AudioCaps and VGGSound, SWIFT outperforms existing coupling methods on nearly all metrics and inference budgets, especially at low NFE, and improves reflow initialization. Our results suggest that few-step generation benefits from straightness consistent with the test-time prior, not straightness alone.

AI use statement

In this work, we used generative AI tools as a coding assistant when implementing the method and evaluation code. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design or give feedback on the research methodology or experiments, generate synthetic data, clean or reformat datasets, support qualitative analysis, or interpret results. Additionally, we used generative AI tools to check grammar and improve the readability of the paper text. The research ideas, the method, and the experimental design are entirely developed by the authors. We have reviewed all AI-assisted work: all AI-assisted code was verified and tested for correctness by the authors, and all AI-edited text was checked by the authors for accuracy and fidelity to the intended meaning. We take responsibility for

the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

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

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650, 2022. doi: 10.1109/ICASSP43922.2022.9746312.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15875–15884, 2025.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *Journal of Machine Learning Research*, 23(176):1–42, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Aram Davtyan, Leello Tadesse Dadi, Volkan Cevher, and Paolo Favaro. Faster inference of flow-based generative models via improved data-noise coupling. *arXiv preprint arXiv:2603.15279*, 2026.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByMVTsR5KQ.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, 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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qgNs5NmQB7.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1695–1704. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/janati20a.html.

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.

Lingkai Kong, Molei Tao, Yang Liu, Bryan Wang, Jinmiao Fu, Chien-Chih Wang, and Huidong Liu. Alignflow: Improving flow-based generative models with semi-discrete optimal transport. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nTCF3QNsIN.

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

Xiquan Li, Junxi Liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen, and Xie Chen. Meanaudio: Fast and faithful text-to-audio generation with mean flows. *arXiv preprint arXiv:2508.06098*, 2025.

Shujian Liao, Hao Ni, Marc Sabate-Vidales, Lukasz Szpruch, Magnus Wiese, and Baoren Xiao. Sig-wasserstein gans for conditional time series generation. *Mathematical Finance*, 34(2):622–670, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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. *arXiv preprint arXiv:2301.12503*, 2023a.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=XVjTT1nw5z.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and qiang liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1k4yZbbDqX.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 48855–48876, 2023.

Manh Luong, Khai Nguyen, Dinh Phung, Gholamreza Haffari, and Lizhen Qu. Unbiased sliced wasserstein kernels for high-quality audio captioning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=83HlJrZigZ.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345. IEEE, 2024.

Alireza Mousavi-Hosseini, Stephen Y. Zhang, Michal Klein, and marco cuturi. Flow matching with semidiscrete couplings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4EGjzT6w80.

Khai Nguyen. An introduction to sliced optimal transport: Foundations, advances, extensions, and applications. *Foundations and Trends® in Computer Graphics and Vision*, 17(3-4):171–391, 2025.

Khai Nguyen and Nhat Ho. Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 36, 2024.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

K R Prajwal, Bowen Shi, Matthew Lee, Apoorv Vyas, Andros Tjandra, Mahi Luthra, Baishan Guo, Huiyu Wang, Triantafyllos Afouras, David Kant, and Wei-Ning Hsu. Musicflow: cascaded flow matching for text guided music generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation l p distance for signal analysis. *Journal of mathematical imaging and vision*, 59(2):187–210, 2017.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37:128118–128138, 2024.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

A Implementation Details

Data preprocessing. For both tasks, audio is downsampled to 16 kHz and the target representation is an 80-bin log-mel spectrogram. For video-to-audio generation, videos are additionally downsampled to 4 FPS before extracting visual features as conditioning inputs.

