LEARNING ACOUSTIC SCENE DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of decomposing a single audio recording into the latent factors that produced it: the dry audio of each source and the room impulse response (RIR) through which it reaches the listener. Existing methods tackle pieces of this problem in isolation, focusing on RIR estimation, audio source separation, or dereverberation, typically under single-source assumptions. We instead pose it as a single inverse problem and present ASCEND, a diffusion-based framework that recovers all components jointly. Our key idea is that a correct decomposition, when re-rendered through the known acoustic forward model, must reproduce the input mixture. We turn this into an inference-time signal via consistency-guided sampling: at every denoising step, we re-render the current estimates, measure the mismatch with the input, and use its gradient to guide sampling. The denoiser is a transformer model with attention that operates both within and across audio source streams. On large-scale simulated benchmarks, ASCEND outperforms task-specific baselines on RIR estimation and dry-source extraction, with gains that transfer zero-shot to real-world recordings. Audio demos are available on Hugging Face.

1 INTRODUCTION

Decomposing a complex observation into the latent factors that produced it is a central pursuit of perception research. In computer vision, this goal has been formalized as *inverse graphics*: recovering geometry, materials, and illumination from a single image under a known forward rendering process. Audio recordings have a similar compositional structure: every waveform we capture is the rendered output of latent acoustic factors that interact through a well-understood physical process. Recovering these factors from a single recording is, in spirit, the auditory analog of inverse graphics.

An acoustic scene can be factored into two sets of latent components: what each source emits (i.e., dry sound or *intrinsic* signal), and how that emission is reshaped by the environment as it propagates to the listener (an *extrinsic* factor jointly determined by the source location, the scene, and the listener). What reaches a microphone is then each emission convolved with its propagation response, summed across all sources, together with any incidental background sound captured at the same time. Formally, this can be written as:

$$y(t) = \sum_{n=1}^{N} (r_n * d_n)(t) + a(t), \tag{1}$$

where $d_n(t)$, $r_n(t)$, and $a(t)$ are time-domain signals (the dry waveform of source n, its RIR, and the ambient sound), and $*$ denotes temporal convolution.

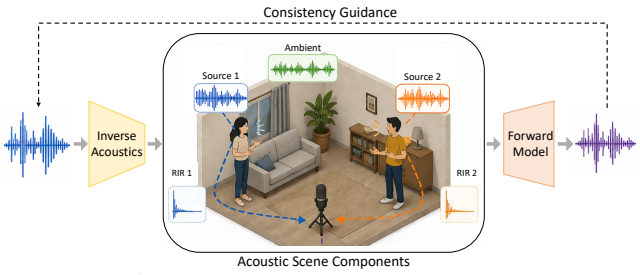

Figure 1: Forward and inverse acoustics.

Inverse acoustics aims to solve the inverse of this process: from a single observed mixture $y(t)$, recover the latent components $\{r_n, d_n\}_{n=1,\ldots,N}$ (Fig. 1). This formulation unifies traditionally distinct tasks including RIR estimation (Steinmetz et al., 2021; Ratnarajah et al., 2023), sound separation (Xu et al., 2025a; Wang et al., 2023), and dereverberation (Lemercier et al., 2023; 2025). Like other inverse problems, inverse acoustics is challenging as it is fundamentally under-constrained: many decompositions can explain the same observed mixture.

No prior method addresses inverse acoustics in its full form. Existing work has advanced individual sub-problems (RIR estimation (Steinmetz et al., 2021; Ratnarajah et al., 2023; Lemercier et al.,

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.

2025), dereverberation (Richter et al., 2023; Lemercier et al., 2023; 2025), and source separation (Xu et al., 2025a; Wang et al., 2023)), but studies them in isolation, typically under single-source (n=1) settings. Feedforward predictive models (Steinmetz et al., 2021; Ratnarajah et al., 2023; 2024; Richter et al., 2023) learn a direct input-to-output mapping for a single sub-task; they cannot enforce reconstruction consistency or scale to multi-source mixtures. Acoustically grounded generative methods (Lemercier et al., 2025; Xu et al., 2025b) do enforce consistency via the forward model, but only with unconditional priors and under restrictive assumptions, e.g., single-source inputs or multi-microphone arrays.

We address this gap with ASCEND, a diffusion-based framework for Acoustic SCENe Decomposition that jointly recovers the latent acoustic components including dry sources and their RIRs from a single multi-source mixture. Our central insight is that, since the forward acoustic model is known, any valid decomposition must reproduce the observed mixture when re-rendered through it. We turn this physical consistency into an inference-time signal: ASCEND pairs a strong conditional diffusion prior over decompositions with an inference-time loop that closes back to the input mixture. Together, these two components address both failure modes above: the conditional prior replaces the unconditional ones used by generative methods, and the inference-time loop brings the forward-model consistency that feedforward predictors lack.

To realize this idea, ASCEND couples a conditional diffusion model with *Consistency-Guided Sampling* (CGS) at inference. Diffusion fits inverse acoustics well: many decompositions explain the same mixture, so we need to model the full posterior over $\{r_n, d_n, a\}$ rather than commit to a single estimate. CGS adds forward-model consistency to the denoising loop. At each diffusion timestep, it recomposes the current dry-source and RIR estimates into a mixture waveform, compares it against the observed input, and backpropagates the reconstruction discrepancy through the forward acoustic model. The resulting consistency gradient pulls the denoising trajectory toward decompositions that reproduce the observation. The acoustic forward model is light (convolutions and summations) and differentiable, so CGS obtains its guidance directly from acoustic reconstruction. The conditional prior is a diffusion transformer tailored to the multi-source structure of the problem. Since the decomposition factors into per-source streams (each source's dry signal and its RIR), we adopt a hybrid global-local attention design: local attention operates within each source's stream to capture per-source dereverberation and RIR structure, while shared global attention operates across streams to model the cross-source cues to solve sound separation. This factorization gives the model an inductive bias matched to inverse acoustic problem.

To train and evaluate ASCEND, we construct a large-scale simulated dataset of realistic multi-source acoustic scenes, pairing high-fidelity RIRs from existing simulation platforms with anechoic speech, vocal songs, and ambient noise, which are common sound sources in the indoor scenario. The resulting benchmark varies substantially in rooms, sources, reverberation, and interference patterns. We further demonstrate that ASCEND achieves zero-shot generalization on real-world recordings.

To summarize, our main contributions are as follows:

- We introduce an inverse acoustics task formulation that unifies multi-source RIR estimation, sound separation, and sound dereverberation as a single inverse problem of acoustic scene decomposition.
- We propose ASCEND, a conditional diffusion framework for inverse acoustics that incorporates Consistency-Guided Sampling to enforce reconstruction consistency during denoising, together with a global-local attention architecture for jointly solving multi-source decomposition.
- We validate ASCEND's performance on large-scale simulated benchmarks of complex acoustic scenes, and show strong zero-shot generalization on real-world recordings.

2 RELATED WORK

Room Impulse Response Estimation. RIR modeling has received increasing attention due to its importance in immersive multimedia. Prior work has explored estimating it from visual observations (Singh et al., 2021), reverberant recordings (Steinmetz et al., 2021; Ratnarajah et al., 2023; Wang & Li, 2025), or audio-visual cues (Ratnarajah et al., 2024). FiNS (Steinmetz et al., 2021) uses an encoder-decoder model to infer RIRs from reverberant speech, Speech2IR (Ratnarajah et al., 2023) adopts a GAN-based formulation, Rec-RIR (Wang & Li, 2025) models the RIR by directly estimating a CTF filter, and AV-RIR (Ratnarajah et al., 2024) combines reverberant speech with visual context. Despite these advances, they largely focus on single-source, noise-free settings and estimate RIRs separately from dry source recovery. As a result, they do not enforce that the predicted RIRs and

sources can reconstruct the observed reverberant mixture during inference. In contrast, our work addresses multi-source RIR estimation in noisy acoustic scenes by jointly estimating sources and RIRs under an explicit forward-consistency constraint.

Speech and Vocal Separation and Dereverberation. A large body of prior work has studied regression-based approaches for speech separation (Wang et al., 2023; 2021; Xu et al., 2025a) and dereverberation (Luo & Mesgarani, 2018; Pandey & Wang, 2019), learning direct mappings from degraded observations to separated or dereverberated signals. More recently, diffusion-based models have shown strong performance in separation (Lutati et al., 2024; Wang et al., 2025), as well as speech and vocal song dereverberation (Richter et al., 2023; Lemercier et al., 2025). However, these approaches primarily focus on directly recovering target signals, do not explicitly estimate RIRs, and therefore cannot enforce consistency between predictions and observations. Other unsupervised methods leverage acoustic constraints for dereverberation or separation. However, they either restrict to a single speaker (Lemercier et al., 2025), or separate sources while preserving reverberation and depend on a microphone array (Wang & Watanabe, 2023; Xu et al., 2025b). These limit their applicability to the inverse acoustics task, where the model needs to handle multi-source sound separation, dereverberation, and RIR estimation with only one single-channel recording.

Inverse Problems and Diffusion Guidance. Our work is connected to a broader family of inverse problems. In computer vision, inverse graphics follows this paradigm by recovering object/scene intrinsics and extrinsics, such as geometry, material, illumination, and pose, from rendered images (Zhang et al., 2023; Geng et al., 2025). Recent diffusion-based methods applies this idea to other problems by coupling generative models with measurement-consistency constraints during sampling (Chung et al., 2023; Xu et al., 2025b). They rely on unconditional priors and refine samples using likelihood gradients induced by the forward operator. In contrast, ASCEND uses a conditional diffusion model that directly learns the posterior over dry sources and RIRs conditioned on the observation, which provides better initialization under ill-posed settings. Moreover, our inference-time guidance also differs from classic classifier guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022): instead of using gradients from an auxiliary classifier, we use gradients from forward acoustic forward modeling.

