WOOSH-MIMIC: CONDITIONING THE WOOSH SOUND EFFECT FOUNDATION MODEL WITH VOCAL MIMICRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-audio generation enables semantic control over generated sounds through natural language. Adding vocal mimicry could provide an intuitive, nonverbal way to convey fine-grained temporal variations in timing, intensity, pitch, rhythm, and spectral characteristics, without describing them in words or manually specifying numerical control trajectories. However, vocal mimics and the sound effects (SFXs) they represent belong to substantially different acoustic domains, and learning the correspondence between them remains underexplored. We propose Woosh-Mimic, a vocal-mimic-conditioned SFX generation method built on the pretrained Woosh framework. Woosh-Mimic combines semantic conditioning from production SFX metadata with temporal acoustic conditioning from vocal mimics and learns from paired vocal-mimic and SFX recordings. To examine the effect of the domain mismatch, we compare a Mimic Model trained with acoustic conditions extracted from vocal mimics against a Reference Model trained with conditions extracted from the corresponding reference SFXs. The Mimic Model achieves lower Fréchet Audio Distance across all seven evaluation splits, whereas the Reference Model more strongly reflects the characteristics of the input vocal mimics in both objective and subjective evaluations. This contrast suggests that matching the conditioning domains between training and inference improves generation quality but does not necessarily improve the reflection of individual vocal mimics. Our findings highlight the potential of vocal mimicry as an expressive complement to language-based semantic control while identifying faithful control across acoustic domains as a central challenge.

1 INTRODUCTION

Recent advances in text-to-audio (TTA) generation have enabled high-quality environmental sounds and sound effects (SFXs) to be generated from natural-language descriptions (Liu et al., 2023; Huang et al., 2023; Ghosal et al., 2023). Text provides intuitive control over the semantic content of a sound. Vocal mimicry could complement this control by allowing users to express time-varying characteristics, such as timing, intensity, pitch, rhythm, and spectral changes, without describing them in words or manually specifying numerical control trajectories (Ekman & Rinott, 2010; Lemaitre et al., 2011; Lemaitre & Rocchesso, 2014; Cartwright & Pardo, 2015; Mehrabi et al., 2017).

Previous studies have conditioned audio generation on time-varying acoustic features to control such characteristics (Chung et al., 2024; García et al., 2025; Novack et al., 2026). These features are often extracted from the target audio during training. If a model trained in this way receives features extracted from a vocal mimic at inference time, the training and inference conditions belong to different acoustic domains. Moreover, a vocal mimic and its intended SFX may express corresponding temporal patterns through substantially different acoustic features. Learning this cross-domain correspondence from paired vocal-mimic and SFX recordings is therefore central to vocal-mimic-conditioned generation.

Semantic conditioning also needs to reflect how SFXs are described in practice. Commercial sound-effect libraries use metadata to describe and organize recordings (Viers, 2008), but their keywords, technical terms, and abbreviations differ from the natural-language captions commonly used in TTA training. Woosh, a pretrained SFX generation framework, identifies this metadata–caption gap (Hadjeres et al., 2026). Building on Woosh, we propose Woosh-Mimic, which combines semantic con-

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.

ditioning from SFX metadata with temporal acoustic conditioning from vocal mimics. We fine-tune the Woosh-CLAP text encoder on SFX metadata and introduce a Mimic Encoder that converts features extracted from vocal mimics into conditioning tokens for Woosh-Flow. Using paired vocal-mimic and SFX recordings, we train the generation model with vocal-mimic-derived conditions, allowing it to learn their cross-domain correspondence while using the same conditioning domain during training and inference.

Our contributions are as follows:

- We propose Woosh-Mimic, combining SFX-metadata-based semantic conditioning and vocal-mimic-based temporal acoustic conditioning in pretrained Woosh.
- We train with paired vocal-mimic and SFX recordings and compare a model trained with vocal-mimic-derived conditions against one trained with reference-SFX-derived conditions to examine the effect of the conditioning domain.
- We evaluate generation quality and vocal-mimic reflection using objective metrics and a listening study, revealing a trade-off in this comparison.

2 Related Work

2.1 Acoustic-Conditioned and Vocal-Mimic-Based Audio Generation

Time-varying acoustic conditions have been used to control when and how sounds evolve. T-Foley conditions Foley generation on sound classes and temporal event features and demonstrates the use of vocal mimicry for temporal control (Chung et al., 2024). Sketch2Sound adapts a pretrained text-to-audio model to follow loudness, brightness, and pitch trajectories extracted from sonic imitations, while FlashFoley extends this approach toward fast, interactive generation (García et al., 2025; Novack et al., 2026). These methods demonstrate the usefulness of vocal input as an acoustic sketch. However, their acoustic conditions are extracted from the target recordings during training; they do not train on conditions extracted from vocal mimics paired with the target sounds. Consequently, they do not explicitly learn cross-domain correspondences between vocal mimics and target SFXs.

Other studies learn from vocal imitations more directly. Okamoto et al. (2024b) proposed a model that synthesizes environmental sounds from vocal imitations and sound-event labels, enabling control over pitch and rhythm. Takizawa et al. (2026b) proposed PronounSE, which learns from paired vocal mimics and SFXs to generate sounds that reflect nuances of the input mimic. Woosh-Mimic instead integrates learning from paired vocal mimics and SFXs into a pretrained SFX generation model and combines it with semantic conditioning from production SFX metadata.

2.2 Woosh and Sound-Effect Metadata Conditioning

Woosh is a pretrained SFX foundation model comprising an audio encoder–decoder (Woosh-AE), a text–audio alignment model (Woosh-CLAP), and conditional audio generators, including the text-to-audio model Woosh-Flow (Hadjeres et al., 2026). Woosh-AE provides the audio latent space in which Woosh-Flow generates, while the Woosh-CLAP text encoder supplies its semantic condition. We build on these publicly available models for vocal-mimic-conditioned SFX generation.

