Local Gaussian Diffusion Posterior Sampling for Unsupervised Multichannel Noisy Speech Separation

Zimo Zhang*, Aditya Arie Nugraha†, Kazuyoshi Yoshii‡†, and Tatsuya Kawahara*

*Graduate School of Informatics, Kyoto University, Kyoto, Japan
†Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan
‡Graduate School of Engineering, Kyoto University, Kyoto, Japan

E-mail: zhang.zimo.46x@st.kyoto-u.ac.jp

***Abstract*—This paper describes unsupervised multichannel speech enhancement and speech separation in the presence of diffuse environmental noise, both formulated under the local Gaussian model (LGM). A recent representative method, ArrayDPS, brought diffusion posterior sampling (DPS) to this setting by treating the mixture as a convolutive process and recovering speech sources jointly with relative room impulse responses estimated by forward convolutive prediction (FCP). This convolutive-mixture formulation, however, cannot represent diffuse noise as an explicit source and decouples the spatial inference from the diffusion update. Both limitations originate from the choice of observation model rather than from the diffusion machinery. To resolve both within a single design decision, we propose MNMF-DPS, which adopts the LGM as the observation backbone for a unified inference procedure. Our method instantiates the FastMNMF2 spatial parameters, an NMF-based noise PSD, and a UDiffSE diffusion speech prior, updates the spatial–noise model after an initial freeze during DPS refinement, and covers both tasks by adjusting only the source counts. Experiments on CHiME-3 and spatialized WSJ0-2mix showed that MNMF-DPS improves SI-SDR over ArrayDPS by 3.93 dB on enhancement and dominates ArrayDPS in 69.7% of paired samples on separation.**

***Index Terms*—Speech enhancement, source separation, diffusion model, local Gaussian model**

I. Introduction

Multichannel speech enhancement and noisy speech separation are central problems in microphone-array signal processing, both naturally formulated under the local Gaussian model (LGM) [1], [2]. The LGM represents each short-time source image as a zero-mean complex Gaussian whose covariance factorizes into a time-varying power spectral density (PSD) and a time-invariant full-rank spatial covariance matrix (SCM), yielding an analytic multichannel likelihood and a Wiener-filter source-image estimator.

Score-based diffusion models [3], combined with the observation through diffusion posterior sampling (DPS) [4], have transformed speech restoration. Single-channel enhancers learn either a conditional noisy–clean score in supervised settings [5] or pair a clean-speech score with an NMF noise model through an expectation–maximization scheme in unsupervised settings [6]. ArrayDPS [7] extended this line to multichannel blind speech separation by treating the mixture as a convolutive process and recovering speech sources jointly with the relative room impulse responses (RIRs), which are point-estimated at each diffusion step by forward convolutive prediction (FCP).

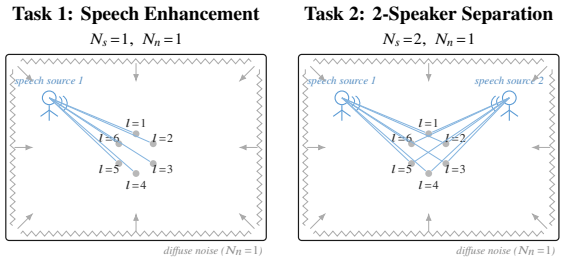


Fig. 1. Target tasks for MNMF-DPS on 6-channel reverberant recordings with diffuse noise: (left) enhancement with $(N_\mathrm{s}, N_\mathrm{n}) = (1, 1)$ and (right) noisy two-speaker separation with $(2, 1)$. The same unsupervised inference is used; only the source counts differ.

This convolutive-mixture formulation, while effective in mildly reverberant settings with localized interferers, leaves important gaps in the modeling. The residual term is modeled as white noise, so diffuse environmental noise has no explicit source in the model, and its energy is implicitly absorbed by the speech estimate through the diffusion prior. The relative RIRs are point-estimated by an auxiliary optimization decoupled from the diffusion update, and this estimation can become unstable under strong reverberation, low SNR, or spatially close sources. Both gaps are structural consequences of the convolutive-mixture formulation, because an explicit noise source and a probabilistic treatment of the spatial parameters have no natural place within this likelihood.

We hypothesize that both gaps could be resolved within a single design decision by replacing the convolutive-mixture observation model with the LGM. Under the LGM, noise sources enter the likelihood as additional zero-mean Gaussian terms with the spatial parameters as free variables of the same likelihood, and the negative log-likelihood reduces to an Itakura–Saito (IS) divergence whose gradient back-propagates to the diffusion state while the spatial and noise parameters are updated by likelihood- based MM steps. We instantiate this design as MNMF-DPS, combining the FastMNMF2 jointly-diagonalizable spatial model [8], an NMF-based noise PSD,

and a UDiffSE diffusion speech prior within one unified inference procedure.