Model Configuration. We adopt two flow matching backbones, one per task. For video-to-audio generation, we use the Frieren model Wang et al. (2024), which conditions generation on visual embeddings extracted by the pre-trained CAVP encoder Luo et al. (2023). Since CAVP produces a sequence of frame-level visual embeddings, we average them into a single vector, which serves as the semantic conditioning vector z^i for computing the condition direction $\tilde{z}^i$ in Eq. equation 7. For text-to-audio generation, we use the MeanAudio model Li et al. (2025), which employs two text encoders: a CLAP text encoder Wu et al. (2023) and a FLAN-T5 encoder Chung et al. (2024). Since the FLAN-T5 encoder captures instruction-level semantics rather than acoustic content, we use the mean-pooled output of the CLAP text encoder as the semantic conditioning vector z^i, as CLAP embeddings are trained to align directly with audio content. The table. 6 shows the implementation details of training the backbone models for baselines and our proposed method.

Table 6: Implementation details for backbone training (left) and reflow training of the Frieren model (right).

Hyperparameter	MeanAudio	Frieren
Learning rate	1e-4	5e-5
Batch size	256	256
Warm-up steps	1,000	1,000
Training iterations	200,000	100,000
Projections L	100	100
Concentration κ	3.0	3.0
Sinkhorn λ	0.05	0.05

Hyperparameter	Frieren (reflow)
Learning rate	5e-5
Batch size	32
Warm-up steps	1,000
Training iterations	50,000

Algorithm 1 SWIFT Audio–Noise Coupling

Require: Minibatch of audio–condition pairs $\{(X_1^i, z^i)\}_{i=1}^B$, noise sequences $\{X_0^j\}_{j=1}^B$, number of projections L, concentration κ, Sinkhorn regularisation λ
Ensure: Coupled pairs $\{\big((X_1^k, z^k), X_0^k\big)\}_{k=1}^B$

// Step 1: Time-augmented path embedding
for $i = 1, \dots, B$ **do**
  $\hat{X}_1^i \leftarrow \big((t_k,\, x_1^{i,k})\big)_{k=1}^T,\ t_k = \frac{k-1}{T-1}$ ▷ Lift audio sequences to $\mathbb{R}^{d+1}$
end for
for $j = 1, \dots, B$ **do**
  $\hat{X}_0^j \leftarrow \big((t_k,\, x_0^{j,k})\big)_{k=1}^T$ ▷ Lift noise sequences to $\mathbb{R}^{d+1}$
end for
// Step 2: Condition directions
for $i = 1, \dots, B$ **do**
  $\tilde{z}^i \leftarrow (0,\, \phi(z^i)) \,/\, \|(0,\, \phi(z^i))\|_2 \in \mathbb{S}^d$ ▷ Lift condition vector to $\mathbb{R}^{d+1}$ and normalize it
end for
// Step 3: Compute ground cost matrix
for $i = 1, \dots, B$ **do**
  Sample $\{\psi_\ell^i\}_{\ell=1}^L \sim \sigma^i = \mathrm{vMF}(\tilde{z}^i, \kappa)$ ▷ Mixed slicing directions
  for $j = 1, \dots, B$ **do**
    $u_\ell \leftarrow \mathrm{sort}\big(\langle \psi_\ell^i, \hat{X}_1^i \rangle\big),\ v_\ell \leftarrow \mathrm{sort}\big(\langle \psi_\ell^i, \hat{X}_0^j \rangle\big)$ ▷ Project and sort
    $\mathbf{C}_{ij} \leftarrow \frac{1}{L} \sum_{\ell=1}^L \|u_\ell - v_\ell\|^2$ ▷ Average 1-D Wasserstein distances
  end for
end for
// Step 4: Sinkhorn optimal coupling
$\pi^* \leftarrow \mathrm{Sinkhorn}(\mathbf{C},\, \lambda)$ ▷ Solve Eq. equation 13
// Step 5: Sample coupled pairs from the optimal plan
$\{\big((X_1^k, z^k), X_0^k\big)\}_{k=1}^B \sim \pi^*$ ▷ Sample audio–noise couplings from the optimal coupling

B Related Work

Flow matching for audio generation. Flow matching Lipman et al. (2022) and rectified flow Liu et al. (2023b) have become the dominant framework for audio generation, replacing diffusion-based models across text-to-speech Mehta et al. (2024); Chen et al. (2025), text-to-audio Hung et al. (2026); Li et al. (2025), and video-to-audio generation Wang et al. (2024); Cheng et al. (2025). These models learn a vector field transporting Gaussian noise to the target audio distribution, with inference cost controlled by the number of function evaluations (NFE). A central challenge is reducing NFE without sacrificing generation quality, which motivates improving the data–noise coupling at training time.

