SVEET: PUBLIC-DATA DESIGN AND MULTI-SCALE FLOW MATCHING FOR SINGING VOICE ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Singing voice enhancement (SVE) restores studio-quality vocals from degraded recordings, but clean singing data are scarce and existing systems often rely on proprietary stems or speech models adapted with limited singing data. We study SVE from a public-data design perspective, systematically varying the composition and scale of public speech and singing corpora while constructing a controlled benchmark for both clean-input preservation and restoration. Building on this data design, we introduce SVEET, a conditional flow-matching enhancer, and SVEET-MS, which adds a degradation-adaptive multi-scale objective for spectrally coherent restoration. Both outperform public zero-shot baselines and a strong discriminative system on singing enhancement, while transferring to speech enhancement with the same checkpoint. On the real-world SingVERSE benchmark, they achieve, to our knowledge, the best perceptual quality among systems with publicly available weights. In a blind listening test, SVEET-MS is preferred over SVEET in 71–76% of decided trials despite only modest differences in objective metrics. Code, data recipes, evaluation protocols, and checkpoints are released.

1 INTRODUCTION

Singing voice enhancement (SVE) aims to recover a studio-quality monophonic vocal from a degraded recording (additive noise, reverberation, clipping, instrumental bleed, band-limiting and analog distortion, often compounded) and must not alter recordings that are already clean, so benchmarks measure the preservation of clean inputs alongside restoration. SVE sits between two mature fields and is served well by neither: speech-enhancement models embed priors of spoken audio and mishandle the wide F_0 range, sustained vowels, vibrato and technique-dependent timbre of singing (Duan et al., 2013; Jiang et al., 2025), while music source separation models isolate stems from a mix but are not trained to remove non-musical degradations. The overarching bottleneck is data: speech enhancement (SE) scales with corpora such as the DNS Challenge (Reddy et al., 2020) and industrial systems such as AnyEnhance use 21,000 hours of proprietary audio (Zhang et al., 2025), whereas clean, publicly usable singing is scarce. Existing SVE systems either train on private studio stems (Zhang et al., 2025) or fine-tune speech restorers on modest singing sets (Jiang et al., 2025); how SVE training data should be composed (how degraded–clean pairs are rendered, in what speech-to-singing ratio, at what volume) has to our knowledge not been studied under a controlled budget. We study this from a public-data design perspective: with 126.4 hours of unique clean singing and public speech as the only clean sources, we render pairs under fixed degradation recipes, sweep composition and volume, train discriminative and flow-matching enhancers on the selected recipe, and release the manifests that reconstruct every pair (licences in Appendix F). Our contributions are:

- **A data recipe from controlled composition sweeps.** A standard discriminative backbone trained on our mixed speech-and-singing recipe substantially outperforms public zero-shot baselines; across the tested speech-to-singing ratios (3:1 to 1:3) performance is nearly flat, and adding rendered volume at a fixed ratio yields diminishing returns, so within the tested range composition and augmentation choices matter more than raw rendered volume.
- **A multi-scale conditional flow-matching enhancer, evaluated under a matched budget.** We adapt conditional flow matching (Lipman et al., 2023) to stereo 44.1 kHz SVE in the complex STFT domain with a mixture-anchored prior, and add a multi-scale, frequency-pooled

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.

vector-field objective whose coarse scales are weighted per training pair by the input degradation level. Trained on the identical data split, both systems exceed the public baselines in fidelity and perceptual quality on the singing benchmark and on two leakage-free speech benchmarks (PESQ against a speech-trained model excepted), match our discriminative system in fidelity and exceed it perceptually, and reach the best perceptual quality among publicly available systems on the real-world SingVERSE benchmark; under a matched budget and optimizer, the multi-scale variant adds a perceptual gain on degraded singing at no cost in overall fidelity and is preferred over the base objective in 71–76% of decided trials of a blind listening test despite near-identical objective scores.

- **A loss-scale pitfall for added objectives.** Reducing an added term by a per-element mean rather than, like the base loss, a per-example sum shrinks its gradient by the number of time–frequency bins (6.5×10^4 here): next to a sum-reduced base loss it then contributes almost nothing, and if the whole objective is reduced this way Adam's ϵ dominates most updates, so the model differs from the baseline through the optimizer rather than the loss geometry. All results here use scale-matched reductions.
- **Release.** Code, data-construction manifests, evaluation partitions, the listening-test application with aggregate results, and checkpoints, all through anonymous links.[1]

2 Related Work

Speech Enhancement and the Generalization Gap. Speech enhancement has progressed rapidly through architectures based on deep filtering, complex spectral masking, dual-path attention, state-space models and MetricGAN-style training (Schröter et al., 2023; Lu et al., 2023; Chao et al., 2024; Fu et al., 2021; Braun et al., 2021). Despite strong results on spoken benchmarks such as VoiceBank-DEMAND (Valentini-Botinhao, 2017), these models embed priors about speech-rate dynamics and vocal-tract resonance that do not transfer to singing: sustained phonation stretches vowels over seconds and speech enhancers misread the tail of a note as stationary noise, while rapid, wide-range pitch trajectories evade frame-based temporal tracking, producing over-attenuation and fluttering artifacts (Jiang et al., 2025). Appendix H.2 reproduces this failure for four public speech enhancers on our benchmark.

Singing Voice Enhancement. Dedicated SVE remains nascent. MBTFNet (Xu et al., 2023) is an early discriminative system; AnyEnhance (Zhang et al., 2025) is a unified generative restoration model covering speech and singing that is trained on 21,000 hours of proprietary audio and whose weights are not public; and SingVERSE (Jiang et al., 2025) recently provided the first real-world benchmark with studio references across 19 acoustic scenarios. Closest to our data study, Jiang et al. (2025) fine-tune VoiceFixer (Liu et al., 2022) and AnyEnhance on mixed speech-and-singing data and find that in-domain singing data improves SVE without hurting speech; we quantify it under a fixed public-data budget with controlled ratio and volume sweeps, study the training objective on top of it, and evaluate on SingVERSE against the strongest system with public weights in their study.

Generative Audio Restoration and Flow Matching. Generative modeling offers an alternative to discriminative regression, whose mean-squared objectives tend toward over-smoothed spectra. Diffusion and score-based enhancers reconstruct clean audio by reversing a stochastic process (Serrà et al., 2022; Welker et al., 2022; Richter et al., 2023; Lemercier et al., 2023; Scheibler et al., 2024), and related models target bandwidth extension and broader restoration (Andreev et al., 2023; Koizumi et al., 2023), at the cost of many sampling steps. Flow matching (Lipman et al., 2023) instead learns a deterministic transport whose near-straight paths can be integrated in few Euler steps; it has been applied to speech enhancement with a mixture-anchored prior (Lee et al., 2025; Wang et al., 2025), which we adopt as our starting point, but not, to our knowledge, to singing enhancement; speech-trained generative enhancers also inherit a speech prior, and, applied zero-shot to singing, FlowSE falls below the unprocessed input in fidelity on both noisy conditions of Table 1. Multi-resolution spectral losses are standard for waveform generators; our multi-scale objective differs in acting on

[1] Code: `https://anonymous.4open.science/r/sveet-0513`; checkpoint, manifests and benchmark audio: `https://huggingface.co/jeanaewaxy/sveet`; audio examples and listening-test application: `https://sveet-demo-31izth.netlify.app`.

the flow-matching vector field itself, and Section 4.4 contrasts it with a waveform-reconstruction auxiliary applied to the same model.

Music Source Separation vs. Audio Enhancement. Music source separation (MSS) architectures, notably the Demucs family (Défossez et al., 2019; Défossez, 2021; Rouard et al., 2023), decouple mixed stems (vocals, drums, bass, other) within clean studio mixes. While effective at separating a vocal from its accompaniment, they are not trained to remove non-musical degradations: environmental noise, low-rate quantization, room resonance or clipping in the underlying vocal pass straight into the vocal stem. We bridge this divide by retraining a high-capacity separation architecture as a joint separation-and-restoration enhancer, and we keep the strongest MSS models as zero-shot baselines throughout.

Data-Centric Audio Machine Learning. The field is moving from raw volume scaling toward curation and composition: in speech enhancement, curated, clean and balanced subsets can outperform models trained on much larger unvetted corpora (Li et al., 2025). We extend this view to singing by relying only on clean public corpora released for research use (VocalSet (Wilkins et al., 2018), OpenCPOP (Wang et al., 2022), M4Singer (Zhang et al., 2022), SingStyle111 (Dai et al., 2023) and GTSinger (Zhang et al., 2024)) and by measuring, rather than assuming, how composition and volume trade off under a fixed budget.

3 METHOD

We formulate singing voice enhancement as a supervised translation task. Let $y \in \mathbb{R}^L$ denote a degraded audio waveform of length L, and $x \in \mathbb{R}^L$ denote its perfectly aligned, studio-quality clean vocal counterpart. We investigate two distinct modeling paradigms trained on identical data distributions: a discriminative source-reconstruction model (f_ϕ) and a generative conditional flow-matching model (v_θ).

3.1 DISCRIMINATIVE BASELINE AND DATA COMPOSITION

Our discriminative baseline instantiates the Hybrid Demucs (`hdemucs`) architecture (Défossez, 2021), which processes signals via parallel time-domain and frequency-domain convolutional paths coupled with BiLSTM layers. We adapt this architecture for a two-source task, separating the input into a vocal target and a residual interference stem, and optimize the standard Demucs objective, an $\mathcal{L}_1$ loss on the time-domain waveform of both stems weighted equally, to decode $\hat{x} = f_\phi(y)$. Given a fixed cap of total training hours, we describe the dataset composition by the volumetric speech-to-singing ratio $\gamma = \mathcal{V}_{\text{speech}}/\mathcal{V}_{\text{singing}}$. All discriminative models share one configuration: 44.1 kHz stereo material supplied as 5 s paired segments and read with the Demucs default 11 s window, batch size 32, Adam at 3×10^{-4}, 8 epochs, EBU R128 loudness normalization of all manifests. We call the discriminative system retrained on our recipe DISCSVE.

