GENERATIVE SOUND MASKING

Anonymous authors
Paper under double-blind review

ABSTRACT

A common way to reduce the salience of disruptive background sounds is to perform sound masking: using a pleasant sound, like pink noise, to conceal the undesirable sounds. We propose a method that learns to mask future background sounds. First, we apply a novel diffusion guidance method to obtain sounds that successfully mask a given background sound, which we use as synthetic training data. Second, we use this data to train a causal diffusion model to generate mask sounds that successfully conceal future background noises. Our online model reduces background-identification accuracy in an audio-language model more than background-independent masking baselines. Additional automated metrics assess the natural-sound content and distributional similarity of the resulting mixtures. Audio examples: `https://king-crimson-requiem.github.io/generative-sound/`.

1 INTRODUCTION

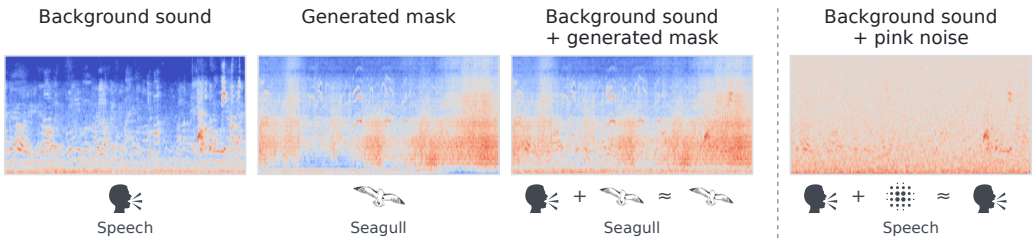

Figure 1: **Generative sound masking.** We *mask* a given background sound (left) by causally generating a sound that conceals it when the two are mixed together (middle). This conceals the background sound better than masking it with pink noise (right). *Please refer to the supplementary material for audio results.*

The world is full of distracting sounds. One's pleasant stay at a hotel, for example, might be interrupted by a loud conversation from a neighboring room, people walking in the hallway, or ambulance sirens emerging from the window. A common solution is *sound masking*: playing loud but pleasing sounds that reduce the saliency of these distracting *background noise*. Popular choices, widely used in practice by white noise machines and wellness apps Calm.com, Inc. (2024); Pigeon (2024); Endel Sound GmbH (2024), include pink noise and recordings of rain and flowing water. This phenomena, however, remains poorly understood: when does (or doesn't) one sound conceal another?

Inspired by this problem, we ask whether it is possible to *generate* sound that successfully masks background noise. Given the background noise observed so far, $\mathbf{x}_{\texttt{bg}}(t)$, we aim to generate a sound $\mathbf{x}_{\texttt{mask}} = f_\theta(\mathbf{x}_{\texttt{bg}}(t))$ that reduces perceptibility of events in future audio $\mathbf{x}_{\texttt{bg}}(t+\Delta_t)$ when both sounds are played at once: $\mathbf{x}_{\texttt{mix}}(t+\Delta_t) = \mathbf{x}_{\texttt{mask}}(t+\Delta_t) + \mathbf{x}_{\texttt{bg}}(t+\Delta_t)$.

This problem requires *causal* generation: we must generate mask sounds now that will be effective when played in the future. Solving this problem thus requires an understanding of which sounds will mask the *distribution* of possible background sounds we will encounter.

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.

One of the major challenges is the lack of "ground truth" mask sounds, making the problem unsuitable for direct supervised learning. Instead, we train a model to subvert the predictions of a pretrained audio diffusion model, which we use as a proxy for human auditory perception. We generate synthetic examples of background/mask pairs using a novel diffusion guidance method. Given a background noise, we sample a new sound using a text prompt that describes a pleasant audio event (e.g., flowing water). In each step of the reverse diffusion process, we impose a constraint that ensures that the generated sound does not significantly change when it is mixed with the background noise. Once we have obtained a large number of these synthetic examples, we train a causal diffusion model to map the background noise to the mask sound.

Sound masking has a variety of applications, such as in assistive technology Park et al. (2024), preserving the privacy by reducing the intelligibility of speech Tung & Shin (2019), and improving sleep Williamson (1992); Stanchina et al. (2005); Nasari et al. (2018). Existing approaches to these problems, however, largely require a human in the loop to manually select a mask sound Park et al. (2024); Zuo et al. (2026). In contrast, we propose to *learn* to perform this masking.

We evaluate masking primarily through an audio-language model that identifies the original background category from the mixture in a four-way forced-choice task. Our online model yields lower identification accuracy than background-independent baselines, including white noise and randomly selected naturalistic sounds. We complement this evaluation with CLAP-based category ranking and distributional similarity metrics.

2 Related Work

Sound and noise masking. One approach to sound masking is to select and combine recorded sounds to mask unwanted noise. Interactive systems allow users to customize mask sounds and recommend combinations based on recorded environmental audio (Park et al., 2024), while automated methods jointly select a mask sound and its playback level to improve the pleasantness of the current soundscape (Watcharasupat et al., 2022). Other work explores generating audio that blends with environmental noise, including personalized music generated through an offline process (Zuo et al., 2026). We bring generative sound masking to an online, causal setting: our model uses past background audio to generate pleasant, naturalistic mask sounds for upcoming sounds, without access to future observations. Another line of work focuses on the design of masking signals. A common method for sound masking is to add pure random noise, such as white or pink noise. More adaptive methods account for the target speech's temporal, spectral, and spatial characteristics (Hui et al., 2021), or shape a mask sound's spectrum so that its masking curve approximates the target speech spectrum (Laffitte et al., 2025). From a perceptual perspective, sound masking involves not only making unwanted sounds less audible, but also how listeners interpret the resulting mixture (Moore, 2012; Yost, 2021), how pleasant they find it, and how it affects their attention (Axelsson et al., 2010; Newbold et al., 2017; Haapakangas et al., 2011). These considerations motivate exploring natural sounds as mask sounds (Haapakangas et al., 2011).

Informational masking. Informational masking occurs when competing sounds interfere with a listener's ability to distinguish or attend to a target sound, beyond the effects of acoustic overlap (Durlach et al., 2003; Kidd et al., 2008). Psychological studies examine how uncertainty about the target and similarity between the target and competing sounds contribute to this difficulty (Neff & Green, 1987; Durlach et al., 2003; Kidd et al., 2008). Other work aims to generate sounds that disrupt automatic speech recognition (ASR) systems (Carlini & Wagner, 2018; Yang et al., 2019; Qin et al., 2019; Du et al., 2020). Predictive attacks generate perturbations for future speech to disrupt automatic speech recognition in real time (Chiquier et al., 2022). Our work instead targets human recognition of diverse environmental sounds using naturalistic mask sounds.

Text-to-audio generation. There have been a variety of recent approaches that can successfully generate sound from text descriptions. These include autoregressive models over discrete audio tokens (Borsos et al., 2023; Kreuk et al., 2023) and diffusion models, such as methods based on latent codes computed from mel spectrograms (Liu et al., 2023a; 2024; Xue et al., 2024; Huang et al., 2023; Hai et al., 2025) or raw waveforms (Evans et al., 2025). These architectures also support audio editing from text instructions (Wang et al., 2023b) and video-to-audio synthesis via latent diffusion or flow matching matching (Luo et al., 2023; Wang et al., 2024; Liu et al., 2023b; Lipman et al., 2023).

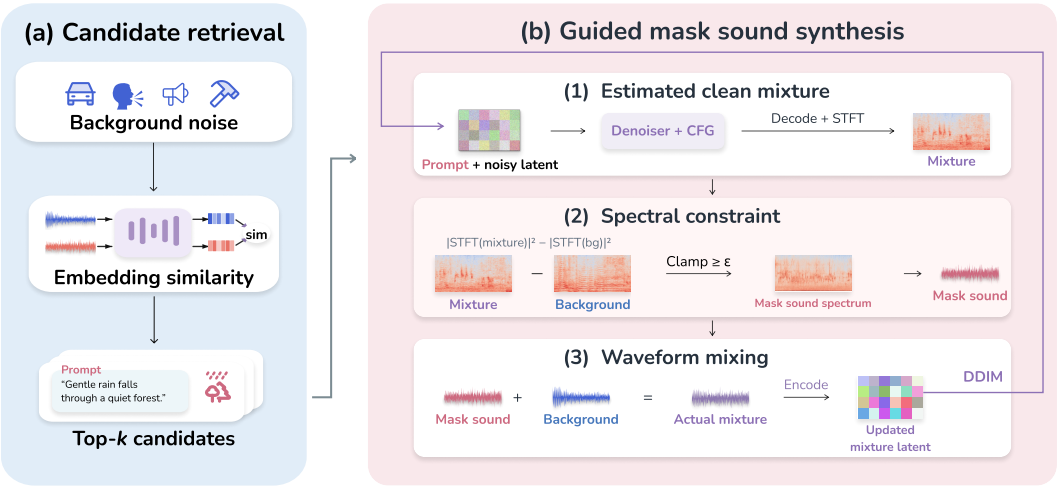

Figure 2: **Generating synthetic sound masking pairs.** (a) We retrieve the top-k audio candidates by embedding similarity to the background and reuse their prompts and random seeds. (b) At each diffusion timestep, we (1) obtain an estimated clean mixture from the noisy latent and prompt, (2) subtract the fixed background power spectrum and reconstruct a mask sound from the clamped residual, and (3) add the mask sound and background waveforms and re-encode the actual mixture. The updated latent representation of the mixture is used in the DDIM update to the next timestep. The background remains fixed.

Our approach is based on diffusion modeling. We generate synthetic sound masking examples by guiding a pretrained diffusion model to generate audio that masks background sounds.

Diffusion guidance methods. Diffusion models (Ho et al., 2020; Song et al., 2021b) and latent diffusion (Rombach et al., 2022) have become standard generative backbones. Conditioning is typically injected through classifier or classifier-free guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Inference-time guidance applies either differentiable objectives (Meng et al., 2022; Poole et al., 2023; Bansal et al., 2023; Song et al., 2023; Chung et al., 2023; Song et al., 2024) or direct manipulation of the estimated clean signal (Lugmayr et al., 2022; Wang et al., 2023a). DDRM (Kawar et al., 2022) uses a pretrained diffusion model as a generative prior for solving linear inverse problems. Multi-constraint composition combines several conditioning signals, either at inference (Liu et al., 2022; Geng et al., 2024; Chen et al., 2024) or through trained architectures (Zhang et al., 2023; Brooks et al., 2023; Saharia et al., 2022). Inspired by these approaches, we design a diffusion guidance method that makes the generated sound resemble the background noise by imposing the spectral content from the background on the generated sound during each reverse diffusion step. We then pair the generated sounds with their corresponding background audio to train a causal masking model, inspired by work in the visual domain that uses generative models to construct paired training data for downstream tasks (Brooks et al., 2023; Yu et al., 2025; Zhang et al., 2021; Wu et al., 2023a).

3 Method

We train a causal model that generates a mask sound for a given background sound. The mask sound is the foreground audio added to the background. Given a background clip $\mathbf{x}_{\texttt{bg}}$, we generate a mask sound $\mathbf{x}_{\texttt{mask}}$ and evaluate the mixture $\mathbf{x}_{\texttt{mix}} = \mathbf{x}_{\texttt{bg}} + \mathbf{x}_{\texttt{mask}}$. We construct synthetic examples of background sounds and effective mask sounds for them, then use these synthetic examples as supervision for our causal generative model after filtering them for quality using automated metrics.