The Woosh study observes that descriptions in professional SFX libraries are shorter, more keyword-like, and more reliant on technical vocabulary and abbreviations than captions in public audio–text datasets (Hadjeres et al., 2026). This difference motivates adapting text conditioning to the metadata used for the SFXs in our study. We therefore fine-tune the publicly available Woosh-CLAP on paired SFX metadata and recordings before using its text encoder to condition Woosh-Flow.

3 Proposed Method

3.1 Overview

Figure 1 illustrates Woosh-Mimic, which adapts pretrained Woosh to generate SFXs conditioned on both SFX metadata and a vocal mimic. Woosh-CLAP, fine-tuned on SFX metadata, provides the

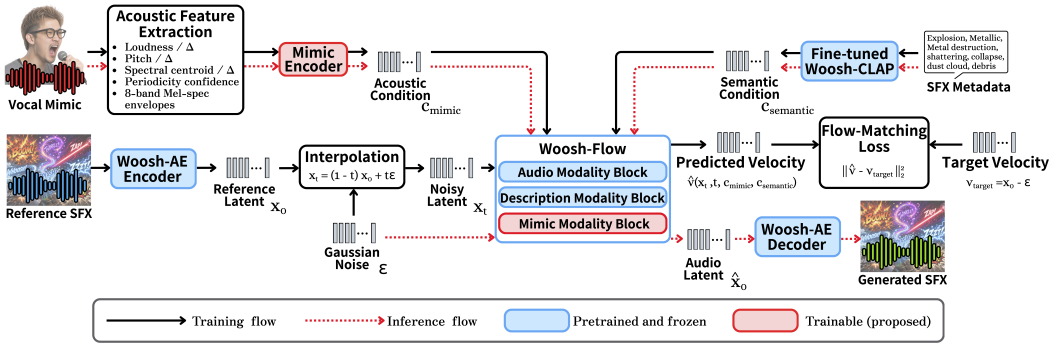

Figure 1: Woosh-Mimic combines semantic conditioning from SFX metadata with temporal acoustic conditioning from vocal mimics.

semantic condition. A Mimic Encoder converts time-varying acoustic features extracted from the vocal mimic into fixed-length conditioning tokens, which are integrated into Woosh-Flow through newly added mimic-specific modality blocks. During training, the pretrained Woosh-AE encodes each reference SFX into a target latent. The Mimic Encoder and mimic-specific blocks are trained on paired vocal mimics and SFXs using a flow-matching objective, while the pretrained Woosh-Flow parameters remain frozen. At inference, Woosh-Flow generates an audio latent from Gaussian noise using both conditions, and Woosh-AE decodes it into an SFX waveform.

3.2 Sound-Effect Metadata Conditioning

Each SFX's metadata provides the semantic condition for Woosh-Mimic. To adapt Woosh-CLAP to this metadata while retaining its pretrained audio representation, we freeze the audio encoder and fine-tune only the text encoder on paired SFX metadata and recordings. Because multiple recordings may share a metadata description, the standard contrastive loss could incorrectly treat recordings with identical metadata as negatives. We therefore use a symmetric audio–text contrastive loss that retains each aligned pair as a positive but masks off-diagonal pairs sharing the same metadata from the negative set; these pairs are not added as positives. The fine-tuned text encoder produces semantic condition tokens for Woosh-Flow and remains frozen during Woosh-Mimic training. Appendix A evaluates its retrieval performance with SFX metadata and general audio captions.

3.3 Vocal-Mimic Acoustic Conditioning

Inspired by methods that use interpretable, time-varying acoustic controls (García et al., 2025; Novack et al., 2026), we represent each vocal mimic with a compact 15-dimensional feature sequence rather than its full Mel spectrogram. The features comprise loudness and its first-order difference; log-frequency pitch and its difference; voicing probability; log spectral centroid and its difference; and eight Mel-band envelopes. The envelopes summarize the time-varying spectral energy distribution across coarse frequency bands. Together, these features describe changes in intensity, pitch, voicing, and spectral properties without using the full spectral detail of the voice.

Let $X_{\text{mimic}} \in \mathbb{R}^{T\times 15}$ denote this sequence, where T is the number of frames. Sequences are padded within each mini-batch, and a binary mask identifies valid frames. As shown in Fig. 2, the Mimic Encoder projects and positionally encodes the feature sequence, processes it with a Transformer encoder, and uses 32 learnable queries for cross-attention pooling. This converts the variable-length sequence into 32 tokens, which are projected to the 1024-dimensional hidden space of Woosh-Flow to form $c_{\text{mimic}} \in \mathbb{R}^{32\times 1024}$. The padding mask is applied during encoding and pooling. The resulting tokens enter Woosh-Flow through the newly added Mimic Modality Block. Further details of acoustic-feature extraction and the Mimic Encoder are provided in Appendix C.

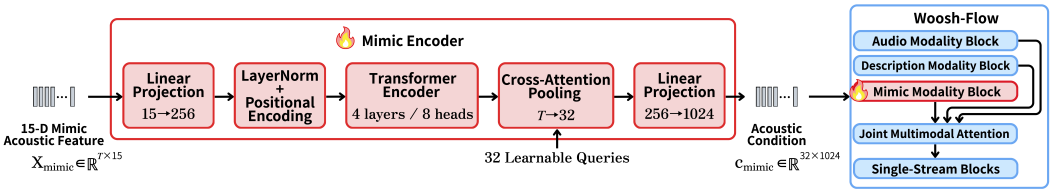

Figure 2: Vocal-mimic conditioning pathway. The Mimic Encoder converts a variable-length 15-dimensional feature sequence into 32 conditioning tokens for Woosh-Flow. Blue blocks are pretrained and frozen; red blocks are trainable.