3.2 CONDITIONAL FLOW MATCHING IN THE COMPLEX STFT DOMAIN

We instantiate conditional flow matching (CFM) (Lipman et al., 2023) in the complex STFT domain, adapting the formulation of Lee et al. (2025) to stereo 44.1 kHz singing audio. Let $x_0 \in \mathbb{C}^{2\times F\times T}$ represent the two-channel complex spectrogram of the clean vocal target, and let $c \in \mathbb{C}^{2\times F\times T}$ represent the complex spectrogram of the degraded mixture input. Each utterance pair is peak-normalized by the maximum absolute amplitude of the degraded waveform prior to the transform (the same factor is re-applied after inference), and the audio is transformed using a Hann-windowed STFT with an FFT size of 510 samples and a hop size of 256 samples, generating $F = 256$ frequency bins. To stabilize numerical bounds, the complex coefficients are amplitude-compressed as $x_{\text{comp}} = \lambda\,|x|^p \exp(i\angle x)$ with $p = 0.5$ and $\lambda = 0.15$.

We define the conditional probability path $p_t(x \mid x_0, c)$ at time $t \in [0, 1]$ as a stochastic linear interpolant that runs from the clean target at $t = 0$ to a mixture-anchored Gaussian prior at $t = 1$. Concretely, the mean interpolates linearly between the clean spectrogram and the degraded mixture, $\mu_t = (1-t)\,x_0 + t\,c$, while the marginal standard deviation grows linearly along the path, $\sigma_t = (1-t)\,\sigma_{\min} + t\,\sigma_{\max}$ with $\sigma_{\min} = 0$ and $\sigma_{\max} = 0.487$. At $t = 1$ this yields the

prior $p_1(x \mid c) = \mathcal{N}(c, \sigma_{\max}^2 I)$, a complex Gaussian centered precisely on the degraded mixture spectrogram; a prior sample is drawn as $x_1 = c + \sigma_{\max} z$ with $z \sim \mathcal{N}(0, I)$. Writing the state at time t as $x_t = \mu_t + \sigma_t z$, the interpolant is

$$x_t = (1 - t)\, x_0 + t\, c + \sigma_t z. \tag{1}$$

The continuous transport path is driven by a time-dependent velocity network $v_\theta(x_t, t \mid c)$, instantiated as an NCSN++ U-Net (Welker et al., 2022) that operates on the stacked real and imaginary parts of the state and conditioning spectrograms (eight real-valued input channels, with four output channels reassembled into the stereo complex-valued vector field), containing 65.7M parameters. The conditional flow-matching objective minimizes the mean-squared error between the network's prediction and the conditional target vector field obtained by differentiating the path, $\frac{\mathrm{d}x_t}{\mathrm{d}t} = (c - x_0) + (\sigma_{\max} - \sigma_{\min})\, z$:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\, x_0,\, z} \left\| v_\theta(x_t, t \mid c) - \left[(c - x_0) + (\sigma_{\max} - \sigma_{\min})\, z \right] \right\|^2. \tag{2}$$

At inference time, enhancement is achieved by initializing the state at the mixture-anchored prior $x_1 = c + \sigma_{\max} z$ and solving the deterministic ordinary differential equation (ODE) $\mathrm{d}x_t/\mathrm{d}t = v_\theta(x_t, t \mid c)$ in reverse time with a first-order Euler solver on the grid $t_N = 1 > \cdots > t_1 = t_\epsilon = 0.03$: the vector field is evaluated at $t_N, \ldots, t_1$ and the final step of size t_ϵ carries the state to $t = 0$ (Lee et al., 2025), so t_ϵ bounds the smallest time at which the network is queried, matching the training range $t \in [t_\epsilon, 1]$. We use $N = 5$, the implementation default fixed before any of our experiments; Appendix B sweeps N post hoc: $N = 5$ sits at the perceptual-quality plateau, and even a single Euler step exceeds every public zero-shot baseline in PESQ. The compressed spectrogram $\hat{x}_0$ is uncompressed and inverted with the inverse STFT. We call the base system SVEET and the variant trained with the objective of Section 3.3 SVEET-MS.

3.3 A Multi-Scale Vector-Field Objective

Equation 2 treats every time–frequency bin of the vector-field error identically, whereas in singing, where energy is concentrated in sharp harmonics and long sustained vowels, errors that are coherent across neighbouring bins (misplaced formant energy, broadband residual noise) are perceptually more damaging than errors that alternate in sign from bin to bin. We therefore add frequency-pooled penalties that re-weight the error toward its spectrally coherent components.

Let $\Delta = v_\theta(x_t, t \mid c) - v^\star$ denote the residual between the predicted vector field and the conditional target $v^\star = (c - x_0) + (\sigma_{\max} - \sigma_{\min})\, z$ of Equation 2, a complex tensor of shape $2 \times F \times T$. For a set of frequency-pooling scales $s \in \mathcal{S} = \{1, 2, 4, 8\}$, we form a coarsened residual by average-pooling Δ along the frequency axis with a non-overlapping kernel of size s (stride s, applied independently to the real and imaginary parts, with the time axis left unchanged), and measure its squared magnitude summed over the pooled grid:

$$\mathcal{L}_{\mathrm{ms}}^{s} = \tfrac{s}{2} \left\| \mathrm{Pool}_s(\Delta) \right\|_2^2, \tag{3}$$

where Pool_1 is the identity, so that $\mathcal{L}_{\mathrm{ms}}^{1}$ is the base CFM loss of Equation 2 as implemented (a per-example sum, with the conventional factor $\frac{1}{2}$), and the factor s compensates for the s-fold smaller pooled grid so that every scale weights the $F \times T$ plane equally. Pooling acts on the complex residual before squaring, so errors that alternate in sign across neighbouring bins cancel and the coarse scales ($s \in \{2, 4, 8\}$, down to 32 bands at $F = 256$) specifically penalize components coherent across 2–8 bins, while $s = 1$ keeps the penalty on the native harmonic grid; our hypothesis is that these coherent components are misplaced formant-scale energy and broadband residual noise, consistent with the residual-noise analysis of Section 4.7. With fixed scale weights the objective is a positive definite quadratic form $\Delta^\top (\sum_s w_s P_s^\top P_s) \Delta$, so its population minimizer is still the conditional expectation of the target field: it changes which errors a finite-capacity network prioritizes, not the field it converges to. The objective acts on the vector field itself and needs no inverse STFT at training time.

Degradation-adaptive weighting. Pooling only helps where there is something coherent to remove: on a clean input the target field reduces to the white-noise term $\sigma_{\max} z$, and the coarse terms then merely dilute the fine-structure penalty that transparency requires. We therefore gate the coarse scales per training pair by the input SNR ρ_i (dB, waveform domain, clean target vs. interference $y - x$) through a soft weight

$$a_i = \Big(1 + \exp\big(- \tfrac{\rho_c - \rho_i}{\tau} \big) \Big)^{-1}, \qquad \rho_c = 15\,\mathrm{dB},\ \tau = 5\,\mathrm{dB}, \tag{4}$$

and combine the scales with dyadic weights $w = (w_1, w_2, w_4, w_8) = (1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8})$, normalized per pair:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_i \left[\frac{\mathcal{L}^1_{\text{ms},i} + a_i \sum_{s>1} w_s \mathcal{L}^s_{\text{ms},i}}{1 + a_i \sum_{s>1} w_s} \right]. \tag{5}$$

Clean and near-clean pairs ($\rho_i \gtrsim 25\,\text{dB}$, $a_i \lesssim 0.12$) are thus trained under an objective that approaches the plain CFM loss (identity pairs, for which $\rho_i = \infty$, use exactly the CFM loss), heavily degraded pairs ($\rho_i \lesssim 5\,\text{dB}$, $a_i \gtrsim 0.88$) under the full multi-scale combination with coefficients $(0.533, 0.267, 0.133, 0.067)$, and every pair contributes with unit total weight. Because a_i depends on the training pair rather than only on the network input (x_t, t, c), the gated objective is a target-dependent re-weighting whose population minimizer is a weighted conditional expectation of the target field, not the plain conditional expectation of the fixed-weight case; this is intentional, since the gate is what keeps the coarse terms off clean inputs, and Section 4.4 estimates its effect empirically. The SNR is free at training time because every pair is rendered from a known clean segment; no estimate is needed at inference.

Loss scale and the optimizer. Equation 5 reduces every scale term, like the base loss, by a per-example sum. Our first implementation instead replaced the base loss by a per-element mean over all scales, including $s{=}1$, which made the whole objective and its gradient $F \times T$ times smaller; at that scale Adam's ϵ dominates the denominator for three quarters of the parameters, and the model differs from the baseline mainly through a different optimizer (Appendix D). All multi-scale results here use the scale-matched reduction.

3.4 Training Data and Schedules

The singing pool holds 126.4 hours of unique clean vocals from GTSinger, M4Singer, SingStyle111, VocalSet and OpenCPOP, and the main recipe draws 209.4 hours of unique clean speech from Hi-Fi TTS (Bakhturina et al., 2021) and VCTK (Yamagishi et al., 2019) (per-corpus hours in Table 14, Appendix H; licences in Appendix F); all hours are unique source audio before segmentation and recipe expansion, recomputable from the released manifests. Because the pool includes the full VCTK corpus, whose speakers p232 and p257 supply the VoiceBank-DEMAND test set, we exclude that benchmark and evaluate speech on the disjoint sets of Section 4.5. Degraded–clean pairs are rendered offline by mixing clean segments with MUSAN noise (Snyder et al., 2015) and MUSDB18 accompaniment (Stöter et al., 2018) under three fixed recipes (Appendix A); every pair is catalogued with its recipe, source clip, offset and gains, so the corpus can be reconstructed exactly.