3.1 Synthetic Data Generation

Figure 2 illustrates candidate retrieval (a) and guided mask sound synthesis (b). We construct synthetic background–mask sound pairs by retrieving audio candidates similar to each background and reusing

their prompts and random seeds for guided mask sound generation. We then filter the generated mixtures and select at most one pair per background.

3.2 Prompt Retrieval via Audio Embedding Similarity

Algorithm 1 Reverse diffusion process with sound masking guidance.

```
def generate_mask(bg_sound, prompt, seed):
  # Gaussian noise from retrieved seed
  z = sample_gaussian_latent(seed)

  # Fixed background power
  bg_power = abs(stft(bg_sound))**2

  # Reverse diffusion process for sound
      mixture
  for t in timesteps:
    # Obtain the estimated clean mixture
    z_clean = estimate_clean(z, t, prompt)
    mixture = decode_waveform(z_clean)
    magnitude = abs(stft(mixture))

    # Spectral subtraction
    mask_power = (magnitude**2 - bg_power)
    mask_power = mask_power.clamp(min=eps)
    mask_sound = reconstruct_waveform(
        mask_power.sqrt())

    # Guide with the actual mixture
    mixture = bg_sound + mask_sound
    z_clean = encode_waveform(mixture)

    # Update using the modified mixture
    z = ddim_step(z, z_clean, t)

  return mask_sound
```

Figure 2(a) shows the retrieval stage, where k denotes the number of retained candidates.

Building the candidate prompt pool. We construct a bank of 23,750 diverse natural-sound descriptions from filtered captions and generated prompts, including AudioCaps 2.0 (Kim et al., 2019). Appendix B details prompt generation and curation.

For each prompt, we generate ten audio clips using AudioLDM2 (Liu et al., 2024), yielding a pool of 237,500 candidates. We index each clip separately and associate it with its prompt and seed.

Retrieving the top examples. Each generated audio clip is a retrieval candidate. We rerank these examples using an approach similar to CLIP reranking (Ramesh et al., 2021) audio-language joint embedding (Dinkel et al., 2024).

For each retrieved audio candidate, we reuse its prompt as the text condition and its random seed to initialize a separate guided generation.

3.3 Mask Guidance

For each retrieved audio candidate, we generate a mask sound by guiding a pretrained audio diffusion model. Our goal is for the mask sound and background, when played together in a mixture, to sound like the natural sound described by the retrieved prompt.

Our approach works by directly generating the *mixture*, $\mathbf{x}_{\texttt{mix}} = \mathbf{x}_{\texttt{mask}} + \mathbf{x}_{\texttt{bg}}$ using diffusion, then extracting the mask sound from it afterwards. We initialize the reverse diffusion process such that it reproduces the retrieved audio candidate, by reusing the prompt and random noise that generated it. During each step of the reverse diffusion process, we impose a constraint that forces the generated sound to contain the spectral energies that are present in $\mathbf{x}_{\texttt{bg}}$. This approach resembles work on solving inverse problems using diffusion priors (Chung et al., 2023; Wang et al., 2023a; Song et al., 2024). Specifically, before each diffusion update, we force the generated sound to contain *at least* the same amount of energy as $\mathbf{x}_{\texttt{bg}}$ in each entry of its spectrogram.

Guiding the diffusion model. At each diffusion timestep, we obtain an estimated clean mixture from the current noisy latent via a one-step prediction (Ho et al., 2020) and the retrieved prompt (Figure 2(b)). We estimate the mask sound by subtracting the fixed background power spectrum from the estimated mixture power and clamping the residual. We then add this mask sound to the background in the waveform domain and use the re-encoded mixture to guide the next DDIM step.

We return the estimated mask sound, after adjusting its loudness so that its mixture with the fixed background reaches the target loudness. Appendix C provides the sampling and implementation details. We describe the sampling procedure in Algorithm 1 and illustrate each step of the reverse diffusion process in Figure 2(b).

3.4 Candidate Filtering and Selection

After generating the sound masks, we select the best-performing ones by finding the ones that pass a series of tests (Figure 3(c)).

First, we test whether sound-to-prompt similarities for two embedding models (PE-AV (Vyas et al., 2026) and LAION-CLAP (Wu et al., 2023b)) exceed a threshold. Second, we use BEATs (Chen et al., 2023) and PaSST (Koutini et al., 2022) to score how much mixing reduces evidence for the background classes and increases evidence for the target sound classes. We require a combined improvement in background suppression and target-class evidence. Third, we check whether the change from background to mixture in the audio embedding aligns with the change from the background description to the target prompt in the text embedding.

For candidates passing these checks, an audio-language judge compares the mixture against the original background and assesses whether the requested natural mask sound is audible and the background source is no longer recognizable. Both conditions are required. We rank accepted candidates by mask sound–prompt similarity. At most one pair is selected per background; backgrounds without a surviving candidate are excluded. Appendix D specifies the score definitions, label construction, acceptance rules, and alternative classification criteria.

4 Training a Masking Model

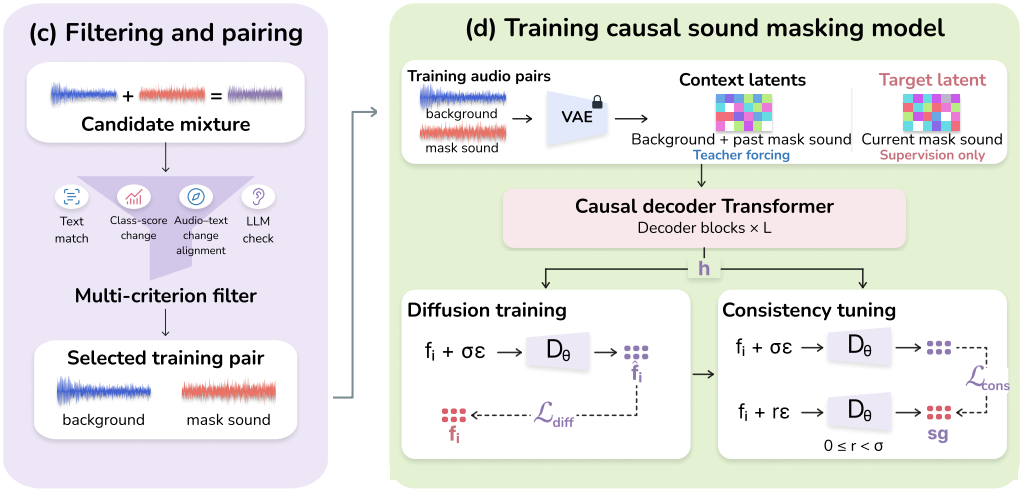

Figure 3: **Training a causal sound masking model.** (c) After guided generation, we filter the candidate mixtures for each background and select at most one background–mask sound pair. (d) A frozen audio VAE encodes the selected pairs into latents. The causal transformer conditions on the available background and past mask sound latents to predict the current mask sound latent with a diffusion head. We first train the diffusion model, then apply consistency tuning for single-step generation.

The offline energy-guided generation method in Sec. 3.3 produces ($\mathbf{x}_{\texttt{bg}}$, $\mathbf{x}_{\texttt{mask}}$) pairs but cannot be used to effectively mask sounds, because: (i) it is not causal, as the full background sound must be available before the generation process; (ii) it can be slow, since every generation requires sampling a lengthy audio from a diffusion model; (iii) like other diffusion guidance methods, the guidance can sometimes fail to successfully steer the generated sound.

We aim for our masking model to run causally on a continuously-observed audio stream, with low latency. We train a model similar to that of Saito et al. (2024). First (Sec. 4.1), we train a causal autoregressive model on these pairs. Second (Sec. 4.2), we distill this model into a single-step consistency model using easy consistency tuning (ECT) Geng et al. (2025) for real-time prediction.

4.1 Causal Sound Masking

We train a latent diffusion model to causally generate the mask $\mathbf{x}_{mask}$ from the background sound $\mathbf{x}_{\texttt{bg}}$ (Figure 3(d)). In other words, at step t, only the past background frames $\mathbf{x}_{\texttt{bg}}(1:t)$ are available, and model must generate $\mathbf{x}_{\texttt{mask}}(t+1)$ before observing $\mathbf{x}_{\texttt{bg}}(t+1)$. Background and previous mask

sound latents are provided as context, with ground-truth previous mask sound latents supplied by teacher forcing during training. As in Saito et al. (2024), we fine-tune a consistency head (based on ECT) that performs one-step audio generation.

Our causal model combines a causal transformer with a diffusion head. The conditioning stream consists of background audio latents, which we interleave with mask sound latents in a single causal context.

Audio latent space. Modeling raw waveforms requires a long transformer context. We therefore use a frozen audio VAE Saito et al. (2024) to compress waveforms into a compact latent sequence (Appendix E).

Causal transformer backbone. The background and masking latents are interleaved in a single sequence as $[\texttt{BOS}, \mathbf{z}_{\texttt{bg}}^{(1)}, \mathbf{z}_{\texttt{mask}}^{(1)}, \mathbf{z}_{\texttt{bg}}^{(2)}, \mathbf{z}_{\texttt{mask}}^{(2)}, \ldots]$, where BOS is a specific latent vector representing the start of the sequence. The causal transformer backbone aggregates the past context into a conditioning vector.

Diffusion head. We estimate each frame with a small diffusion head, formulated as in Karras et al. (2024), to avoid compounding errors during autoregressive generation.

Training. We train the causal model using unconditional pretraining followed by conditional finetuning on the paired data; Appendix E provides the training procedure, augmentation, and implementation details.

4.2 Real-Time Sound Masking

The model in Sec. 4.1 is causal but still runs multi-step diffusion per latent frame, which is too slow for real-time deployment. We post-train our model using Easy Consistency Tuning (ECT) Geng et al. (2025). ECT replaces the diffusion head's iterative denoising with a single learned mapping from a noisy mask sound latent to an estimate of its clean counterpart. We run ECT distillation on the same paired data while keeping every other component unchanged. The resulting model predicts $\mathbf{x}_{\texttt{mask}}(t)$ within $1-4$ number of function evaluations (NFEs) while preserving the causal inference. This enables the model to perform real-time sound masking.

5 Experiments

We evaluate the quality of the generated masks for a variety of different background sounds.

5.1 Data and Generation Configuration

For background sounds, we use 48,840 background clips from SONYC (Cartwright et al., 2020), TAU (Mesaros et al., 2019), UrbanSound8K (Salamon et al., 2014), BABYCRY (Bonafos et al., 2026), and NOTSOFAR-1 (Vinnikov et al., 2024). Audio is mono at 16 kHz with duration 10.24 s. We retrieve 15 candidates per background from the 237,500-clip pool and select at most one pair after filtering. Appendices C and F give generation settings and data details. **Model implementation.** Appendix E details the architecture and training schedules. Quantitative results for our model are obtained by predicting the full latent sequence and then decoding it in a single pass with the (non-causal) SoundCTM VAE. Note that the same VAE supports per-frame causal decoding for online deployment.

5.2 Evaluation Setup

Evaluation protocol. We evaluate all methods on 3,919 background clips randomly sampled from the paired dataset. We use a common audio duration across methods and normalize background and mixture audio to -30 and -23 LUFS, respectively, before computing metrics. Additional evaluation details are provided in Appendix F.