3.4 Multimodal Integration into Woosh-Flow

Woosh-Flow processes audio latents and text conditions using multimodal and single-stream transformer blocks (Hadjeres et al., 2026). We add c_{mimic} as a third modality for conditioning, distinct from the audio latents being generated. As shown in Fig. 2, a Mimic Modality Block is added to each multimodal block. Its queries, keys, and values are concatenated with those of the audio and description modalities along the sequence dimension for joint attention. The mimic tokens also participate in the subsequent single-stream blocks.

Because the 32 pooled mimic tokens do not correspond one-to-one to frames of the input vocal mimic, we assign them a shared Rotary Position Embedding (RoPE) position (Su et al., 2024), following Woosh's treatment of description tokens (Hadjeres et al., 2026). Thus, the mimic tokens provide acoustic conditioning without imposing a framewise alignment with the audio latents.

3.5 Training Strategy

Woosh-Mimic is trained on paired vocal mimics, SFX metadata, and reference SFXs. The pretrained Woosh-AE encodes each reference SFX into a target latent x_0; its reconstruction performance is evaluated in Appendix B. The metadata and vocal mimic provide the semantic and acoustic conditions described in Secs. 3.2 and 3.3, respectively.

Given the target latent x_0, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and a time step $t \in [0, 1]$, we construct a linear flow-matching path (Lipman et al., 2023):

$$x_t = (1-t)x_0 + t\epsilon, \qquad v_{\text{target}} = x_0 - \epsilon. \tag{1}$$

Here, v_{target} is the reverse-time velocity toward the reference latent, following the convention of the released Woosh-Flow implementation (Hadjeres et al., 2026). The model predicts this velocity from x_t under both conditions and minimizes the mean squared error over valid reference-latent frames.

Woosh-AE, the fine-tuned Woosh-CLAP, and pretrained Woosh-Flow parameters remain frozen. Only the Mimic Encoder, the newly added mimic-specific Modality Blocks, and a learnable null-mimic condition are optimized. During training, the mimic condition is occasionally replaced by the null condition while the semantic condition is retained. Optimization settings and other training details are provided in Appendix C.

4 Experimental Setup

4.1 Datasets

We use four datasets comprising 27,191 vocal-mimic–SFX pairs: KLDataset, ESC-50-Voice, Explosion Dataset, and Explosion Test. Table 1 summarizes their splits.

KLDataset pairs newly recorded vocal mimics with SFXs from six commercial libraries: Krotos Battle Bundle, Krotos Whooshes, Krotos Magic, SoundMorph Robotic Lifeforms 1, SoundMorph Energy, and Paul Stoughton Pulse Energy Weapons. One speaker recorded one mimic per reference SFX. We use the libraries' metadata descriptions where available; otherwise, we construct descriptions from directory names, file names, tags, and available sound descriptions. The dataset

Table 1: The number of reference SFXs is counted within each split; for Explosion Dataset, duplicate source recordings are grouped into canonical references.

Dataset	Split	Pairs	Ref. SFXs	Metadata	Speakers
KLDataset	Train	6,146	6,146	443	1
	Val (CLAP-unseen)	946	946	60	1
	Test (CLAP-seen)	407	407	407	1
	Test (CLAP-unseen)	1,055	1,055	58	1
ESC-50-Voice	Train	7,440	1,240	31	6
	Val (unseen speaker)	1,240	1,240	31	1
	Test (unseen speaker)	1,240	1,240	31	1
Explosion Dataset	Train	7,306	1,740	1,075	3
	Validation	811	643	504	3
Explosion Test	Test	600	100	15	6
Total		**27,191**	–	–	–

is split into training, validation, and test sets. The test set is divided according to whether its metadata descriptions were used to fine-tune Woosh-CLAP (*CLAP-seen* or *CLAP-unseen*); all validation metadata are CLAP-unseen.

ESC-50-Voice (Okamoto et al., 2024a) contains vocal mimics of ESC-50 environmental sounds (Piczak, 2015) recorded by eight speakers. We use the ESC-50 sound-event class labels as metadata. Six speakers are assigned to training, one to validation, and one to testing. The reference sounds occur across these splits, but the validation and test speakers are absent from training; these splits therefore assess generalization to unseen speakers rather than unseen reference sounds.

Explosion Dataset contains 8,117 vocal-mimic–SFX pairs recorded by three speakers. After grouping duplicate source recordings, these pairs correspond to 1,745 distinct reference SFXs. We manually created comma-separated pseudo-metadata from the audio and its directory and file names. Training and validation are split by vocal-mimic–SFX pair: 638 of the 643 canonical reference SFXs in validation also occur in training. This validation split therefore evaluates held-out pairs, not generalization to unseen reference SFXs.

For an independent evaluation of explosion sounds, we use the publicly available Explosion Test (Takizawa et al., 2026a), comprising 100 reference SFXs and 600 vocal mimics from six speakers absent from training. We organize its metadata into 15 descriptions, each beginning with "Explosion," to provide consistent semantic conditions.

4.2 Compared Models

To examine the training–inference acoustic-domain gap, we compare two variants with the same architecture, training data, semantic condition, objective, and optimization settings. The proposed *Mimic Model* uses acoustic features extracted from the paired vocal mimics during training. The *Reference Model* instead uses the same types of features extracted from the corresponding reference SFXs, following a self-conditioning setup. At inference, both models receive features from the same input vocal mimic and the same SFX-metadata condition. Thus, the Reference Model encounters a change in acoustic-conditioning domain between training and inference, whereas the Mimic Model does not.

For overlength training examples, reference-derived features are cropped together with the target audio latent, while vocal-mimic features are retained in full because the two recordings are not assumed to be frame-aligned. This comparison therefore assesses the practical effect of training with vocal-mimic-derived rather than reference-SFX-derived conditions, including this source-dependent handling of temporal context.