Data–noise coupling for flow matching. Under the independent coupling $\pi = q_1 \otimes q_0$, learned flows are curved because trajectories from different source–target pairs cross in the interior of the integration interval Liu et al. (2023b); Pooladian et al. (2023). Replacing the independent coupling

with an optimal transport (OT) plan straightens probability paths and reduces the required NFE by 30–60% on image benchmarks Pooladian et al. (2023); Tong et al. (2024). LOOM-CFM Davtyan et al. (2026) extends minibatch OT by recycling locally optimal assignments across training steps, further reducing inference cost. Recently, AlignFlow Kong et al. (2026) and SD-FM Mousavi-Hosseini et al. (2026) address the scalability bottleneck of minibatch OT by pre-computing a semi-discrete OT potential over the full training set, replacing the per-batch Sinkhorn solve with a cheap maximum inner-product search at training time; however, both methods still rely on a Euclidean ground cost that ignores the temporal structure and semantic conditioning of audio sequences. The reflow procedure Liu et al. (2023b) iteratively straightens paths by training on source–target pairs generated by a previously trained model. However, all of these methods compute the ground cost between sequences via an isotropic Euclidean distance, which ignores both the temporal ordering of frames and the semantic conditioning signal. A recent analysis Cheng & Schwing (2025) demonstrates that OT with a condition-agnostic ground cost can perform worse than independent coupling in conditional generation settings, directly motivating our work.

Optimal transport for sequential and audio data. Computing meaningful distances between ordered sequences is a long-standing challenge in optimal transport. Lifting a signal to a probability measure on its graph — the set of (time, feature) pairs — and computing OT between such measures is established in the Transportation L^p framework Thorpe et al. (2017). For sequences, time augmentation — prepending a normalised time coordinate to each frame before computing OT — is the standard device for injecting temporal ordering, and underpins the Sig-Wasserstein family of models for time series generation Liao et al. (2024); Chevyrev & Oberhauser (2022). Sliced Wasserstein distances Bonneel et al. (2015); Nguyen et al. (2021) make OT computationally tractable on high-dimensional empirical measures via random 1-D projections, and have recently been applied to audio sequences for cross-modal similarity Luong et al. (2026). Our work is the first to combine time-augmented path embeddings with condition-directed sliced projections as a *ground cost for data–noise coupling* in flow matching, enabling semantically informed, sequence-aware transport plans for conditional audio generation.

C Details of evaluation metrics

Fréchet Audio Distance (FAD). Conceptually related to FID Heusel et al. (2017), FAD measures audio similarity through an alternative feature extractor. Specifically, we adopt a variant grounded in PANNs Kong et al. (2020) embeddings. Since PANNs models are pretrained on the large-scale AudioSet benchmark Gemmeke et al. (2017), this metric offers strong robustness for assessing audio fidelity.