Baselines. We compare against three background-independent masking strategies: white noise, random excerpts from seven nature recordings obtained from Calm Sound (Calm Sound, 2024), and random clips from a pool of approximately 150,000 unconditional naturalistic sounds. The

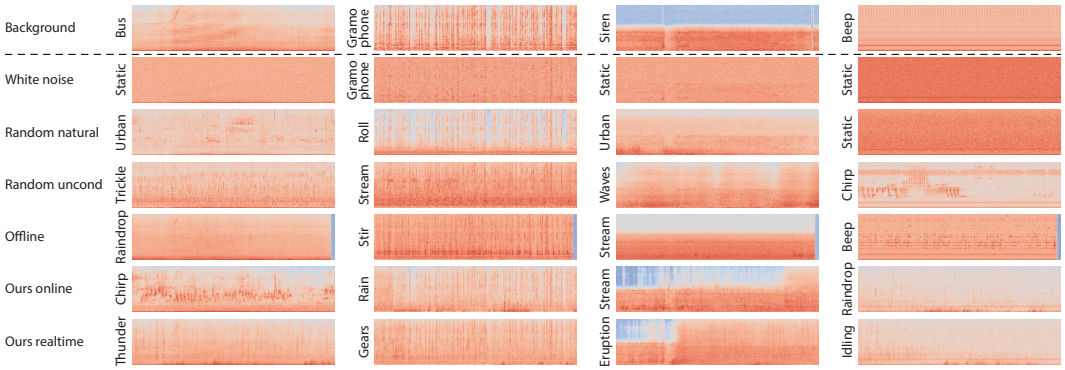

Figure 4: **Qualitative comparison of masking methods.** Each column shows a representative background clip (top row) and the mixture obtained by adding the mask sound from each method (rows below). The label on the left of every spectrogram is the top-1 CLAP zero-shot classification category. Energetic baselines (*White noise*) push the mixture toward broadband classes such as *Static*; random-natural and random-unconditional sometimes result in natural categories. Our online and real-time models maintain a balance between masking ability and naturalism.

nature recordings approximate the audio served by white-noise machines and wellness apps (Pigeon, 2024; Endel Sound GmbH, 2024). These comparisons assess the benefit of generating mask sounds conditioned on the background rather than playing a fixed or independently sampled sound. We also evaluate our offline energy-guided generation (Sec. 3.3).

Offline oracle baselines. We finetune three pretrained generative audio models on the same $(\mathbf{x}_{\texttt{bg}}, \mathbf{x}_{\texttt{mask}})$ pairs: Stable Audio Open (Evans et al., 2025), AudioLDM2 (Liu et al., 2024), and Frieren (Wang et al., 2024). Each model accepts a background audio embedding as conditioning input while preserving the rest of its base architecture, and is trained to predict the full $\mathbf{x}_{\texttt{mask}}$. These models observe the complete background clip and serve as non-causal reference points for our causal and real-time models. Their different backbones and latent spaces also allow us to assess whether the generated training pairs are useful across architectures. Appendix E provides the training details.

Our models. We evaluate both our causal autoregressive diffusion model (Sec. 4.1) and its ECT-distilled variant (Sec. 4.2). Both generate masking frames conditioned on the available background and previously generated masking frames; the distilled variant reduces the sampling cost to support real-time generation.

5.3 Metrics

We focus on LLM-based background identification to assess masking performance, complemented by metrics of natural-sound content and distributional similarity in the resulting mixture.

Forced-choice identification. We test whether an audio-language model can identify the background category from the mixture. We pose this as a four-way forced-choice task: given $\mathbf{x}_{\texttt{mix}}$, the model selects the category description that best matches the detectable sound in the clip. The four options are the GT background category plus three random distractors, drawn from a curated 474-leaf subset of the AudioSet ontology (Gemmeke et al., 2017). The GT background category is assigned by CLAP zero-shot classification: we encode $\mathbf{x}_{\texttt{bg}}$ with the CLAP audio encoder (Elizalde et al., 2023) and take the leaf with highest similarity as the GT label. Clips whose GT lands in one of the 33 natural leaves (defined by human expert, including rain, ocean waves, bird chirping etc.) are excluded since the $\mathbf{x}_{\texttt{bg}}$ already belongs to the target distribution. This removes $\sim$17% of the test set and leaves 3,024 test clips. Our primary identification metric uses Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025). Lower accuracy means that the model identifies the background less often.

Audio naturalism. Our second metric evaluates how often the mixture sound is recognized not as the original background category but as a generic natural soundscape, i.e., how close the mixture has been pulled toward the natural region in the embedding space. Using the same 474-leaf taxonomy and LLM-generated descriptions as above, we encode each $\mathbf{x}_{\texttt{mix}}$ with CLAP and rank all 474 leaves

Table 1: **Main results.** Qwen3-Omni forced-choice identification accuracy (↓) is our primary masking metric for the five deployable methods. Naturalism (↑) and distributional similarity (↓) metrics are defined in Sec. 5.3. † Offline methods; * trivial wins from sampling the naturalistic reference pool.

	Forced-choice (↓)	*Audio naturalism* (↑)			*Distributional similarity* (↓)		
Method	**LLM**	**Nat.%@1**	**Nat.%@5**	**ΔRank**	**FAD**	**MMD**	**KL-P.**
Trivial baselines							
White noise	0.58	13.33	63.86	58.80	0.78	0.51	1.75
Random natural	0.55	53.01	83.86	60.57	0.45	0.31	0.73
Random uncond.	0.47	77.74*	96.26*	87.62*	0.12*	0.18*	0.09*
Energy-guided generation							
Energy-guided†	–	**69.51**	**91.83**	76.85	**0.27**	**0.27**	0.48
Offline oracles							
Stable Audio†	–	54.20	84.26	68.41	0.39	0.33	0.43
AudioLDM2†	–	51.03	84.82	68.87	0.33	0.30	**0.42**
Frieren†	–	45.54	80.32	62.80	0.50	0.38	0.62
Ours							
Online model	**0.43**	32.31	83.30	**80.46**	0.52	0.41	0.84
Real-time model	0.45	26.26	71.79	68.55	0.59	0.41	0.90

by cosine similarity. Next we compute three metrics on the 3,024-clip subset sames as above: (i) Nat.%@1 and (ii) Nat.%@5: percentage of mixture sounds whose top-1 (top-5) similar category is in the 33-leaf natural subset; and (iii) $\Delta\text{Rank}(\text{bg})$: the mean rank drop of the GT background category in mixture sound's similarity ranking, and higher drop means original background is harder to detect.

Distributional similarity. We also measure how closely the generated mixtures match a reference distribution of naturalistic audio. We compare each method's mixture distribution with this reference using three metrics: (i) FAD (Kilgour et al., 2019) and (ii) MMD (Gretton et al., 2012) computed in the CLAP audio embedding, and (iii) KL-PaSST (Koutini et al., 2022) computed in the PaSST softmax prediction. The reference distribution is a fixed pool of ∼150k naturalistic clips. All three metrics are computed on the 3,919-clip test set; lower is better.

5.4 Results

Masking performance and audio quality. Figure 4 and Table 1 show that our online model achieves the lowest LLM background-identification accuracy among deployable methods (0.43), compared with 0.58 for white noise, and the highest background rank drop among non-trivial methods. A larger background rank drop does not necessarily imply higher natural-category hit rates: the background can be outranked by categories outside the designated natural-sound subset. Its mixtures are closer to the naturalistic reference distribution than those of white noise, but farther from it than those of offline energy-guided generation and the oracle models, which observe the full background. Random unconditional audio scores best on natural-sound classification and distributional similarity because it is drawn from the naturalistic reference pool, yet masks less effectively than our online model in forced-choice identification. The real-time model uses single-step inference, with higher background-identification accuracy and slightly larger distributional distances than the online model. Appendix F.4 provides the detailed comparisons.

Ablation Study

Table 2 shows that increasing CFG from 2 to 5 worsens background identification accuracy and rank drop for both models, although several audio-quality metrics improve. We therefore use CFG $= 2$ for all main results, prioritizing masking effectiveness. Appendix F.4 gives the per-metric analysis.

6 Conclusion

In this paper, we proposed a model that learns to generate mask sounds conditioned on past background audio. We introduce an energy-guided diffusion method that generates synthetic training pairs for this problem. We then use this data to train causal masking models. Our online model achieves lower LLM background-identification accuracy than background-independent masking baselines and

Table 2: **CFG ablation.** Effect of classifier-free guidance scale on the online and real-time masking models. Bold = best CFG value within each model. Column groupings and metric directions match Tab. 1.

	Forced-choice (↓)	*Audio naturalism* (↑)			*Distributional similarity* (↓)		
CFG	**LLM**	**Nat.%@1**	**Nat.%@5**	**△Rank**	**FAD**	**MMD**	**KL-P.**
Online model							
2	**0.43**	32.31	**83.30**	**80.46**	0.52	0.41	0.84
3	0.47	35.32	80.75	73.11	0.48	**0.39**	0.78
4	0.47	**35.52**	76.92	66.76	**0.47**	**0.39**	0.74
5	0.49	34.66	74.24	63.12	0.48	**0.39**	**0.72**
Real-time model							
2	**0.45**	26.26	71.79	**68.55**	0.59	0.41	0.90
3	0.49	29.43	75.07	65.56	0.52	0.39	0.83
4	0.49	32.87	**77.05**	61.94	0.49	0.38	0.77
5	0.48	**34.19**	76.26	60.69	**0.47**	**0.37**	**0.72**

the largest background-class rank drop among non-trivial methods. The real-time variant retains much of the masking benefit while enabling single-step inference.

Limitations. One limitation is that our approach relies on being able to generate sounds that are statistically similar to the distracting background noises, which for many sounds many not always be possible. Second, our method is that it depends on the capabilities of an existing audio generative model, thus limiting the potential performance.

AI Use Statement

We used generative AI models to construct the synthetic training data described in the paper: large language models assisted with generating and filtering text descriptions, and AudioLDM2 generated candidate audio and synthetic sound-masking examples. We also used an AI assistant to draft and edit the paper text and help organize and edit the appendix.

Reproducibility Statement

The main text describes our sound-masking method and training framework. The appendix provides additional details on candidate-pool construction, energy-guided data generation and filtering, model architectures, training schedules, and evaluation procedures to support reproducibility.

REFERENCES

Östen Axelsson, Mats Nilsson, and Birgitta Berglund. A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*, 128:2836–2846, 2010. doi: 10.1121/1.3493436.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 843–852, June 2023.

Guillem Bonafos, Lény LEGO, David Cognée, Jessica Audouard, Aurélie De Vito, Jérémy Rouch, Rémi Emonet, Hugues PATURAL, and Nicolas Mathevon. Babycry-ujm-axa: A sound library of cries from 1,000 full-term babies aged between 1 and 3 days, March 2026. URL https://doi.org/10.5281/zenodo.19205929.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023.

Calm Sound. Calm sound: Free relaxing nature sound recordings, 2024. URL https://www.calmsound.com/.

Calm.com, Inc. Calm: Meditation, sleep, relaxation, 2024. URL https://www.calm.com/.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018. doi: 10.1109/SPW.2018.00009.

Mark Cartwright, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, et al. Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context. *arXiv preprint arXiv:2009.05188*, 2020.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 85045–85073. Curran Associates, Inc., 2024. doi: 10.52202/079017-2700. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9a6f6e0d6781d1cb8689192408946d73-Paper-Conference.pdf.

