LEVERAGING CARNATIC LIVE RECORDINGS FOR SINGING VOICE SEPARATION USING REGRESSION-GUIDED LATENT DIFFUSION

Genís Plaja-Roglans **Xavier Serra** **Martín Rocamora**
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{genis.plaja, xavier.serra, martin.rocamora}@upf.edu

ABSTRACT

Diffusion models have demonstrated potential to separate individual sources from music mixtures in a generative fashion, enabling a new solution for this challenging problem. However, existing works require clean multi-stem data, which is scarce for several repertoires, consequently compromising generalization. We explore the potential of generative modeling to perform weakly-supervised singing voice separation for Carnatic Music, a music repertoire for which large quantities of multi-stem recordings with bleeding between sources have been collected from live performances. We pre-train a latent diffusion model to perform preliminary vocal separation conditioning on the corresponding mixture. Then, using a regressive model which is separately trained on a clean, smaller, and out-of-domain dataset, we estimate the level of bleeding in the preliminary separations and use that information to guide the diffusion model toward generating cleaner samples. The objective and perceptual evaluations show the potential of the proposed generative system for Carnatic vocal separation. Code, weights, and further materials are available online. [1]

1. INTRODUCTION

Denoising diffusion probabilistic models (DDPM) are a class of generative systems that are emerging as an alternative solution for audio inverse problems such as enhancement [1], upsampling [2], and even source separation [3–5]. Music source separation (MSS) is the task of estimating the individual elements in a musical mixture [6]. Because of their conditioning flexibility and generative potential, DDPM are considered a promising solution for MSS [7]. While competitive diffusion separation systems exist [5, 8], these focus on instrumental music.

Large training data is key for DDPM [1, 9, 10], however, gathering clean, multi-stem data is challenging [11]. While large multi-stem collections recorded in live shows exist [12–16], these come with *source bleeding*: the other sources, room response, and other interferences leak into the individual stems. Regularly training an MSS model on such data often results in suboptimal performance [17].

In this work, we aim to leverage the inherent domain knowledge in a large collection of live multi-stem tracks with bleeding while still targeting clean separation. Carnatic Music, which is mostly enjoyed live [18], presents an interesting case of study. Prior work targeted the same objective for Carnatic vocals [17] and violin [19]. However, [17] relies on a complex heuristic compromising generalization and efficiency, while [19] uses a large, clean but private in-domain dataset. Clean Carnatic multi-stem data exist [20], but only for a small collection of 5 concerts.

We propose a generative approach to this problem, relying on latent diffusion models (LDM) [21]: the generative diffusion process operates on a compact, pre-learnt audio representation, enhancing efficiency and learning capacity [10]. We pre-train an LDM to generate signing vocals with source bleeding conditioned on music mixtures [22]. In parallel, we train a regressor to estimate the bleeding ratio in vocal signals using open, clean, non-Carnatic multi-stem data. We then refine the pre-trained LDM using a loss penalization term based on the bleeding predictions aiming at generating cleaner vocals. Inspired by gradient guidance for diffusion models [23, 24], we subsequently propose *regression-based bleeding level guidance*: we steer the gradients of the bleeding estimator to inform the diffusion sampler toward the direction for cleaner separation.

Non-generative MSS systems that transform or mask time-frequency representations normally rely on access to all stems, assuming these combine linearly to the reference mixture [25]. Leveraging generative flexibility we consider two added challenges: **(1)** access to the mixture and the corresponding vocal stem with bleeding only [26], and **(2)** the mixture alone has undergone non-linear processing.

We prioritize efficiency using a compact latent space, at the expense of signal quality and a significant penalization on separation metrics, a known problem for the evaluation of generative models [27, 28]. Nevertheless, our system achieves, without the need for clean, in-domain, multi-stem samples, competitive objective generation quality and perceptual separation preference over the baselines.

2. BACKGROUND

2.1 Latent diffusion

Let $X \in \mathbb{R}^{F \times D}$ be a latent embedding with feature size F and time dimension $D = \frac{T}{cf}$, where T is audio length

[1] https://github.com/genisplaja/ldm-carnatic-separation

© G. Plaja-Roglans, X. Serra, and M. Rocamora. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** G. Plaja-Roglans, X. Serra, and M. Rocamora, "Leveraging Carnatic live recordings for singing voice separation using regression-guided latent diffusion", in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

and c_f the compression factor of a certain latent encoder $E : x \in \mathbb{R}^{1\times T} \rightarrow X \in \mathbb{R}^{F\times D}$. In this work, we rely on a latent forward diffusion process defined by a Markov chain of T steps that converts a latent embedding $X \sim p(X)$, into a sample of Gaussian noise $\epsilon \in \mathbb{R}^{F\times D}$. The intermediate steps of this transformation are computed as [29]:

$$X_{\sigma_t} = \alpha_{\sigma_t} X_{\sigma_0} + \beta_{\sigma_t}\epsilon, \tag{1}$$

where $\sigma_t \in [0, 1]$ is a noise schedule of T values to control the transformation, while we define $\alpha_{\sigma_t} := cos(\phi_t)$ and $\beta_{\sigma_t} := sin(\phi_t)$, where $\phi_t := \frac{\pi}{2}\sigma_t$. Note also that $X_{\sigma_0} = X$. A model is then trained to revert this process, approximating the data distribution $p(X)$ by learning to map Gaussian samples to observations $X \sim p(X)$.