4.3 Evaluation Metrics

Generation Quality. We measure distributional similarity between generated and reference SFXs using Fréchet Audio Distance (FAD) (Kilgour et al., 2019), with embeddings extracted by PANNs Wavegram-Logmel (Kong et al., 2020). Lower FAD indicates closer distributions in this embedding space; it does not measure how faithfully an individual output reflects its input vocal mimic. FAD is computed separately for each validation and test split.

Vocal-Mimic Reflection. We evaluate how well the generated SFXs reflect the time-varying acoustic characteristics expressed in the input vocal mimic using objective measures and a subjective listening test.

For the objective evaluation, we compare trajectories of loudness, spectral centroid, and Mel envelope extracted from each vocal mimic and generated SFX. For each characteristic, we compute the Pearson correlation coefficient and a dynamic time warping (DTW) distance. Each trajectory is standardized independently. For Pearson correlation, trajectories are resampled to a common normalized-time grid; for DTW, the distance is normalized by the length of the warping path and multiplied by -1. Thus, higher values indicate greater similarity for all six metrics. The Mel envelope is a one-dimensional trajectory obtained by averaging the Mel representation across frequency bins. Because both models generate SFXs from the same vocal-mimic inputs, their results are compared using two-sided paired t-tests for each metric and evaluation split, with the Holm procedure (Holm, 1979) applied across all 42 comparisons.

In the subjective listening test, participants hear an original vocal mimic and the SFXs generated from it by the Mimic and Reference models, presented as Audio A and Audio B. They rate which SFX better reflects the characteristics expressed in the vocal mimic on a five-point scale, with the midpoint indicating no preference. Model assignment to Audio A or Audio B is counterbalanced across participants and stimuli. Ten vocal-mimic inputs are selected from each of the seven validation and test splits, yielding 70 stimuli. Each stimulus receives 20 valid ratings, for a total of 1,400 ratings. Participants are recruited through Lancers, a Japanese crowdsourcing platform. Each participant evaluates 20 trials and must listen to the vocal mimic and both generated SFXs at least once before responding.

We analyze the five-point responses with a cumulative link mixed model to assess the effect of presentation position. The model includes presentation position as a fixed effect and participant and stimulus as crossed random intercepts. To assess overall model preference, we exclude equal-preference responses and fit a logistic mixed-effects model to the binary preference for the Mimic or Reference Model, with the same fixed and random effects. We test the null hypothesis of equal preference odds by comparing this model with one whose preference intercept is fixed at zero using a one-degree-of-freedom likelihood-ratio test. We report the odds ratio for preferring the Reference Model over the Mimic Model and its 95% profile-likelihood confidence interval.

5 Results

5.1 Generation Quality

Table 2 shows the FAD results for the Mimic Model and Reference Model across the seven validation and test splits. The Mimic Model achieved lower FAD than the Reference Model on all seven splits, indicating that its generated SFXs have feature distributions closer to those of the reference SFXs. The difference was particularly large on ESC-50-Voice, where the FAD decreased from 41.956 to 26.683 on the validation split and from 39.133 to 25.257 on the test split. On the Explosion Dataset validation split, the Mimic Model also reduced the FAD from 7.281 to 4.231. The Mimic Model also achieved a lower FAD on Explosion Test, with 31.053 compared with 32.831 for the Reference Model. Overall, training with acoustic conditions extracted from vocal mimics resulted in consistently lower FAD across all evaluation splits.

5.2 Vocal-Mimic Reflection

Objective Evaluation. Table 3 summarizes the objective vocal-mimic reflection results over all 6,299 samples from the seven validation and test splits. The Reference Model achieved higher mean

Table 2: FAD results for the Mimic and Reference Models across the validation and test splits. Lower values are better. The best result for each split is shown in bold.

Dataset / Split	Mimic ↓	Reference ↓
KLDataset Val (CLAP-unseen)	**12.869**	13.738
KLDataset Test (CLAP-unseen)	**10.265**	12.146
KLDataset Test (CLAP-seen)	**10.517**	12.467
ESC-50-Voice Val (unseen speaker)	**26.683**	41.956
ESC-50-Voice Test (unseen speaker)	**25.257**	39.133
Explosion Dataset Val	**4.231**	7.281
Explosion Test	**31.053**	32.831

Table 3: Objective results for vocal-mimic reflection over all 6,299 samples from the seven validation and test splits. Values are reported as mean ± standard deviation. Higher values are better for all metrics.

Characteristic	Model	Corr. ↑	−DTW ↑
Loudness	Mimic	0.389 ± 0.400	-0.264 ± 0.135
	Reference	$\mathbf{0.491 \pm 0.383}$	$\mathbf{-0.227 \pm 0.118}$
Spectral centroid	Mimic	0.136 ± 0.307	-0.437 ± 0.212
	Reference	$\mathbf{0.187 \pm 0.324}$	$\mathbf{-0.376 \pm 0.169}$
Mel envelope	Mimic	0.363 ± 0.409	-0.260 ± 0.142
	Reference	$\mathbf{0.478 \pm 0.397}$	$\mathbf{-0.215 \pm 0.121}$

values than the Mimic Model for all six metrics, including both Pearson correlation and negative DTW distance for loudness, spectral centroid, and Mel envelope. For example, the loudness correlation increased from 0.389 for the Mimic Model to 0.491 for the Reference Model, while the Mel-envelope correlation increased from 0.363 to 0.478. A similar trend was observed for the remaining metrics.

Two-sided paired t-tests were conducted between the Mimic Model and Reference Model for each metric and evaluation split, and the Holm procedure (Holm, 1979) was applied to correct for multiple comparisons across the 42 comparisons. The Reference Model achieved significantly higher values in 40 of the 42 comparisons. The only comparisons without a significant difference were the spectral-centroid correlation and negative DTW distance on the Explosion Dataset validation split.

These results show that the Reference Model achieved higher similarity between the temporal acoustic trajectories of the input vocal mimics and generated SFXs according to the objective vocal-mimic reflection metrics. In contrast, the FAD results in Sec. 5.1 consistently favored the Mimic Model across all evaluation splits, showing different trends between generation quality and vocal-mimic reflection measured by the acoustic features used in this evaluation.