The main contribution of this work is to formulate an LGM-based DPS framework that jointly optimizes the spatial, NMF noise, and diffusion parameters within a single reverse-SDE loop, replacing the FCP-based RIR point estimation used in prior multichannel DPS. We show that the explicit noise source within the LGM prevents environmental noise from leaking into the speech estimate, an effect confirmed directly by an ablation that removes the noise source. We further demonstrate that the same model handles both enhancement and noisy two-speaker separation by adjusting only the source counts, improving over ArrayDPS on enhancement and showing paired separation dominance under matched preprocessing.

II. Related Work

MNMF-DPS draws on LGM-based multichannel BSS with its progressively deepening source priors, single-channel diffusion priors combined with explicit noise models for unsupervised enhancement, and the recent extension of DPS to multichannel speech separation.

A. LGM-based multichannel BSS

The LGM has been the standard probabilistic backbone for multichannel BSS for over a decade, with continual refinement of its spatial model. After the full-rank SCM was established for reverberant mixing [1], auxiliary-function-based IVA [9] provided efficient frequency-wise demixing updates under statistical independence, multichannel NMF [10] jointly modeled the spatial covariance and spectral structure, and ILRMA [11] unified the two for determined mixtures. FastMNMF2 [8] accelerated inference by jointly diagonalizing the source SCMs while retaining their full-rank nature, and was extended to joint dereverberation [12].

Subsequent work replaced NMF with deeper generative priors within the same LGM backbone. A multichannel variational autoencoder (MVAE) [13] used a conditional VAE as a speech PSD model and was accelerated by FastMVAE2 [14], while a deep speech prior was coupled with FastMNMF for semi-supervised enhancement [15]. Normalizing flows [16] and neural full-rank covariance analysis [17] further increased the expressiveness of the source and spatial models. These methods established the LGM as a flexible host for deep generative speech priors through alternating inference, yet predate diffusion-based priors.

B. Single-channel diffusion priors for speech restoration

Diffusion models have become a powerful tool for single-channel speech enhancement, initially in supervised settings. SGMSE+ [5] uses an OUVE-SDE framework with a conditional score trained on noisy–clean pairs, and StoRM [18] added stochastic regeneration of a discriminative output to improve robustness under mismatched noise.

Unsupervised single-channel approaches instead combine a clean-speech diffusion prior with an explicit noise model through alternating inference. UDiffSE [6] couples a clean-speech score model with an NMF noise covariance through an expectation–maximization scheme in which a DPS E-step samples clean speech and an M-step updates the noise parameters. An audio-visual variant adds video conditioning [19]. These methods integrate a diffusion speech prior with an explicit noise model in a single-channel setting but cannot exploit multichannel spatial information.

C. Multichannel diffusion posterior sampling

ArrayDPS [7] brought DPS to multichannel blind speech separation by treating the mixture as a convolutive process and recovering speech sources jointly with relative RIRs point-estimated by FCP at each diffusion step. It is the most directly comparable prior work to ours and serves as our main multichannel baseline in Section IV. Two concurrent variants apply similar DPS guidance to refine the outputs of pretrained discriminative models [20], [21], and thereby assume the availability of strong supervised discriminative front-end estimates.

Our work differs from ArrayDPS in the choice of observation model rather than in the diffusion machinery. By adopting the LGM, we obtain an explicit noise source, probabilistic inference of the spatial parameters, and a gradient coupling between the diffusion state and the spatial–noise model as consequences of a single design decision. MNMF-DPS thereby inherits the analytically tractable multichannel likelihood from the LGM-based BSS line, the explicit noise modeling from the single-channel diffusion line, and the diffusion-based speech prior from the multichannel DPS line, integrating all three within a single unsupervised inference loop.

III. Proposed Method

MNMF-DPS realizes the LGM-based design motivated in Section I by coupling a FastMNMF2 jointly-diagonalizable spatial model, an NMF-based noise PSD, and a UDiffSE diffusion speech prior within a single LGM likelihood. As shown in Fig. 2 and Algorithm 1, inference has three stages: FastMNMF2 warm-up, DPS refinement, and final refit with multichannel Wiener filtering (MWF). Within the DPS refinement stage, Phase A freezes the FastMNMF2 state for the first D reverse steps, and Phase B updates the spatial, calibration, and noise parameters by MM using the evolving diffusion-derived speech PSD. The same model and procedure cover both enhancement and noisy two-speaker separation by adjusting only the source counts N_{s} and N_{n}.