Let $v_{\sigma_t} \in \mathbb{R}^{1\times D}$ be the *velocity* objective, which corresponds to the inner variable of the diffusion process which tracks the transformation between X_{σ_0} and X_{σ_T}. The objective v_{σ_t} is formally computed as:

$$v_{\sigma_t} = \alpha_{\sigma_t}\epsilon - \beta_{\sigma_t} X_{\sigma_0}, \tag{2}$$

and estimated by neural network m with parameters θ:

$$\hat{v}_{\sigma_t} = m_\theta(X_{\sigma_t}, \sigma_t, C) \tag{3}$$

Network m_θ is the generative LDM. Input $C \in \mathbb{R}^{F\times D}$ represents the conditioning signal. Diffusion systems may be trained unconditionally to sample random observations from approximated $\hat{p}(X)$, while instructions from various modalities (e.g., text prompts [9], audio signals [5], and more) can be injected to the posterior to modify the generation trajectory. However, our work focuses on a well-defined inverse problem. As a result, we inject C during both training and inference, architecturally optimizing the system to tailor the diffusion trajectory relating the conditioning signal and the generator target $\hat{X}$. Let $\mathbb{E}$ denote expectation. The diffusion loss objective is defined as [29]:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t\sim[0,T],\sigma_t,X_{\sigma_t}}\left[||\hat{v}_{\sigma_t} - v_{\sigma_t}||_2^2\right] \tag{4}$$

2.2 Sampling process

The sampling process progressively models a sample pertaining to the approximated distribution $\hat{p}(X)$ by denoising a random sample of Gaussian noise. Previous works in audio generation have relied on the Denoising Diffusion Implicit Models (DDIM) sampler, achieving satisfactory compromise between sampling steps and generation quality [30]. In DDIM sampling [29], the inference process is performed using arbitrary T, and it is initiated at $\sigma_t = 1$. A given sampling step t is composed of a set of operations:

We first run a forward pass with model m_θ as defined in Eqn (3). Using predicted velocity $\hat{v}_{\sigma_t}$, we can compute $\hat{X}_{\sigma_0}$, which corresponds to the estimated target sample at $t = 0$, and $\hat{\epsilon}_{\sigma_t} \in \mathbb{R}^{1\times D}$ which is Gaussian noise at step t:

$$\hat{X}_{\sigma_0} = \alpha_{\sigma_t} X_{\sigma_t} - \beta_{\sigma_t}\hat{v}_{\sigma_t} \tag{5}$$

$$\hat{\epsilon}_{\sigma_t} = \beta_{\sigma_t} X_{\sigma_t} + \alpha_{\sigma_t}\hat{v}_{\sigma_t} \tag{6}$$

Note that, for $t \approx T$, i.e. at an early stage of the sampling process, predicted $\hat{X}_{\sigma_0}$ is expected to be noisy, limitedly consistent with signal C, while at $t \approx 0$, it approximates further to the final, refined separation. For $t > 0$, the input for the next sampling step is formally defined as:

$$\hat{X}_{\sigma_{t-1}} = \alpha_{\sigma_{t-1}}\hat{X}_{\sigma_0} + \beta_{\sigma_{t-1}}\hat{\epsilon}_{\sigma_t} \tag{7}$$

Finally, $\hat{X}_{\sigma_0}$ is decoded to the original domain using decoder $E' : X \in \mathbb{R}^{F\times D} \rightarrow x \in \mathbb{R}^{1\times T}$. Encoder E and decoder E' are normally pre-trained and kept frozen.

3. METHOD

Let $\mathcal{A}$ and $\mathcal{B}$ represent musical repertoires or domains which differ on instrumentation, concepts, and practices. In our work, $\mathcal{A}$ corresponds to Carnatic Music and $\mathcal{B}$ to Western radio music (e.g. pop, rock, hip-hop, and related).

3.1 Latent encoder

We use Music2Latent v1 [31] (M2L), which is a neural codec based on a consistency model [32]. Both M2L encoder and decoder are depicted in red in Figure 1. M2L compresses signals sampled at 48kHz down to 12Hz, and produces 64-dimensional codes with 0 mean and deviation 1. The significant compression of M2L enables the development of our work in an environment with limited computational resources. M2L is trained using MTG-Jamendo dataset [33], which includes numeros tracks for repertoire $\mathcal{B}$, and 90 recordings tagged as *indian*. It also includes $\approx$ 2k vocal tracks, and $\approx$ 2k tracks with violin. We are unaware of the number of recordings mixing these sources.

The enormous compression rate of M2L comes at a cost: authors report -3.85dB of reconstruction SI-SDR, a standard separation metric, and perceptible artifacts often arise in the reconstructions. While official code to train or fine-tune M2L is not available, we rely on the open pre-trained model, prioritizing its compression and feature learning capabilities to study the effectiveness of latent diffusion for weakly-supervised MSS. Moreover, the M2L compression rate enables us to perform our LDM study with very limited computational resources, yet training a model with the size on par with the literature [34–36].

3.2 Latent diffusion for separation

Model m_θ is a 1D attention U-Net with skip-connections. It is depicted in green in Fig. 1. It is composed of n residual blocks which include two 1D convolutional layers, each preceded by GroupNorm [37] and SiLU activation [38]. A pre-defined number of blocks include time-wise self-attention to learn the relationship between different time steps and enrich context, which is crucial for MSS.