Mia Chiquier, Chengzhi Mao, and Carl Vondrick. Real-time neural voice camouflage. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=qj1IZ-6TInc.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. 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. 7480–7512. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. In *Interspeech 2024*, pp. 547–551, 2024. doi: 10.21437/Interspeech.2024-246.

Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, ASIA CCS '20, pp. 357–369. Association for Computing Machinery, 2020. doi: 10.1145/3320269.3384733.

Nathaniel I. Durlach, Christine R. Mason, Jr. Kidd, Gerald, Tanya L. Arbogast, H. Steven Colburn, and Barbara G. Shinn-Cunningham. Note on informational masking (l). *The Journal of the Acoustical Society of America*, 113(6):2984–2987, 2003. doi: 10.1121/1.1570435.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Endel Sound GmbH. Endel: Ai-generated soundscapes for focus, sleep, and relaxation, 2024. URL https://endel.io/.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888461.

Jort F. Gemmeke, Daniel P. W. 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, 2017. doi: 10.1109/ICASSP.2017.7952261.

Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24154–24163, June 2024.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and J Zico Kolter. Consistency models made easy. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xQVxo9dSID.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Annu Haapakangas, E. Kankkunen, Valtteri Hongisto, Petra Virjonen, David Oliva, and Esko Keskinen. Effects of five speech masking sounds on performance and acoustic satisfaction. implications for open-plan offices. *Acta Acustica united with Acustica*, 97:641–655, 2011. doi: 10.3813/AAA.918444.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. EzAudio: Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer. In *Interspeech 2025*, pp. 4233–4237, 2025. doi: 10.21437/Interspeech.2025-1137.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.379. URL https://aclanthology.org/2020.findings-emnlp.379/.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation, 2023. URL https://arxiv.org/abs/2305.18474.

C.T. Justine Hui, Moeto Ikuta, Mochinobu Obata, Yusuke Hioka, and Takayuki Arai. Spatially target adaptive speech masking – a pilot study on masking effect and annoyance. *Applied Acoustics*, 182: 108214, 2021. ISSN 0003-682X. doi: https://doi.org/10.1016/j.apacoust.2021.108214. URL https://www.sciencedirect.com/science/article/pii/S0003682X2100308X.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, June 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23593–23606. Curran Associates, Inc., 2022. doi: 10.52202/068431-1714. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/95504595b6169131b6ed6cd72eb05616-Paper-Conference.pdf.

Gerald Kidd, Christine R. Mason, Virginia M. Richards, Frederick J. Gallun, and Nathaniel I. Durlach. Informational masking. In William A. Yost, Arthur N. Popper, and Richard R. Fay (eds.), *Auditory Perception of Sound Sources*, pp. 143–189. Springer US, Boston, MA, 2008. ISBN 978-0-387-71305-2. doi: 10.1007/978-0-387-71305-2_6.

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 Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17022–17033. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient Training of Audio Transformers with Patchout. In *Interspeech 2022*, pp. 2753–2757, 2022. doi: 10.21437/Interspeech.2022-227.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CYK7RfcOzQ4.

Pierre Laffitte, {Sankha Subhra} Bhattacharjee, {Jesper Rindom} Jensen, and {Mads Græsbøll} Christensen. Using the masking curve of a masker signal to approximate the spectral shape of a target speech signal. In *Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum / EuroNoise 2025*, Forum Acusticum, pp. 125–132. European Acoustics Association (EAA), June 2025. URL https://www.fa-euronoise2025.org/fa-2025. Forum Acusticum 2025, FA2025 ; Conference date: 22-06-2025 Through 26-06-2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew 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, 2023a.

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

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 423–439, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19790-1.

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.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11451–11461, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01117. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01117.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48855–48876. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/98c50f47a37f63477c01558600dd225a-Paper-Conference.pdf.

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification, 2018. URL https://arxiv.org/abs/1807.09840.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE 2019)*, 2019.

Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

Maryam Nasari, Tahereh Najafi Ghezeljeh, and Hamid Haghani. Effects of nature sounds on sleep quality among patients hospitalized in coronary care units: A randomized controlled clinical trial. *Nursing and Midwifery Studies*, 7(1):18–23, 2018.

Donna L Neff and David M Green. Masking produced by spectral uncertainty with multicomponent maskers. *Perception & Psychophysics*, 41(5):409–415, 1987.

Joseph W. Newbold, Jacob Luton, Anna L. Cox, and Sandy J. J. Gould. Using nature-based soundscapes to support task performance and mood. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017. doi: 10.1145/3027063.3053214. URL https://doi.org/10.1145/3027063.3053214.

Anna Y Park, Andy Jin, Jeremy Zhengqi Huang, Jesse Carr, and Dhruv Jain. Masksound: Exploring sound masking approaches to support people with autism in managing noise sensitivity. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, 2024. doi: 10.1145/3663548.3675656. URL https://doi.org/10.1145/3663548.3675656.

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

Stéphane Pigeon. mynoise: Custom soundscapes for focus, relaxation, and sleep, 2024. URL https://mynoise.net/.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5231–5240. PMLR, 2019. URL https://proceedings.mlr.press/v97/qin19a.html.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393379. doi: 10.1145/3528233.3530757. URL https://doi.org/10.1145/3528233.3530757.

Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. SoundCTM: Uniting score-based and consistency models for text-to-sound generation. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. URL https://openreview.net/forum?id=MZT5hVsMOH.

Koichi Saito, Julian Tanke, Christian Simon, Masato Ishii, Kazuki Shimada, Zachary Novack, Zhi Zhong, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. Soundreactor: Frame-level online video-to-audio generation, 2025. URL https://arxiv.org/abs/2510.02110.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, pp. 1041–1044, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330633. doi: 10.1145/2647868.2655045. URL https://doi.org/10.1145/2647868.2655045.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 7624–7654, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1f09e1ee5035a4c3fe38a5681cae5815-Paper-Conference.pdf.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32483–32498. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23k.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Michael L. Stanchina, Muhanned Abu-Hijleh, Bilal K. Chaudhry, Carol C. Carlisle, and Richard P. Millman. The influence of white noise on sleep in subjects exposed to ICU noise. *Sleep Medicine*, 6(5):423–428, 2005. ISSN 1389-9457. doi: 10.1016/j.sleep.2004.12.004.

Yu-Chih Tung and Kang G. Shin. Exploiting sound masking for audio privacy in smartphones. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, pp. 257–268, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367523. doi: 10.1145/3321705.3329799. URL https://doi.org/10.1145/3321705.3329799.

Alon Vinnikov, Amir Ivry, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Peer, Xiong Xiao, Benjamin Martinez Elizalde, Naoyuki Kanda, Xiaofei Wang, Shalev Shaer, Stav Yagev, Yossi Asher, Sunit Sivasankaran, Yifan Gong, Min Tang, Huaming Wang, and Eyal Krupka. NOTSOFAR-1 Challenge: New Datasets, Baseline, and Tasks for Distant Meeting Transcription. In *Interspeech 2024*, pp. 5003–5007, 2024. doi: 10.21437/Interspeech.2024-1788.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30172–30182, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Vyas_Pushing_the_Frontier_of_Audiovisual_Perception_with_Large-Scale_Multimodal_Correspondence_CVPR_2026_paper.html.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=mRieQgMtNTQ.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 128118–128138. Curran Associates, Inc., 2024. doi: 10.52202/079017-4068. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e7384de302bef52319b5067c3115bbfb-Paper-Conference.pdf.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. Audit: Audio editing by following instructions with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 71340–71357. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e1b619a9e241606a23eb21767f16cf81-Paper-Conference.pdf.

Karn N. Watcharasupat, Kenneth Ooi, Bhan Lam, Trevor Wong, Zhen-Ting Ong, and Woon-Seng Gan. Autonomous in-situ soundscape augmentation via joint selection of masker and gain. *IEEE Signal Processing Letters*, 29:1749–1753, 2022. ISSN 1558-2361. doi: 10.1109/lsp.2022.3194419. URL http://dx.doi.org/10.1109/LSP.2022.3194419.

JW Williamson. The effects of ocean sounds on sleep after coronary artery bypass graft surgery. *American Journal of Critical Care*, 1(1):91–97, 1992.

Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1206–1217, October 2023a.

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, 2023b. doi: 10.1109/ICASSP49357.2023.10095969.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4700–4712, 2024. doi: 10.1109/TASLP.2024.3485485.

Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing audio adversarial examples using temporal dependency. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1g4E3C9t7.

William Yost. *Fundamentals of Hearing: An Introduction: Fifth Edition*. Brill, Leiden, The Netherlands, 2021. ISBN 978-90-04-50193-5. doi: 10.1163/9789004501935. URL https://brill.com/view/title/24208.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135, June 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10145–10155, June 2021.

Chi Zuo, Martin B. Møller, Pablo Martínez-Nuevo, Huayang Huang, Yu Wu, and Ye Zhu. BNMusic: Blending environmental noises into personalized music. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=IIgVYnadfR.

Appendix

This appendix provides notation, candidate-pool construction, details of energy-guided generation and candidate selection, model training settings, and additional evaluation details and analysis.

CONTENTS

A Notations **17**

B Candidate Pool Construction **17**

B.1 Prompt Filtering 17

B.2 Prompt Generation 19

B.3 Audio Candidate Generation 22

C Energy-Guided Generation Algorithm **22**

C.1 Spectral Subtraction and Phase Analysis 22

C.2 Matched Reconstruction Correction 23

C.3 Loudness Normalization and Calibration 23

D Candidate Filtering and Selection **24**

D.1 Audio Representations and Label Sets 24

D.2 Acceptance and Selection 25

D.3 Alternative Classification Criteria 26

E Model Architecture and Training **27**

E.1 Offline Oracle Baselines 27

E.2 Causal Masking Model 27

E.3 Consistency Fine-Tuning 28

F Evaluation Details **28**

F.1 Data and Evaluation Configuration 28

F.2 Audio Quality and Distributional Similarity 28

F.3 Masking Effectiveness 29

F.4 Additional Results Analysis 29

A NOTATIONS

Table 3 summarizes the symbols used in the synthetic data generation and candidate selection. Bold symbols denote vectors or tensors; lowercase $\mathbf{x}$ denotes a waveform and uppercase $\mathbf{X} = \mathrm{STFT}(\mathbf{x})$ its complex spectrum. Intermediate predictions are recomputed at each diffusion step; their step indices are omitted for readability.

Table 3: Notation for synthetic data generation and candidate selection.