3 METHOD

We first describe the problem definition of inverse acoustics, our consistency-guided sampling strategy, and finally our conditional transformer with global-local attention design.

3.1 PROBLEM DEFINITION

We consider a single-channel reverberant observation $y \in \mathbb{R}^T$ captured in an acoustic scene with N sound sources. The n-th source emits a dry sound signal d_n, which reaches the microphone through a source-specific RIR r_n. The observation also contains an ambient background component a and a residual term η accounting for measurement noise and model mismatch. The observation is given by:

$$y = \mathcal{A}(x) + \eta = \sum_{n=1}^{N} r_n * d_n + a + \eta. \tag{2}$$

Given only the observed mixture y, our goal is to recover the underlying acoustic scene decomposition x, including the dry source signals $\{d_n\}_{n=1}^N$ and their corresponding RIR $\{r_n\}_{n=1}^N$ for all N sound sources. Viewing all acoustic components together as a structured acoustic scene $x = \big(\{d_n, r_n\}_{n=1}^N, a\big)$ that construct the observation y, we refer to this task as *inverse acoustics*.

3.2 CONSISTENCY GUIDED SAMPLING

CGS uses the acoustic forward model to evaluate the consistency of an intermediate clean-scene estimate with the observed mixture. In class-conditioned diffusion, classifier guidance (Dhariwal & Nichol, 2021) tilts the diffusion score from pure posterior using the gradient of $p_\phi(c \mid x_t)$, with a guidance weight w controlling the fidelity-diversity tradeoff:

$$\tilde{s}(x_t, c) = \nabla_{x_t} \log p(x_t \mid c) + w \nabla_{x_t} \log p_\phi(c \mid x_t), \tag{3}$$

where x_t denotes the noisy sample at diffusion timestep t, c is the class condition, and $\tilde{s}(x_t, c)$ is the modified score the diffusion model samples from. This up-weights data with high likelihood to c, and the same principle underlies classifier-free guidance (Ho & Salimans, 2022).

In our inverse acoustics setting, we encourage reconstruction consistency through the forward acoustic model (Eq. 2). Given the forward process, valid predictions must reproduce the observed signal. The forward model therefore acts as a natural "classifier" of physical plausibility in the acoustic scene.

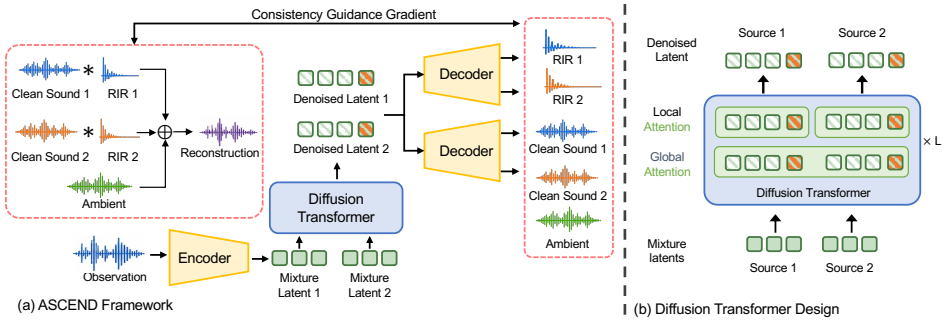

Figure 2: Overall pipeline for multi-source RIR and source estimation.

Intuitively, CGS recomposes the estimated acoustic components via the forward acoustic model and compares the resulting reconstruction against the observed mixture y. The gradient of this discrepancy guides sampling toward decompositions that better reconstruct the observation.

To motivate this guidance, we consider augmenting the learned conditional score with a consistency likelihood gradient:

$$\tilde{s}(x_t, y) = \underbrace{\nabla_{x_t} \log p(x_t \mid y)}_{\text{learned conditional score}} + \underbrace{\lambda \nabla_{x_t} \log p(y \mid x_t)}_{\text{consistency likelihood}}, \tag{4}$$

where λ is the consistency guidance scale. The first term is estimated by our conditional diffusion model, while the second term favors agreement with the forward acoustic model.

Directly evaluating $\log p(y \mid x_t)$ is intractable because the acoustic forward model relates the measurement y to the clean decomposition x_{clean}, not to an arbitrary noisy state x_t. Following the plug-in approximation used in Diffusion Posterior Sampling (Chung et al., 2023), we evaluate the likelihood at the denoiser's current clean estimate:

$$\nabla_{x_t} \log p(y \mid x_t) \approx \nabla_{x_t} \log p(y \mid \hat{x}_{\text{clean}}), \tag{5}$$

where $\hat{x}_{\text{clean}}$ denotes the conditional clean estimate predicted from (x_t, y). More details in App. A.

Specifically, the forward model output $\mathcal{A}(\hat{x}_{\text{clean}})$ is calculated by each acoustic component prediction extracted from the single estimated clean state $\hat{x}_{\text{clean}}$ via the forward acoustic model: $\mathcal{A}(\hat{x}_{\text{clean}}) = \left(\sum_{n=1}^{N} \hat{r}_{\text{clean}}^{(n)} * \hat{d}_{\text{clean}}^{(n)} + \hat{a}_{\text{clean}}\right)$. We consider time-domain and frequency-domain losses on the reconstructed mixture $\hat{y}_t = \mathcal{A}(\hat{x}_{\text{clean}})$:

$$\mathcal{L}_{\text{time}} = \|y - \hat{y}_t\|_2^2, \qquad \mathcal{L}_{\text{freq}} = \ell_{\text{STFT}}(y, \hat{y}_t), \tag{6}$$

where ℓ_{STFT} multi-resolution STFT loss measures discrepancies in spectral domain (Steinmetz & Reiss, 2020). We compute the two gradients and combine their normalized versions with a controllable weight w:

$$g_t^{\text{time}} = \nabla_{x_t} \mathcal{L}_{\text{time}}, \quad g_t^{\text{freq}} = \nabla_{x_t} \mathcal{L}_{\text{freq}}, \quad g_t^{\text{CGS}} = (1-w) g_t^{\text{time}} + w g_t^{\text{freq}}. \tag{7}$$

This combination gives a practical likelihood-inspired consistency direction. $w = 0$ reduces the consistency likelihood to the time-domain squared-error objective: $\nabla_{x_t} \log p(y \mid \hat{x}_{\text{clean}}) \propto -\nabla_{x_t} \mathcal{L}_{\text{time}}$, which is under standard isotropic Gaussian residual model. The w controllable frequency-domain loss provides an additional consistency criterion for the same reconstructed mixture.

The derivations above are presented in the data space x for clarity, while our model operates in a compressed latent space z. During CGS, the estimated clean latent $\hat{\mathbf{z}}_{\text{clean}}$ first pass through the VAE decoder $\mathcal{D}$ to reconstruct the audio waveforms $\hat{x}_{\text{clean}} = \mathcal{D}(\hat{\mathbf{z}}_{\text{clean}})$ before applying the physical forward model $\mathcal{A}(\hat{x}_{\text{clean}})$. Loss gradients are taken with respect to $\mathbf{z}_t$.

We apply this consistency direction within diffusion and rectified-flow sampling. Both formulations provide an intermediate clean-scene estimate, while their prediction parameterizations and sampling updates differ. Our latent diffusion model formulation predicts the velocity $v_\theta(z_t, y)$ (Salimans & Ho, 2022). At each denoising step t, we obtain the clean estimate $\hat{z}_0$ from the velocity prediction: $\hat{z}_0 = \sqrt{\bar{\alpha}_t}\, z_t - \sqrt{1-\bar{\alpha}_t}\, v_\theta(z_t, y)$. In a standard latent DDIM sampler (Song et al., 2021) with 50 steps, we update the sample to the previous timestep $t-1$ as follows, with a gradient term that nudges the DDIM update toward decompositions whose decoded components better reconstruct the observation. We provide the pseudo-code for our implementation in Alg. 1.

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{z}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \left(\sqrt{1-\bar{\alpha}_t}\, z_t + \sqrt{\bar{\alpha}_t}\, v_\theta(z_t, y) - \lambda g_t^{\text{CGS}}\right) \tag{8}$$

Algorithm 1 Consistency Guided Sampling (CGS) for Diffusion

Require: Observed Mixture y, Conditional Diffusion Model DiT_θ, VAE Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$, Forward acoustic model $\mathcal{A}$, Noise schedule $\bar{\alpha}_{1:T}$, Guidance scale $\lambda > 0$, consistency weight w

$\mathbf{z}_y \leftarrow \mathcal{E}(y)$ ▷ Encode the observed mixture condition

$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Initialize noisy latents for $\{\mathbf{z}_{d,n}, \mathbf{z}_{r,n}\}_{n=1}^{N}$ and $\mathbf{z}_a$

for $t = T$ **to** 1 **do**

  $\hat{\mathbf{v}}_t \leftarrow \mathrm{DiT}_\theta(\mathbf{z}_t, \mathbf{z}_y, t)$ ▷ Predict diffusion velocity

  $\hat{\mathbf{z}}_0 \leftarrow \sqrt{\bar{\alpha}_t}\,\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\mathbf{v}}_t$ ▷ Estimate clean latents

  $\hat{\boldsymbol{\epsilon}}_t \leftarrow \sqrt{1-\bar{\alpha}_t}\,\mathbf{z}_t + \sqrt{\bar{\alpha}_t}\,\hat{\mathbf{v}}_t$ ▷ Estimate noise

  $\hat{y}_t \leftarrow \mathcal{A}(\mathcal{D}(\hat{\mathbf{z}}_0))$ ▷ Reconstruct the mixture

  $\mathcal{L}_{\mathrm{time}} \leftarrow \|y - \hat{y}_t\|_2^2$, $\mathcal{L}_{\mathrm{freq}} \leftarrow \ell_{\mathrm{freq}}(y, \hat{y}_t)$

  $\mathbf{g}_t^{\mathrm{CGS}} \leftarrow (1-w)\nabla_{\mathbf{z}_t}\mathcal{L}_{\mathrm{time}} + w\nabla_{\mathbf{z}_t}\mathcal{L}_{\mathrm{freq}}$ ▷ Weighted consistency directions

  $\mathbf{z}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\,\hat{\mathbf{z}}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\left(\hat{\boldsymbol{\epsilon}}_t - \lambda_t \mathbf{g}_t^{\mathrm{CGS}}\right)$ ▷ Guided DDIM update

end for

return $\mathcal{D}(\hat{\mathbf{z}}_0)$ ▷ Return the final decoded acoustic components

In the rectified flow formulation (Lipman et al., 2023; Liu et al., 2022b), sampling proceeds from Gaussian noise at $t = 0$ to data at $t = 1$. The straight-line training path $z_t = (1-t)\epsilon + tz_1$ has target velocity $z_1 - \epsilon$, giving the clean estimate $\hat{z}_1 = z_t + (1-t)\hat{v}_t$. CGS decodes this estimate and evaluates the same reconstruction losses defined above. We apply the resulting consistency direction after an Euler step $z_{t+\Delta t} = z_t + \Delta t\,\hat{v}_t - \lambda_t g_t^{\mathrm{CGS}}$. Thus, the acoustic consistency computation is shared across both formulations, while the clean-estimate parameterization and base sampling update differ. Details for flow matching are provided in the Appendix C.5 with pseudo code in Alg. 2.