To down and upsample the features at each level in the U-Net, we add an extra layer with kernel size $k\times k$, k being the time compression or expansion factor. When $k > 1$, for downsampling we double the feature channels, while halving them for the upsampling blocks.

The time-step σ_t is projected into a 1024-channeled random Fourier feature embeddings, which are processed through a 3-layer multi-layer perceptron (MLP) with GELU activations. The resulting embedding is incorporated into the model via FiLM layers.

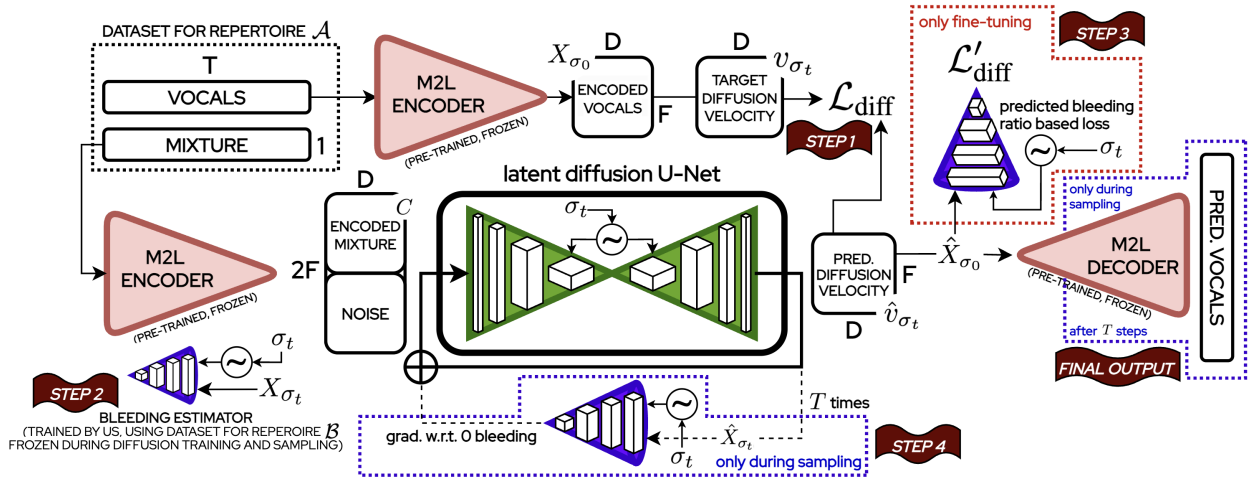

Figure 1. Diagram of the proposed system. The LDM is first trained to generate encoded vocals with bleeding. Next, we fine-tune using the bleeding ratio loss. Finally, during sampling, we use the bleeding predictions to compute the gradients towards less bleeding and modify the generation trajectory on that direction. The order of development steps is indicated.

Several mechanisms to inject conditioning signals in diffusion U-Nets exist [1, 30, 34]. We find the best quality-efficiency compromise on concatenating, over the feature channels, the conditioning signal C and X_{σ_t} [39]. Previous latent diffusion work using M2L embeddings has relied on this mechanism [34]. Even if the M2L embeddings are 2D, we employ 1D convolutional layers to effectively capture temporal dependencies in the compressed representation, processing each feature vector independently without imposing artificial spatial correlations.

The network is trained relying on the objective in Eq. 4 using corresponding pairs of vocal stems with bleeding X_{σ_0} and mixture C, both encoded using E [22].

3.3 Bleeding level estimator

The glass ceiling of the separation LDM presented in Section 3.2 is established at the inherent source bleeding in the training data for domain $\mathcal{A}$. However, the network may still be trained to map from mixture to the corresponding vocals with bleeding, leveraging domain knowledge [17].

Prior work has shown that a separator model trained using only data with source bleeding can be fine-tuned towards cleaner outputs by steering a *bleeding estimator* network [19], which predicts the ratio of bleeding in the preliminary separations, while the non-optimal separator is optimized to minimize this ratio. Building on this insight, we hypothesize that estimating bleeding ratios is less prone to severe generalization errors compared to MSS. This allows us to leverage the knowledge embedded in a pre-trained separator for repertoire $\mathcal{A}$, while fine-tuning using the bleeding estimator trained using repertoire $\mathcal{B}$, bypassing access to clean multi-stem data for repertoire $\mathcal{A}$.

3.3.1 Regression-based bleeding level guidance

In an attempt to guide the pre-trained LDM to generate cleaner vocals, we leverage a regression model to guide the diffusion process using the level of source bleeding.

Similarly to [19], we aim to introduce a bleeding estimator model to guide the separation system toward reducing the bleeding. Leveraging the flexibility of the sampling process of diffusion models, we propose *regression-based bleeding level guidance* (RG), which is inspired by classifier guidance (CG) [23], a technique to enhance quality and control in diffusion image generation. CG leverages a pre-trained classifier to estimate the class to which X_{σ_t} belongs to. Using the gradients obtained w.r.t. X_{σ_t}, we may modify the sampling trajectory targetting better the desired class. Prior studies have also relied on gradients from external networks to tune diffusion sampling [24, 40].