Symbol	Meaning
$\mathbf{x}_{\mathtt{bg}}$	Background waveform, fixed at -30 LUFS.
$\mathbf{x}_{\mathtt{mask}}$	Gain-calibrated additive masking waveform.
$\mathbf{x}_{\mathtt{mix}}$	Physical mixture $\mathbf{x}_{\mathtt{bg}} + \mathbf{x}_{\mathtt{mask}}$.
$\mathbf{X}$, $\lvert\mathbf{X}\rvert$, $\lvert\mathbf{X}\rvert^2$	Complex STFT, elementwise magnitude, and power of $\mathbf{x}$.
$\hat{\mathbf{X}}_{\mathtt{mix}}$	Complex STFT of the predicted mixture waveform.
$\tilde{\mathbf{x}}_{\mathtt{mask}}$	Mask sound synthesized from the magnitude proposal, before gain calibration.
g	Positive linear amplitude gain applied only to the mask sound.
$\mathbf{x}_{\mathtt{ref}}$, $\mathbf{X}_{\mathtt{ref}}$	Gain-matched reconstruction reference and its STFT.
$\hat{\mathbf{z}}_0$, $\hat{\mathbf{z}}_0^{\mathtt{adj}}$	Latent representations of the estimated clean mixture before and after spectral guidance.
c, $c_{\mathtt{bg}}$	Retrieved text prompt and text describing inferred background labels.
STFT	Short-time Fourier transform.
$E(\lvert\mathbf{X}\rvert)$	Log-mel conversion followed by the VAE encoder.
$f^{\mathrm{PE}}_{\mathrm{audio}}$, $f^{\mathrm{PE}}_{\mathrm{text}}$	PE-AV audio and text embedding functions.
d	Cosine similarity between the audio and text semantic-change directions.
$\mathcal{C}_{\mathtt{bg}}$, $\mathcal{C}_{\mathtt{tgt}}$	Fixed background and target class sets; superscript `noise` restricts background labels to unwanted noise.
h, $p_h(j \mid \mathbf{x})$	Classifier identity and its predicted probability for class j given waveform $\mathbf{x}$.
$w_{h,j}$	Normalized background-confidence weight for class j under classifier h.
D_h, G_h, I	Background suppression, target gain, and combined improvement.
k	Number of classes used in the top-k naturalness aggregation.
$e_j(g)$	K-weighted mean-square value of mixture block j at gain g.
$e_{\mathtt{bg},j}$, $e_{\mathtt{mask},j}$	K-weighted block mean squares of background and uncalibrated mask sound.
$e_{\mathtt{cross},j}$	Block mean of their samplewise product; may be negative.
$\mathrm{Re}(\cdot)$, $(\cdot)^*$	Real part and complex conjugation.

Role subscripts carry over from waveforms to spectra. A hat denotes a prediction or proposal; `adj` and `ref` mean adjusted and reconstruction reference, respectively. Spectral operations are elementwise, whereas $|\mathcal{C}|$ denotes set cardinality. Indices are local: Algorithm 1 iterates over timesteps t, i, j enumerate classes in the classifier formulas, and j enumerates loudness blocks.

B CANDIDATE POOL CONSTRUCTION

The bank in Sec. 3.2 contains 23,750 prompts: 13,200 descriptions from caption filtering and generation, 9,441 accepted descriptions from 10,000 additional generated candidates, and 1,109 AudioCaps 2.0 captions. The AudioCaps captions pass semantic review and have ROUGE-L below 0.7 against previously accepted prompts.

B.1 PROMPT FILTERING

The keyword filter and language-model instructions below specify the natural-sound criteria used to construct the initial 13,200-description bank.

Keyword filter. Before invoking the LLM stage, we apply a fast regex-based keyword filter that rejects any caption whose lowercased text contains a word-boundary-anchored, case-insensitive

match against the blocklist below. Matching is implemented as a single Python regular expression; both singular and plural forms of each term are included.

- Vehicles and transit: `traffic`, `vehicle`, `car`, `bus`, `truck`, `taxi`, `automobile`, `ambulance`, `aircraft`, `airplane`, `plane`, `jet`, `train`, `railway`, `railroad`, `rail`, `motorcycle`, `motorbike`, `scooter`, `speedboat`, `boat`, `ship`, `snowmobile`.
- Engines and vehicle parts: `engine`, `motor`, `motorized`, `siren`, `horn`, `brake`, `braking`, `tire`, `tyre`, `squeal`, `squealing`, `screech`, `screeching`.
- Roads and urban infrastructure: `freeway`, `highway`, `road`, `street`, `intersection`, `crosswalk`, `roadway noise`, `urban noise`.
- Machinery and industrial: `mechanism`, `machinery`, `machine`, `mechanical`, `industrial`, `construction`.
- Human voice, speech, and singing: `speak`, `speaking`, `speech`, `talk`, `talking`, `say`, `saying`, `sing`, `singing`, `singer`, `song`, `voice`, `vocalize`, `vocalizing`, `vocal`, and `{person, people, man, woman, human}` followed by `{speak, talk, say, sing, voice}`.

LLM filter prompt (Qwen2.5-7B-Instruct). We prompt Qwen2.5-7B-Instruct with the instruction below for each candidate caption and retain only those for which the model returns `YES`. The placeholder `{prompt}` is replaced with the candidate caption at inference time.

```
You are a strict but context-aware audio-scene classifier.

Task:
- Determine whether the given audio description represents a PURE,
  PLEASANT, and NATURALLY BALANCED natural soundscape suitable for
  ambient or immersive background use

Answer:
- Answer ONLY with "YES" or "NO"
- Do NOT explain

Definition:

A valid natural soundscape must:
- Contain ONLY natural sounds from nature (Biophony or Geophony)
- Feel emotionally safe, non-threatening, and non-aggressive
- Allow natural variation, layering, and organic change
- Be suitable for ambient or immersive listening, even if moderately lively
- Not be chaotic, harsh, or disruptively attention-grabbing

Natural soundscapes do NOT need to be silent or minimal:
- They may include multiple natural sound sources
- They may include gentle dynamic change
- They must remain comfortable and non-threatening

ACCEPTABLE SOUNDS (YES):

Biophony (non-human living organisms, calm or neutral):
- Birdsong, birds chirping, or birds calling
- Overlapping bird calls or call-and-response patterns
- Insects buzzing or chirping, individually or as a natural chorus
- Frogs croaking calmly near water
- Soft, natural animal movement sounds without distress or aggression

Geophony (natural physical phenomena, non-extreme):
- Wind moving through trees, grasses, or leaves with natural fluctuations
- Light to moderate rain, including brief changes in intensity
- Distant or gentle thunder that feels atmospheric
- Streams, creeks, or gently flowing water with evolving texture
- Ocean waves with steady or breathing-like motion
- Calm natural fire sounds such as a steady campfire crackle
```

```
Multiple acceptable sounds may coexist naturally.

UNACCEPTABLE SOUNDS (IMMEDIATE "NO"):

Anthrophony (human-related sounds):
- Human voice, speech, talking, or singing
- Laughing, crying, shouting, or whispering
- Any implied human activity or human presence

Artificial or mechanical sounds:
- Machinery, engines, vehicles, or traffic
- Construction, industrial, or urban noise
- Electronic sounds or synthesized effects
- Recording artifacts or technical descriptors

Natural but unsuitable sounds:
- Aggressive, threatening, or predatory animal vocalizations
- Sounds expressing pain, fear, alarm, or distress
- Extreme or violent natural events intended to evoke danger
- Overly harsh, chaotic, or overwhelming natural sounds unsuitable
  for ambience

Mixed scenes:
- Any combination of acceptable natural sounds with any unacceptable
  sound

Meaningless or non-descriptive (IMMEDIATE "NO"):
- Descriptions that do NOT describe actual sound events
- Examples: "No sound events are listed.",
            "A sound of a chair is not being used.",
            "No instrument is playing.",
            "No effects added.",
            "No sounds are listed.",
            "There is no background sound.",
            "No sound events are described/happening."
- Any phrase that states the absence of sounds, lists nothing, or
  describes non-events (e.g. something "is not being used") must be
  answered "NO"

Decision rule:
- If ANY unacceptable sound is mentioned, answer "NO"
- Otherwise, answer "YES"

Audio description:
- "{prompt}"
```

B.2 Prompt Generation

To diversify descriptions, we sample compatible combinations of environments, sound sources, secondary layers, intensities, times, listener perspectives, and writing styles. Structural validation, semantic review, and novelty checks remove unsuitable or redundant descriptions.

LLM generation prompt (GPT-5.6 Sol). To broaden coverage of natural categories beyond what the filtered captions provide, we additionally prompt GPT-5.6 Sol with the instruction below to synthesize $4{,}000$ new audio descriptions. The model is asked to emit one description per line; we collect outputs across multiple sampling runs.

```
You are an audio prompt generator.

Task:
- Generate short audio descriptions that represent PURE, PLEASANT, and
  IMMERSIVE natural soundscapes suitable for ambient background use
```

```
Core principles:
- ONLY include natural sounds from nature (Biophony or Geophony)
- The soundscape should feel naturally balanced, soothing, and
  emotionally safe
- The soundscape may include gentle natural variation, evolving layers,
  and organic rhythm
- The description must be suitable for immersive natural ambience

Forbidden elements:
- Do NOT include humans, human voices, or human activity
- Do NOT include artificial, mechanical, industrial, or urban sounds
- Do NOT include recording artifacts or technical terms
- Do NOT include aggressive animals, predator behavior, distress,
  pain, or threat
- Avoid extreme or violent natural events
- Do NOT generate meaningless or non-descriptive prompts: every line
  must describe actual, concrete natural sounds. Forbidden examples:
  "No sound events are listed.", "A sound of a chair is not being
  used.", "No instrument is playing.", "No effects added.",
  "No sounds are listed.", "There is no background sound." -- reject
  any phrase that states the absence of sounds or describes non-events
  (e.g. something "is not being used")

Allowed sound elements:

Biophony (calm or neutral):
- Birdsong or birds chirping, including light natural overlap or
  call-and-response
- Insects buzzing or chirping, individually or as a gentle chorus
- Frogs croaking calmly near water
- Soft, non-intrusive animal movement sounds

Geophony (non-extreme):
- Wind moving through trees, grasses, or leaves with natural fluctuation
- Light to moderate rain, including brief changes in intensity
- Distant or gentle thunder that feels atmospheric
- Streams, creeks, or gently flowing water with evolving texture
- Ocean waves with slow, breathing-like motion
- Calm natural fire sounds such as a steady campfire crackle

Examples of valid audio descriptions:
- Light thunder rolls far in the distance as a gentle rainfall taps
  steadily across leaves.
- Soft ocean waves rise and fall while seabirds call faintly above
  the shore.
- A flowing creek murmurs over stones as insects hum quietly in the
  surrounding grass.
- Wind moves through tall trees in slow pulses while birds exchange
  gentle calls.
- Evening frogs croak calmly near still water beneath a soft chorus
  of insects.

Output format:
- One audio description per line
- Each description must be exactly one sentence
- Plain text only

Style guidelines:
- Keep descriptions vivid and grounded in natural realism
- Allow multiple natural sound sources to coexist harmoniously
- Vary wording, imagery, and sentence structure
- Avoid repetitive phrasing or rigid templates

Generation requirement:
```

```
- Generate 500 distinct audio descriptions
```

Controlled scene composition. We separate the choice of an acoustic scene from its linguistic realization. A seeded sampler first draws an environment and time, then selects a primary sound, an optional secondary layer, intensity, and listening perspective from fixed attribute pools. The language model expresses this specification as a caption while preserving its sound sources and their relationships. This gives explicit control over scene composition without prescribing a fixed sentence template.

Compatibility constraints. Attributes are sampled conditionally rather than independently. Habitat tags restrict sound sources to compatible environments, while cross-attribute rules constrain time, intensity, perspective, and secondary layers. For example, a dawn chorus is restricted to periods around dawn, ocean waves require a marine environment, and stream and wave sources are not paired with an overhead perspective. Secondary layers that duplicate the primary source or conflict with it are excluded. These rules constrain the sampled specification; semantic review subsequently checks the generated description for implausible combinations.