3.3 Diffusion Transformer with Global–Local Attention

To empower the CGS mechanism described above, a highly capable conditional generative model is required. We introduce a specialized latent diffusion transformer (DiT) with a global–local attention mechanism. This design explicitly captures both the cross-source separation cues and source-specific acoustic structures, implicitly organizing the learning process into tightly coupled sub-tasks: sound separation, dereverberation, and RIR estimation.

Our DiT operates in a compressed latent space to efficiently model the mixture and predict the constituent components. We train VAEs (Evans et al., 2025) to encode all acoustic components. We encode the observed input mixture $y \in \mathbb{R}^T$ into a latent representation $\mathbf{z}_y \in \mathbb{R}^{M_s \times D}$, where M_s is the number of tokens used for source/mixture latent and D is the VAE latent feature dimension. Similarly, the acoustic components: dry signals d_n, RIRs r_n, and ambient background a are encoded into latents $\mathbf{z}_{d,n}$, $\mathbf{z}_{r,n}$, and $\mathbf{z}_a$, respectively. The target acoustic scene is comprised of $2N+1$ distinct components: the dry signals and RIRs for N sources, plus the ambient background. We denote their corresponding clean latents as the set: $\{\mathbf{z}_{d,n}, \mathbf{z}_{r,n}\}_{n=1,2,\ldots,N} \cup \{\mathbf{z}_a\}$. Crucially, while the dry speech latents $\mathbf{z}_{d,n}$ and ambient latent $\mathbf{z}_a$ share the same token length M_s as the mixture, the RIR latents $\mathbf{z}_{r,n}$ represent time-invariant RIR filters and possess a different token length, M_r.

Latent Preparation and Conditioning. At a given diffusion timestep t, our DiT takes the set of $2N+1$ independent noisy latents $\mathbf{z}_{d,n,t}$, $\mathbf{z}_{r,n,t}$, and $\mathbf{z}_{a,t}$ and maps them into a unified transformer hidden space of dimension D_{dit}. To condition the generation on the observed mixture, we concatenate the noisy dry source latent $\mathbf{z}_{d,n,t}$ and ambient latent $\mathbf{z}_{a,t}$ with the mixture condition latent $\mathbf{z}_y$ along the feature dimension. Learned MLPs project these concatenated tokens to the transformer dimension. Because the RIR latents have a mismatched token length M_r, they are not directly concatenated with $\mathbf{z}_y$. Instead, each RIR latent $\mathbf{z}_{r,n,t}$ is mapped to the transformer hidden dimension through a separate MLP. These projected components are then collected into a combined token sequence $\mathbf{z}_t$ containing $(N+1)M_s + NM_r$ total tokens, which serves as the input to the transformer blocks. Unlike a standard DiT that processes a single token stream, our diffusion transformer explicitly partitions its processing into multiple output streams representing the acoustic scene.

Alternating global and local attention. Our model needs to jointly perform three tasks, including RIR estimation, speech dereverberation, and source separation. Since the former two require modeling within one source, and the later separation task requires separating N different sources within a mixture, we propose alternate between using global-across-source attention and local-within-source

attention. Our DiT is built from L stacked blocks, each composed of a *global attention* layer followed by *local attention* layers (Fig. 2). Let $\mathbf{H}_t^{(0)} = \mathbf{z}_t + \mathrm{PE}(\mathbf{z}_t)$ be the initial complete token sequence after adding positional embeddings. The ℓ-th global attention layer computes: $\tilde{\mathbf{H}}_t^{(\ell)} = \mathrm{MSA}_{\text{global}}(\mathbf{H}_t^{(\ell-1)})$. Global attention applies multi-head self-attention (MSA) over all RIR, source, and ambient tokens in the combined sequence. This enables unconstrained information exchange, allowing the model to perform source separation by cross-referencing different sources and capture global room acoustics. The global output $\tilde{\mathbf{H}}_t^{(\ell)}$ is then partitioned back into N distinct slots. For each source n, the slot $\tilde{\mathbf{H}}_t^{(\ell,n)}$ isolates exactly $M_s + M_r$ tokens that belongs to one sound source. The ambient stream is treated as a separate slot with M_s tokens. We apply a shared *local* self-attention module independently to each source slot: $\mathbf{H}_t^{(\ell,n)} = \mathrm{MSA}_{\text{local}}(\tilde{\mathbf{H}}_t^{(\ell,n)})$.

Within each local module, attention is strictly restricted to the tokens belonging to a single source-RIR pair. This encourages each slot to specialize in a coherent source-RIR pair, while global attention handles cross-source separation. The outputs from all local modules are concatenated back into a unified sequence $\mathbf{H}_t^{(\ell)}$ and passed to the next block.

Model Outputs and v-Prediction Objective. After L global-local blocks, our DiT decodes the final hidden states using lightweight linear heads to predict the diffusion velocity v_θ for each scene component. We following the standard v-prediction framework (Lin et al., 2024), the ground-truth velocity targets for source n are defined as $\mathbf{v}_{d,n,t}$, $\mathbf{v}_{r,n,t}$, and $\mathbf{v}_{a,t}$. The training objective is the expected mean squared error over all timesteps and acoustic components.

Source Slots and Evaluation. Our model outputs a set of source slots, where each slot contains a dry-source latent and its paired RIR latent. During training, we directly supervise each slot with its input ordering for their corresponding source-RIR pair. At test time, the output slots are interpreted as an unordered set. For quantitative evaluation, we first match predicted dry sources to ground-truth dry sources using the standard permutation assignment used in source separation evaluation. The RIR contained in each matched slot is then evaluated against the RIR of the corresponding ground-truth source. This gives a consistent assignment for both dry-source and RIR metrics while avoiding artificial penalties from arbitrary slot permutations.

4 Evaluation

4.1 Dataset

Simulated Dataset. We construct large-scale simulated datasets using two acoustic simulators: the ray-tracing-based SoundSpaces 2.0 (Chen et al., 2020; 2022; Xia et al., 2018) and the ray-wave hybrid Treble simulator (Treble Technologies, 2026). Both simulators generate realistic room impulse responses (RIRs) in indoor environments with various acoustic properties. For each scene, we sample multiple source locations that share the same receiver location, yielding sets of source-specific RIRs for a fixed microphone. This property is crucial for our inverse acoustics setting, where a single mixture is formed by multiple sources propagating to the same receiver. In contrast, existing real-world RIR datasets typically do not provide large-scale collections of RIRs with a fixed receiver and a large number of source locations in the same room, making them difficult to use for our inverse acoustics task. We sample 6-second clean speech utterances from LibriSpeech (Panayotov et al., 2015) and ambient noise from WHAM! (Wichern et al., 2019). We also sample 6-second clean song tracks from five singing voice corpora (Choi et al.; Zhang et al., 2024; Sharma et al., 2021; Duan et al., 2013; Liu et al., 2022a). Each clean source is convolved with its corresponding source-specific RIR, and the resulting reverberant source signals are summed with ambient noise to synthesize either overlapping speech mixtures or speech-song mixtures. The resulting simulated datasets are named IA-SoundSpaces and IA-Treble. IA-SoundSpaces contains 550k training mixtures (8k test mixtures), while IA-Treble contains 280k training mixtures (6k test mixtures). Test mixtures are synthesized with unseen RIRs, clean speech, clean singing vocals, and ambient noise. Additional dataset details are provided in App. B.1.

Real-World Dataset. To evaluate sim-to-real transfer, we collect a real-world test dataset by playing LibriSpeech test-set utterances through a loudspeaker in five rooms with diverse acoustic properties, including an auditorium, meeting rooms, and classrooms. Reverberant speech is recorded using a Logitech microphone. To construct two-source mixtures, we sum pairs of recorded reverberant utterances captured at the same microphone location but played from different loudspeaker positions. RIR measurements are also performed. The resulting real-world evaluation set contains approximately 4 hours of audio. Further details are provided in App. D.