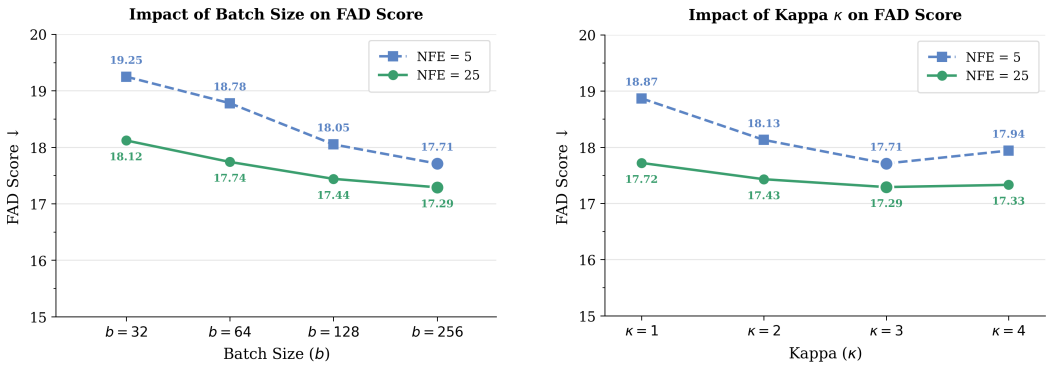
Kullback–Leibler Divergence (KL). We employ KL divergence to estimate the acoustic similarity between generated and ground-truth audio. Concretely, we measure the divergence between the multi-label class probability distributions predicted by a PANNs model Kong et al. (2020) over both sets of samples.

Inception Score (IS). IS is an established metric for evaluating generative models, jointly capturing sample quality and diversity — the former reflecting the clarity and distinctiveness of individual audio events, and the latter reflecting variation across generated outputs Donahue et al. (2019); Liu et al. (2023a).

CLAP Score. The Contrastive Language–Audio Pretraining (CLAP) model maps audio and text into a shared embedding space to capture cross-modal alignment. We use the CLAP Score — defined as the cosine similarity between embeddings from the pretrained CLAP encoders Wu et al. (2023) — to quantify the semantic correspondence between generated audio and its conditioning text prompt, with higher values indicating stronger alignment.

D Additional ablation studies

We train the MeanAudio model for $50,000$ iteration steps to ablate the impact of batch size and the concentration parameter κ. According to the ablation studies, we train the final model with the batch



(a) Ablation study regarding the impact of batch size.

(b) Ablation study regarding the impact of κ.

Figure 3: An ablation study regarding the impact of different batch sizes and values of κ at two inference budget: NFE = 5 and NFE = 25.

Table 7: Training time per iteration (seconds) for a batch size of 256.

Method	Frieren	MeanAudio
Baseline	2.31	0.736
MinibatchOT	2.49	0.948
LOOM-CFM	2.41	0.921
C2OT	2.72	1.221
SWSD (Ours)	2.64	1.113

size $b = 256$ and the concentration parameter $\kappa = 3$ for computing the SWSD in the SWIFT training framework.

Impact of batch size. As shown in Figure 3a, increasing the batch size consistently reduces the FAD score under both inference budgets. The most substantial gains occur in the lower range — moving from $b = 32$ to $b = 128$ yields significant improvements at both NFE = 5 ($19.25 \rightarrow 18.05$) and NFE = 25 ($18.12 \rightarrow 17.44$) — while the marginal benefit of scaling from $b = 128$ to $b = 256$ diminishes notably. This suggests that $b = 256$ offers a favorable trade-off between computational cost and generation quality, and we adopt it as our default.

Impact of κ. Figure 3b reveals that the choice of κ has a pronounced effect at $\kappa = 1$ but stabilizes quickly beyond $\kappa = 2$. At NFE = 5, FAD drops sharply from 18.87 at $\kappa = 1$ to 17.71 at $\kappa = 3$, and then the FAD increases a bit to 17.94 at $\kappa = 4$. A similar pattern holds at NFE = 25. This indicates that the model is robust to κ once a minimum threshold is exceeded, and we set $\kappa = 3$ as the default to ensure stable performance across inference budgets.