Sentence structure and opening words. We sample uniformly from seven structures: a subject–verb–object statement, a participial opening, a fronted adverbial clause, two independent clauses joined by a semicolon, two or three short sensory fragments, a locative prepositional opening, and a sentence with a subordinate clause. For participial, adverbial, and prepositional openings, the sampler also selects the required opening word or phrase. Their pools contain 12, 5, and 10 choices, respectively, such as *Sweeping*, *When*, and *Beneath*. The model must begin with the selected expression. Controlling both the structure and its opening is intended to avoid concentrating varied scenes into a few repeated sentence patterns.

Lexical style and length. Each specification independently receives one of twelve coarse lexical styles, including field-listening notes, nature-journal prose, meditation-app copy, documentary narration, and factual environmental logs. Style controls wording only: it must not introduce a narrator, observer, or human activity into the soundscape. We also sample concise, standard, and detailed length preferences with probabilities 0.4, 0.3, and 0.3, targeting approximately 6–9, 10–14, and 15–20 words. These ranges are soft preferences; natural phrasing and complete acoustic descriptions take priority. Only lengths outside 5–24 words are rejected by the length check.

Caption validation and novelty. The model returns one caption per specification, paired with its identifier. Structural checks enforce a single sentence, the word-count limits, required openings, and checkable punctuation or clause constraints, and reject prohibited content. Sensory fragments are separated by semicolons under one final period. To control lexical redundancy, each structurally valid candidate must have ROUGE-L F1 below 0.7 against the reference descriptions and previously accepted candidates; accepted captions are immediately added to this comparison set. A separate semantic review, in batches of at most 200, checks whether captions describe pleasant, concrete, physically plausible natural sounds. This review retains 9,441 of the 10,000 generated candidates.

Attribute-conditioned generation prompt. We generate captions in batches of 200 specifications using the following instruction. Each specification supplies the sampled scene attributes, lexical style, sentence structure (including any required opener), and soft length preference.

Prompt for attribute-conditioned caption generation
Generate exactly one English soundscape caption for every specification below.
Do not use tools or inspect files. Answer directly.
Hard rules:

- Describe only pleasant, concrete, actually audible natural sounds.
- No humans, voices, observers, occupations, camps, shelters, boats, equipment, machinery, urban sounds, danger, predators, distress, silence, or absent sounds.
- Lexical style controls wording only and must never appear as a character.
- Preserve the biome, primary sound, secondary layer when present, intensity, time, and listener perspective naturally. Do not invent another salient source.
- Obey the requested sentence structure. Treat length as a soft preference: natural phrasing and complete audio semantics take priority over an exact count. Only avoid extreme captions below 5 or above 24 words.
- Produce one grammatical sentence per caption. For minimalist fragments, use semicolon-separated fragments under one final period.
- For a required opening word or phrase, begin with it exactly.
- Make every caption independently worded. Avoid templates, near-duplicates, and repeated multiword phrases across the batch.
- Return only the JSON object required by the supplied schema. Use every given id exactly once, with no extra ids, and keep the ids in the supplied order.

Specifications:
{One entry per id: lexical style; biome; primary sound; optional secondary layer; intensity; time; perspective; sentence structure; soft length preference.}

B.3 Audio Candidate Generation

For each curated prompt, we generate ten mono audio clips at 16 kHz with duration 10.24 s using AudioLDM2 Liu et al. (2024) and CFG scale 40. The 23,750 prompts yield 237,500 candidates, each indexed separately with its prompt and seed.

Audio generation seeds. After curation, a deterministic sampler draws ten distinct integer seeds from $[1, 10^6]$ for each prompt. These seeds specify the audio realizations used in the retrieval pool and are separate from the random seeds used to sample caption attributes.

C Energy-Guided Generation Algorithm

Sampling settings. We use 100 DDIM steps with classifier-free guidance scale 3.5. The normalized background is supplied as `bg_sound` to Algorithm 1, and its STFT power is precomputed as `bg_power`.

Reconstruction operators. The reconstruction operators accept magnitude-only proposals as well as STFT magnitudes. All synthesized waveforms are cropped to the same background playback interval before addition or spectral analysis. We use deterministic VAE encoding and a HiFiGAN vocoder Kong et al. (2020).

C.1 Spectral Subtraction and Phase Analysis

Spectral proposal. Let $\hat{\mathbf{X}}_{\texttt{mix}}$ be the STFT of the waveform decoded from the latent representation of the estimated clean mixture, and let $\mathbf{X}_{\texttt{bg}}$ be the fixed background STFT. The proposed mask magnitude is

$$\hat{\mathbf{A}}_{\texttt{mask}} = \sqrt{\max\Big(|\hat{\mathbf{X}}_{\texttt{mix}}|^2 - |\mathbf{X}_{\texttt{bg}}|^2, \epsilon\Big)}, \qquad \epsilon = 10^{-10},$$

where the operations are elementwise. We reconstruct a mask sound waveform from this magnitude using log-mel conversion and the vocoder.

Cross-term decomposition. At a fixed time–frequency bin, let $X_{\texttt{bg}}$, $X_{\texttt{mask}}$, and $X_{\texttt{mix}}$ be the scalar entries of the corresponding complex STFTs $\mathbf{X}_{\texttt{bg}}$, $\mathbf{X}_{\texttt{mask}}$, and $\mathbf{X}_{\texttt{mix}}$. Waveform addition gives $X_{\texttt{mix}} = X_{\texttt{bg}} + X_{\texttt{mask}}$, so

$$|X_{\texttt{mix}}|^2 = |X_{\texttt{bg}}|^2 + |X_{\texttt{mask}}|^2 + 2\,\mathrm{Re}(X_{\texttt{bg}} X_{\texttt{mask}}^*). \tag{1}$$

Here Re denotes the real part and $*$ denotes complex conjugation. The cross-term is positive for constructive interference and negative for destructive interference. Direct power subtraction recovers mask sound power only if this term vanishes.

Approximation and correction. A zero expected cross-term would require additional assumptions about phase and its dependence on magnitude. Magnitude-only vocoding does not enforce those assumptions: the mask sound magnitude itself depends on the background, and the vocoder is a learned nonlinear mapping. Even an expectation of zero would not imply zero cross-term in each realized time-frequency bin. We therefore use spectral subtraction only to propose a mask sound, then synthesize it, add the unchanged background in the waveform domain, and re-analyze this physical sum. The adjusted latent and loudness measurement see the realized cross-term. This procedure neither guarantees absence of local destructive interference nor establishes perceptual masking without evaluation.

C.2 Matched Reconstruction Correction

In practice, the waveform primitives introduce reconstruction mismatch through log-mel conversion, vocoding, and VAE encoding. We correct for this using a matched reference. Let $\hat{\mathbf{z}}_0$ denote the original output of `estimate_clean`, let $\hat{\mathbf{X}}_{\texttt{mix}}$ be the STFT of $\texttt{decode_waveform}(\hat{\mathbf{z}}_0)$, and let g be the gain selected by `match_loudness` (Appendix C.3). We obtain the matched reference $\mathbf{x}_{\texttt{ref}}$ by applying `reconstruct_waveform` to $|\hat{\mathbf{X}}_{\texttt{mix}}|$ and scaling the result by g. With $\mathbf{X}_{\texttt{mix}}$ and $\mathbf{X}_{\texttt{ref}}$ denoting the STFTs of the actual mixture and matched reference, respectively, the updated latent representation of the estimated clean mixture passed to `ddim_step` is

$$\hat{\mathbf{z}}_0^{\texttt{adj}} = \hat{\mathbf{z}}_0 + E(|\mathbf{X}_{\texttt{mix}}|) - E(|\mathbf{X}_{\texttt{ref}}|). \tag{2}$$

Here $\hat{\mathbf{z}}_0$ is the original latent representation of the estimated clean mixture, and E denotes LogMel followed by the VAE encoder, so that

$$E(|\,\mathrm{STFT}(\mathbf{x})|) = \texttt{encode_waveform}(\mathbf{x}).$$

This matched difference compensates for shared reconstruction and gain effects while steering the latent toward the physical mixture. This updated latent representation then enters the ordinary DDIM update Song et al. (2021a).

C.3 Loudness Normalization and Calibration

Normalization settings. We measure integrated perceptual loudness in LUFS (loudness units relative to full scale). Before generation, we normalize the background waveform $\mathbf{x}_{\texttt{bg}}$ to -30 LUFS and keep it fixed throughout sampling. The target loudness of the actual background–mask sound mixture is -23 LUFS.

Calibration objective and timing. For an uncalibrated mask sound $\tilde{\mathbf{x}}_{\texttt{mask}}$, `match_loudness` returns $g^\star \tilde{\mathbf{x}}_{\texttt{mask}}$, where

$$g^\star = \arg\min_{g>0} \left|\mathrm{LUFS}(\mathbf{x}_{\texttt{bg}} + g\tilde{\mathbf{x}}_{\texttt{mask}}) + 23\right|.$$

Only the mask sound amplitude is adjusted. We calibrate the mask sound before waveform mixing at each sampling step and once more for the final output.

Gain search. For the fixed background $\mathbf{x}_{\texttt{bg}}$ and uncalibrated mask sound $\tilde{\mathbf{x}}_{\texttt{mask}}$, linear K-weighting permits each loudness block energy of $\mathbf{x}_{\texttt{bg}} + g\tilde{\mathbf{x}}_{\texttt{mask}}$ to be evaluated as

$$e_j(g) = e_{\texttt{bg},j} + 2g\, e_{\texttt{cross},j} + g^2 e_{\texttt{mask},j}.$$

Table 4: Target-category mapping for classification. Entries use AudioSet display names; multiple categories contribute the union of their labels. This mapping defines prompt targets, not the desirable-natural grouping used by the alternative criterion.

Prompt category	**AudioSet target labels**
Rain and thunder	Rain; Raindrop; Rain on surface; Thunderstorm; Thunder
Wind	Wind; Wind noise (microphone)
Birds	Bird; Bird vocalization, bird call, bird song; Chirp, tweet
Insects	Insect; Cricket; Mosquito; Fly, housefly
Coast and ocean	Ocean; Waves, surf
Fresh water	Water; Stream; Waterfall; Gurgling
Ice, snow, and geology	Crack; Crackle; Rumble
Frogs	Frog; Croak
Fire	Fire; Crackle

Here $e_{\mathtt{bg},j}$ and $e_{\mathtt{mask},j}$ are the mean squared K-weighted samples of the two unscaled signals in block j; $e_{\mathtt{cross},j}$ is the mean of their samplewise product and can be negative. We apply BS.1770 absolute and relative gates to these cached statistics. Because the active gate set changes with g, loudness need not vary monotonically with gain. The search starts from the mask sound level, expands its range, and refines promising candidates. The selected gain is checked on the actual float32 waveform sum; additional waveform-based refinement resolves gate transitions down to 10^{-6} dB. The target tolerance is 0.03 LU. If the search does not meet this tolerance, it returns the nearest finite evaluated result with its measured error.

Waveform output. The output is the mask sound waveform reconstructed in step (2) of the final iteration. Although the implementation subsequently completes waveform mixing and the DDIM update, these final operations do not alter the saved mask sound waveform. After the final calibration, we store the background and mask sound as unclipped float32 waveforms so that their sum preserves the calibrated mixture, and record loudness diagnostics separately.