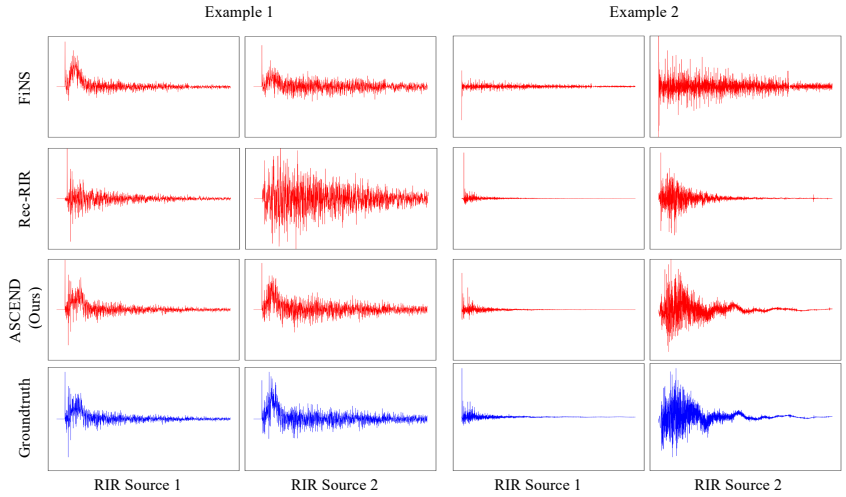

Figure 3: Visualization of the multi-source RIR predictions. We show two example acoustic scenes. Each scene contains two predicted RIRs (in red) from two sources, with ground truth (in blue).

Separator	RIR Model	IA-SoundSpaces				IA-Treble			
		EDT↓	T60↓	C50↓	STFT↓	EDT↓	T60↓	C50↓	STFT↓
TIGER	BuDDy	184	22.5	3.40	2.80	163	34.9	3.37	2.57
	FiNS	124	11.8	2.02	1.26	167	29.6	2.38	1.60
	Rec-RIR	143	10.0	2.48	1.51	214	23.3	3.54	1.83
TFGridNet	BuDDy	186	23.1	3.34	2.80	161	34.6	3.34	2.55
	FiNS	107	11.3	1.83	1.26	162	28.7	2.30	1.58
	Rec-RIR	105	8.71	1.90	1.48	189	24.3	3.76	1.73
ASCEND w/o CGS		66.2	7.74	1.08	0.84	57.4	14.8	1.07	1.29
ASCEND w/ CGS		**56.6**	**6.37**	**1.01**	**0.81**	**55.8**	**14.4**	**1.04**	**1.28**

Table 1: Quantitative results on multi-source RIR estimation.

4.2 METRICS AND BASELINES

RIR Metrics. We evaluate predicted RIRs using standard acoustic metrics (Luo et al., 2022; Su et al., 2022; Lan et al., 2024; Wang et al., 2024). Specifically, we measure the accuracy of energy decay and temporal structure using early decay time (EDT), reverberation time (T60), clarity (C50), and multi-resolution STFT error. Details are in App. B.1.

Sound Extraction Metrics. We evaluate dry sound extraction for both speech and singing vocals. We evaluate the anechoic sounds on multi-resolution STFT error to evaluate signal similarity. For speech, we report an intrusive reference-based metric Perceptual Evaluation of Speech Quality Improvement (PESQi) (Rix et al., 2001), multi-resolution STFT error and the non-intrusive DNS-MOS (Reddy et al., 2021) for perceptual quality and naturalness. For singing vocals, we report multi-resolution STFT error and latent-distribution KL divergence.

RIR Baselines. Since no existing one-stage method estimates multiple source-specific RIRs from a multi-source mixture, we compare against cascaded pipelines. We first separate the reverberant mixture into single-source reverberant signals by training two strong separation models, TIGER (Xu et al., 2025a) and TF-GridNet (Wang et al., 2023). We then estimate each source RIR with single-source RIR estimators FiNS (Steinmetz et al., 2021), Rec-RIR (Wang & Li, 2025), and BUDDy (Lemercier et al., 2025), which we have trained them and evaluated them on these separate reverberant sounds.

Sound Extraction Baselines. For direct comparison, we train TIGER and TF-GridNet to estimate anechoic speech and singing vocals directly from the reverberant mixture. We also evaluate cascaded pipelines that first separate the mixture into single-source reverberant signals, and then apply single-source generative dereverberation models, SGMSE (Richter et al., 2023) and STORM (Lemercier et al., 2023), that are trained and evaluated on the single-source reverberant sound.

4.3 RESULTS OF INVERSE ACOUSTICS

Multi-source RIR Estimation. Tab. 1 reports multi-source RIR estimation results. ASCEND consistently outperforms all cascaded baselines across both datasets and all metrics. Even without CGS, our conditional DiT substantially improves over pipelines that first separate reverberant sources

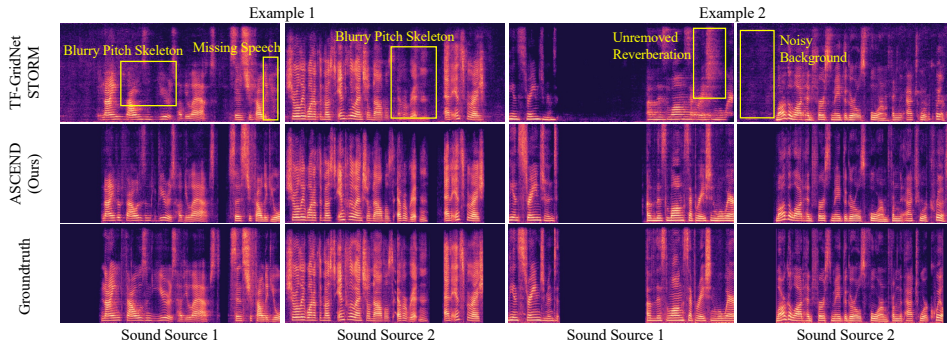

Figure 4: Visualization of the estimated anechoic speech for two examples.

Separator	Dereverb	IA-SoundSpaces					IA-Treble				
		Speech Metrics			Song Metrics		Speech Metrics			Song Metrics	
		PESQi↑	STFT↓	MOS↑	STFT↓	KL↓	PESQi↑	STFT↓	MOS↑	STFT↓	KL↓
TIGER	–	0.071	2.760	2.86	2.748	1.41	0.043	2.746	2.70	2.562	1.26
	SGMSE	0.329	2.153	2.91	2.043	0.53	0.319	2.194	3.03	2.024	0.45
	STORM	0.375	2.017	3.08	1.918	0.47	0.327	2.172	3.10	2.001	0.43
TFGridNet	–	0.588	2.088	2.98	2.178	0.57	0.451	2.107	2.97	2.049	0.48
	SGMSE	0.659	1.922	3.21	1.764	0.39	0.524	2.013	3.26	1.770	0.36
	STORM	0.679	1.837	3.37	1.654	0.37	0.503	2.009	3.25	1.769	0.37
ASCEND w/o CGS		0.716	1.432	**3.62**	1.439	0.16	0.583	1.615	**3.57**	1.626	0.21
ASCEND w/ CGS		**0.741**	**1.397**	3.59	**1.410**	**0.15**	**0.644**	**1.578**	3.56	**1.579**	**0.20**

Table 2: Quantitative results on the multi-source anechoic sound estimation tasks.

and then apply single-source RIR estimators, demonstrating the benefit of jointly modeling source-RIR pairs. CGS further improves all metrics, showing that enforcing the acoustic forward model during sampling yields RIRs that better explain the observed mixture. On IA-SoundSpaces, ASCEND reduces the best baseline errors 105 ms to 56.6 ms for EDT, and 1.83 dB to 1.01 dB for C50. On the more challenging IA-Treble, the gap is larger: the best baseline has 23.3% T60 error and 2.30 dB C50 error, while ASCEND achieves 14.6% and 1.04 dB. These results show that joint decomposition with explicit forward consistency is more effective than estimating reverberant sources and RIRs in separate stages. Qualitative examples in Fig. 3 show that ASCEND captures finer temporal structure.

Multi-source Anechoic Sound Source Estimation. Tab. 2 reports dry sound extraction results for speech and singing vocals. In a single-stage setting, the end-to-end TIGER baseline yields only marginal PESQi and weak STFT similarity, while single-stage TF-GridNet provides a much stronger baseline, with better perceptual quality and signal similarity across both datasets due to its stronger time-frequency modeling. Connecting with a generative model (SGMSE or STORM) consistently exhibits clear gains compared to a single-stage method in quality, signal fidelity, and naturalness in speech, as well as signal and distribution similarity in vocal songs. ASCEND outperforms both direct separation models and cascaded separation-dereverberation pipelines on nearly all metrics. On IA-SoundSpaces, ASCEND improves the best baseline PESQi and reduce speech and vocal song STFT error strongly. The flow matching variant in Table 3 obtains even more significant gains. On IA-Treble, it improves baseline PESQi with a larger margin while constantly reducing vocal song and speech STFT error. CGS improves most intrusive, reconstruction and distribution metrics, including PESQi, STFT and KL, while DNSMOS remains largely comparable, suggesting that forward consistency improves signal fidelity without sacrificing perceptual quality. Qualitative examples in Fig. 4 show cleaner anechoic speech with less residual reverberation and fewer cross-source artifacts.

User Study. We conduct a listening study (32 participants, 20 utterances) rating quality, intelligibility, and similarity to GT on a 5-point MOS scale (see App. E for full details). ASCEND outperforms both baselines on all three dimensions for both speech and vocal songs. Our method improving over TFGridNet+STORM by around 0.7 points on MOS, confirming consistent gains.