Training overhead. Table 7 reports the per-iteration training time at batch size 256 using two NVIDIA L40S GPUs. SWIFT introduces a moderate overhead over the baseline, increasing training time by 14% and 51% on Frieren and MeanAudio respectively, compared to 8% and 29% for MinibatchOT, 4% and 25% for LOOM-CFM, and 18% and 66% for C2OT. Notably, SWIFT is cheaper than C2OT, the other condition-aware coupling, on both backbones (2.64 vs. 2.72 s and 1.113 vs. 1.221 s per iteration), while outperforming it on every metric–setting combination in Table 1. The additional cost of SWIFT over MinibatchOT and LOOM-CFM stems from two sources: sampling L projection directions from the semantic-aware slicing measure σ_κ^i and computing the $B \times B$ SWSD cost matrix via Monte Carlo quantile comparison; the Sinkhorn solve is shared with all OT-based baselines. Importantly, this overhead is incurred only at training time; at inference, SWSD imposes no additional cost over the backbone model since the coupling is not used during generation. Given the consistent improvements in FAD, KL, IS, and CLAP score reported in Table 1, the additional training

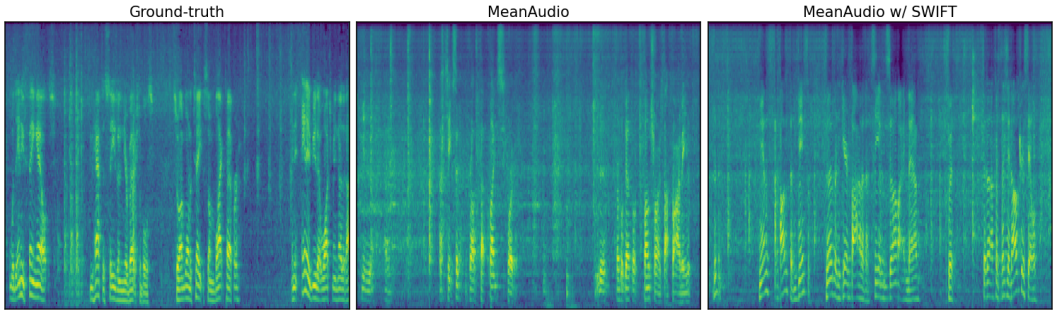
cost of SWSD represents a favourable trade-off between computational overhead and generation quality.

E LIMITATIONS AND FUTURE WORK.

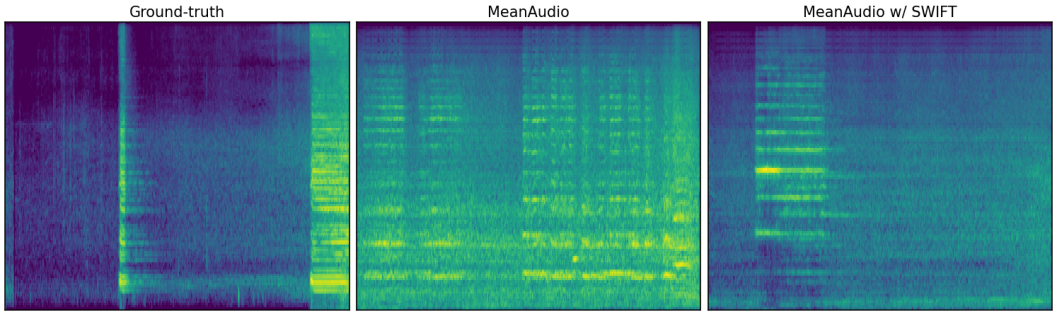
The current formulation projects high-dimensional conditioning vectors via truncation or mean pooling, which may discard semantically relevant dimensions when $d_z \gg d$; replacing this with a jointly learned linear projection is a natural extension. SWSD is evaluated on fixed-length sequences of equal frame rate; extending it to variable-length audio or multi-resolution representations is an open direction. Finally, while SWSD improves coupling quality within the standard minibatch OT framework, combining it with offline semi-discrete OT solvers Kong et al. (2026); Mousavi-Hosseini et al. (2026) could further reduce the variance of the coupling estimate at large dataset scales.