We use a bleeding estimator, in blue in Fig. 1, to predict the amount of source bleeding in an individual stem, represented by a floating point value $b \in [0, 1]$, where 0 represents no bleeding and 1 the mixture. Using a clean multi-stem but out-of-domain dataset for repertoire $\mathcal{B}$, we train a neural network r_ϕ to perform this task. Since r_ϕ is meant to be integrated within the iterative diffusion process, the bleeding prediction input is expected to be $\hat{X}_{\sigma_t}$, which is infused with Gaussian noise following Eq. 1. Therefore, the training input of r_ϕ is an M2L-encoded vocal stem with bleeding (with ratio b), corrupted using Eq. 1 from the diffusion formulation. Note that the M2L codes have shown competitive performance in several downstream tasks [31]. The model is trained using L2 loss:

$$\mathcal{L}_{r_\phi}(X_{\sigma_t}, \sigma_t, b) = \|r_\phi(X_{\sigma_t}, \sigma_t) - b\|_2^2 \qquad (8)$$

Bleeding is expected to stay consistent along an audio sample, thus the model estimates a single $\hat{b}$ per each input. Diffusion time-step σ_t is also injected to the regressor to provide information of the current noise level [23, 24].

Regression-guidance for diffusion sampling. We incorporate RG in the diffusion sampling algorithm by steering the gradients from the bleeding predictor. We predict the bleeding of the input diffusion forward variable X_{σ_t} at each sampling step t, and calculate the gradients that point toward the direction of our target: 0 bleeding [40]. The ex-

tracted gradients are used to guide predicted velocity $\hat{v}_{\sigma_t}$, following the formulation below:

$$W_{\text{guid}} = \eta \cdot 10^2 \cdot \frac{1}{t-1} \cdot {\sigma_t}^2 \quad (9)$$

$$\hat{v}_{\sigma_t}^{guid} = \hat{v}_{\sigma_t} + W_{\text{guid}} \cdot \nabla_{X_{\sigma_t}} |0 - r_\phi(X_{\sigma_t}, \sigma_t)| \quad (10)$$

The gradients are normalized using per-sample L2 normalization, ensuring stable guidance. The guidance level is manually controlled by η, and is also dynamically scaled to provide less guidance in the beginning of the sampling process where $X_{\sigma_t} \approx \epsilon$, and strengthen the guidance effect on the intermediate sampling steps [24].

LDM bleeding-aware fine-tuning. We incorporate a penalization loss term to penalize the pre-trained LDM using the level of bleeding in the predictions. For time-steps with low noise exposure ($\sigma_t < 0.6$), the frozen bleeding estimator predicts the bleeding ratio before and after a denoising step, denoted as $\hat{b}_{\text{pre}} = r_\phi(X_{\sigma_t})$ and $\hat{b}_{\text{post}} = r_\phi(\hat{X}_{\sigma_0})$, respectively. A max-margin hinge term $max\big(0,\ \hat{b}_{\text{post}} - \hat{b}_{\text{pre}} + m\big)$ with margin $m = 0.05$, ensures that the model must reduce bleed by at least m, otherwise it is penalized. The penalization term is further weighted by $(1-\sigma_t)^2$ to amplify its impact at later and perceptually clearer steps. Overall, the fine-tuning loss becomes:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda \cdot (1-\sigma_t)^2 \cdot max\big(0,\ \hat{b}_{\text{post}} - \hat{b}_{\text{pre}} + m\big) \quad (11)$$

We use parameter λ to control the balance between penalization term and diffusion loss, encouraging consistent reduction in bleed while maintaining generation fidelity.

3.3.2 Network details

Bleeding estimator r_ϕ is based on a stack of dilated convolutions which are regularized using GroupNorm, SiLU activations, and dropout. The output embedding is batch-normalized and then σ_t, which is processed using the same step embedder than the LDM, is injected via summation. The resultant vector is then passed through a bidirectional LSTM, capturing temporal dependencies. To model global temporal relationships, we use multi-head self-attention over the sequence. Finally, we apply global average pooling across the time dimension and use sigmoid-activated linear layer to produce a single scalar corresponding to predicted bleed score ($\in [0, 1]$).

4. EXPERIMENTS

4.1 Experimental setup

4.1.1 LDM separation pre-training

The LDM U-Net is 7 levels deep. The feature channels per layer are set as: $\{128, 256, 256, 512, 512, 1024, 2048\}$, input channels being $64 * 2$–being the channel-wise concatenation of the M2L codes and the model input X_{σ_t}. However, the last convolutional layer outputs a 64-channeled signal, corresponding to the generated embedding to decode. The training input context are 1052000 samples ($\approx$ 24s), which is compressed to 2048 samples using M2L. Time compression factors of the LDM are set to $\{1, 2, 1, 2, 1, 2, 1\}$, factor 1 representing no compression, thus we reach 128 samples in the bottleneck, trying not to over-compress the information.

The first U-Net level is composed of 1 block, while the deepest includes 4. The rest are composed of 2 blocks. The four deepest levels of the U-Net include time-wise self-attention with 8 heads, aiming at enriching context.

The LDM network has $\approx$ 365M trainable parameters, and it is trained using the 168 multi-stem recordings from Saraga Carnatic–15 recordings are kept for validation, each of them from a different concert. We use ADAM optimizer with learning rate $1 * 10^{-5}$, and use a linear warm-up stage using a cosine scheduler with a initial rate of $1.6 * 10^{-6}$. We reach 500k training steps in two weeks in an 8GB GPU.