DiscSVE and the composition sweeps were trained on eight NVIDIA H200 GPUs (8 epochs at batch 32, about 180,000 updates). The two flow-matching systems were trained from scratch under identical conditions on eight NVIDIA A100 GPUs: the same 720,000-pair manifest, Adam at 10^{-4}, weight EMA with decay 0.999, batch 64, random 256-frame (1.5 s) crops, and a fixed 47-epoch budget (528,750 updates). We report one training run per system (the SVEET-MS run was chosen from two training seeds by validation SI-SDR), evaluated at its final checkpoint without checkpoint selection, and every reported number averages three inference runs with independent prior samples. "Matched budget" therefore always refers to the two flow-matching objectives; DiscSVE shares the data but not the training budget.

4 Experiments

4.1 Evaluation Protocol

We construct a 44.1 kHz benchmark of four manually audited conditions that jointly probe clean preservation and restoration. The two *Clean* conditions are transparency tests: the input is the clean reference itself, so an ideal system passes it through unaltered, and any change is a cost. *Acapella (Clean)* contains pristine, isolated studio vocals of a-cappella style spanning low and high registers, falsetto, belting, and rapid articulation; *Singing w/ Acc (Clean)* contains isolated vocal stems taken from commercial-grade pop/rock productions, i.e. the vocal material that normally sits inside an accompanied mix. The two *Noisy* conditions degrade the same 39 source clips, so the 78 benchmark clips form 39 source clusters, which all statistics below respect. *Acapella*

(Noisy) adds broadband environmental noise, sensor saturation, low-frequency rumble, and synthetic room convolutions, measuring denoising and dereverberation; notably, our training recipes contain no reverberation (Appendix A), so dereverberation performance here reflects out-of-distribution generalization. *Singing w/ Acc (Noisy)*, the hardest condition, simulates poor live recordings in which instrumental bleed, heavy reverberation, and background noise co-occur. Source material is drawn from held-out manifests of MIR-1K (Hsu & Jang, 2010), commercial vocal packs (Vandalism Sounds, 2026), and manually double-filtered OpenSinger segments (Huang et al., 2021); the degradation sources used to construct the noisy conditions are fully disjoint from the MUSAN and MUSDB18 material used at training time. The benchmark comprises 78 manually audited clips (20/20/19/19 across the four conditions). We report SI-SDR, PESQ (wideband, after polyphase resampling to 16 kHz), and the Production Quality (PQ) axis of the Audiobox-Aesthetics model (Tjandra et al., 2025); all systems in Tables 1 and 8 are scored by one unified pipeline on the identical benchmark version (Appendix G tables use an earlier, internally consistent revision). Baselines are the music-separation systems HTDemucs and HDemucs and the flow-matching speech enhancer FlowSE; public speech-enhancement models collapse on the noisy singing conditions (screening in Appendix G).

4.2 Data Composition Sweeps

We freeze the HDemucs backbone and sweep the composition (full tables and discussion in Appendix G). At a fixed 1000-hour budget, speech-to-singing ratios from 3:1 to 1:3 give nearly identical results (overall SI-SDR spread 1.4 dB, PESQ spread 0.08, single seed); we adopt $\gamma = 1{:}1$, the strongest recipe on the hardest condition (Singing w/ Acc (Noisy): 17.78 dB, 3.12 PESQ), noting that this selection used an earlier revision of the benchmark rather than a separate validation set (Section 5). Removing the volumetric cap does not help at 1:1 (1000h $\rightarrow$ 1562.6h: -0.17 dB) but lifts the 1:2 recipe by $+1.1$ dB when combined with source-position shifts, so within the tested range composition and augmentation matter more than rendered volume; the sweep has no speech-only or singing-only recipe and the fixed epoch count gives larger manifests more updates, so we claim no general scaling law.

4.3 Main Results

Data recipes strengthen discriminative backbones. On the singing benchmark (Table 1), DiscSVE improves over its zero-shot counterpart by $+3.00$ dB overall SI-SDR and $+0.11$ PESQ (Acapella (Clean): 46.46 $\rightarrow$ 51.85 dB).

Flow matching matches the discriminative system in fidelity and exceeds it perceptually. On the identical data (matched data, not matched compute; Section 3.4), SVEET reaches 34.71 dB overall SI-SDR ($+1.87$ dB over DiscSVE, carried by the clean conditions), 3.84 PESQ and 7.45 PQ; on the noisy conditions the discriminative system keeps a small fidelity edge (Acapella (Noisy) 12.88 vs. 11.55 dB; Singing w/ Acc (Noisy) 15.29 vs. 15.22 dB) while the generative system is clearly ahead in PESQ ($+0.08$, $+0.33$) and PQ ($+0.33$, $+0.33$).

The multi-scale objective adds perceptual gains where restoration is needed. SVEET-MS, trained under the identical budget and optimizer, obtains the best overall SI-SDR (35.10 dB), PESQ (3.87) and PQ (7.48) of all systems. Relative to SVEET the gains concentrate on the degraded conditions: Acapella (Noisy) PESQ rises from 2.74 to 2.84 and PQ from 7.33 to 7.39, Singing w/ Acc (Noisy) PQ from 7.39 to 7.45, and on the hardest condition SVEET-MS attains the best SI-SDR (15.33 dB) and PQ (7.45) of all enhancement systems. The clean conditions are not degraded: PESQ and PQ stay within 0.02 of SVEET, and clean-input transparency improves (Acapella (Clean) 59.17 vs. 56.80 dB); the one fidelity cost is Acapella (Noisy) SI-SDR (-0.6 dB), so the overall SI-SDR gain is carried by the clean conditions rather than by restoration. Under a paired bootstrap clustered by the 39 source clips the PQ improvement is significant (95% CI $[+0.007, +0.052]$), whereas the SI-SDR ($+0.4$ dB) and PESQ ($+0.03$) differences lie within their intervals; SVEET-MS leads in 7 of the 12 condition cells, and the other five differ by at most 0.02 (PESQ, PQ) or 0.3 dB except for the Acapella (Noisy) SI-SDR cell above. Paired bootstrap intervals clustered by source confirm the ordering: both flow-matching systems exceed every public baseline on all three metrics (CIs exclude zero), and SVEET-MS exceeds DiscSVE on PESQ and PQ, whereas its $+2.3$ dB SI-SDR advantage, carried by the high-variance clean conditions, has an interval that includes zero.

Table 1: Singing benchmark (78 clips: 20/20/19/19 per subgroup; one unified pipeline). Cells: SI-SDR (dB; "SDR") / PESQ-WB / PQ; *Overall* = unweighted mean over the four subgroups. **Bold**: best enhancement system per column; "—": SI-SDR undefined because the unprocessed input equals the reference. The flow-matching rows share data, optimizer and a 47-epoch budget, use the final checkpoint, and average three inference runs. Significance (paired bootstrap CIs clustered by the 39 source clips) is discussed in the text.

	Overall			Acapella (Clean)			Acapella (Noisy)			Singing w/ Acc (Clean)			Singing w/ Acc (Noisy)		
System	SDR	PESQ	PQ	SDR	PESQ	PQ	SDR	PESQ	PQ	SDR	PESQ	PQ	SDR	PESQ	PQ
Unprocessed (noisy input)	—	3.08	7.40	—	4.64	7.64	−2.60	1.52	6.95	—	4.64	7.48	3.12	1.51	7.51
Public baselines (zero-shot)															
HTDemucs (zero-shot) (Rouard et al., 2023)	27.53	3.54	7.28	43.74	4.52	7.66	8.40	2.14	6.87	44.51	4.54	7.47	13.47	2.98	7.11
HDemucs (zero-shot) (Défossez, 2021)	29.84	3.63	7.26	46.46	4.55	7.64	9.73	2.26	6.82	48.33	4.59	7.40	14.83	3.13	7.18
FlowSE (zero-shot) (Lee et al., 2025)	8.25	2.25	7.02	18.59	3.08	**7.67**	−5.75	1.36	6.37	18.21	2.92	7.37	1.96	1.64	6.69
Ours: discriminative															
DISCSVE (1:1, 1000 h)	32.84	3.74	7.30	51.85	**4.59**	7.65	**12.88**	2.66	7.00	51.36	4.60	**7.49**	15.29	3.12	7.06
Ours: conditional flow matching															
SVEET	34.71	3.84	7.45	56.80	4.58	7.65	11.55	2.74	7.33	**55.27**	**4.61**	7.45	15.22	**3.45**	7.39
SVEET-MS	**35.10**	**3.87**	**7.48**	**59.17**	**4.59**	7.64	10.94	**2.84**	**7.39**	54.97	4.60	7.46	**15.33**	3.43	**7.45**

4.4 Ablation: Objective Variants at a Matched Budget

Table 8 (Appendix E) trains five objective variants on top of the identical SVEET recipe, budget and optimizer, and evaluates all of them at their final checkpoint with three inference runs: the scale-matched multi-scale objective with $\mathcal{S} = \{1, 2, 4\}$ or $\{1, 2, 4, 8\}$, each with and without the adaptive gate (SVEET-MS is the gated $\{1, 2, 4, 8\}$ cell), and a Demucs-style waveform-$\mathcal{L}_1$ auxiliary on the clean estimate recovered analytically from the predicted vector field, with its weight calibrated so that it contributes about a quarter of the gradient norm. Three observations follow. First, the gate governs fidelity: without it both scale sets lose clean-input transparency (Acapella (Clean) 53.5 and 51.4 vs. 56.8 dB) and 1.6 and 2.7 dB of overall SI-SDR, whereas with it both recover (34.3 and 35.1 dB); the ungated variants keep the perceptual gain (PESQ 3.89, PQ 7.47–7.48), match SVEET on SingVERSE (PQ 6.46 vs. 6.45), and the ungated $\{1, 2, 4, 8\}$ variant reaches a DNS PESQ of 2.66, above the speech-trained DeepFilterNet3 (2.58), so the perceptual gain is a property of the multi-scale objective itself and the gate is the price of transparency. Second, with the gate the perceptual gain survives only with the coarsest scale: gated $\{1, 2, 4\}$ falls back to SVEET on PESQ (3.83) and loses 0.40 PQ on SingVERSE, whereas gated $\{1, 2, 4, 8\}$ keeps PESQ 3.87 and gives back only 0.05 PQ there; SVEET-MS is thus the one cell that loses neither fidelity nor perception, which is why it is the reported system. Third, the x_0-space reconstruction auxiliary behaves like its discriminative ancestor: strongest in fidelity at its validation-selected checkpoint (Appendix C) but worse perceptually on every set outside the singing benchmark (SingVERSE PQ -0.21, EARS-WHAM PQ -0.53), so the perceptual gain of the multi-scale objective is specific to it, not a generic effect of a noise-free auxiliary target. All variants share the fast early convergence discussed in Appendix D; every cell is a single run (the SVEET-MS cell is the better of two seeds by validation SI-SDR, Section 3.4).