Subjective Evaluation. A total of 1,400 responses were collected from 70 participants. Table 4 shows the distribution of responses across the five preference levels. When the five-point responses were grouped according to model preference, the Reference Model was preferred in 713 responses (50.9%), while the Mimic Model was preferred in 467 responses (33.4%). The two models were rated as equally preferable in 220 responses (15.7%). After excluding the equal-preference responses, the Reference Model accounted for 60.4% of the preferences, compared with 39.6% for the Mimic Model.

The logistic mixed-effects analysis accounting for repeated responses from participants and repeated evaluations of stimuli showed a significant preference for the Reference Model over the Mimic Model (OR = 1.69, 95% CI [1.24, 2.31], $p < 0.001$). The cumulative link mixed model found no statistically significant effect of A/B presentation position (OR = 0.83, 95% CI [0.68, 1.01], $p = 0.059$).

Table 4: Distribution of the five-point responses in the subjective evaluation of vocal-mimic reflection.

Response	Count	Percentage
Reference clearly preferred	229	16.4%
Reference slightly preferred	484	34.6%
Equal preference	220	15.7%
Mimic slightly preferred	364	26.0%
Mimic clearly preferred	103	7.4%
Total	1,400	100.0%

6 DISCUSSION

Generation Quality and Training–Inference Condition Gap. The Mimic Model achieved lower FAD than the Reference Model across all seven evaluation splits. Since the two models use the same architecture and training configuration and differ only in the source of the acoustic condition during training, one possible explanation for this result is the difference in the acoustic-condition domain between training and inference. The Reference Model is trained with acoustic features extracted from the reference SFXs, whereas it receives features extracted from vocal mimics at inference time. In contrast, the Mimic Model is trained with vocal-mimic-derived acoustic features and therefore receives the same type of acoustic condition during both training and inference. Reducing this training–inference condition gap may allow the model to better adapt to generating SFXs from vocal-mimic-derived conditions, resulting in a generated SFX distribution closer to that of the reference SFXs. However, FAD measures the similarity between the distributions of generated and reference SFXs and does not directly measure how well the characteristics expressed in each vocal mimic are reflected in the corresponding generated SFX. The improvement in FAD should therefore be distinguished from the ability to follow the characteristics of individual vocal mimics.

Vocal-Mimic Reflection and Cross-Domain Correspondence. Vocal-mimic reflection showed a different trend from generation quality, with the Reference Model outperforming the Mimic Model in both the objective and subjective evaluations. The objective metrics quantify alignment of acoustic-feature trajectories, rather than the full range of perceptual nuances conveyed by vocal mimicry. The subjective preference provides complementary evidence that listeners also perceived a difference between the models. One possible explanation is the difference in the correspondence between the acoustic condition and the target SFX during training. The Reference Model is trained to generate a reference SFX from acoustic features extracted from that same SFX. Consequently, temporal changes in the conditioning features, such as loudness and spectral characteristics, directly correspond to those of the target SFX. This direct correspondence may make it easier for the model to learn to reflect the temporal variation of the acoustic condition in the generated audio. Although the acoustic condition is extracted from a vocal mimic at inference time, this learned ability to follow the conditioning features may generalize to vocal-mimic-derived features to some extent. In contrast, the Mimic Model is trained to generate a reference SFX from acoustic features extracted from its corresponding vocal mimic. Because vocal mimics and SFXs belong to different acoustic domains, their acoustic features do not necessarily correspond directly. Furthermore, even when imitating the same SFX, how characteristics such as intensity, temporal structure, pitch variation, and spectral properties are expressed can vary depending on the speaker and the manner of imitation. The degree to which a vocal mimic reproduces the characteristics of the original SFX may also vary. The Mimic Model must therefore learn a more complex and variable correspondence between vocal mimics and SFXs than the Reference Model. This increased difficulty may introduce greater variability in how the characteristics of individual vocal mimics are reflected in the generated SFXs, resulting in lower vocal-mimic reflection. These findings suggest that matching the acoustic-condition domain between training and inference does not necessarily lead directly to stronger reflection of vocal-mimic characteristics. Generating high-quality SFXs from vocal-mimic conditions and reflecting the characteristics expressed in individual vocal mimics are therefore distinct aspects that should both be considered when designing vocal-mimic-conditioned SFX generation models.

Limitations and Future Directions. In this study, the temporal characteristics of vocal mimics are represented using a 15-dimensional acoustic feature sequence describing loudness, pitch, periodicity, and spectral characteristics. However, the goal of vocal-mimic reflection is not to directly match the absolute acoustic feature values or trajectories between a vocal mimic and a generated SFX. Because vocal mimics and SFXs belong to different acoustic domains, characteristics such as absolute pitch and spectral properties do not necessarily need to be identical. Instead, variations expressed through the vocal mimic, such as intensity, attack and decay, temporal structure, and timbral changes, should be appropriately reflected in the generated SFX. The present results do not establish that the 15-dimensional representation itself limits the reflection performance of the Mimic Model; the difficulty of learning the complex correspondence between vocal mimics and SFXs may also contribute to the observed results. At the same time, this study does not determine which acoustic characteristics of vocal mimicry are most important for conveying the intended characteristics of an SFX. Analyzing the relationship between individual acoustic features and perceived characteristics of vocal mimics may help identify more effective representations for vocal-mimic-based audio generation. Another direction is to introduce an auxiliary objective that more explicitly encourages correspondence between the temporal characteristics expressed in a vocal mimic and those of the generated SFX. In addition, the strong reflection performance of the Reference Model suggests a possible two-stage training strategy: first learning to reflect temporal acoustic characteristics using conditions extracted from the reference SFXs, and then fine-tuning the model using vocal-mimic-derived conditions. Such a strategy may allow the model to first acquire the ability to follow acoustic conditions and subsequently adapt this ability to the more complex correspondence between vocal mimics and SFXs. Finally, although Woosh-Mimic combines acoustic conditioning from vocal mimics with semantic conditioning from SFX metadata, the present evaluation primarily focused on how the source of the acoustic condition used during training affects generation quality and vocal-mimic reflection, and did not independently evaluate the influence of the metadata condition on the generated SFXs. It therefore remains unclear how well the semantic content specified by the metadata is reflected in the generated SFXs and how the semantic condition from metadata interacts with the acoustic condition from vocal mimics. Future work should evaluate the effect of metadata conditioning by varying the presence and content of the metadata condition and investigate how semantic and vocal-mimic conditions jointly affect the generated SFXs.