4.1.2 Bleeding estimator guidance

Artificial bleeding dataset. The bleeding estimator r_ϕ is trained using musdb18hq [41], corresponding to domain $\mathcal{B}$. We artificially create the bleeding following the pipeline described in the SDX 2023 bleeding challenge [25]. Let S_i be a given source stem. The accompaniment A is the weighted sum of the non-vocal sources: $A = \sum_{i=1}^{N} w_i S_i$, where $w_i \sim U(0, 1)$ are independently sampled random mixing weights. We randomly filter the sources using the following categorical distribution [19, 25]:

$$S_i = \begin{cases} \text{BPF}(S_i, f_c^{\text{low}}, f_c^{\text{high}}) & \text{with } p = 0.4 \\ \text{HPF}(S_i, f_c) & \text{with } p = 0.4 \\ S_i & \text{with } p = 0.2 \end{cases}$$

where we use band pass filter (BPF) with low cutoff frequency $f_c^{\text{low}} \sim U(200, 600)$ Hz and high cutoff frequency $f_c^{\text{high}} \sim U(8\text{k}, 10\text{k})$ Hz, and high-pass filter (HPF) with $f_c \sim U(900, 9\text{k})$ Hz. The order of the filters is also randomly sampled from $\sim U(3, 8)$. Next, a bleeding ratio $b \sim U(0, 1)$ is sampled and used to compute the reference mixture $M = S_v + b \cdot A$. We normalize M to prevent clipping and confine all values between $[-1, 1]$.

Model details. The input artificial mix M is encoded using M2L and merged with Gaussian noise using Eq. 1, therefore, the input channel size is 64. We use five dilated convolutions with ratios {1, 2, 4, 8, 8}. The number filters of the convolutional stack and also the size of the hidden state of the LSTM are set to 512. The multi-head attention mechanism is configured with 8 heads.

Training scheme. The training context for the bleeding predictor is the same as that of the LDM. The bleeding estimator model totals $\approx$ 1M parameters. We use ADAM optimizer with a learning rate $4 * 10^{-4}$, and train until convergence. Subsequently, the pre-trained LDM is fine-tuned for 10k steps using $\lambda = 50$, using learning rate $1 * 10^{-6}$.

4.1.3 Diffusion sampling parameters

We sample using overlapping segments of $\approx$ 24s with $\approx$ 5s hop, which are subsequently combined using overlap-add. If sampling with $T = 64$ on a single TITAN X 8GB GPU, our system separates audio at an average speed of ≈ 0.4x the duration of the track.

4.2 Datasets

Saraga Carnatic (Domain $\mathcal{A}$, training data). This is a collection containing 168 real multi-stem recordings (totaling $\approx$ 60h of music), including vocals, violin, and percussion instruments, missing only the tanpura, which provides the tonic from which an entire performance is built. Carnatic Music is mostly enjoyed live, therefore, to ensure ecological validity, Saraga is recorded in live performances, collecting the stems from the mixer. However, this has a drawback: the microphone of a given source captures, in the background, the other sources.

musdb18hq (Domain $\mathcal{B}$, for RG). It is one of the most established open datasets for MSS. It includes 100 training and 50 testing multi-stem tracks split in vocals, bass, drums, and others. It represents a limited set of styles mostly confined in Western commercial music.

Sanidha ($\mathcal{A}$, evaluation data). It is the only available open collection of clean multi-stem recordings for Carnatic Music [20]. After some exploration on this dataset, which is composed of 5 concerts, we discard 1 having bleeding in the vocal stem leaked through the singer headphones. While Sanidha has not been yet shown a potential dataset for training over Saraga, we employ it for testing, enabling more reliable objective evaluation for this repertoire.

4.3 Evaluation metrics

The objective evaluation of generative systems for audio inverse problems is challenging [1]. In MSS, traditional definitions for source-to-distortion ratio (SDR), the standard separation metric, have been reported to often misrepresent the perceptual quality of separations [42, 43]. Moreover, SDR is significantly penalized by potential subtle differences and phase mismatches commonly present when evaluating fully-generative models. For these reasons, SDR is being less used in prior generative separation work. In the case of LDM, given the added phase reconstruction mismatch introduced by the latent encoder, not even scale-invariant SDR (SI-SDR), present in various generative separation systems [5, 26], is being used [8, 44].

Therefore, we rely on alternative audio quality measures that have been employed in prior work on latent diffusion for generation [34, 35] and source separation [5, 8, 44]: log-spectral distance (LSD) [35] and log mel-spectrogram L2 error [43]. These metrics are phase-independent and may be more appropriate for generative systems. We also report perceptual evaluation of speech quality (PESQ) [45], aiming at measuring intelligibility.

To assess the quality of the generated signals without relying on matching audio pairs, we report the Fréchet Audio Distance (FAD) [46]. Model outputs with higher quality and lesser interferences should report lower FAD. The FAD is often computed on short chunks. However, in addition to diversity, context is important [46]. Carnatic renditions often feature prolonged improvisational segments, such as *alapana* and *tanam*, which can span several minutes. For these reasons, we split the samples into 1-minute chunks. We discard the chunks with $> 25\%$ of silence, which results in ≈ 150 testing samples.