Ablation Studies. We ablate two key components of ASCEND: CGS and the global-local attention. Results are summarized in Tab. 3. Using either time-domain CGS or time and frequency domain CGS consistently improves both RIR estimation and anechoic speech extraction over standard sampling, showing that explicit forward acoustic consistency is effective for this inverse problem. The addition of frequency domain CGS provide better RIR estimation metrics, while providing trade-off

	Variant	RIR Estimation			Anechoic Sound		
		EDT ↓	T60 ↓	C50 ↓	PESQi ↑	Speech-STFT ↓	Song-STFT ↓
CGS loss	none	66.2 (0.6)	7.74 (0.06)	1.08 (0.01)	0.716 (0.004)	1.432 (0.003)	1.439 (0.004)
	$\mathcal{L}_{\text{time}}$	58.0 (0.5)	6.75 (0.05)	**1.00** (0.01)	**0.749** (0.004)	1.416 (0.003)	1.421 (0.005)
	$\mathcal{L}_{\text{time}} + \mathcal{L}_{\text{freq}}$	**56.6** (0.5)	**6.37** (0.05)	1.01 (0.01)	0.741 (0.004)	**1.397** (0.003)	**1.410** (0.005)
CGS scale	$\lambda = 0.01$	62.5 (0.6)	7.29 (0.05)	1.04 (0.01)	0.730 (0.004)	1.407 (0.003)	**1.412** (0.005)
	$\lambda = 0.02$	59.8 (0.5)	6.74 (0.05)	1.03 (0.01)	0.740 (0.004)	1.398 (0.003)	**1.406** (0.005)
	$\lambda = 0.04$	**56.6** (0.5)	**6.37** (0.05)	**1.01** (0.01)	**0.741** (0.004)	**1.397** (0.003)	1.410 (0.005)
Attention	Self-attention	71.1 (0.7)	8.14 (0.07)	1.25 (0.02)	0.655 (0.004)	1.485 (0.003)	1.491 (0.004)
	Global-local	**66.2** (0.6)	**7.74** (0.06)	**1.08** (0.01)	**0.716** (0.004)	**1.432** (0.003)	**1.439** (0.004)
Flow matching	w/o CGS	66.6 (0.6)	7.69 (0.06)	1.07 (0.01)	0.757 (0.004)	1.451 (0.003)	1.468 (0.005)
	w/ CGS	**58.6** (0.5)	**6.38** (0.05)	**0.97** (0.01)	**0.798** (0.004)	**1.413** (0.004)	**1.420** (0.005)

Table 3: Ablation results for CGS and global-local attention. CGS improves both RIR estimation and anechoic speech extraction, while global-local attention outperforms standard self-attention. Gray numbers are the bootstrapped standard deviation on the test mixtures within a same random seed.

in other anechoic sound metrics. The inclusion of bootstrapped standard deviation shows that all improvements from using CGS exceed a 95% confidence interval (2 standard deviation). We also vary the CGS guidance scale λ and observe small trade-offs across metrics, with $\lambda = 0.04$ giving the best overall performance. Finally, replacing global-local attention with standard self-attention degrades performance under the same total number of attention layers and computation. This suggests that global attention helps capture cross-source separation cues, while local attention preserves source-specific dereverberation and RIR estimation structure. We also evaluate the CGS on the flow matching objective, we find that CGS constently improves all metrics on flow matching, showing the value of CGS to various score-based frameworks.

Real-world Results. We further evaluate sim-to-real transfer on real recordings. All models are trained only on simulated data and tested on real recordings without fine-tuning. As shown in Tab. 4, ASCEND achieves the best results on both RIR estimation and anechoic speech extraction. For speech extraction, ASCEND outperforms both the single-stage TFGridNet baseline and the strongest cascaded TFGridNet+STORM baseline. For RIR estimation, it also substantially improves over TFGridNet+FiNS. These gains suggest that the proposed unified generative formulation captures acoustic propagation more faithfully and transfers better to real environments.

Effect of number of sources. We analyze the effect of the assumed number of sources by varying it from 1 to 3 in Fig. 5. Performance drops substantially from one to two sources, reflecting the added difficulty of separating overlapping sources. In contrast, moving from two to three sources causes only a modest degradation, indicating that ASCEND remains capable as scene complexity increases.

Method	RIR Estimation			Anechoic Sound		
	EDT	T60	C50	PESQi	DNSMOS	STFT
TFGridNet	–	–	–	0.348	3.02	2.59
TFGridNet + STORM	–	–	–	0.418	3.33	2.26
TFGridNet + FiNS	336	64.3	3.79	–	–	–
ASCEND	**64**	**22.0**	**1.22**	**0.483**	**3.68**	**1.74**

Table 4: Real-world results without fine-tuning. – means a method does not support the corresponding task.

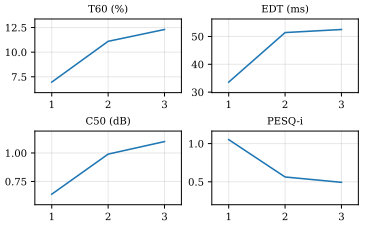

Figure 5: Effect of source number.

5 DISCUSSION

Limitation and Future work. Our forward acoustic model assume linear-time invariant (LTI). Important directions include extending our framework to dynamic acoustic scenes with moving sources and time-varying environments, as well as scaling to open-domain sound sources beyond speech. Another promising avenue is handling variable and unknown numbers of sources, making the model more robust to real-world settings.

Conclusion. We presented ASCEND, a diffusion-based framework for inverse acoustics that jointly decomposes a reverberant mixture into dry sources and their source-specific RIRs for multiple sound sources under a unified generative formulation. Beyond the immediate task, inverse acoustics opens a broader direction toward spatial auditory intelligence, where machines can infer, reconstruct, and reason about the physical world through sound. As generative models and multimodal perception systems continue to advance, such capabilities may enable acoustically aware digital twins, immersive AR/VR experiences, embodied robotic perception, and richer audio-visual scene understanding.

AI USE STATEMENT

ChatGPT was used to assist with language editing and manuscript preparation, as well as with code verification and unit tests. All AI-assisted suggestions and outputs were reviewed and verified by the authors, who take responsibility for the final content of the paper, including its text, claims, code, and experimental results. The proposed method itself is a conditional diffusion model with Consistency-Guided Sampling (CGS), as described in the Method and Implementation Details sections. Diffusion-based generation is also used as part of the proposed methodology to produce the reconstructed audio outputs presented in the accompanying audio demo.

ETHICS STATEMENT

Our user study involved only listening to audio samples, including speech and singing vocals produced by multiple methods. Under the applicable institutional guidelines, the study did not require IRB review. Participants were recruited on campus, were informed of the study procedure, and could withdraw from the study at any time. The study used a blind, randomized listening protocol in which participants evaluated quality, intelligibility, and similarity to ground truth using a five-point MOS scale. All datasets used in this work were obtained and used in accordance with their applicable licenses.

This work advances methods for decomposing complex acoustic scenes into their constituent sound sources and room acoustics. Potential beneficial applications include improving hearing aids and assistive-listening technologies, enhancing speech communication in noisy environments such as teleconferencing and emergency-response settings, supporting acoustic forensics and archival audio restoration, and enabling more accurate modeling for architectural acoustics, virtual and augmented reality, and spatial-audio rendering. At the same time, improved speech separation and recovery methods may raise privacy concerns if applied to recordings of private conversations. Our method requires access to the underlying audio recording and provides capabilities related to those of existing speech-separation systems. We therefore encourage its use only in settings with appropriate consent, data governance, and privacy protections.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, model architecture, training objectives, inference procedure, dataset construction, evaluation metrics, and experimental settings in the main paper and appendices. We also included pseudocode for Consistency-Guided Sampling. We additionally report uncertainty estimates for the ablation experiments using bootstrapped standard deviations. The source code and supporting implementation will be released publicly upon acceptance.

REFERENCES

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 17–36. Springer, 2020.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022.

Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. Children's song dataset for singing voice research.

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.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural*

Information Processing Systems, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt.

Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9, 2013. doi: 10.1109/APSIPA.2013.6694316.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. ISSN 01621459. URL http://www.jstor.org/stable/23239562.

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. IEEE, 2025.

Chen Geng, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. Birth and death of a rose. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 26102–26113, June 2025.

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

Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YCKuXkw6UL.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2724–2737, July 2023. ISSN 2329-9290. doi: 10.1109/TASLP.2023.3294692. URL https://doi.org/10.1109/TASLP.2023.3294692.

Jean-Marie Lemercier, Eloi Moliner, Simon Welker, Vesa Välimäki, and Timo Gerkmann. Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models, 2025. URL https://arxiv.org/abs/2408.07472.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5404–5411, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Jinglin Liu, Chengxi Li, Yi Ren, Zhiying Zhu, and Zhou Zhao. Learning the beauty in songs: Neural singing voice beautifier. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022b. URL https://arxiv.org/abs/2209.03003.

Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022.

Yi Luo and Nima Mesgarani. Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network. In *Interspeech 2018*, pp. 342–346, 2018. doi: 10.21437/Interspeech.2018-2290.

Shahar Lutati, Eliya Nachmani, and Lior Wolf. Separate and diffuse: Using a pretrained diffusion model for better source separation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UXALv0lJZS.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6875–6879, 2019. doi: 10.1109/ICASSP.2019.8683634.

Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards improved room impulse response estimation for speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094770.

Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27164–27175, 2024.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, 2021. URL https://arxiv.org/abs/2010.15258.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023. doi: 10.1109/TASLP.2023.3285241.

Antony W. Rix, John G. Beerends, Mike Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2:749–752 vol.2, 2001. URL https://api.semanticscholar.org/CorpusID:5325454.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Bidisha Sharma, Xiaoxue Gao, Karthika Vijayan, Xiaohai Tian, and Haizhou Li. Nhss: A speech and singing parallel database. *Speech Commun.*, 133(C):9–22, October 2021. ISSN 0167-6393.

Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 286–295, October 2021.

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*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Christian J. Steinmetz and Joshua D. Reiss. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020.

Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. Filtered noise shaping for time domain room impulse response estimation from reverberant speech. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 221–225, 2021. doi: 10.1109/WASPAA52581.2021.9632680.

Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022.

Treble Technologies. Treble software development kit. https://www.treble.tech/software-development-kit, 2026. Accessed: 2026-05-04.

Helin Wang, Jiarui Hai, Dongchao Yang, Chen Chen, Kai Li, Junyi Peng, Thomas Thebaud, Laureano Moro Velazquez, Jesus Villalba, and Najim Dehak. Solospeech: Enhancing intelligibility and quality in target speech extraction through a cascaded generative pipeline, 2025. URL https://arxiv.org/abs/2505.19314.

Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11790–11799, 2024.

Pengyu Wang and Xiaofei Li. Rec-rir: Monaural blind room impulse response identification via dnn-based reverberant speech reconstruction in stft domain, 2025. URL https://arxiv.org/abs/2509.15628.