7 Conclusion

We proposed Woosh-Mimic, an SFX generation method that combines semantic conditioning from SFX metadata with acoustic conditioning based on time-varying features extracted from vocal mimics. Woosh-CLAP was adapted to SFX metadata, and a Mimic Encoder was introduced to convert vocal-mimic acoustic features into fixed-length conditioning tokens that are integrated into the pretrained Woosh-Flow as an additional modality. We compared the Mimic Model, trained with acoustic conditions extracted from vocal mimics, with the Reference Model, trained with conditions extracted from the corresponding reference SFXs, and evaluated both generation quality and vocal-mimic reflection.

The Mimic Model achieved lower FAD than the Reference Model across all seven evaluation splits, suggesting a distribution-level benefit from training with vocal-mimic-derived acoustic conditions. In contrast, the Reference Model achieved better vocal-mimic reflection in both the objective and subjective evaluations. These results suggest that generating high-quality SFXs from vocal-mimic conditions and reflecting the temporal characteristics expressed in individual vocal mimics are distinct challenges. Future work will investigate methods for more explicitly learning the complex correspondence between vocal mimics and SFXs to improve both generation quality and vocal-mimic reflection.

AI USE STATEMENT

We used OpenAI's ChatGPT-5.6 Sol to check and improve the grammar and readability of the English manuscript. The research ideas were developed entirely by the authors. All AI-assisted language edits were reviewed by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

The collection of vocal-mimic recordings for KLDataset and the subjective listening study were approved by the ethics committee of the authors' institution in accordance with its regulations for research involving human participants. The approved recording protocol covered participant consent, permission to publicly release the recordings, and voice-data privacy. Participants in the listening study provided informed consent and received compensation determined on an hourly-rate basis using an estimated completion time of 10–15 minutes. Before using the commercial sound effects in KLDataset, we shared the research details with Krotos and asked whether the sound libraries could be used for the described research. Krotos responded that the described use was consistent with its End User License Agreement (EULA).[1] Upon acceptance, we plan to publicly release the newly recorded vocal mimics and mapping files specifying their correspondence to the reference recordings in the purchased sound libraries. The commercial reference sound-effect recordings themselves will not be redistributed. The release will respect participant consent and applicable licensing restrictions.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sec. 4 describes the datasets, data splits, model-training conditions, evaluation metrics, and statistical analyses used in this study. The architecture and training procedure of Woosh-Mimic are described in Sec. 3. Appendix C provides further implementation details, including acoustic-feature extraction, model configurations, optimization settings, and inference procedures. Additional evaluations of the fine-tuned Woosh-CLAP and pretrained Woosh-AE are reported in Appendices A and B, respectively. Upon acceptance, we plan to release the newly recorded vocal mimics, and mapping files specifying their correspondence to the commercial reference sound effects, subject to participant consent and applicable licensing restrictions. The commercial reference sound-effect recordings themselves will not be redistributed.

REFERENCES

Mark Cartwright and Bryan Pardo. Vocalsketch: Vocally imitating audio concepts. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 43–46, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702387. URL https://doi.org/10.1145/2702123.2702387.

Yoonjin Chung, Junwon Lee, and Juhan Nam. T-foley: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 6820–6824, 2024. ISSN 1520-6149. doi: 10.1109/ICASSP48485.2024.10447380. Publisher Copyright: ©2024 IEEE.; 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 ; Conference date: 14-04-2024 Through 19-04-2024.

Inger Ekman and Michal Rinott. Using vocal sketching for designing sonic interactions. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, pp. 123–131, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301039. doi: 10.1145/1858171.1858195. URL https://doi.org/10.1145/1858171.1858195.

Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888184.

[1] https://www.krotosaudio.com/terms-conditions/

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 3590–3598, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612348. URL https://doi.org/10.1145/3581783.3612348.

Gaëtan Hadjeres, Marc Ferras, Khaled Koutini, Benno Weck, Alexandre Bittar, Thomas Hummel, Zineb Lahrichi, Hakim Missoum, Joan Serrà, and Yuki Mitsufuji. Woosh: A sound effects foundation model, 2026. URL https://arxiv.org/abs/2604.01929.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

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. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Guillaume Lemaitre and Davide Rocchesso. On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, 135(2):862–873, 02 2014. ISSN 0001-4966. doi: 10.1121/1.4861245. URL https://doi.org/10.1121/1.4861245.

Guillaume Lemaitre, Arnaud Dessein, Patrick Susini, and Karine Aura. Vocal imitations and the identification of sound events. *Ecological Psychology*, 23(4):267–307, 2011. doi: 10.1080/10407413.2011.617225. URL https://doi.org/10.1080/10407413.2011.617225.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *Proceedings of the International Conference on Machine Learning*, pp. 21450–21474, 2023.

Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, 2014. doi: 10.1109/ICASSP.2014.6853678.

Adib Mehrabi, Simon Dixon, and Mark B. Sandler. Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid. *The Journal of the Acoustical Society of America*, 141 (2):783–796, 02 2017. ISSN 0001-4966. doi: 10.1121/1.4974825. URL https://doi.org/10.1121/1.4974825.