4.5 Generalization to Speech Enhancement

To verify whether our singing-aware training recipes compromise generalization on standard speech data, we evaluate all systems zero-shot on two public speech test sets whose material is fully disjoint from our training corpora. VoiceBank-DEMAND (Valentini-Botinhao, 2017) is excluded because its clean test speakers (VCTK p232, p257) are in our training pool; we adopt the *DNS 2020* synthetic no-reverb test set (Reddy et al., 2020) (150 utterances, 16 kHz, Graz clean speech) and the *EARS-WHAM v2* test partition (Richter et al., 2024; Wichern et al., 2019) (886 segments, 48 kHz full band, anechoic speech with real WHAM! noise), built with the official script and released with our manifests.

As Table 7 (Appendix E) shows, the conditional flow-matching systems transfer to speech markedly better than the music-separation baselines and our discriminative system on both test sets and across all four metrics (paired Wilcoxon tests, $p < 10^{-20}$ throughout). Against DeepFilterNet3 (Schröter et al., 2023), a full-band model trained on speech, SVEET is ahead by $+2.0$/$+1.9$ dB SI-SDR and $+0.85$/$+0.57$ PQ on DNS/EARS-WHAM (CIs exclude zero) while PESQ is a tie (-0.04/0.00, CIs

include zero). On DNS, SVEET attains 18.32 dB SI-SDR and 2.54 PESQ-WB, $+4.5$ dB and $+0.57$ PESQ over the strongest music-separation or discriminative baseline; on the harder full-band EARS-WHAM benchmark the advantage over those baselines persists ($+1.3$ dB SI-SDR and $+0.36$ PESQ for SVEET-MS). No-reference production quality reveals a qualitative split: the music-separation baselines and DISCSVE *degrade* PQ relative to the unprocessed input (5.08–5.57 vs. 5.89 on DNS) and DeepFilterNet3 improves it moderately (6.29), whereas the generative systems improve it by up to $+1.25$ (7.14). The multi-scale objective, designed for degraded singing, behaves neutrally on speech: relative to SVEET it is slightly lower on DNS (SI-SDR and PESQ within their intervals, ESTOI -0.005 and PQ -0.03 at the edge of significance), significantly higher on EARS-WHAM in SI-SDR ($+0.36$ dB), PESQ ($+0.04$) and ESTOI, and lower in PQ there (-0.09; paired bootstrap over utterances). A single checkpoint trained for singing thus transfers to leakage-free speech without speech-specific tuning; we do not claim speech-enhancement state of the art.

4.6 REAL-WORLD GENERALIZATION: THE SINGVERSE BENCHMARK

Our singing benchmark is constructed from synthetic degradations; to probe fully in-the-wild behaviour we additionally evaluate zero-shot on *SingVERSE* (Jiang et al., 2025), a public benchmark of 3,929 real-world noisy singing recordings with time-aligned studio references spanning 19 acoustic scenarios at 44.1 kHz. Because the noisy clips are genuine re-recordings, waveform-level reference metrics are uninformative (clock drift leaves even the unprocessed input at strongly negative SI-SDR); following the benchmark's own protocol we therefore report non-intrusive perceptual metrics (DNSMOS, PQ) and a WavLM-based speaker-similarity score (SIM) against the clean reference.

Three observations emerge from Table 9 (Appendix E). First, the flow-matching family is the only one whose PQ exceeds that of the unprocessed input, and it leads the strongest public baseline by $+0.36$ DNSMOS-OVRL (VoiceFixer) and $+0.37$ PQ (HDemucs); to our knowledge our systems are the strongest with publicly released weights on this benchmark (VoiceFixer, the best such system in the benchmark's own evaluation, scores 2.23 and 5.72 under our protocol), and, trained on 126.4 hours of public singing data, they land within 0.13 DNSMOS-OVRL of the closed-source AnyEnhance. The two flow-matching variants lie within 0.05 PQ and 0.03 DNSMOS-OVRL of each other (the base objective marginally but significantly ahead on PQ under a paired bootstrap over the 3,929 clips, the multi-scale objective marginally ahead on DNSMOS), so the perceptual gain of the multi-scale objective on our synthetic noisy conditions does not carry over to these recordings. Second, the benchmark's quality–content trade-off reproduces: the separation baselines and VoiceFixer preserve speaker identity best (SIM 0.83), while the generative enhancers trade some similarity for perceptual quality. Third, the gains concentrate in genuinely noisy scenes, while scenes with intentional ambience (concerts, KTV) see small negative deltas (per-scenario numbers in Appendix E).

4.7 LISTENING TEST

To complement the objective metrics with human judgments we ran a blind best–worst-scaling test. Each trial presents the labelled clean reference and the outputs of the four systems (HDemucs, DISCSVE, SVEET, SVEET-MS) in anonymized positions, and asks for the best and the worst output, with a "no audible difference" option intended for the clean conditions. The fixed 30-trial set comprises 13 singing clips from the benchmark (3/4/3/3 across the four conditions), 13 speech clips from EARS-WHAM (7) and DNS (6), and four attention probes in which one system is replaced by the clean reference (should win) or the unprocessed input (should lose); listeners with three or more probe errors are excluded, a pre-declared criterion. Thirty listeners completed the test with headphones; none was excluded (104 of 120 probe responses correct), yielding 780 scored trials. "No audible difference" trials carry no votes and are excluded from the denominator; on the decided trials we compute, per listener, the share chosen best minus the share chosen worst, and report the mean over listeners as the best-minus-worst (BWS) score with listener-bootstrap 95% CIs and three pre-declared, Holm-corrected contrasts.

Table 2 shows a clear ordering: both flow-matching systems are preferred over the public baseline and our discriminative system by a wide margin, and SVEET-MS is preferred over SVEET in every domain (76% of decided noisy-singing trials, 71% of speech trials; Table 6). The size of this preference contrasts with the objective picture of Sections 4.3–4.5, where the two systems differ by a few hundredths in PESQ and PQ on the singing benchmark and by at most 0.4 dB, 0.05 PESQ

Table 2: Blind best–worst-scaling listening test (30 listeners, 780 scored trials). BWS is the best-minus-worst share per system with listener-bootstrap 95% CIs; contrasts are Holm-corrected. Per-domain head-to-head counts are given in Table 6 (Appendix E).

System	BWS	95% CI
SVEET-MS	$\mathbf{+0.653}$	$[+0.555, +0.754]$
SVEET	$+0.250$	$[+0.169, +0.331]$
DISCSVE	-0.437	$[-0.477, -0.394]$
HDemucs	-0.467	$[-0.503, -0.431]$
Contrast	**ΔBWS**	**95% CI**
SVEET-MS vs. HDemucs (public baseline)	$+1.121$	$[+1.020, +1.220], p < 10^{-3}$
SVEET-MS vs. DISCSVE (discriminative)	$+1.091$	$[+0.955, +1.223], p < 10^{-3}$
SVEET-MS vs. SVEET (multi-scale objective)	$+0.404$	$[+0.227, +0.583], p < 10^{-3}$

and 0.1 PQ on speech. An acoustic comparison of the excerpts points to the residual noise floor: on the speech clips the 10th-percentile frame energy (20 ms frames) of SVEET-MS lies 3.5 dB below that of SVEET (3.2 vs. 6.7 dB above the reference), a difference that intrusive and aesthetic metrics barely register but listeners do. On the clean conditions listeners reported no audible difference in 70% (Acapella) and 82% (Singing w/ Acc) of trials, and where they decided, SVEET-MS drew the fewest "worst" votes (12%).

5 Discussion and Limitations

The multi-scale objective re-prioritizes errors rather than changing the target (fixed weights: unchanged minimizer; gated: target-dependent re-weighting), and its footprint matches that: modest objective gains but a much larger listening preference on degraded singing, neutrality on speech, and slight conservatism on real recordings, where all generative systems trade some speaker similarity for perceptual quality (Table 9).

Limitations and Future Scope. First, the sweeps use a single seed and no speech-only or singing-only control, and the flow-matching comparison reports one run per system (SVEET-MS being the better of two seeds on validation SI-SDR); benchmark SI-SDR is sensitive to a few clean-input clips whose pass-through score varies by several dB across checkpoints (Appendix C), so small SI-SDR differences between systems should not be over-read, whereas PESQ and PQ are far less affected. Second, the data recipe was selected on an earlier revision of the benchmark rather than on a separate validation set; the flow-matching comparison, ablations and listening test involve no such selection. Third, DISCSVE is a matched-data, not a matched-compute, baseline, and every ablation cell is a single run. Further limitations (model capacity, fixed STFT resolution, distribution-level metrics) are listed in Appendix H.