D Candidate Filtering and Selection

We detail the five-stage cascade in Sec. 3.4. The first four stages determine acceptance; the fifth ranks only accepted candidates. A high ranking score cannot compensate for a failed acceptance test. All numerical gates use strict inequalities, so equality with a threshold does not pass.

D.1 Audio Representations and Label Sets

Scoring inputs. The filter operates on the generated mask sound $\mathbf{x}_{\mathtt{mask}}$, fixed background $\mathbf{x}_{\mathtt{bg}}$, physical mixture $\mathbf{x}_{\mathtt{mix}}$, and exact retrieved prompt c. Filtering does not change their loudness or regenerate the mixture. PE-AV and LAION-CLAP receive audio resampled to 48 kHz through their respective processors. BEATs receives 16-kHz audio with padding masks; PaSST receives 32-kHz audio center-cropped to ten seconds, or symmetrically zero-padded when shorter. The audio–text encoders are PE-AV Small and LAION-CLAP HTSAT-unfused. Their normalized embeddings give cosine similarities. PaSST class probabilities are sigmoid-transformed logits; BEATs supplies its checkpoint's class probabilities. These multilabel scores are not normalized to sum to one across classes.

Canonical classes and target labels. All event predictions are aligned to the same 527 AudioSet classes by class identifiers, not by native output position. Let $\mathcal{C}_{\mathtt{tgt}}$ denote the target label set. We first use the prompt's category metadata; when metadata is empty, deterministic keyword matching supplies categories, whose mapped label sets are combined. Table 4 gives the category-to-label mapping. A candidate with no resolved target category cannot pass the classification stage. Predictions on retrieved prompt audio do not replace the intended textual target.

Background-label fusion. Background labels are inferred once from the original background and kept fixed across its 15 candidates. We combine LAION-CLAP, BEATs, and PaSST using within-audio ranks, avoiding direct averaging of their differently scaled raw scores. For CLAP, the class text is "the sound of" followed by the AudioSet display name. For each model m, convert its 527 class scores to ordinal ranks,

$$r_m(j \mid \mathbf{x}) = \frac{\mathrm{rank}_m(j \mid \mathbf{x})}{526}, \qquad \mathrm{rank}_m(j \mid \mathbf{x}) \in \{0, \ldots, 526\}, \tag{3}$$

where larger scores have higher ranks. Average these ranks equally across the three models. The three highest-scoring labels over all 527 classes form $\mathcal{C}_{\mathtt{bg}}$; their ordered display names, joined by "and", form $c_{\mathtt{bg}}$ for audio–text change alignment scoring. Separately, the three highest-scoring labels within the unwanted-noise group form $\mathcal{C}_{\mathtt{bg}}^{\mathtt{noise}}$ for classification. The application-specific grouping contains 38 desirable natural sounds, 264 unwanted noises, and 225 ambiguous or excluded labels. Rank fusion selects labels only: the classification improvement below uses the original class probabilities, not these ranks.

D.2 Acceptance and Selection

Stage 1: mask sound–text alignment. Let s_{PE} and s_{CLAP} be the mask sound–prompt cosine similarities:

$$s_m = \cos(f^m_{\mathrm{audio}}(\mathbf{x}_{\mathtt{mask}}), f^m_{\mathrm{text}}(c)), \qquad m \in \{\mathrm{PE}, \mathrm{CLAP}\}. \tag{4}$$

The candidate passes only if $s_{\mathrm{PE}} > \tau_{\mathrm{PE}}$ and $s_{\mathrm{CLAP}} > \tau_{\mathrm{CLAP}}$. These scores assess the generated mask sound itself. Mixture–prompt similarities are auxiliary measurements and do not enter this gate or the final ranking.

Stage 2: background suppression and target-score increase. We measure paired changes between the original background and the physical mixture. For classifier $h \in \{\mathrm{BEATs}, \mathrm{PaSST}\}$, let $p_h(j \mid \mathbf{x})$ denote its predicted probability for class j given waveform $\mathbf{x}$. We weight the fixed unwanted-background labels by their confidence on the original background:

$$w_{h,j} = \frac{p_h(j \mid \mathbf{x}_{\mathtt{bg}})}{\sum_{i \in \mathcal{C}_{\mathtt{bg}}^{\mathtt{noise}}} p_h(i \mid \mathbf{x}_{\mathtt{bg}})}, \qquad j \in \mathcal{C}_{\mathtt{bg}}^{\mathtt{noise}}. \tag{5}$$

If the denominator is zero, use uniform weights on this set. Define background suppression D_h, target gain G_h, and their combined score I as

$$D_h = \sum_{j \in \mathcal{C}_{\mathtt{bg}}^{\mathtt{noise}}} w_{h,j} \left[p_h(j \mid \mathbf{x}_{\mathtt{bg}}) - p_h(j \mid \mathbf{x}_{\mathtt{mix}}) \right], \tag{6}$$

$$G_h = \frac{1}{|\mathcal{C}_{\mathtt{tgt}}|} \sum_{j \in \mathcal{C}_{\mathtt{tgt}}} \left[p_h(j \mid \mathbf{x}_{\mathtt{mix}}) - p_h(j \mid \mathbf{x}_{\mathtt{bg}}) \right], \tag{7}$$

$$I = \frac{1}{2} \sum_{h \in \{\mathrm{BEATs}, \mathrm{PaSST}\}} \frac{D_h + G_h}{2}. \tag{8}$$

The classification gate is $I > 0.05$. This is a threshold on the final average of four terms, not on each classifier or component separately. Positive D_h indicates reduced background evidence; positive G_h indicates increased target evidence. A negative component is permitted if the combined score passes. Both class sets must be nonempty and remain fixed between background and mixture. Scores use raw probability changes, not percentage changes, ranks, or standardized similarities.

Stage 3: audio–text change alignment. Let $f^{\mathrm{PE}}_{\mathrm{audio}}$ and $f^{\mathrm{PE}}_{\mathrm{text}}$ denote the unit-normalized PE-AV audio and text embedding functions, respectively. The text c is the retrieved prompt, and $c_{\mathtt{bg}}$ describes background labels inferred from LAION-CLAP, BEATs, and PaSST. We form the audio and text embedding changes as

$$\boldsymbol{\Delta}_{\mathrm{audio}} = f^{\mathrm{PE}}_{\mathrm{audio}}(\mathbf{x}_{\mathtt{mix}}) - f^{\mathrm{PE}}_{\mathrm{audio}}(\mathbf{x}_{\mathtt{bg}}), \tag{9}$$

$$\boldsymbol{\Delta}_{\mathrm{text}} = f^{\mathrm{PE}}_{\mathrm{text}}(c) - f^{\mathrm{PE}}_{\mathrm{text}}(c_{\mathtt{bg}}). \tag{10}$$

We measure their alignment by cosine similarity, evaluated with a numerical stabilizer as

$$d = \frac{\mathbf{\Delta}_{\text{audio}}^{\mathsf{T}} \mathbf{\Delta}_{\text{text}}}{\|\mathbf{\Delta}_{\text{audio}}\|_2 \|\mathbf{\Delta}_{\text{text}}\|_2 + 10^{-12}}. \tag{11}$$

The candidate passes if $d > \tau_d$. Subtraction takes place in the embedding space, not between audio waveforms. A zero difference gives $d = 0$ under this numerical convention. All three deterministic stages must pass before audio-language validation.

Stage 4: audio-language validation. The judge receives the original background as a reference, the physical mixture, and the requested prompt. It returns two Boolean decisions: $J_{\texttt{match}}$, whether the requested natural mask sound is audible in the mixture, and $J_{\texttt{hidden}}$, whether the original background source is no longer recognizable. Acceptance requires both decisions to be true. An absent decision leaves the candidate pending and ineligible for selection; it is not treated as a positive answer. Thus the eligible set for a background is

$$\begin{aligned} \mathcal{S} = \{a : s_{\text{PE},a} > \tau_{\text{PE}},\ s_{\text{CLAP},a} > \tau_{\text{CLAP}},\ I_a > 0.05, \\ d_a > \tau_d,\ J_{\texttt{match},a} = J_{\texttt{hidden},a} = 1\}. \end{aligned} \tag{12}$$

Here a indexes the retrieved candidates. The classifier and embedding gates are complementary to, rather than substitutes for, the two perceptual judgments.

Stage 5: ranking the survivors. For each similarity model m, let μ_m and $\sigma_m > 0$ be the mean and standard deviation of mask sound–prompt similarities on the held-out calibration split. Define

$$u_m = \frac{s_m - \mu_m}{\sigma_m}, \qquad Q = \frac{\beta_{\text{PE}} u_{\text{PE}} + \beta_{\text{CLAP}} u_{\text{CLAP}}}{\beta_{\text{PE}} + \beta_{\text{CLAP}}}, \tag{13}$$

with nonnegative weights and positive total weight. We select the candidate in $\mathcal{S}$ with maximum Q, breaking exact ties by lower retrieval rank. No pair is selected when $\mathcal{S}$ is empty. Normalization uses fixed calibration statistics, not statistics recomputed from each background's survivors. The u_m are standardized scores, not diffusion latents.

Calibration parameters. The classification threshold is fixed at 0.05. The remaining acceptance thresholds τ_{PE}, τ_{CLAP}, and τ_d, and the ranking statistics and weights $(\mu_m, \sigma_m, \beta_m)$, must be selected on a held-out labeled validation split and fixed before evaluation. All required scores must be finite. The judge checkpoint and instruction, as well as the score-model checkpoints and audio preprocessing, must remain fixed for a filtering configuration.

D.3 Alternative Classification Criteria

The following criteria replace Stage 2 rather than adding further mandatory gates. They retain the same mask sound–prompt similarity, audio–text change alignment, audio-language validation, and final-ranking stages.

Target-over-background margin. For each classifier, define

$$M_h = \max_{j \in \mathcal{C}_{\texttt{tgt}}} p_h(j \mid \mathbf{x}_{\texttt{mix}}) - \max_{j \in \mathcal{C}_{\texttt{bg}}} p_h(j \mid \mathbf{x}_{\texttt{mix}}). \tag{14}$$

This alternative requires both $M_{\text{BEATs}} > \tau_{\text{BEATs}}$ and $M_{\text{PaSST}} > \tau_{\text{PaSST}}$. Each term is a maximum within a class set. Unlike Eq. 8, the margin compares target and background evidence in the same mixture, without measuring their changes relative to the original background. This criterion may become more suitable as generative models produce the intended natural sounds more consistently and audio classifiers become better calibrated on generated mixtures. Whereas relative improvement can admit a mixture in which the background remains prominent, a positive target-over-background margin requires target-class evidence to exceed background-class evidence in the final mixture. Its effectiveness depends on reliable comparisons between class scores and should be validated before use.