Dataset	L2 Loss↓	Avg. $\hat{b}$ mix	Avg. $\hat{b}$ voc.
musdb18hq	0.054	0.89 ± 0.18	0.05 ± 0.09
Sanidha	0.098	0.80 ± 0.23	0.08 ± 0.18
Saraga	-	0.97 ± 0.07	0.25 ± 0.29

Table 1. **Assessing the bleeding regressor.** Ideally, the avg. $\hat{b}$ should be ≈ 1 for mix, and ≈ 0 for vocals, except for Saraga, whose vocal stems have inherent, real bleeding.

To complement the objective measures we run a perceptual test with human listeners. We conduct a preference-based experiment [1]. We split the separations into chunks of $\approx$ 15 seconds, discarding unvoiced regions, and randomly select an instance for each rendition. Using the mixture as reference, we select 6 examples from the pool, including diversity of music sections and singer gender.

The participants are shown several unlabeled and randomly ordered pairs of samples of our model against other systems. We introduce comparisons between non-generative models to prevent the participants from getting familiar with model-specific artifacts.

4.4 Compared systems

We compare against the multi-source diffusion model (MSDM) [5] for separation. Since no weights for vocals are available, and the system is designed for clean multi-stem data, we train it using musdb18hq, following the instructions in the repository. We do not compare against existing LDM separators since these are not optimized for vocals [8] or code or weights are not yet available [44].

While not directly comparable, since these are non-generative and mask-based, we evaluate *cold-diff* [17] and the *mixer* model from [19], the baseline systems addressing the same task for the Carnatic study case. Both models are directly used through the `compIAM` package [47]. Also, to provide a performance bound for our model, we evaluate the M2L-reconstructed vocal stems in Sanidha.

We perform an ablation study on the bleeding fine-tuning (FT), regression guidance level (RG$^\eta$) for levels $\eta \in [0, 5, 10, 20]$, and sampling steps $T \in [32, 64]$. The non fine-tuned model is trained for $\approx$ 10k more steps for a fairer comparison with the FT model. For the perceptual test we use $T = 32$, mid-guidance $\eta = 10$.

5. RESULTS

5.1 Evaluating the bleeding predictor

Evaluating the bleeding predictor is complex, since no music datasets with real, annotated bleeding exist. However, we perform two sanity checks. First, we compute the model L2 loss on artificial bleeding mixtures using musdb18hq and Sanidha. Second, we compute the average bleeding ratio on Saraga mixtures and vocal stems. To simulate the actual application of the bleeding predictor, inputs are noised using Eq. 1, uniformly sampling σ_t values per-example. We predict the bleeding for 500 randomly sampled and voiced 12s-excerpts per each dataset.

Model	T	FAD↓	LogMel L2↓	LSD↓	PESQ↑
M2L	–	0.281	2.95	1.22	2.73
cold-diff [17]	8	**0.515**	15.79	2.29	1.39
mixer [19]	–	0.648	13.26	1.77	1.22
MSDM [5]	150	0.791	12.52	2.00	**1.78**
Proposed					
no FT	32	0.637	16.74	1.80	1.15
FT	32	0.593	16.02	1.75	1.17
FT-RG^5	32	0.587	13.36	1.68	<u>1.22</u>
FT-RG^{10}	32	<u>0.579</u>	12.89	1.66	1.18
FT-RG^{20}	32	0.626	13.44	1.67	1.19
no FT	64	0.642	16.78	1.79	1.16
FT	64	0.602	16.10	1.74	1.16
FT-RG^5	64	0.600	13.41	1.68	1.21
FT-RG^{10}	64	0.595	12.61	1.65	1.19
FT-RG^{20}	64	0.623	<u>**12.31**</u>	<u>**1.64**</u>	1.16

Table 2. **Objective evaluation of diverse systems on audio and vocal quality measures.** Arrow ↓ indicates lower is better, ↑ otherwise. In bold, we indicate the best score among all systems. We underline the best scores of the ablation. See further ablation results in the companion repo.[1]

See the results in Table 1. The regressor generalizes quite satisfactorily to the Carnatic domain. The system also discriminates real Carnatic mixtures ($\hat{b} \approx 1$) from vocal stems with bleeding ($\hat{b} = 0.25 \pm 0.29$). The high standard deviation in the predicted bleeding ratio for Saraga may be explained by the high variance in accompaniment presence in different sections of a Carnatic rendition.

5.2 Objective evaluation

See the objective evaluation in Table 2. We observe a tangible improvement for our model when using the bleeding guidance during sampling, finding the sweet spot on RG^{10} for $T = 32$, and RG^{20} for $T = 64$, despite the metrics not always correlate. The bleeding fine-tuning loss term provides a more moderate improvement.

Our generative system outperforms the baselines on the spectral assessment metrics, while ranking second on FAD, only outperformed by *cold-diff*, a non-generative system. In terms of PESQ, our system scores the lowest, only leveling the *mixer* model when using $T = 32$ and $\eta = 10$.

The performance across metrics for our system suggests that stronger guidance further cleans and brings the generation closer to the target signal overall. However, this results in a trade-off: stronger interference removal comes at the cost of degraded vocal quality. While our system shows competitive overall quality, especially when guided, it reports lower PESQ than MSDM, the generative baseline, suggesting that MSDM generations have further intelligibility but also stronger interference. Note however that MSDM does not generate encoded latents but directly waveforms, potentially accumulating less phase discrepancy. Nevertheless, the general low PESQ scores for all models may be explained by the fact that this metric is for speech and it does not assume potential interferences, while it is unclear how it characterizes the extremely common and strong vocal ornaments in Carnatic Music.