6 Conclusion

Studying SVE from a public-data design perspective, we sweep the composition and volume of public speech-and-singing corpora and arrive at a reproducible recipe that lifts a discriminative backbone well above zero-shot baselines and is insensitive to the speech-to-singing ratio over the tested range. A conditional flow-matching enhancer trained on the same data matches that backbone in fidelity and clearly exceeds it perceptually; under a matched budget and optimizer, a degradation-adaptive multi-scale objective adds perceptual gains on degraded singing at no cost in overall fidelity and is preferred by listeners over the base objective in nearly three quarters of decided trials, a preference the objective metrics only hint at. Loss-reduction mismatches can suppress auxiliary contributions or, when they rescale the entire objective, alter Adam updates; a single checkpoint covers speech and singing. Neither the recipe nor the gated multi-scale objective is specific to singing, and extending both to other low-resource restoration domains, adding reverberation to the training recipes, and ablating the degradation recipes on the discriminative backbone are natural next steps.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, for drafting and refactoring experiment-orchestration and evaluation scripts, and for checking numerical consistency between result files and the manuscript. We have not used generative AI tools for generating research ideas, for producing or altering experimental results, or for writing the core model code, and other disclosure categories are not applicable to this work. We have reviewed all AI-assisted work; AI-drafted scripts were verified against independent recomputations of the reported numbers by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The listening test was conducted with adult volunteers who took part anonymously under a self-chosen identifier, could stop at any time, and were informed that only their responses and browser type would be stored; no other personal data were collected; individual responses are not distributed and only aggregate results are released. All training and evaluation corpora are public research datasets used under their respective licences (Appendix F); we release manifests and synthetic degradations rather than redistributing source audio whose licence does not permit it.

REPRODUCIBILITY STATEMENT

Section 3.4 and Appendix A specify the data pools, recipes and rendering parameters; the released manifests record recipe, source clip, offset and gains for every training pair so the corpus can be regenerated. Sections 3.2–3.3 give all model and objective hyperparameters, and Section 3.4 the training schedules; Appendix C reports checkpoint-selection and endpoint variability. The evaluation protocol, benchmark composition and statistical procedures are described in Section 4.1 and the listening-test design in Section 4.7; the benchmark audio for the public-source items, evaluation scripts, listening-test application and aggregate listening results, and checkpoints are provided through the anonymous links in the footnote to Section 1.

REFERENCES

Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Dmitry Vetrov. HiFi++: A unified framework for bandwidth extension and speech enhancement. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10097255.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-Fi multi-speaker English TTS dataset. In *Proc. Interspeech*, pp. 2776–2780, 2021. doi: 10.21437/Interspeech.2021-1599.

Sebastian Braun, Hannes Gamper, Chandan K. A. Reddy, and Ivan Tashev. Towards efficient models for real-time deep noise suppression. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 656–660, 2021. doi: 10.1109/ICASSP39728.2021.9413580.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. An investigation of incorporating Mamba for speech enhancement. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 302–308, 2024. doi: 10.1109/SLT61566.2024.10832332.

Shuqi Dai, Yuxuan Wu, Siqi Chen, Roy Huang, and Roger B. Dannenberg. SingStyle111: A multilingual singing dataset with style transfer. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 765–773, Milan, Italy, 2023.

Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proc. ISMIR Workshop on Music Source Separation*, 2021.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain, 2019. URL `https://arxiv.org/abs/1911.13254`.

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 *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–9, 2013. doi: 10.1109/APSIPA.2013.6694316.

Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. MetricGAN+: An improved version of MetricGAN for speech enhancement. In *Proc. Interspeech*, pp. 201–205, 2021. doi: 10.21437/Interspeech.2021-599.

Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, 2010. doi: 10.1109/TASL.2009.2026503.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-Singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proc. ACM International Conference on Multimedia (ACM MM)*, pp. 3945–3954, 2021. doi: 10.1145/3474085.3475437.

Shaohan Jiang, Junan Zhang, Yunjia Zhang, Jing Yang, Fan Fan, and Zhizheng Wu. SingVERSE: A diverse, real-world benchmark for singing voice enhancement, 2025. URL https://arxiv.org/abs/2509.20969.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Yu Zhang, Wei Han, Ankur Bapna, and Michiel Bacchiani. Miipher: A robust speech restoration model integrating self-supervised speech and text representations. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2023. doi: 10.1109/WASPAA58266.2023.10248089.

Seonggyu Lee, Sein Cheong, Sangwook Han, and Jong Won Shin. FlowSE: Flow matching-based speech enhancement. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888274.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2724–2737, 2023. doi: 10.1109/TASLP.2023.3294692.

Chenda Li, Wangyou Zhang, Wei Wang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Yihui Fu, Marvin Sach, Zhaoheng Ni, Anurag Kumar, Tim Fingscheidt, Shinji Watanabe, and Yanmin Qian. Less is more: Data curation matters in scaling speech enhancement. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11434638.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: A unified framework for high-fidelity speech restoration. In *Proc. Interspeech*, pp. 4232–4236, 2022. doi: 10.21437/Interspeech.2022-11026.

Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra. In *Proc. Interspeech*, pp. 3834–3838, 2023. doi: 10.21437/Interspeech.2023-1441.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. URL https://arxiv.org/abs/2106.04624.

Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The INTERSPEECH 2020 deep noise suppression

challenge: Datasets, subjective testing framework, and challenge results. In *Proc. Interspeech*, pp. 2492–2496, 2020. doi: 10.21437/Interspeech.2020-3038.

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.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Proc. Interspeech*, pp. 4873–4877, 2024. doi: 10.21437/Interspeech.2024-153.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096956.

Robin Scheibler, Yusuke Fujita, Yuma Shirahata, and Tatsuya Komatsu. Universal score-based speech enhancement with high content preservation. In *Proc. Interspeech*, pp. 1165–1169, 2024. doi: 10.21437/Interspeech.2024-138.

Hendrik Schröter, Alberto N. Escalante-B., Tobias Rosenkranz, and Andreas Maier. DeepFilterNet: Perceptually motivated real-time speech enhancement. In *Proc. Interspeech*, pp. 2008–2009, 2023.

Joan Serrà, Santiago Pascual, Jordi Pons, R. Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion, 2022. URL https://arxiv.org/abs/2206.03065.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus, 2015. URL https://arxiv.org/abs/1510.08484.

Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 293–305, 2018. doi: 10.1007/978-3-319-93764-9_28.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic assessment for speech, music and sound. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11434623.

Cassia Valentini-Botinhao. Noisy speech database for training speech enhancement algorithms and TTS models. University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017. URL https://doi.org/10.7488/ds/2117.

Vandalism Sounds. Vandalism sounds vocal sample packs. https://vandalism-sounds.com/, 2026. Commercial audio sample library, accessed 2026.

Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source Chinese popular song corpus for singing voice synthesis. In *Proc. Interspeech*, pp. 4242–4246, 2022. doi: 10.21437/Interspeech.2022-48.

Ziqian Wang, Zikai Liu, Xinfa Zhu, Yike Zhu, Mingshuai Liu, Jun Chen, Longshuai Xiao, Chao Weng, and Lei Xie. FlowSE: Efficient and high-quality speech enhancement via flow matching. In *Proc. Interspeech*, pp. 4858–4862, 2025. doi: 10.21437/Interspeech.2025-1745.

Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. In *Proc. Interspeech*, pp. 2928–2932, 2022. doi: 10.21437/Interspeech.2022-10653.

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. In *Proc. Interspeech*, pp. 1368–1372, 2019. doi: 10.21437/Interspeech.2019-2821.

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. VocalSet: A singing voice dataset. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 468–474, Paris, France, 2018.

Weiming Xu, Zhouxuan Chen, Zhili Tan, Shubo Lv, Runduo Han, Wenjiang Zhou, Weifeng Zhao, and Lei Xie. MBTFNet: Multi-band temporal-frequency neural network for singing voice enhancement. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023. doi: 10.1109/ASRU57964.2023.10389670.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

Junan Zhang, Jing Yang, Zihao Fang, Yuancheng Wang, Zehua Zhang, Zhuo Wang, Fan Fan, and Zhizheng Wu. AnyEnhance: A unified generative model with prompt-guidance and self-critic for voice enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 3085–3098, 2025. doi: 10.1109/TASLPRO.2025.3587393.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, and Zhou Zhao. M4Singer: A multi-style, multi-singer and musical score provided Mandarin singing corpus. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6914–6926. Curran Associates, Inc., 2022. doi: 10.52202/068431-0501.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 1117–1140. Curran Associates, Inc., 2024. doi: 10.52202/079017-0034.

A Data-Construction Recipes

Training mixtures are synthesized from clean segments according to

$$y = x + \alpha m + \beta n, \tag{6}$$

where x is the clean vocal or speech segment, m denotes instrumental interference sampled from MUSDB18, n denotes additive noise sampled from MUSAN, and α, β are per-example gain factors.

Every clean segment is expanded by all three fixed recipes (implemented as `00_clean`, `01_A_noise_basic`, and `06_B_music_only` in the released codebase); the resulting pool is then sampled to meet the target speech-to-singing ratio and hour budget of each configuration, so the per-recipe counts in a given manifest need not be exactly equal (for the 1:1, 1000 h manifest they are 175,700 / 258,421 / 285,879 clean / noise / music pairs). In all cases the decomposition vocals + interference = mixture is preserved exactly, so training targets remain sample-aligned with their inputs.

Clean pass-through (α=β=0). The mixture equals the clean target. Including these identity pairs trains the model to leave already-clean inputs unaltered, directly supporting the clean-preservation requirement of Section 4.1.

Additive noise (α=0, β>0). One or more independent MUSAN noise layers are summed onto the segment, with per-layer gains (in dB relative to the clean signal) drawn uniformly from domain-specific ranges: speech segments receive a single layer at $[-30, -14]$ dB, while singing segments from the expanded multi-lingual partition receive one to four layers at $[-40, -2]$ dB, exposing the model to substantially heavier contamination on the singing side. Singing segments originating from the four standalone corpora predate this heavier profile and are degraded with the same single-layer $[-30, -14]$ dB setting as speech; the released manifests record which profile each pair received.

Instrumental interference ($\alpha>0$, $\beta=0$)**.** A MUSDB18 accompaniment excerpt is drawn from a random offset of a random track and mixed under the segment. Each pair is assigned to one of three loudness tiers (sparse, medium, or dense) with probabilities of approximately 25%, 35%, and 40%, and the accompaniment gain (in dB relative to the clean signal) is then drawn uniformly from the tier-specific range in Table 3. For singing, the upper bound of $+2$ dB in the dense tier deliberately allows the accompaniment to exceed the vocal level, mirroring the hardest evaluation condition.