F QUALITATIVE EXAMPLES

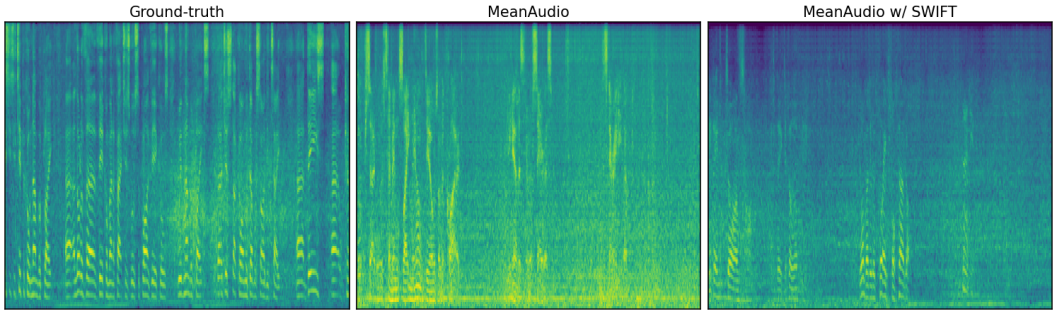
We attached a HTML file which contains 10 triplet samples at a low inference budget(NFE = 5) for qualitative evaluation. Please open this file to see the qualitative differences between our training method and the base model.



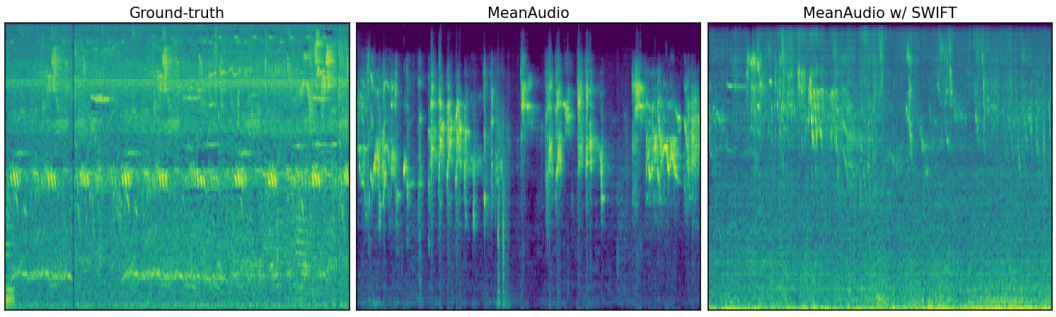
Prompt: "A man is talking, and food is frying."



Prompt: "A distant horn ring as a vehicle approaches."

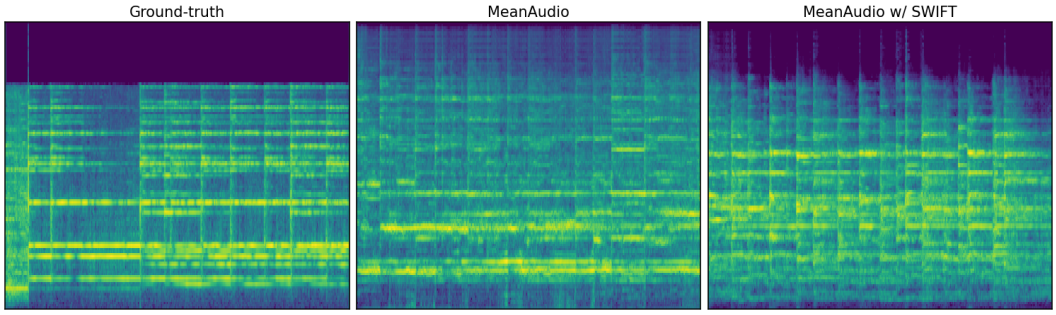


Prompt: "A man talking as a vehicle drives by as rain patters on a hard surface."