Model	Quality (%)		Interference (%)	
	Ours	Other	Ours	Other
cold-diff [17]	5.0	95.0	97.50	2.50
mixer [19]	15.0	85.0	100.0	0.0
MSDM [5]	20.0	80.0	100.0	0.0

Table 3. Perceptual evaluation results showing the percentage of participants who preferred our system or baseline models in terms of quality and interference removal.

The results suggest that an LDM can be trained towards generating separated complex sources such as vocals, while the proposed guidance method contributes to a cleaner generation that gets closer to the target signal. The objective metrics support the expected behavior of the proposed approach, although we hypothesize potential stronger improvement if future efforts are done on fine-tuning the latent encoder and refining the bleeding estimator, as well as scaling the network up (closely related LDM relies on > 500M parameters [34]). These may contribute to a refine the source quality of the generated vocals.

5.3 Perceptual assessment

A total of 20 participants took the test. Interestingly, the participant agreement is remarkable. The results are reported in Table 3. The perceptual assessment is significantly clear: our system leads in interference removal but is not able to reach the source quality of the baselines. These results agree with the objective metrics, which suggest that the overall quality and cleanliness of our generations are competitive, however, the fidelity and intelligibility of the generated vocals leave room for improvement.

6. CONCLUSIONS

We present a deep generative model to address weakly-supervised singing voice separation for Carnatic Music, leveraging pairs of in-domain mixture and vocal stems which have source bleeding because these are recorded in real live performances. We propose to train a latent diffusion model to generate vocals with bleeding conditioned on the corresponding mixture. We then guide the pre-trained generative system toward producing cleaner samples using a bleeding ratio predictor. Our system achieves competitive scores for generation quality measures, and outperforms the baselines in terms of interference removal in a preference listening test. While the proposed framework shows promise, we envision extensive future work, especially to refine the vocal quality, which is sub-optimally ranked in the results. Tailoring the latent encoder to Carnatic, improving the bleeding estimator, and performing multi-source separation are potential future research lines.

We believe that the flexibility, conditioning, and guidance capabilities of DDPM may enable approaches to tackle separation in non-optimal contexts, or even improve performance on ideal conditions. This has potential for separation of repertoires that are not commonly recorded in studios, considering underrepresented instruments, and prioritizing particular aspects such as interference removal.

7. ETHICS STATEMENT

While this work deals with generative modeling, the system is trained using fully-open data, while we address a purely inverse problem. No copyright implications should be involved. The model is architecturally developed not to generate unseen music recordings. The results of the perceptual test were included in this work with the participants' permission, whose identities remain anonymous. No personal of sensitive data from the participants is collected and/or distributed.

8. ACKNOWLEDGEMENTS

This work is supported by IA y Música: Cátedra en Inteligencia Artificial y Música (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA, and IMPA: Multimodal AI for Audio Processing (PID2023-152250OB-I00), funded by the Ministry of Science, Innovation and Universities of the Spanish Government, the Agencia Estatal de Investigación (AEI) and co-financed by the European Union.

9. REFERENCES

[1] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arxiv:2206.03065*, 2022.

[2] J. Lee and S. Han, "NU-wave: A diffusion probabilistic model for neural audio upsampling," in *Annual Conf. of the Int. Speech Communication Assoc. (INTERSPEECH),* Brno, Czech Republic, 2021, pp. 2698–2702.

[3] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Singapore, 2022.

[4] C.-Y. Yu, E. Postolache, E. Rodolà, and G. Fazekas, "Zero-shot duet singing voices separation with diffusion models," in *Sound Demixing Workshop (SDX)*, 2023.

[5] G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà, "Multi-source diffusion models for simultaneous music generation and separation," in *12th Int. Conf. on Learning Representations,* Viena, Austria, 2024.

[6] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *18th Int. Society for Music Information Retrieval Conf. (ISMIR),* Suzhou, China, 2017, pp. 745–751.

[7] S. Araki, N. Ito, R. Haeb-Umbach, G. Wichern, Z.-Q. Wang, and Y. Mitsufuji, "30+ years of source separation research: Achievements and future challenges," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Hyderabad, India, 2025.

[8] T. Karchkhadze, M. Izadi, and S. Dubnov, "Simultaneous music separation and generation using multi-track latent diffusion models," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Hyderabad, India, 2025.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *33th Advances in Neural Information Processing Systems (NeurIPS),* Online, 2020, pp. 6840–6851.

[10] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *Int. Conf. on Machine Learning (ICML),* Vienna, Austria, 2024.

[11] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[12] T. Prätzlich, M. Müller, B. Bohl, and J. Veit, "Freischütz Digital: Demos of audio-related contributions," in *Demos and Late Breaking News of the Int. Society for Music Information Retrieval Conf. (ISMIR),* Málaga, Spain, 2015.

[13] O. Mayor, Q. Llimona, M. Marchini, P. Papiotis, and E. Gómez, "repoVizz: a framework for remote storage, browsing, annotation, and exchange of multimodal data," in *ACM Int. Conf. on Multimedia,* Barcelona, Spain, 2013.