Table 3: Accompaniment gain ranges (dB, relative to the clean signal) for the instrumental-interference recipe. Percentages give the share of pairs assigned to each loudness tier. The heavier singing profile applies to the expanded multi-lingual singing partition; pairs from the four standalone corpora use the speech ranges (see text).

Domain	Sparse ($\sim$25%)	Medium ($\sim$35%)	Dense ($\sim$40%)
Speech	$[-24, -16]$	$[-16, -10]$	$[-10, -4]$
Singing	$[-34, -16]$	$[-24, -8]$	$[-18, +2]$

B ODE Step-Count Sweep

Table 4 sweeps the number of Euler steps N for both flow-matching systems on the singing benchmark (78 clips; the $N=5$ column is the three-run average of Table 1). Signal fidelity peaks at $N=3$ and *decreases* with longer integration, while perceptual quality (PESQ, ESTOI) rises and plateaus from $N=5$ on, a compact illustration of the generative–faithful trade-off along the transport trajectory: more integration steps allow the model to generate more and adhere to the input less. $N = 5$, fixed a priori, sits at the PESQ plateau at less than a third of the function evaluations of $N=16$; a single Euler step (real-time factor ≈ 0.09 on an A100, measured on 16 kHz speech) already exceeds every public zero-shot baseline of Table 1 in PESQ (DiscSVE, at 3.74, is not exceeded) and lies within 1.7 dB of the strongest public baseline in SI-SDR. The multi-scale system matches or exceeds SVEET in PESQ at every step count.

Table 4: Effect of the Euler step count N on the singing benchmark (78 clips), reported as SI-SDR (dB) / PESQ-WB / ESTOI with the subgroup-mean convention of Table 1. $N=5$ averages three inference runs; the other columns are single runs. **Bold** marks the best value per row.

System	Metric	$N=1$	2	3	5	8	16	32
SVEET	SI-SDR	28.18	33.95	**35.02**	34.71	34.60	34.66	34.63
	PESQ	3.72	3.77	3.83	**3.84**	3.83	3.83	3.83
	ESTOI	0.879	0.888	0.892	**0.895**	**0.895**	0.894	0.893
SVEET-MS	SI-SDR	28.86	33.53	**35.22**	35.10	34.61	34.34	34.40
	PESQ	3.73	3.78	3.83	**3.87**	**3.87**	3.84	3.83
	ESTOI	0.884	0.891	0.895	0.899	**0.901**	0.898	0.898

C Checkpoint Selection and Endpoint Variability

All main-text numbers use the final checkpoint of each run. For robustness we also report, for every flow-matching variant, the checkpoint with the highest SI-SDR on the 100-pair validation subset (Table 5; validation never involves the benchmark). The two conventions agree for SVEET-MS within noise, but they can differ by more than 3 dB for a single run (36.5 vs. 33.2 dB for the auxiliary-loss run): the benchmark mean is dominated by a handful of clean-input clips whose pass-through SI-SDR (50–60 dB) fluctuates from epoch to epoch, while the per-clip median difference between neighbouring checkpoints stays within a few tenths of a dB. Perceptual metrics are far less sensitive to this (PESQ within 0.05, PQ within 0.04 across the two conventions), which is why we treat the fixed final checkpoint as the primary protocol and report the selected checkpoints only here.

Table 5: Final checkpoint vs. validation-selected checkpoint for each flow-matching variant (singing benchmark, three inference runs, *Overall* as in Table 1).

	Final checkpoint (epoch 46)			Validation-selected		
Objective	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ
CFM (SVEET)	34.71	3.84	7.45	33.69 (ep. 41)	3.89	7.48
+ multi-scale $\{1, 2, 4\}$, no gate	33.14	3.89	7.47	33.73 (ep. 41)	3.88	7.46
+ multi-scale $\{1, 2, 4, 8\}$, no gate	32.04	3.89	7.48	33.07 (ep. 40)	3.88	7.47
+ adaptive multi-scale $\{1, 2, 4\}$	34.31	3.83	7.47	34.07 (ep. 31)	3.81	7.44
+ adaptive multi-scale $\{1, 2, 4, 8\}$ (SVEET-MS)	35.10	3.87	7.48	34.82 (ep. 44)	3.85	7.48
+ x_0 waveform-$\mathcal{L}_1$ auxiliary	33.22	3.86	7.45	36.48 (ep. 45)	3.88	7.46

D Loss Scale and Adam: Measurement

Two implementations. The base CFM loss is implemented, following Lee et al. (2025), as $\frac{1}{2}$ times the sum of the squared residual over the $2 \times F \times T$ complex entries of each example, averaged over the batch. In our first implementation the multi-scale objective *replaced* this loss, and every scale term, including $s{=}1$, was reduced by a per-element mean over batch, channel, frequency and time; the whole training objective, and hence its gradient, was therefore $F \times T \approx 6.5 \times 10^4$ times smaller than the base loss. The scale-matched implementation of Equation 5 instead reduces each scale by a per-example sum (with the factor $s/2$) and averages over the batch, so the $s{=}1$ term coincides with the base loss and the coarse terms enter at a comparable scale. The same mismatch has the opposite effect when a mean-reduced term is *added* to a sum-reduced base loss: a waveform-$\mathcal{L}_1$ auxiliary with weight 0.1 reduced this way has a gradient norm below 10^{-5} of the base gradient and leaves training essentially unchanged, which is why the auxiliary of Section 4.4 is sum-reduced and weighted by gradient norm. Adam is scale-invariant only while its second-moment estimate dominates ϵ in the denominator $\sqrt{\hat{v}_t} + \epsilon$; the following measurements test whether that held.

Optimizer state. Adam's update is $\Delta\theta = -\eta\, \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$, so the loss scale matters exactly when $\sqrt{\hat{v}_t}$ is not large compared with ϵ. We read the bias-corrected second-moment estimate $\hat{v}_t$ from the saved optimizer state of four complete 47-epoch runs ($528{,}750$ updates, $\epsilon = 10^{-8}$, no gradient clipping or weight decay, full precision). For the two runs trained with the mean-reduced multi-scale objective (the earlier $\{1, 2, 4\}$ and $\{1, 2, 4, 8\}$ models), the median of $\sqrt{\hat{v}_t}$ over all 65.7M parameters is 5.6×10^{-10} and 1.2×10^{-9}, 74% and 76% of the parameters have $\sqrt{\hat{v}_t} < \epsilon$, and the denominator fraction $\sqrt{\hat{v}_t}/(\sqrt{\hat{v}_t} + \epsilon)$, the ratio of the actual to the nominal Adam step, has mean 0.25–0.26 and median 0.05–0.11. For the two runs trained with sum-reduced objectives (SVEET and SVEET-MS), the median of $\sqrt{\hat{v}_t}$ is 2.4–3.5×10^{-5}, 0.1–0.2% of the parameters lie below ϵ, and the denominator fraction has mean 0.98–0.99 and median 1.00.

Gradient probe. On a fixed batch of eight 256-frame examples and a converged checkpoint, the gradient of the mean-reduced multi-scale objective has cosine similarity 0.97 with the gradient of the base CFM loss but is smaller in norm by a factor of 1.1×10^{-5}, close to the loss-scale factor $1/(F \cdot T) = 1.5 \times 10^{-5}$; reducing the multi-scale terms by per-example summation restores the scale while leaving the direction unchanged.

Convergence on the validation set. Figure 1 tracks both flow-matching systems on a fixed 100-pair validation subset (never used for any decision on the benchmark) after every epoch. The multi-scale objective yields a usable enhancer after a single epoch (23.6 dB SI-SDR vs. -2.9 dB for the base objective at the same point); the base objective catches up by epoch 2, the two alternate until epoch 19, and from epoch 20 onward the multi-scale system is ahead at every epoch, by 0.7 dB on average over epochs 20–46. In PESQ the multi-scale system lags by up to 0.32 during epochs 3–9, stays within 0.12 until epoch 17, and the two remain within 0.03 of each other thereafter. We attribute the early-convergence effect to the pooled terms averaging out the irreducible white-noise component $\sigma_{\max} z$ of the regression target, which lowers the variance of the gradient signal in the coarse scales; the same fast start appears when a noise-free target is supplied by other means: the x_0-space auxiliary of Section 4.4 reaches 18.5 dB on the benchmark after its first epoch (vs. -4.1 dB for the base objective at the same point), even though its final perceptual behaviour is very different.

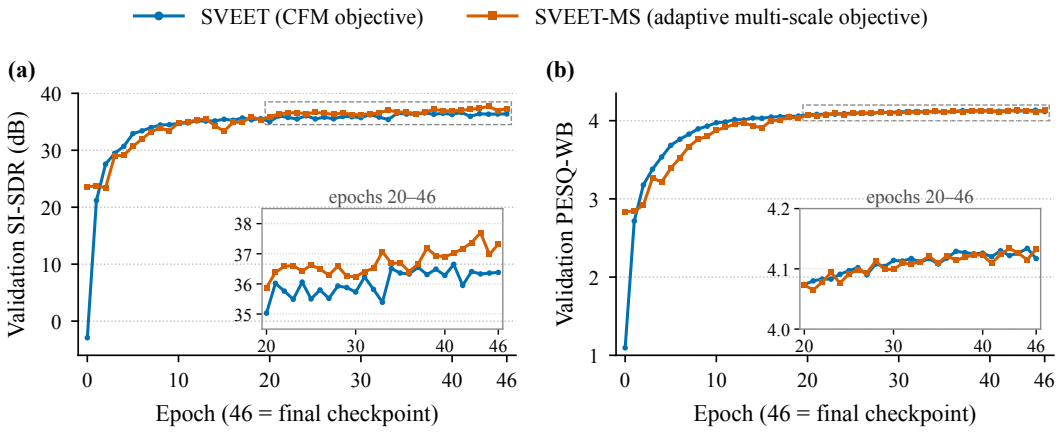

Figure 1: Validation-set trajectories of SVEET and SVEET-MS (100 held-out pairs, $N{=}5$ Euler steps, one inference run per epoch): (a) SI-SDR, (b) PESQ-WB. Epoch index 0 is the checkpoint after the first epoch and 46 the final checkpoint reported throughout; the insets enlarge epochs 20–46 (dashed boxes). Both systems share data, optimizer and schedule.