Zachary Novack, Koichi Saito, Zhi Zhong, Takashi Shibuya, Shuyang Cui, Julian McAuley, Taylor Berg-Kirkpatrick, Christian Simon, Shusuke Takahashi, and Yuki Mitsufuji. Flashfoley: Fast interactive sketch2audio generation. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15497–15501, 2026. doi: 10.1109/ICASSP55912.2026.11465146.

Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryotaro Nagase, Takahiro Fukumori, and Yoichi Yamashita. Esc-50-voice: Dataset of vocal imitation for environmental sound in esc-50, may 2024a. URL https://doi.org/10.5281/zenodo.11385662.

Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryotaro Nagase, Takahiro Fukumori, and Yoichi Yamashita. Environmental sound synthesis from vocal imitations and sound event labels. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 411–415, 2024b. doi: 10.1109/ICASSP48485.2024.10446906.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press, 2015. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Riki Takizawa, Shigeyuki Hirai, Asako Kanezaki, and Hitoshi Suda. Pronounse evaluation dataset for explosion sound synthesis, may 2026a. URL https://doi.org/10.5281/zenodo.20283840.

Riki Takizawa, Shigeyuki Hirai, Asako Kanezaki, and Hitoshi Suda. Sound effect synthesis from language-independent mimicry: a case study on explosion sounds. *APSIPA Transactions on Signal and Information Processing*, 15(1):464–487, 08 2026b. ISSN 2048-7703. doi: 10.1108/ATSIP-02-2026-0010. URL https://doi.org/10.1108/ATSIP-02-2026-0010.

Ric Viers. *The Sound Effects Bible: How to Create and Record Hollywood Style Sound Effects*. Michael Wiese Productions, Studio City, CA, USA, 2008.

A Evaluation of Woosh-CLAP Fine-Tuning

A.1 Retrieval with SFX Metadata

To assess whether fine-tuning Woosh-CLAP on SFX metadata improves the alignment between SFX recordings and their metadata descriptions, the publicly available pretrained Woosh-CLAP, denoted as *Pretrained*, is compared with a model whose text encoder is fine-tuned on SFX metadata, denoted as *Fine-tuned*. The audio encoder is kept frozen during fine-tuning, as described in Sec. 3.2.

Retrieval performance is evaluated in both directions. In text-to-audio (T2A) retrieval, each unique metadata description is used as a query to retrieve its associated SFX recordings. A retrieval is regarded as correct if at least one SFX associated with the query metadata appears within the top-K results. In audio-to-text (A2T) retrieval, each SFX recording is used as a query to retrieve its associated metadata description. Following the evaluation protocol of Woosh (Hadjeres et al., 2026), recall at ranks 5 and 10 (R@5 and R@10) is reported for both retrieval directions.

Table 5 summarizes the results. Fine-tuning improves all four retrieval metrics on the KLDataset validation and test splits. For example, on the KLDataset validation split with unseen metadata, T2A R@5 increases from 0.367 to 0.700, and A2T R@5 increases from 0.253 to 0.496. Similar improvements are observed on the Explosion Dataset validation split. On ESC-50-Voice, T2A performance remains unchanged, while A2T performance improves slightly, indicating that retrieval performance based on environmental-sound class labels is preserved after fine-tuning.

The results on Explosion Test differ from the general trend. Fine-tuning decreases both T2A metrics and A2T R@5, although A2T R@10 increases from 0.310 to 0.360. Overall, fine-tuning Woosh-CLAP on SFX metadata substantially improves metadata–audio alignment on KLDataset and the Explosion Dataset validation split, although the improvement does not generalize uniformly across metadata domains.

Table 5: SFX-metadata retrieval performance of the publicly available pretrained Woosh-CLAP and the model fine-tuned on SFX metadata. Following the Woosh evaluation protocol, recall at ranks 5 and 10 is reported for text-to-audio (T2A) and audio-to-text (A2T) retrieval. Higher values are better. The best result for each split is shown in bold.

Dataset / Split	Model	T2A		A2T	
		R@5	R@10	R@5	R@10
KLDataset Val (CLAP-unseen)	Pretrained	0.367	0.433	0.253	0.358
	Fine-tuned	**0.700**	**0.750**	**0.496**	**0.635**
KLDataset Test (CLAP-seen)	Pretrained	0.140	0.226	0.147	0.248
	Fine-tuned	**0.619**	**0.781**	**0.740**	**0.867**
KLDataset Test (CLAP-unseen)	Pretrained	0.362	0.466	0.183	0.201
	Fine-tuned	**0.810**	**0.879**	**0.489**	**0.541**
Explosion Dataset Val	Pretrained	0.099	0.155	0.113	0.181
	Fine-tuned	**0.416**	**0.565**	**0.400**	**0.513**
Explosion Test	Pretrained	**0.600**	**0.800**	**0.260**	0.310
	Fine-tuned	0.333	0.533	0.190	**0.360**
ESC-50-Voice Test	Pretrained	**1.000**	**1.000**	0.923	0.923
	Fine-tuned	**1.000**	**1.000**	**0.931**	**0.931**

Table 6: Retrieval performance on 874 available environmental sound–caption pairs from the AudioCaps test set. Following the Woosh evaluation protocol, recall at ranks 5 and 10 is reported for text-to-audio (T2A) and audio-to-text (A2T) retrieval. Higher values are better. The best result for each metric is shown in bold.

Model	T2A		A2T	
	R@5	R@10	R@5	R@10
Pretrained	**0.375**	0.502	0.254	0.383
Fine-tuned	0.366	**0.515**	**0.356**	**0.507**

A.2 Retrieval with General Audio Captions

To examine whether fine-tuning on SFX metadata affects retrieval performance for general natural-language descriptions of environmental sounds, the Pretrained and Fine-tuned models are evaluated on the AudioCaps test set (Kim et al., 2019). The evaluation uses 874 available environmental sound–caption pairs. In T2A retrieval, each caption is used as a query to retrieve its paired environmental sound, whereas in A2T retrieval, each environmental sound is used as a query to retrieve its paired caption. Following the evaluation protocol of Woosh (Hadjeres et al., 2026), R@5 and R@10 are reported for both retrieval directions.