Zhong-Qiu Wang and Shinji Watanabe. UNSSOR: Unsupervised neural speech separation by leveraging over-determined training mixtures. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=T5h69frFF7.

Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux. Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3476–3490, November 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129363. URL https://doi.org/10.1109/TASLP.2021.3129363.

Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094992.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments, 2019. URL https://arxiv.org/abs/1907.01160.

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Mohan Xu, Kai Li, Guo Chen, and Xiaolin Hu. TIGER: Time-frequency interleaved gain extraction and reconstruction for efficient speech separation. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=rzx3vcvlzj.

Zhongweiyang Xu, Xulin Fan, Zhong-Qiu Wang, Xilin Jiang, and Romit Roy Choudhury. ArrayDPS: Unsupervised blind speech separation with a diffusion prior. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=1Dq4rW1Oy4.

Yu Zhang, Changhao Pan, Wenxiang Guo, Ruiqi Li, Zhiyuan Zhu, Jialei Wang, Wenhao Xu, Jingyu Lu, Zhiqing Hong, Chuxin Wang, Lichao Zhang, Jinzheng He, Ziyue Jiang, Yuxin Chen, Chen Yang, Jiecheng Zhou, Xinyu Cheng, and Zhou Zhao. GTSinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=z64azPC6Nl.

Yunzhi Zhang, Shangzhe Wu, Noah Snavely, and Jiajun Wu. Seeing a rose in five thousand ways. In *CVPR*, 2023.

A CONSISTENCY LIKELIHOOD APPROXIMATION

In Section 3.2, we introduce a consistency likelihood term $\nabla_{x_t} \log p(y \mid x_t)$ to guide the conditional diffusion sampling process toward the measurement subspace. Because the forward acoustic model (Eq. 2) defines the measurement y only as a function of the clean acoustic scene x_0, this likelihood cannot be evaluated directly at an arbitrary intermediate noise scale $t > 0$.

To yield a tractable objective, we follow the alternating projection principles of Diffusion Posterior Sampling Chung et al. (2023) and adapt them for our conditional framework. The derivation proceeds in three distinct steps: marginalization, expectation approximation, and conditional substitution.

Marginalization and Conditional Independence. We first express the intractable likelihood $p(y \mid x_t)$ by marginalizing over all possible clean states x_0:

$$p(y \mid x_t) = \int p(y, x_0 \mid x_t) dx_0 = \int p(y \mid x_0, x_t) p(x_0 \mid x_t) dx_0$$

According to our physical forward model $y = \mathcal{A}(x_0) + \eta$, the measurement y is entirely determined by the clean state x_0 and the Gaussian measurement noise η. Consequently, once x_0 is known, y is conditionally independent of the intermediate noisy state x_t. This Markovian property simplifies the integral:

$$p(y \mid x_0, x_t) = p(y \mid x_0)$$

Substituting this back into the integral, we can rewrite the likelihood as an expectation over the reverse diffusion transition distribution $p(x_0 \mid x_t)$:

$$p(y \mid x_t) = \int p(y \mid x_0) p(x_0 \mid x_t) dx_0 = \mathbb{E}_{x_0 \sim p(x_0 \mid x_t)}[p(y \mid x_0)]$$

Expectation Approximation. Calculating the exact expectation of a highly non-linear measurement likelihood over a complex posterior $p(x_0 \mid x_t)$ is computationally prohibitive. Following Chung et al. (2023), we employ a first-order Taylor expansion to push the expectation inside the likelihood function. This allows us to approximate the expectation of the likelihood with the likelihood of the expectation:

$$p(y \mid x_t) \simeq p(y \mid \mathbb{E}_{x_0 \sim p(x_0 \mid x_t)}[x_0]) = p(y \mid \mathbb{E}[x_0 \mid x_t])$$

Taking the gradient of the log-likelihood yields the intermediate approximation:

$$\nabla_{x_t} \log p(y \mid x_t) \simeq \nabla_{x_t} \log p(y \mid \mathbb{E}[x_0 \mid x_t])$$

The Conditional Expectation Substitution. In standard unconditional diffusion models, the term $\mathbb{E}[x_0 \mid x_t]$ is analytically estimated using Tweedie's formula Efron (2011) alongside an unconditional noise prediction network. However, our proposed framework utilizes a conditional diffusion transformer parameterized by $v_\theta(x_t, y)$. Therefore, our network naturally computes the conditional expectation, $\mathbb{E}[x_0 \mid x_t, y]$. We approximate the unconditional expectation with our model's conditional prediction:

$$\mathbb{E}[x_0 \mid x_t] \approx \mathbb{E}[x_0 \mid x_t, y] := \hat{x}_0$$

This approximation is practical for three key reasons. Firstly, this ensures manifold consistency. During inference, our base generative sampling trajectory is already strongly constrained to the conditional manifold $p(x_t \mid y)$. Evaluating an unconditional expectation at a state x_t that doesn't lie in a y-conditioned subspace. Secondly, using a conditional prediction yields a smaller variance. The conditional posterior $p(x_0 \mid x_t, y)$ has lower entropy than the unconditional posterior $p(x_0 \mid x_t)$.

By utilizing the structural condition y, $\mathbb{E}[x_0 \mid x_t, y]$ provides a sharper, more accurate estimate of the true clean signal x_0. Finally, this also offers gradient stability. Because $\hat{x}_0$ is a more accurate estimate of the ground truth, the subsequent likelihood evaluation $p(y \mid \hat{x}_0)$ exhibits lower variance, leading to a more stable and informative consistency gradient $\nabla_{x_t} \log p(y \mid \hat{x}_0)$ during the reverse sampling steps.

By substituting this conditional expectation $\hat{x}_0$ into our approximation, we arrive at the final, computationally tractable gradient utilized in our Consistency Guided Sampling (CGS) method:

$$\nabla_{x_t} \log p(y \mid x_t) \simeq \nabla_{x_t} \log p(y \mid \hat{x}_0)$$

B Implementation details

B.1 Dataset and Metrics Details

Details of Simulated Dataset. We construct two large-scale simulated datasets, IA-SoundSpaces and IA-Treble, using two complementary acoustic simulators. For IA-SoundSpaces, we use the ray-tracing-based SoundSpaces 2.0 simulator Chen et al. (2020; 2022), which provides realistic RIR simulation in Gibson indoor environments Xia et al. (2018). For IA-Treble, we use Treble SDK [1], a ray-wave hybrid acoustic simulator that combines geometric acoustics with wave-based simulation. Compared with purely geometric simulators, the hybrid simulator better captures low-frequency wave effects, yielding more accurate RIRs in acoustically complex indoor scenes.

We include 33 training rooms and 11 testing rooms. For each simulated room, we sample multiple receiver locations (around 100). For each receiver location, we sample multiple source locations (around 50). The source-receiver distance is randomly sampled between 0.5 m and 15 m. This sampling strategy gives multiple source-specific RIRs that share the same receiver location, matching the physical setup of a single microphone receiving sound from multiple sources. Such fixed-receiver, multi-source RIR sets are crucial for our inverse acoustics task, since the observed mixture is formed by multiple dry sources propagating through different RIRs to the same microphone. We clip all RIRs to 0.5 seconds. We truncate all RIRs to 0.5 s to obtain a fixed representation while retaining nearly all energy in our datasets. Specifically, across our datasets, the average energy density within the initial 0.5 s is around 40000 times (on IA-Soundspaces) and 1200 times (on IA-Treble) larger than that between 0.5 s and the end of the RIR.

We synthesize reverberant mixtures by pairing the simulated RIRs with clean audio sources. For speech, we sample around 6-second utterances from LibriSpeech Panayotov et al. (2015). To construct two-speaker mixtures, we combine utterances from different speakers with a random temporal overlap ratio between 50% and 100%. The relative SNR between the two clean utterances is randomly sampled between -7 dB and 7 dB. We also sample 6-second clean vocal tracks from five singing voice corpora Choi et al.; Zhang et al. (2024); Sharma et al. (2021); Duan et al. (2013); Liu et al. (2022a) to construct speech-vocal mixtures. Ambient background noise is sampled from WHAM! Wichern et al. (2019), with the noise SNR randomly sampled between 0 dB and 15 dB relative to the source loudness.

Each clean source is convolved with its corresponding source-specific RIR to produce a reverberant source signal. The final observed mixture is obtained by summing the reverberant source signals and the ambient noise in the temporal domain:

$$y = \sum_{n=1}^{N} r_n * d_n + a.$$

The resulting mixtures are approximately 6 seconds long. For each sample, we store the observed reverberant mixture, the clean dry sources, the source-specific RIRs, and the ambient component, enabling supervised training and evaluation for joint source extraction and RIR estimation.

The resulting datasets are named IA-SoundSpaces and IA-Treble. IA-SoundSpaces contains 550k training mixtures (8k test mixtures), while IA-Treble contains 280k training mixtures (6k test mixtures). The test split is constructed to ensure that the evaluated RIRs, geometries, speech, vocal song, sound sources and speakers are not seen during training.

[1] https://www.treble.tech/software-development-kit

RIR Metrics. We evaluate predicted RIRs using standard acoustic metrics that characterize both temporal structure and energy decay behavior.