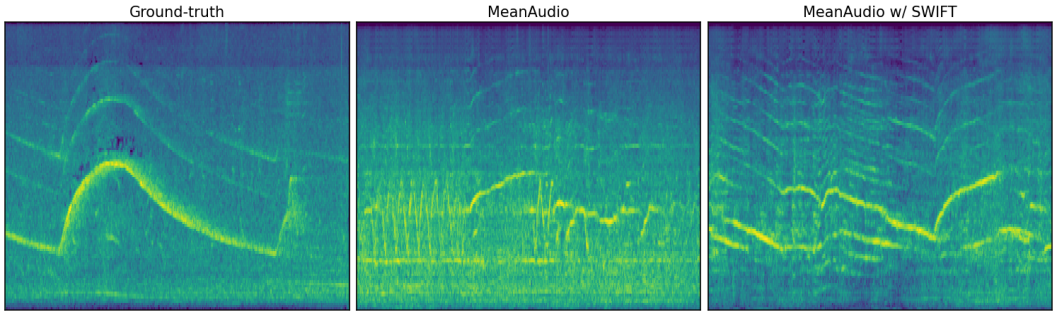


Prompt: "A group of birds call and chirp for a while."

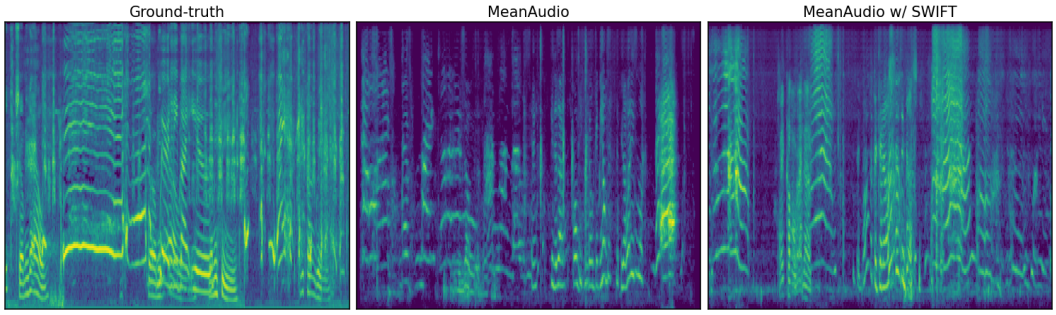
Figure 4: Mel-spectrogram comparisons between Ground-truth, MeanAudio, and MeanAudio trained with the SWIFT method across four audio samples. Each row corresponds to a different text prompt. All these synthetic audio are generated with NFE $= 5$



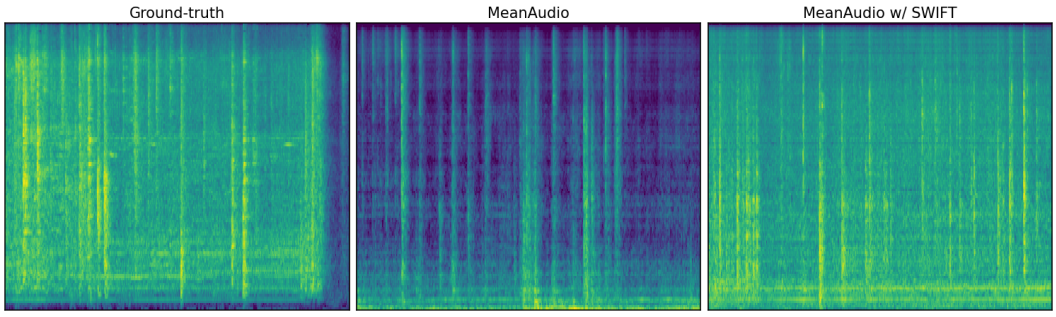
Prompt: "Chiming of loud bells."



Prompt: "A siren with people talking faintly in the background."



Prompt: "A child cries and a man and woman speaking with a duck quacking."



Prompt: "Rattling and banging with faint humming."

Figure 5: Mel-spectrogram comparisons between Ground-truth, MeanAudio, and MeanAudio trained with the SWIFT method across four audio samples. Each row corresponds to a different text prompt.All these synthetic audio are generated with NFE $= 5$