E Ablation and SingVERSE Tables

Tables 8, 7, 9 and 6 support Sections 4.4, 4.5, 4.6 and 4.7. On SingVERSE, per-scenario perceptual gains concentrate in genuinely noisy scenes (for SVEET-MS, $+2.0$ PQ over the input in arcades, $+0.9$ in restaurants and $+0.7$ on roadsides), whereas scenes whose ambience is intentional (concerts -0.5, KTV -0.3) see negative deltas: the models remove reverberation and crowd atmosphere that listeners may value, a real-world mirror of the clean-preservation requirement.

Table 6: Listening test: decided trials in which SVEET-MS ranked above vs. below SVEET, per domain. The counts are descriptive; inference rests on the listener-bootstrap contrasts of Table 2.

Domain	above / below	share
Singing, noisy conditions	124 / 40	76%
Singing, clean conditions	31 / 11	74% (76% “no difference”)
Speech (DNS, EARS-WHAM)	250 / 102	71%

Table 7: Zero-shot speech enhancement on two leakage-free public test sets: DNS 2020 synthetic no-reverb (150 utterances, 16 kHz) and EARS-WHAM v2 test (886 segments, 48 kHz). PESQ-WB is computed after polyphase resampling to 16 kHz; PQ is the Production Quality axis of Audiobox-Aesthetics. **Bold** marks the best system per column.

	DNS 2020 (no reverb)				EARS-WHAM v2 (test)			
System	SI-SDR	PESQ	ESTOI	PQ	SI-SDR	PESQ	ESTOI	PQ
Unprocessed (noisy)	9.23	1.58	0.810	5.89	5.35	1.23	0.640	5.70
HTDemucs (zero-shot)	13.47	1.96	0.857	5.56	8.93	1.46	0.704	5.73
HDemucs (`hdemucs_mmi`, zero-shot)	13.27	1.97	0.864	5.57	10.76	1.59	0.748	5.58
DiscSVE	13.84	1.54	0.804	5.08	11.27	1.46	0.742	5.45
DeepFilterNet3 (zero-shot, speech-trained) (Schröter et al., 2023)	16.30	**2.58**	0.902	6.29	10.30	1.91	0.726	5.77
SVEET	**18.32**	2.54	**0.919**	**7.14**	12.17	1.91	0.791	**6.34**
SVEET-MS	18.26	2.50	0.914	7.11	**12.53**	**1.95**	**0.797**	6.25

F Data Licences and Benchmark Sources

All training corpora are public; several are restricted to non-commercial research use, which is why we describe our budget as one of *public research corpora* rather than of permissively licensed data. Table 10 lists the licence of each corpus as stated by its distributor at the time of writing. We redistribute no source audio: the release contains manifests (source file, offset, recipe and

Table 8: Objective variants trained on the identical recipe, optimizer and 47-epoch budget, evaluated at the final checkpoint (three inference runs); the four multi-scale rows form a 2×2 over scale set and gate. Singing-benchmark columns follow the convention of Table 1; PQ is the Audiobox-Aesthetics production-quality axis. **Bold** marks the best value per column.

	Singing benchmark			SingVERSE	DNS 2020		EARS-WHAM	
Objective	SI-SDR	PESQ	PQ	PQ	PESQ	PQ	PESQ	PQ
CFM (SVEET)	34.71	3.84	7.45	6.45	2.54	**7.14**	1.91	6.34
+ multi-scale $\{1, 2, 4\}$, no gate	33.14	**3.89**	7.47	**6.46**	2.58	7.06	**1.99**	**6.36**
+ multi-scale $\{1, 2, 4, 8\}$, no gate	32.04	**3.89**	**7.48**	**6.46**	**2.66**	7.09	1.96	6.34
+ adaptive multi-scale $\{1, 2, 4\}$	34.31	3.83	7.47	6.05	**2.66**	6.91	1.90	6.24
+ adaptive multi-scale $\{1, 2, 4, 8\}$ (SVEET-MS)	**35.10**	3.87	**7.48**	6.40	2.50	7.11	1.95	6.25
+ x_0 waveform-$\mathcal{L}_1$ auxiliary	33.22	3.86	7.45	6.24	2.19	6.75	1.73	5.81

Table 9: Zero-shot evaluation on the real-world SingVERSE benchmark (3,929 clips, 19 scenarios). DNSMOS-OVRL and PQ are non-intrusive; SIM is the cosine similarity of WavLM speaker embeddings against the studio reference. VoiceFixer (Liu et al., 2022), the strongest system with publicly released weights evaluated by Jiang et al. (2025), is run under our protocol. For context we quote the DNSMOS-OVRL and SIM reported by Jiang et al. (2025) for AnyEnhance (Zhang et al., 2025), an industrial system trained on 21,000 hours of proprietary audio whose weights are not public (protocol re-implementations may differ). **Bold** marks the best among our evaluated systems.

System	DNSMOS-OVRL	PQ	SIM
Unprocessed (noisy)	1.28	6.09	0.750
HTDemucs (zero-shot)	1.82	6.04	0.833
HDemucs (zero-shot)	1.93	6.08	**0.834**
DISCSVE	1.92	5.65	0.807
VoiceFixer (zero-shot) (Liu et al., 2022)	2.23	5.72	0.832
SVEET	2.56	**6.45**	0.747
SVEET-MS	**2.59**	6.40	0.758
AnyEnhance (as reported)	*2.72*	—	*0.815*

gains for every pair), the rendering code, our synthetic degradation parameters and the evaluation partitions, so that anyone who obtains the corpora under their own licences can reconstruct the training set exactly. The singing benchmark draws its vocal material from MIR-1K, OpenSinger and a commercial vocal-sample library (Vandalism Sounds; Vandalism Sounds, 2026); the library's audio cannot be redistributed, so these items appear in the release only by benchmark identifier (`data/benchmark_items.csv` in the code repository), while degraded inputs, references and system outputs are released for the MIR-1K and OpenSinger items, and we additionally evaluate all systems on the fully public SingVERSE benchmark.

Table 10: Licences of the corpora used in this work (as stated by their distributors).

Corpus	Role	Licence	Redistribution of audio
GTSinger	singing (train)	CC BY-NC-SA 4.0	non-commercial, share-alike
M4Singer	singing (train)	CC BY-NC-SA 4.0	non-commercial, share-alike
SingStyle111	singing (train)	CC BY 4.0	permitted with attribution
VocalSet	singing (train)	CC BY 4.0	permitted with attribution
OpenCPOP	singing (train)	CC BY-NC-ND 4.0	non-commercial, no derivatives
Hi-Fi TTS	speech (train)	CC BY 4.0	permitted with attribution
VCTK	speech (train)	CC BY 4.0	permitted with attribution
MUSAN	noise (train)	CC BY 4.0	permitted with attribution
MUSDB18	accompaniment (train)	research use, per-track	not permitted
DNS 2020 / EARS / WHAM!	speech evaluation	per source / CC BY-NC 4.0 / CC BY-NC 4.0	per distributor
SingVERSE	singing evaluation	CC BY 4.0	permitted with attribution
MIR-1K	singing evaluation	research use	released per source terms
OpenSinger	singing evaluation	CC BY-NC-SA 4.0	released per source terms
Vandalism Sounds	singing evaluation	commercial	identifier only, no audio

G Baseline Screening and Data-Composition Tables

Table 11 screens public speech-enhancement systems zero-shot on the singing benchmark (Appendix H.2); Tables 12 and 13 give the full data-composition sweeps of Section 4.2. These tables were computed on an earlier, internally consistent revision of the benchmark and evaluation pipeline; only relative comparisons within each table are used.

Model	Overall			Acapella (Clean)			Acapella (Noisy)			Singing w/ Acc (Clean)			Singing w/ Acc (Noisy)		
	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ
DeepFilterNet3 (Schröter et al., 2023)	**10.60**	2.62	**7.21**	**21.90**	3.57	**7.73**	0.00	**1.61**	**6.90**	19.61	3.54	**7.52**	0.90	1.77	6.70
MP-SENet (Lu et al., 2023)	7.08	**2.68**	7.17	20.17	**3.67**	7.59	−11.00	1.42	6.88	**22.32**	**4.01**	7.40	−3.17	1.63	6.81
SpeechBrain MetricGAN+ (Fu et al., 2021; Ravanelli et al., 2021)	−5.95	1.46	6.43	−3.58	1.40	6.81	−13.12	1.12	5.85	−0.57	2.05	6.72	−6.52	1.25	6.32
NSNet2 (Braun et al., 2021)	8.30	2.49	7.15	14.98	3.27	7.60	**1.65**	1.59	6.76	13.28	3.28	7.40	**3.28**	**1.83**	**6.82**

Table 11: Screening of public *speech*-enhancement systems applied zero-shot to the paired singing evaluation set, reported per subgroup as SI-SDR (dB) / PESQ / PQ. *Overall* is the unweighted mean of each metric across the four subgroups. **Bold** marks the best value in each column. Outputs are polyphase-resampled to 44.1 kHz for evaluation (DeepFilterNet3 runs natively at 48 kHz, the others at 16 kHz). These systems are not carried into the main comparison; the numbers were obtained on an earlier revision of the benchmark and are internally consistent, whereas every system in Table 1 is scored under the final unified pipeline.