- **Early Decay Time (EDT).** EDT measures the early decay behavior of the RIR. It is computed from the initial decay slope of the energy decay curve, using the first 10 dB of decay and extrapolating to a 60 dB decay time. We report EDT error in milliseconds. Lower EDT error indicates that the predicted RIR better captures the perceived reverberance associated with the early decay region.
- **Reverberation Time (T60).** T60 measures the time required for the sound energy to decay by 60 dB. In practice, we estimate T60 from the energy decay curve by fitting the decay slope between -5 dB and -25 dB, and then extrapolating this slope to a 60 dB decay time. We compute T60 for both the predicted and ground-truth RIRs and report the relative error in percentage. Lower T60 error indicates that the predicted RIR more accurately matches the overall reverberation time.
- **Clarity (C50).** C50 measures the ratio between early-arriving energy and late reverberant energy in an RIR. Specifically, it compares the energy arriving within the first 50 ms to the energy arriving after 50 ms, and is reported in dB. Lower C50 error indicates that the predicted RIR better matches the early-to-late energy balance of the ground-truth RIR.
- **Multi-resolution STFT Error.** In addition to hand-crafted acoustic parameters, we compute a multi-resolution STFT error between the predicted and ground-truth RIR waveforms. This metric compares time-frequency structure across multiple STFT window sizes, capturing discrepancies in both fine temporal details and broader spectral decay patterns. Lower STFT error indicates closer waveform-level agreement between the predicted and ground-truth RIRs.

B.2 Baseline Details

We compare ASCEND with both direct and cascaded baselines for multi-source RIR estimation and anechoic sound extraction. We show the introduction as follow.

TF-GridNet. TF-GridNet Wang et al. (2023) is a strong time-frequency domain speech separation model that combines local spectral-temporal modeling with global sequence modeling. We use it as a reverberant source separator in cascaded pipelines. For the direct sound extraction baseline, we also train TF-GridNet to predict anechoic speech or singing vocals directly from the reverberant mixture. We use the ESPnet implementation `https://github.com/espnet/espnet/blob/master/espnet2/enh/separator/tfgridnet_separator.py` and train it on our datasets.

TIGER. TIGER Xu et al. (2025a) is a recent speech separation model that we use as another strong separator baseline. In the cascaded pipelines, TIGER first separates the input reverberant mixture into single-source reverberant signals, which are then passed to downstream RIR-estimation or dereverberation models. For direct anechoic sound extraction, we train TIGER to estimate dry speech or singing vocals directly from the reverberant mixture. We use the official implementation from `https://github.com/JusperLee/TIGER` and train it on our simulated datasets.

FiNS. FiNS Steinmetz et al. (2021) is a single-source blind RIR estimation method based on filtered noise shaping. Given a reverberant speech signal, FiNS directly estimates the corresponding time-domain RIR by predicting early reflections and shaping the late reverberation tail. Since FiNS assumes a single reverberant source, we use it after the reverberant source separation stage. We use the implementation `https://github.com/kyungyunlee/fins` for training and inference.

Rec-RIR. Rec-RIR Wang & Li (2025) is a monaural blind RIR identification method based on the convolutive transfer function approximation in the STFT domain. It estimates room-acoustic filters using neural blocks that model both narrow-band and cross-band structure. We evaluate Rec-RIR in a cascaded pipeline: a separator first estimates each reverberant source, and Rec-RIR then estimates the corresponding source-specific RIR. We use the official PyTorch implementation from `https://github.com/Audio-WestlakeU/Rec-RIR` for training and inference.

BUDDy. BUDDy Lemercier et al. (2025) is a diffusion-based method for single-channel blind dereverberation and RIR estimation. It performs posterior sampling with a speech prior and a

reverberation model to jointly infer the clean speech and room-acoustic parameters for a single reverberant source. Because BUDDy is not designed for multi-source mixtures, we use it as a single-source RIR-estimation baseline after separating the reverberant mixture with TIGER or TF-GridNet. Each separated reverberant source is processed independently by BUDDy to obtain its estimated RIR. We use the official implementation from `https://github.com/sp-uhh/buddy`.

SGMSE. SGMSE Richter et al. (2023) is a score-based generative model for speech enhancement and dereverberation in the complex STFT domain. It models clean speech generation through a stochastic reverse process conditioned on a degraded observation. We use SGMSE as a single-source generative dereverberation baseline in cascaded sound extraction pipelines. The input mixture is first separated into single-source reverberant signals using TIGER or TF-GridNet, and SGMSE is then applied to recover the corresponding dry source from the single-source signals. We use the official implementation from `https://github.com/sp-uhh/sgmse` for training and inference.

STORM. STORM Lemercier et al. (2023) is a stochastic regeneration model for speech enhancement and dereverberation. It uses a diffusion-based generative process to regenerate clean speech from degraded reverberant observations. We use STORM as another single-source dereverberation baseline in the cascaded sound extraction setting. After reverberant source separation with TIGER or TF-GridNet, STORM is applied to each separated source to estimate the corresponding anechoic signal. We use the official implementation from `https://github.com/sp-uhh/storm` for training and inference.

C ASCEND DETAILS

We implement ASCEND as a latent diffusion model operating over the acoustic scene components described in Sec. 3. The model contains four main parts: a sound VAE for dry sources and ambient sound, an RIR VAE for room impulse responses, a conditional diffusion transformer for joint acoustic scene generation, and diffusion or flow matching training and sampling details.

C.1 VAE DETAILS

Sound VAE. We use a time-frequency domain VAE Wang et al. (2025) to compress audio waveforms that have the same duration as the input mixture, including observed mixture, dry speech and dry singing vocals. Each waveform is encoded into a sequence of latent tokens with 128 channels at a temporal resolution of 50 Hz, and the decoder maps this latent sequence back to the waveform domain.

The VAE is optimized with reconstruction, adversarial, feature-matching, and latent-regularization objectives:

$$\mathcal{L}_{\text{sound_VAE}} = \lambda_{\text{stft}}\mathcal{L}_{\text{MR-STFT}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{FM}} + \lambda_{\text{kl}}\mathcal{L}_{\text{KL}}.$$

For waveform reconstruction, we use a perceptually weighted multi-resolution STFT objective $\mathcal{L}_{\text{MR-STFT}}$ computed with window sizes $[1280, 640, 320, 160, 80, 40, 20]$ and hop sizes $[320, 160, 80, 40, 20, 10, 5]$. The adversarial term $\mathcal{L}_{\text{adv}}$ is computed using five convolutional discriminators in the style of EnCodec, with 64 mel bins, window sizes $[1280, 640, 320, 160, 80]$, and hop sizes $[320, 160, 80, 40, 20]$. To improve training stability and preserve perceptual details, we add a feature-matching objective $\mathcal{L}_{\text{FM}}$ between discriminator activations. The KL penalty $\mathcal{L}_{\text{KL}}$ encourages a regularized latent space.

The loss weights are $\lambda_{\text{stft}} = 1.0$, $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{fm}} = 5.0$, and $\lambda_{\text{kl}} = 1 \times 10^{-4}$. We train the VAE with AdamW using a batch size of 32, an autoencoder learning rate of 1.5×10^{-4}, a discriminator learning rate of 3×10^{-4}, and weight decay of 1×10^{-3}.

The ambient sound VAE follows the same training and model configuration but is trained separately.

RIR VAE. Since RIRs are short acoustic filters with temporal statistics that differ from speech, vocals, and ambient waveforms, we train a separate VAE for RIR representation learning with the same architecture. The decoder maps this latent sequence back to the time-domain RIR.

We optimize the RIR VAE using waveform reconstruction, multi-resolution STFT, energy-decay, and KL regularization losses:

$$\mathcal{L}_{\text{RIR_VAE}} = \lambda_{\text{stft}}\mathcal{L}_{\text{MR-STFT}}(r, \hat{r}) + \lambda_{\text{edc}}\mathcal{L}_{\text{EDC}}(r, \hat{r}) + \lambda_{\text{kl}}\mathcal{L}_{\text{KL}}.$$

Here, $\mathcal{L}_{\text{MR-STFT}}$ encourages time-frequency agreement between the predicted and ground-truth RIRs, while $\mathcal{L}_{\text{EDC}}$ compares their energy decay curves to better preserve reverberation structure and late-decay behavior. The KL term regularizes the RIR latent space.

The RIR duration is 0.5 seconds, and RIRs are normalized using it standard deviation. We use loss weights $\lambda_{\text{stft}} = 1.0$, $\lambda_{\text{edc}} = 0.1$, and $\lambda_{\text{kl}} = 1 \times 10^{-4}$. The RIR VAE is trained with AdamW using batch size 32 and learning rate 1.5×10^{-4}. After training, both the sound VAE and RIR VAE are frozen and used to encode training targets and decode generated latents during consistency-guided sampling.

C.2 Consistency Guidance Details

Scale ambiguity is inherent to this inverse problem. Multiplying a dry source by a constant and dividing its RIR by the same constant produces the same reverberant signal. Therefore, the absolute gains of the dry source and RIR cannot be recovered separately from a single signal.

During VAE and diffusion-model training, the dry sources and RIRs are normalized to unit standard deviation. The model therefore learns normalized dry-source signals and normalized RIR structure, rather than absolute source loudness or propagation gain. During CGS inference and reconstruction, we also estimate maximum-likelihood-estimate (MLE) optimal per-slot gains between the reconstructed signals and the observed mixture before computing the consistency loss to make the loss and its gradient less sensitive to scale. This prevents an amplitude mismatch from dominating the gradient. We also train a separate network to predict ambient for rescaling issues.

Our evaluation also follows the same principle. The RIR metrics are scale invariant, and PESQi are generally unaffected by constant scaling factors. For the reported waveform-level STFT errors, we align each prediction to the ground truth with an optimal scalar before computing the error.

C.3 Diffusion Transformer

Our conditional diffusion transformer operates on the latent scene representation on the VAE latent space. At diffusion timestep t, Gaussian noise is added independently to each target latent component. The transformer receives the noisy component latents and the mixture condition latent, and predicts the velocity target for every component. The projected dry-source, RIR, and ambient tokens are concatenated into a single sequence with $(N + 1)M_s + NM_r$ tokens.

The diffusion transformer contains $L = 8$ global-local blocks with 16 total self-attention layers. For each attention layer, the hidden dimension is $D_{\text{dit}} = 1024$ with 16 attention heads.