[14] E. Gómez, M. Grachten, A. Hanjalic, J. Janer, S. Jordà, C. F. Julià, C. Liem, A. Martorell, M. Schedl, and G. Widmer, "PHENICX: Performances as Highly Enriched and Interactive Concert Experiences," in *Proc. of the 10th Sound and Music Computing Conf. (SMC),* Stockholm, Sweden, 2013.

[15] A. Srinivasamurthy, S. Gulati, R. C. Repetto, and X. Serra, "Saraga: Open datasets for research on indian art music," *Empirical Musicology Review*, vol. 16, no. 1, pp. 85–98, 2021.

[16] A. Shankar, G. Plaja-Roglans, T. Nuttall, M. Rocamora, and X. Serra, "Saraga Audiovisual: a large multimodal open data collection for the analysis of Carnatic Music," in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

[17] G. Plaja-Roglans, M. Miron, A. Shankar, and X. Serra, "Carnatic singing voice separation using cold diffusion

on training data with bleeding," in *24th Int. Society for Music Information Retrieval Conf. (ISMIR),* Milano, Italy, 2023.

[18] G. Plaja-Roglans, T. Nuttall, L. Pearson, X. Serra, and M. Miron, "Repertoire-specific vocal pitch data generation for improved melodic analysis of carnatic music," *Transactions of the International Society for Music Information Retrieval*, 2023.

[19] A. Shankar, S. Schweinitz, G. Plaja-Roglans, X. Serra, and M. Rocamora, "Disentangling overlapping sources: Improving vocal and violin source separation in carnatic music," in *Workshop on Indian Music Analysis and Generative Applications (WIMAGA) in ICASSP*, 2025.

[20] V. V. Krishnan, N. Alben, A. A. Nair, and N. Condit-Schultz, "Sanidha: A studio quality multi-modal dataset for carnatic music," in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Conf. on Computer Vision and Pattern Recognition (CVPR),* New Orleans, Louisiana, USA, 2021.

[22] G. Plaja-Roglans, Y.-N. Hung, X. Serra, and I. Pereira, "Efficient and fast generative-based singing voice separation using a latent diffusion model," in *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN),* Rome, Italy, 2025.

[23] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *35th Conf. on Neural Information Processing Systems (NeurIPS 2021),* Online, 2021.

[24] A. Bansal, H.-M. Chu, A. Schwarzschild, R. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," in *Proc. of the 2th International Conference on Learning Representations*, 2024.

[25] G. Fabbro, S. Uhlich, C.-H. Lai, W. Choi, M. Martínez-Ramírez, W. Liao, I. Gadelha, G. Ramos, E. Hsu, H. Rodrigues *et al.*, "The Sound Demixing Challenge 2023: Music Demixing Track," *Transactions of the Int. Society for Music Information Retrieval*, 2023.

[26] G. Zhu, J. Darefsky, F. Jiang, A. Selitskiy, and Z. Duan, "Music source separation with generative flow," *IEEE Signal Processing Letters*, vol. 29, p. 2288–2292, 2022. [Online]. Available: http://dx.doi.org/10.1109/LSP.2022.3219355

[27] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Content based singing voice extraction from a musical mixture," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Online, 2020, pp. 781–785.

[28] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "WARP-Q: quality prediction for generative neural speech codecs," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Online, 2021, pp. 401–405.

[29] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. on Learning Representations (ICLR),* Online, 2021.

[30] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," 2023.

[31] M. Pasini, S. Lattner, and G. Fazekas, "Music2latent: Consistency autoencoders for latent audio compression," in *25th Int. Society for Music Information Retrieval Conf. (ISMIR),* San Francisco, USA, 2024.

[32] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023.

[33] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: http://hdl.handle.net/10230/42015

[34] J. Nistal, M. Pasini, C. Aouameur, M. Grachten, and S. Lattner, "Diff-A-Riff: Musical Accompaniment Co-creation via Latent Diffusion Models," in *25th Int. Society for Music Information Retrieval Conf. (ISMIR),* San Francisco, USA, 2024.

[35] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Int. Conf. on Machine Learning (ICML),* Honolulu, Hawaii, 2023.

[36] J.-S. Hwang, S.-H. Lee, and S.-W. Lee, "Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models," *arXiv preprint arXiv:2306.06814*, 2023.

[37] Y. Wu and K. He, "Group normalization," *arXiv:1803.08494*, 2018.

[38] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks*, vol. 107, pp. 3–11, 2018.

[39] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *34th Conf. in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 12 438–12 448.

[40] Y. Guo, H. Yuan, Y. Yang, M. Chen, and M. Wang, "Gradient guidance for diffusion models: An optimization perspective," in *Proc. of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[41] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18 - a corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[42] E. Cano, D. Fitzgerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *24th European Signal Processing Conf. (EUSIPCO),* Budapest, Hungary, 2016, pp. 1758–1762.

[43] E. Gusó, J. Pons, S. Pascual, and J. Serrà, "On loss functions and evaluation metrics for music source separation," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 306–310, 2022.

[44] Y. Chae and K. Lee, "Mge-ldm: Joint latent diffusion for simultaneous music generation and source extraction," *arXiv:2505.23305*, 2025.

[45] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2001, pp. 749–752.

[46] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Seoul, South Korea. IEEE, 2024, pp. 1331–1335.

[47] Genís Plaja-Roglans and Thomas Nuttall and Xavier Serra, "compiam," 2024. [Online]. Available: https://mtg.github.io/compIAM/