Model	Speech:Singing	Overall			Acapella (Clean)			Acapella (Noisy)			Singing w/ Acc (Clean)			Singing w/ Acc (Noisy)		
		SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ
Ours, fixed budget†	1:1	33.56	**3.74**	7.30	52.07	**4.60**	7.65	12.91	**2.66**	7.00	51.46	4.60	7.49	**17.78**	**3.12**	7.06
Ours, fixed budget	1:2	**33.89**	3.72	**7.35**	**54.08**	4.59	**7.66**	**13.06**	2.64	**7.12**	51.24	4.58	**7.51**	17.16	3.05	7.11
Ours, fixed budget	2:1	32.51	3.67	7.33	50.64	4.58	7.65	11.38	2.55	7.01	51.82	4.60	7.49	16.20	2.95	7.16
Ours, fixed budget	1:3	33.41	3.68	7.32	52.77	4.59	7.65	12.14	2.64	6.97	**51.95**	4.59	7.49	16.76	2.91	7.18
Ours, fixed budget	3:1	33.31	3.66	7.34	54.01	4.58	7.65	11.81	2.52	7.00	50.86	**4.61**	7.49	16.54	2.94	**7.22**

Table 12: Fixed-budget ratio sweep (1000 h recipe-expanded). *Overall* is the unweighted mean across the four subgroups. **Bold** marks the best value in each column; † marks the selected 1:1 configuration. Differences across ratios are small (overall SI-SDR spread 1.4 dB, PESQ spread 0.08) and single-seed. Sweep tables were computed on an earlier, internally consistent revision of the benchmark; only relative comparisons within these tables are used for recipe selection.

Model	Configuration	Overall			Acapella (Clean)			Acapella (Noisy)			Singing w/ Acc (Clean)			Singing w/ Acc (Noisy)		
		SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ	SI-SDR	PESQ	PQ
Ours, 1:1†	1000 h	33.56	**3.74**	7.30	52.07	4.60	7.65	12.91	2.66	7.00	51.46	4.60	7.49	**17.78**	**3.12**	7.06
Ours, 1:1	1562.6 h max-data	33.39	3.70	7.29	52.88	4.59	7.65	13.00	2.62	7.00	51.03	4.59	7.49	16.64	2.99	7.03
Ours, 1:2	1000 h	33.89	3.72	**7.35**	54.08	4.59	**7.66**	**13.06**	2.64	**7.12**	51.24	4.58	**7.51**	17.16	3.05	7.11
Ours, 1:2	1316 h max-data	34.43	3.72	7.31	54.25	4.60	7.65	12.42	2.64	7.01	54.49	4.62	7.50	16.54	3.03	7.09
Ours, 1:2	max-data + shifts	**35.01**	3.73	7.32	**55.18**	**4.61**	7.65	12.55	**2.67**	7.03	**55.65**	**4.63**	7.49	16.65	3.01	**7.12**

Table 13: Effect of scaling training data from 1000 hours to ∼1500 hours under fixed ratio. *Overall* is the unweighted mean across the four subgroups. **Bold** marks the best value in each column; † marks our primary selected 1:1 configuration.

H Additional Details Moved from the Main Text

This appendix collects details that were condensed in the main text to meet the page limit; it contains no results beyond those in the main text.

H.1 Training corpora

Table 14 lists the unique clean source audio that the main 1:1 recipe draws from each corpus, obtained by deduplicating source files and their durations in the released manifests, i.e. before segmentation and recipe expansion. The full speech pool measures 332.3 h (Hi-Fi TTS 291.5 h, VCTK 40.8 h), of which the 1:1 manifest uses 209.4 h; the singing side uses 126.4 h. GTSinger contributes eight of its nine language partitions.

Table 14: Unique clean source audio used by the main recipe, per corpus (hours, from the released manifests).

Corpus	Domain	Hours
GTSinger (Zhang et al., 2024)	singing	76.7
M4Singer (Zhang et al., 2022)	singing	27.8
SingStyle111 (Dai et al., 2023)	singing	11.1
VocalSet (Wilkins et al., 2018)	singing	6.0
OpenCPOP (Wang et al., 2022)	singing	4.8
Singing total		126.4
Hi-Fi TTS (Bakhturina et al., 2021)	speech	178.6
VCTK (Yamagishi et al., 2019)	speech	30.8
Speech total		209.4

H.2 Baseline screening and data-composition sweeps

We begin by screening open-source enhancement models zero-shot on the singing benchmark, which motivates the choice of baselines carried into the main comparison. A clear split emerges between domains. Music source separation models (HTDemucs, HDemucs) perform comparatively well because their architectures are trained to isolate vocal energy (their numbers under the final unified pipeline appear in Table 1), yet they lack the capacity to remove non-musical degradations and leave substantial residual artifacts. Conversely, leading speech-enhancement architectures (NSNet2, DeepFilterNet3, MP-SENet, MetricGAN+) generalize poorly to singing audio (Table 11, Appendix G), trailing the music-separation systems by a wide margin in overall SI-SDR and, in the noisy subgroups, frequently collapsing to near-zero or negative reconstruction fidelity. Three of these systems operate natively at 16 kHz (DeepFilterNet3 runs at 48 kHz); all outputs were resampled to 44.1 kHz with a high-order polyphase pipeline before scoring. While resampling introduces an unavoidable penalty from the missing content above 8 kHz for the 16 kHz systems, their low SI-SDR across the full band reflects a more fundamental failure to track singing contours. The speech enhancers treat sustained high-register notes as acoustic noise, cutting out whole vocal blocks and inducing severe spectral fluttering. We therefore retain the two music-separation systems, together with the flow-matching speech enhancer FlowSE, as the baselines for the remainder of the paper.

To isolate the impact of data composition under a fixed budget, we freeze the HDemucs backbone and run a controlled sweep across the composition space (tables in Appendix G). We first fix the total volumetric training budget at exactly 1000 hours and adjust the speech-to-singing ratio $\gamma \in \{3{:}1,\ 2{:}1,\ 1{:}1,\ 1{:}2,\ 1{:}3\}$. As shown in Table 12, performance is remarkably flat across ratios: the overall SI-SDR spread is only 1.4 dB and the PESQ spread only 0.08, all single-seed. The 1:2 split attains the nominal best overall SI-SDR (33.89 dB), with the balanced 1:1 configuration close behind at 33.56 dB; this gap stems almost entirely from Acapella (Clean) (+2.01 dB), where absolute SI-SDR values are high and all recipes already preserve the signal well. We instead select by the hardest condition, Singing w/ Acc (Noisy), in which instrumental bleed, reverberation, and background noise co-occur: there the balanced recipe is the strongest of all ratios (17.78 dB SI-SDR and 3.12 PESQ), and we therefore adopt $\gamma = 1{:}1$ (500h speech / 500h recipe-expanded singing); we note that this selection used an earlier revision of the benchmark itself rather than a separate validation set, a limitation we return to in Section 5. To test whether further gains stem from the ratio or from simple data accumulation, we remove the volumetric cap and scale up using all available public speech and singing data under identical augmentation recipes. Both sides are expanded from their respective base pools of unique source audio (a 332.3h speech pool and the curated singing pool of Section 3.4) by segmentation and the three recipes, so the hours quoted for each configuration denote rendered training volume rather than distinct source material. At the 1:1 ratio, expanding from 1000h to 1562.6h (a $\sim$56% increase, $\sim$781h per side) does not help and in fact costs a marginal -0.17 dB SI-SDR (33.56 $\rightarrow$ 33.39; Table 13). Scaling is not uniformly inert, however: at the 1:2 ratio, added volume plus source-position shifts lifts overall SI-SDR from 33.89 dB to 35.01 dB. We read this as evidence that, within the tested range and under our source-data constraint, composition and augmentation choices matter more than raw rendered volume, while volume scaling at a fixed balanced ratio yields diminishing returns; the sweep does not include speech-only or singing-only recipes, and

the fixed 8-epoch schedule means larger manifests also receive proportionally more updates, so we do not claim a general scaling law. We therefore fix the 1:1 balanced, 1000h configuration as our foundational data recipe and carry it forward to the main enhancer comparison.

H.3 Extended discussion

Within the tested range, composition and augmentation choices matter more than rendered volume. On matched data the flow-matching enhancer matches the discriminative backbone in fidelity and exceeds it perceptually; under a matched budget and optimizer the multi-scale objective adds a perceptual gain on degraded singing at no cost in overall fidelity, is neutral on speech and slightly conservative on real recordings, where all generative systems trade some speaker similarity for perceptual quality (Table 9). With fixed weights the multi-scale loss leaves the population minimizer unchanged, and with the gate it is a target-dependent re-weighting; in both cases it re-prioritizes errors rather than changing the target, consistent with modest objective gains and a much larger listening preference.

H.4 Further limitations

Listening-test scoring (full text). Trials answered "no audible difference" carry no votes and are excluded from the scoring denominator; on the remaining (decided) trials we compute, per listener, the share in which a system was chosen best minus the share in which it was chosen worst, and report the mean over listeners as the best-minus-worst (BWS) score, with 95% confidence intervals from a bootstrap over listeners and three pre-declared contrasts corrected by Holm's procedure. The per-domain counts of Table 6 are descriptive; because listeners judge repeated clips, the listener-bootstrap intervals on the contrasts are the inferential statement.

Seed and endpoint variability (full text). Our data sweeps use a single initialization seed and no speech-only or singing-only control, and the flow-matching comparison reports one run per system (SVEET-MS being the better of two seeds on validation SI-SDR); benchmark SI-SDR is sensitive to a handful of clean-input clips whose pass-through score varies by several dB from checkpoint to checkpoint (Appendix C), so SI-SDR differences of a few tenths of a dB between systems should not be over-read, whereas PESQ and PQ are far less affected.

Additional limitations. A larger-capacity variant of the multi-scale system that we trained used the mean-reduced objective discussed in Section 3.3 and is therefore not comparable; whether the gains persist at larger capacity is left open. Our multi-scale loss relies on fixed STFT window choices; exploring dynamic, learnable multi-resolution filterbanks represents a promising front for singing resonance modeling. Finally, distribution-level metrics such as Fréchet Audio Distance would complement the per-clip and subjective evaluations reported here.