Naturalness and original-background suppression. We group the 527 AudioSet classes into sets $\mathcal{C}_{\texttt{nat}}$ of 38 desirable natural sounds, $\mathcal{C}_{\texttt{noise}}$ of 264 unwanted noises, and 225 ambiguous or excluded labels. Using the ordinal ranks r_h defined above, let $T_k(r, \mathcal{C})$ be the mean of the largest $\min(k, |\mathcal{C}|)$ values within $\mathcal{C}$. Natural dominance is

$$V_h = T_k(r_h(\cdot \mid \mathbf{x}_{\texttt{mix}}), \mathcal{C}_{\texttt{nat}}) - T_k(r_h(\cdot \mid \mathbf{x}_{\texttt{mix}}), \mathcal{C}_{\texttt{noise}}). \tag{15}$$

Select the three highest fused background ranks within $\mathcal{C}_{\texttt{noise}}$ to form $\mathcal{C}_{\texttt{bg}}^{\texttt{noise}}$, holding this set fixed when scoring mixtures. Background suppression is

$$B_h = \frac{1}{|\mathcal{C}_{\texttt{bg}}^{\texttt{noise}}|} \sum_{j \in \mathcal{C}_{\texttt{bg}}^{\texttt{noise}}} \left[r_h(j \mid \mathbf{x}_{\texttt{bg}}) - r_h(j \mid \mathbf{x}_{\texttt{mix}}) \right]. \tag{16}$$

This alternative requires $V_h > \tau_{V,h}$ and $B_h > \tau_{B,h}$ for both classifiers. The top-k aggregation compares equally sized subsets rather than summing over groups of unequal size; $k = 3$ uses their three strongest labels. The noise-restricted background set differs from the unrestricted set used in the target-over-background margin and audio–text change alignment text.

E Model Architecture and Training

Model training uses ∼1,200 h of paired $(\mathbf{x}_{\texttt{bg}}, \mathbf{x}_{\texttt{mask}})$ audio, with backgrounds sampled from TAU Urban Acoustic Scenes 2019 Mesaros et al. (2018; 2019), SONYC-UST-V2 Cartwright et al. (2020), and WavCaps Mei et al. (2024), all at 16 kHz mono.

E.1 Offline Oracle Baselines

We finetune three pretrained generative audio models on the same $(\mathbf{x}_{\texttt{bg}}, \mathbf{x}_{\texttt{mask}})$ data pairs: Stable Audio Evans et al. (2025), AudioLDM2 Liu et al. (2024) and Frieren Wang et al. (2024). Each model is adapted to accept a background audio embedding as conditioning input while preserving the rest of the base architecture, and is trained to predict the full $\mathbf{x}_{\texttt{mask}}$.

The three models are architecturally diverse: AudioLDM2 is a UNet latent diffusion model that runs on 2D mel-spectrogram-domain VAE latents; Stable Audio is a Diffusion Transformer (DiT) Peebles & Xie (2023) running on 1D waveform-domain VAE latents; and Frieren is a feed-forward transformer trained with rectified flow matching Liu et al. (2023b); Lipman et al. (2023) on 1D mel-spectrogram-domain VAE latents. They share no common backbone or latent space.

E.2 Causal Masking Model

The online model couples a frozen SoundCTM Saito et al. (2024) VAE (64-D latents at 43 Hz) with a 12-layer LLaMA-style backbone (512 model dimension, 8 attention heads) and an EDM2 Karras et al. (2024) diffusion head (1024 channels, 10 residual blocks); inference uses 50 sampling steps with CFG $= 2.0$.

At 16 kHz, a 10s waveform contains 160,000 samples. The frozen VAE's 43 Hz latent rate keeps the transformer context short, while its 64D channel size keeps the diffusion head's target distribution low-dimensional and easier to fit.

We use QK-norm Henry et al. (2020); Dehghani et al. (2023) in the transformer backbone. Without QK-norm, we observe that the gradient norm diverges as training proceeds to ∼30k steps, resulting in model collapse. Adding QK-norm significantly stabilizes training and enables longer training runs. During diffusion-head training, we reuse the same conditioning vector across multiple noise estimates for efficiency.

Two-stage training. We first pretrain the model on unpaired audio to model the unconditional audio distribution, then initialize conditional finetuning from this checkpoint using the 1,200 h of paired data and standard classifier-free guidance (CFG) training. Preliminary listening observations when training from scratch on the paired data suggested a trade-off: low CFG scales produced natural-sounding audio with a weak response to the background, whereas higher scales strengthened the response but introduced audible artifacts. We hypothesize that the paired data distribution may

be too narrow for CFG to steer within a realistic region, motivating unconditional pretraining. This motivation is based on preliminary observations rather than a quantitative ablation.

We train with AdamW under bf16 mixed precision in two stages: 30k steps of unconditional pretraining on ∼7000 h of WavCaps Mei et al. (2024), learning rate 2×10^{-5}, cosine schedule, followed by 65k steps of conditional finetuning on the paired data (label dropout rate $= 0.1$, learning rate 1×10^{-5}, constant) with prompt-balanced sampling and temporal split-and-merge augmentation. During unconditional pretraining, we set the label dropout rate to 1.0, zeroing every background hidden state so that the backbone models the unconditional audio distribution. Following Diff-Foley Luo et al. (2023), temporal split-and-merge augmentation concatenates random crops of paired clips. Prompt-balanced smoothing addresses the long-tail distribution of prompt retrieval. For example, thunder categories are retrieved disproportionately often because they resemble urban background noise that is mainly low-frequency. Uniform sampling would make the model overfit to these categories. All training runs use a per-device batch size of 24 on 32 NVIDIA A40 GPUs.

A related causal-transformer and diffusion-head architecture is used in SoundReactor Saito et al. (2025) for online video-to-audio generation. Quantitative results for our model are obtained by predicting the full latent sequence and then decoding it in a single pass with the (non-causal) SoundCTM VAE. Note that the same VAE supports per-frame causal decoding for online deployment.

E.3 Consistency Fine-Tuning

The real-time variant is obtained by Easy Consistency Tuning Geng et al. (2025) for 40k steps with the backbone frozen. The resulting model uses 1–4 function evaluations per masking frame while preserving causal conditioning.

F Evaluation Details

We evaluate two complementary aspects of sound masking: the quality and naturalistic character of the resulting audio, and the extent to which it conceals the background event. All metrics below evaluate the physical mixture rather than the mask sound in isolation.

F.1 Data and Evaluation Configuration

The background collection contains 48,840 prepared clips: 18,510 from SONYC, 14,400 from TAU, 5,930 from UrbanSound, 5,000 from BABYCRY, and 5,000 from NOTSOFAR. Retrieving 15 prompt–seed pairs per background defines 732,600 candidate configurations; filtering selects at most one pair per background. Evaluation clips are mono at 16 kHz and truncated to 159,754 samples, the natural output length of the online model's VAE decoder. Background and mixture audio are normalized to -30 and -23 LUFS before feature extraction.

For forced-choice identification, each AudioSet leaf is rendered as a one-sentence description generated by Qwen3.5-9B. The options are the CLAP-assigned background category plus three random distractors. Qwen3-Omni-30B-A3B-Instruct Xu et al. (2025) receives the mixture and answers a single letter A/B/C/D.

F.2 Audio Quality and Distributional Similarity

We compare each method's mixtures with the fixed reference pool of approximately 150,000 naturalistic AudioLDM2 clips described in Section 5. Lower distributional distances indicate greater similarity to this reference pool; they do not directly measure listener comfort or masking effectiveness.

Fréchet Audio Distance (FAD). FAD (Kilgour et al., 2019) compares Gaussian approximations to the reference and evaluated audio embedding distributions through their means and covariances. We compute it using CLAP audio embeddings (Wu et al., 2023b; Elizalde et al., 2023), rather than the VGGish features used in the original metric. A lower score indicates closer first- and second-order feature statistics.

Maximum Mean Discrepancy (MMD). MMD (Gretton et al., 2012) compares the two CLAP embedding distributions using a radial basis function kernel, without fitting Gaussian distributions. We subsample at most 2,000 embeddings from each set with random seed 42. For $k(u, v) = \exp(-\|u - v\|^2/(2\sigma^2))$, σ^2 is the median pairwise squared distance in the pooled subsample, bounded below by 10^{-8}. We compute the unbiased estimate of squared MMD, clamp it to zero if negative, and report its square root. Lower values indicate more similar embedding distributions.

KL-PaSST. We use PaSST (Koutini et al., 2022) to obtain sound-event class predictions. We average the prediction vectors separately over the evaluated and reference sets, add 10^{-10} to each component, and normalize each vector to sum to one. Denoting these distributions by p and q, respectively, we report $D_{\mathrm{KL}}(p\|q) = \sum_c p_c \log(p_c/q_c)$. Lower values indicate more similar aggregate sound-event distributions. This is a dataset-level comparison, not a distance between matched audio pairs.

Natural-category hit rate. We rank the 474 AudioSet leaf categories by cosine similarity between the mixture's CLAP audio embedding and the category text embeddings. Nat.%@1 and Nat.%@5 report the percentage of mixtures with at least one designated natural-sound category among the top one or five predictions, respectively. Higher values indicate that natural sounds are more prominent in the classifier's interpretation of the mixture. As in Table 1, backgrounds whose reference category is itself natural are excluded from this analysis. These scores describe semantic content and do not establish perceptual pleasantness.

F.3 Masking Effectiveness

Background-category rank drop. We assign each background its highest-scoring CLAP leaf category before mixing and retain this label when evaluating every masking method. Let $r_{\texttt{mix}}$ be its one-based rank among the 474 categories after mixing. We report the mean $\Delta\mathrm{Rank} = r_{\texttt{mix}} - 1$ over the same non-natural-background subset. The score ranges from 0 to 473; a higher value means that more categories outrank the original background category after masking. This measures suppression of classifier evidence for the background, rather than human identification directly.

Forced-choice background identification. We evaluate an audio-language model on four-way background identification. Accuracy is the fraction of responses that select the correct background category. Lower accuracy indicates stronger masking for the corresponding evaluator, with 25% accuracy under uniform random guessing. Reduced machine recognition does not necessarily imply reduced recognition by human listeners.

F.4 Additional Results Analysis

Masking performance and audio quality. Our online model achieves the lowest LLM forced-choice identification accuracy (0.43) among deployable methods, and the highest ΔRank (80.46) among all non-trivial methods. White noise has a higher LLM identification accuracy (0.58), indicating that it suppresses model recognition less effectively under this protocol. On ΔRank our model exceeds the offline energy-guided generation (76.85) and the three offline oracles (62.80–68.87), despite their access to the full background.

On Nat.%@1, Nat.%@5, FAD, MMD, and KL-PaSST, our model is better than white noise but behind the offline energy-guided generation and the offline oracles. These methods observe the entire background, and the oracle models are trained on the selected synthetic pairs.

A random clip drawn from the 150k naturalistic pool wins across natural-sound classification and distributional similarity metrics ($77.74/96.26/87.62$ and $0.12/0.18/0.09$, marked with *). Such a clip already lies in the natural distribution: it is drawn from the reference set used for distributional comparison and is naturalistic by curation. However, this baseline has higher LLM forced-choice identification accuracy than our online model (0.47 vs. 0.43). The online model thus learns a non-trivial conditional function that masks backgrounds more effectively than the deployable baselines while maintaining lower distributional distances than white noise. The real-time model is marginally behind the online model on every reported metric, as expected from single-step distillation; its goal is real-time latency rather than surpassing the online model.

CFG ablation. As CFG rises from 2 to 5, forced-choice accuracy worsens from 0.43 to 0.49 for the online model and from 0.45 to 0.48 for the real-time model. Background rank drop falls from 80.46 to 63.12 and from 68.55 to 60.69, respectively. FAD, MMD, KL-PaSST, and Nat.%@1 improve modestly at higher CFG (e.g., online FAD $0.52 \rightarrow 0.47$ and KL-PaSST $0.84 \rightarrow 0.72$), while Nat.%@5 changes inconsistently. We use CFG $= 2$ because masking is the primary objective and this setting performs best on forced-choice identification and background rank drop.