Table 6 shows the results. For T2A retrieval, fine-tuning slightly decreases R@5 from 0.375 to 0.366, while increasing R@10 from 0.502 to 0.515. In contrast, A2T retrieval improves substantially: R@5 increases from 0.254 to 0.356, and R@10 increases from 0.383 to 0.507. These results indicate that fine-tuning on SFX metadata largely preserves the T2A retrieval performance of the pretrained model for general natural-language captions while improving its A2T retrieval performance. Thus, no substantial degradation in general environmental sound–language alignment is observed after adaptation to the SFX-metadata domain.

B Evaluation of the Pretrained Woosh-AE

Because Woosh-Mimic generates SFXs in the latent space of the pretrained Woosh-AE and decodes them into waveforms using its decoder, the reconstruction performance of Woosh-AE is evaluated on the SFX collections used in this study. Each reference SFX is encoded and reconstructed using

Table 7: Reconstruction performance of the pretrained Woosh-AE. FAD is computed between the original and reconstructed SFX distributions. For KLDataset, FAD is computed separately for each constituent commercial sound library. Lower values are better.

Dataset	Sound library	FAD ↓
KLDataset	Krotos Battle Bundle	0.058
	Krotos Whooshes	0.162
	SoundMorph Robotic Lifeforms 1	0.306
	Krotos Magic	0.379
	Paul Stoughton Pulse Energy Weapons	0.699
	SoundMorph Energy	0.994
Explosion Dataset	–	0.145
Explosion Test	–	0.520
ESC-50-Voice	–	0.270

the pretrained Woosh-AE without additional fine-tuning. Reconstruction performance is measured using Fréchet Audio Distance (FAD) (Kilgour et al., 2019) between the distributions of the original and reconstructed SFXs. Audio embeddings are extracted using PANNs Wavegram-Logmel (Kong et al., 2020), following the FAD evaluation in the main experiments. Lower FAD indicates a smaller distributional difference between the original and reconstructed SFXs.

For KLDataset, FAD is computed separately for each of its six constituent commercial sound libraries rather than after pooling all recordings. Table 7 shows the reconstruction results. FAD is below 1.0 for all evaluated datasets and sound libraries, ranging from 0.058 for Krotos Battle Bundle to 0.994 for SoundMorph Energy. These results indicate that the pretrained Woosh-AE introduces only a small distributional change when encoding and reconstructing the SFXs considered in this study. We therefore use the pretrained Woosh-AE without additional fine-tuning. Because FAD measures similarity between distributions, these results do not directly quantify the reconstruction fidelity of individual SFX recordings.

C Implementation Details

Acoustic features and Mimic Encoder. Vocal-mimic recordings are converted to mono at 48 kHz and truncated to a maximum duration of 12 s. Acoustic features are extracted with a 960-sample window and a 480-sample hop using 128 Mel bins. Pitch and voicing probability are estimated using pYIN (Mauch & Dixon, 2014) with a 1,920-sample frame length and a pitch range from C2 to C7. The 15-dimensional representation consists of loudness, log-frequency pitch, log spectral centroid, their first-order deltas, voicing probability, and eight Mel-band envelopes. The features are standardized using statistics computed from the training set.

The Mimic Encoder first projects the acoustic features to 256 dimensions and applies sinusoidal positional encoding, followed by a four-layer Transformer encoder with eight attention heads and a feed-forward dimension of 1,024. Cross-attention pooling with 32 learnable queries converts the variable-length sequence into 32 fixed-length tokens, which are subsequently projected to the 1,024-dimensional hidden space of Woosh-Flow. Dropout is set to 0.1. Padding masks are applied during both Transformer encoding and cross-attention pooling.

Training. Training was performed on a single NVIDIA H200 GPU. Woosh-Mimic is trained for 100 epochs with a batch size of 32 using Adam with a peak learning rate of 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and zero weight decay. A linear warm-up of 1,000 steps is followed by cosine learning-rate decay. Training uses bfloat16 mixed precision and gradient clipping with a maximum norm of 1.0.

Reference audio latents and fine-tuned Woosh-CLAP text tokens are precomputed. The pretrained Woosh-AE, the fine-tuned Woosh-CLAP, and all pretrained parameters of Woosh-Flow are kept frozen. Only the Mimic Encoder, the newly added mimic-specific Modality Blocks, and the learnable null-mimic condition are optimized. For each training sample, the entire mimic condition is

replaced with the learned null condition with a probability of 0.1, while the text condition is retained. Overlength reference latents are randomly cropped during training and center-cropped during validation, while the complete vocal-mimic feature sequence is retained. Padding positions in the reference latent are excluded from the flow-matching loss. The checkpoint with the lowest loss on the KLDataset validation split is selected for evaluation.

Inference. Generation uses both SFX-metadata and vocal-mimic conditioning. The classifier-free guidance convention is defined as

$$v_g = v_{\text{text,mimic}} + g\left(v_{\text{text,mimic}} - v_{\text{text,null}}\right), \tag{2}$$

where the text condition is shared between the two branches and the mimic condition is replaced with the learned null condition in the latter. Thus, $g = 0$ corresponds to ordinary conditional generation, which is used in all experiments. Flow integration is performed using the adaptive Dormand–Prince solver (`dopri5`) with relative and absolute tolerances of 10^{-3}. The generated latent length is determined by the input vocal-mimic feature length. If the feature sequence exceeds the maximum sequence length supported by Woosh-Flow, it is deterministically center-cropped. The resulting latent is decoded into a waveform using the pretrained Woosh-AE decoder.