The model is trained with the standard v-prediction objective. For each clean latent $\mathbf{z}_0$ and Gaussian noise $\boldsymbol{\epsilon}$, we construct

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \mathbf{v}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon} - \sqrt{1 - \bar{\alpha}_t}\mathbf{z}_0.$$

The DiT predicts $\hat{\mathbf{v}}_t = v_\theta(\mathbf{z}_t, \mathbf{z}_y, t)$ and is optimized with an MSE loss over all dry-source, RIR, and ambient latents. We train the diffusion transformer for 300 epochs using the AdamW optimizer with learning rate 5×10^{-5}, weight decay 1×10^{-4}, batch size 128, and diffusion timesteps 1000. During inference, we use DDIM sampling with 50 denoising steps.

C.4 Diffusion Training and Sampling Objective

We train the diffusion variant using the v-prediction objective. The schedule is rescaled to enforce zero terminal signal-to-noise ratio (Lin et al., 2024).

For each clean scene latent $\mathbf{z}_{\text{clean}}$, we sample t uniformly and independent Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noisy state and velocity target are

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_{\text{clean}} + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon},$$
$$\mathbf{v}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{\epsilon} - \sqrt{1 - \bar{\alpha}_t}\,\mathbf{z}_{\text{clean}}.$$

The DiT predicts $\hat{\mathbf{v}}_t = \mathrm{DiT}_\theta(\mathbf{z}_t, \mathbf{z}_y, t)$, conditioned on the observed mixture latent $\mathbf{z}_y$. We compute MSE separately for each component stream and sum the losses:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_{\text{clean}}, \mathbf{z}_y, \boldsymbol{\epsilon}, t}\left[\mathrm{MSE}(\hat{\mathbf{v}}_{a,t}, \mathbf{v}_{a,t}) + \sum_{n=1}^{N}\left(\mathrm{MSE}(\hat{\mathbf{v}}_{d,n,t}, \mathbf{v}_{d,n,t}) + \mathrm{MSE}(\hat{\mathbf{v}}_{r,n,t}, \mathbf{v}_{r,n,t})\right)\right].$$

Each MSE averages over the entries of its corresponding stream.

At inference, we use deterministic DDIM sampling and 50 steps. Starting from a standard Gaussian latent, each step converts the predicted velocity into clean-latent and noise estimates:

$$\hat{\mathbf{z}}_{\text{clean}} = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\mathbf{v}}_t,$$
$$\hat{\boldsymbol{\epsilon}}_t = \sqrt{1-\bar{\alpha}_t}\,\mathbf{z}_t + \sqrt{\bar{\alpha}_t}\,\hat{\mathbf{v}}_t.$$

The base DDIM update is

$$\mathbf{z}_{\text{prev}} = \sqrt{\bar{\alpha}_{\text{prev}}}\,\hat{\mathbf{z}}_{\text{clean}} + \sqrt{1-\bar{\alpha}_{\text{prev}}}\,\hat{\boldsymbol{\epsilon}}_t,$$

where $\bar{\alpha}_{\text{prev}}$ corresponds to the next selected timestep. CGS augments this base update with the consistency correction described in Sec. 3.2 and Alg. 1.

C.5 Flow Matching Extensions

Algorithm 2 Consistency Guided Sampling (CGS) for Flow Matching

Require: Observed Mixture y, Conditional Flow Model DiT_θ, VAE Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$, Forward acoustic model $\mathcal{A}$, Time grid $0 = t_0 < t_1 < \cdots < t_K = 1$, Guidance scale $\lambda > 0$, consistency weight w

$\mathbf{z}_y \leftarrow \mathcal{E}(y)$ ▷ Encode the observed mixture condition
$\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Initialize latents for $\{\mathbf{z}_{d,n}, \mathbf{z}_{r,n}\}_{n=1}^{N}$ and $\mathbf{z}_a$ ($t = 0$ is noise)
for $k = 0$ **to** $K - 1$ **do**
  $t \leftarrow t_k, \quad \Delta t \leftarrow t_{k+1} - t_k$
  $\hat{\mathbf{v}}_t \leftarrow \mathrm{DiT}_\theta(\mathbf{z}_t, \mathbf{z}_y, t)$ ▷ Predict flow velocity
  $\mathbf{z}_{t+\Delta t} \leftarrow \mathbf{z}_t + \Delta t\,\hat{\mathbf{v}}_t$ ▷ Euler step
  $\hat{\mathbf{z}}_1 \leftarrow \mathbf{z}_t + (1-t)\,\hat{\mathbf{v}}_t$ ▷ Estimate clean latents
  $\hat{y}_t \leftarrow \mathcal{A}(\mathcal{D}(\hat{\mathbf{z}}_1))$ ▷ Reconstruct the mixture
  $\mathcal{L}_{\text{time}} \leftarrow \|y - \hat{y}_t\|_2^2$, $\mathcal{L}_{\text{freq}} \leftarrow \ell_{\text{freq}}(y, \hat{y}_t)$
  $\mathbf{g}_t^{\text{CGS}} \leftarrow (1-w)\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{time}} + w\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{freq}}$ ▷ Weighted consistency directions
  $\mathbf{z}_{t+\Delta t} \leftarrow \mathbf{z}_{t+\Delta t} - \lambda_t \mathbf{g}_t^{\text{CGS}}$ ▷ Guided Euler update
end for
return $\mathcal{D}(\mathbf{z}_1)$ ▷ Return the final decoded acoustic components

We implement the flow-matching variant using a rectified-flow objective. For each clean scene latent $\mathbf{z}_1$, we sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a logit-normal timestep $t = \mathrm{sigmoid}(u)$, where $u \sim \mathcal{N}(0, 1)$. The intermediate state and regression target are

$$\mathbf{z}_t = (1-t)\boldsymbol{\epsilon} + t\mathbf{z}_1, \qquad \mathbf{v} = \mathbf{z}_1 - \boldsymbol{\epsilon}.$$

Thus, $t = 0$ denotes noise and $t = 1$ denotes data. We compute the velocity-prediction MSE separately for each component stream and sum the resulting losses. Each MSE averages over the entries of its corresponding stream.

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_y, \boldsymbol{\epsilon}, t}\left[\mathrm{MSE}(\hat{\mathbf{v}}_{a,t}, \mathbf{v}_a) + \sum_{n=1}^{N}\left(\mathrm{MSE}(\hat{\mathbf{v}}_{d,n,t}, \mathbf{v}_{d,n}) + \mathrm{MSE}(\hat{\mathbf{v}}_{r,n,t}, \mathbf{v}_{r,n})\right)\right].$$

At inference, we initialize $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and integrate from noise to data using $K = 50$ Euler steps on the uniform grid $t_k = k/K$, with $\Delta t = 1/K$. The base update is $\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \Delta t\,\hat{\mathbf{v}}_t$. Sampling is deterministic given the initial latent and mixture condition. For CGS, the straight-line parameterization provides the clean estimate $\hat{\mathbf{z}}_1 = \mathbf{z}_t + (1-t)\hat{\mathbf{v}}_t$. We decode this estimate to evaluate the reconstruction losses and subtract the resulting consistency correction from the Euler update:

$$\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \Delta t\,\hat{\mathbf{v}}_t - \lambda_t \mathbf{g}_t^{\text{CGS}}.$$

Algorithm 2 summarizes the sampling procedure.

D REAL-WORLD DATASET COLLECTION.

We evaluate sim-to-real generalization in real-world dataset. As no existing dataset contains a large corpus of multiple RIRs at different source locations, but at the same receiver locations across different rooms, we collecta dn construct a real-world acoustic evaluation dataset. We use a single microphone and a single loudspeaker in real indoor environments. All recordings are collected in five rooms with diverse acoustic characteristics, including classrooms, meeting rooms, and a large auditorium. We use a Logitech Yeti Nano microphone as the receiver and a Bose SoundLink Revolve II speaker as the sound source emitter. During data collection, for each loudspeaker position, we first play clean speech signals sampled from the LibriSpeech Panayotov et al. (2015) and five singing voice corpora Choi et al.; Zhang et al. (2024); Sharma et al. (2021); Duan et al. (2013); Liu et al. (2022a) test set, consistent with the speech sources used in our simulated dataset. The reverberant recordings captured by the microphone therefore contain the room-dependent propagation effects corresponding to each source location. To construct multi-source mixtures, we combine reverberant recordings captured at the same microphone location but generated from different loudspeaker positions within the same room. This setup enables controlled construction of multi-source acoustic mixtures while preserving realistic room acoustics and shared receiver conditions. To obtain ground-truth RIRs, we additionally perform explicit RIR measurements for every speaker-microphone pair. Specifically, we play a 5-second linear chirp signal spanning frequencies from 50 Hz to 20 kHz through the loudspeaker and record the response using the microphone. The recorded chirp response is then deconvolved with the original excitation signal to recover the corresponding RIR. These measured RIRs are used as ground truth for evaluating real-world RIR estimation performance. We record at 48kHz and then resample to 16kHz for evaluation in accordance to our simulated dataset.

E ADDITIONAL EXPERIMENT

User Study Details. We recruit 32 participants (graduate students and researchers with normal hearing) and present them with 20 randomly selected test utterances. Listeners evaluate audio samples from ASCEND and the two strongest baselines in a blind, randomized order. Participants rate each sample on three dimensions using a 5-point MOS scale (1 = poor, 5 = excellent): (1) **Quality**: overall audio quality and absence of artifacts; (2) **Intelligibility**: how clearly the speech content can be understood; (3) **Similarity to GT**: perceived closeness to the ground-truth clean speech.

Method	Sound Metrics		
	Quality ↑	Intelligibility ↑	Sim. to GT ↑
TFGridNet	2.66±0.34	2.70±0.41	2.50±0.28
TFGridNet + STORM	2.89±0.40	2.93±0.35	2.79±0.40
ASCEND (ours)	**3.55**±0.34	**3.64**±0.33	**3.55**±0.32

Table 5: User study MOS results on a 1–5 scale, reported as mean ± participant standard deviation. We report perceptual metrics, including quality, intelligibility, and similarity to ground truth.