A. Problem specification

We consider multichannel speech enhancement and noisy speech separation in the STFT domain under the LGM. Let $\mathbf{s}_{nft} \in \mathbb{C}^M$ denote the M-channel image of source $n \in [1, N]$ at frequency bin $f \in [1, F]$ and time frame $t \in [1, T]$, with channel index $m \in [1, M]$ and $N = N_{\mathrm{s}} + N_{\mathrm{n}}$ sources, and let $(N_{\mathrm{s}}, N_{\mathrm{n}}) = (1, 1)$ for enhancement and $(2, 1)$ for noisy two-speaker separation. Each source image follows a zero-mean complex Gaussian

$$\mathbf{s}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0},\, \lambda_{nft}\, \mathbf{R}_{nf}), \tag{1}$$

where $\lambda_{nft} \in \mathbb{R}_+$ is the time-varying PSD and $\mathbf{R}_{nf} \in \mathbb{S}_+^M$ is the time-invariant SCM. The observed STFT mixture $\mathbf{X} \triangleq$

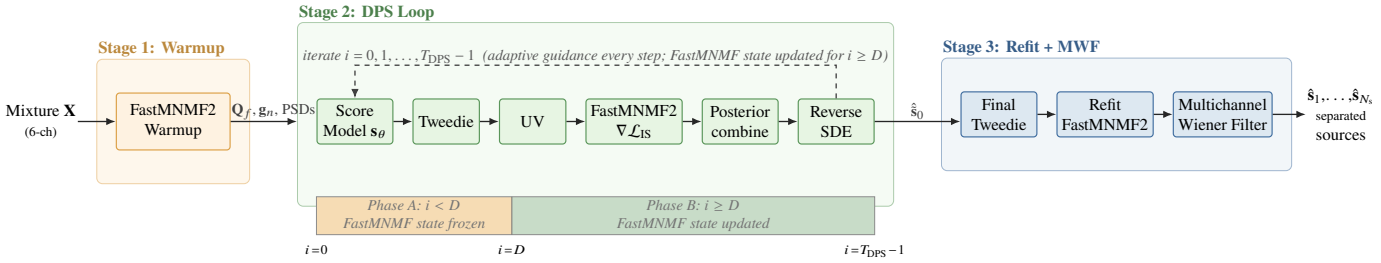


Fig. 2. Three-stage MNMF-DPS inference pipeline. Stage 1 (warm-up) initializes the FastMNMF2 spatial parameters and noise NMF factors from the observation. Stage 2 is a DPS loop with adaptive prior–likelihood guidance at every reverse SDE step; the band under the loop indicates that the FastMNMF state ($\mathbf{Q}$, $\mathbf{G}$, $\mathbf{U}_{\text{s}}$, $\mathbf{V}_{\text{s}}$, $\mathbf{W}_{\text{n}}$, $\mathbf{H}_{\text{n}}$) is frozen at warm-up values for the first D steps (Phase A) and jointly updated thereafter (Phase B). Stage 3 refits FastMNMF2 with the converged diffusion-derived PSD and applies the multichannel Wiener filter to produce $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{N_{\text{s}}}$.

$\{\mathbf{x}_{ft}\}_{f,t}$ with $\mathbf{x}_{ft} = \sum_n \mathbf{s}_{nft}$ is again a zero-mean complex Gaussian with covariance $\sum_n \lambda_{nft}\,\mathbf{R}_{nf}$, and each source image is recovered by MWF

$$\hat{\mathbf{s}}_{nft} = \lambda_{nft}\,\mathbf{R}_{nf}\left(\sum_{n'=1}^{N} \lambda_{n'ft}\,\mathbf{R}_{n'f}\right)^{-1}\mathbf{x}_{ft}. \tag{2}$$

Inference therefore reduces to estimating the PSDs and SCMs of all sources from the observed mixture.

B. Model formulation

The SCMs of all sources follow the jointly-diagonalizable parameterization of FastMNMF2 [8]:

$$\mathbf{R}_{nf} = \mathbf{Q}_f^{-1}\,\mathrm{diag}(\mathbf{g}_n)\,\mathbf{Q}_f^{-\mathsf{H}}, \tag{3}$$

where $\mathbf{Q}_f \in \mathbb{C}^{M\times M}$ is a frequency-dependent diagonalizer shared across sources and $\mathbf{g}_n \in \mathbb{R}_+^M$ is a source-specific frequency-independent spatial vector. Sharing $\mathbf{Q}_f$ reduces spatial parameters and stabilizes inference while still admitting full-rank SCMs that represent diffuse noise and non-pointwise sources.

For each noise source $n > N_{\text{s}}$, the PSD is parameterized by NMF with K_n basis vectors

$$\lambda_{nft} = \textstyle\sum_{k=1}^{K_n} w_{nfk}\,h_{nkt}, \tag{4}$$

with nonnegative w_{nfk} and h_{nkt}. The noise PSD is thereby estimated adaptively within the likelihood and remains agnostic to unseen noise types.

For each speech source $n \leq N_{\text{s}}$, the PSD is provided by a pretrained UDiffSE [6] score model operating in a compressed complex STFT domain. The single-channel speech coefficient $s_{nft} \in \mathbb{C}$ is compressed as

$$\tilde{s}_{nft} = \beta\,|s_{nft}|^{\alpha}\,e^{\jmath\angle s_{nft}}, \tag{5}$$

with $(\alpha, \beta) = (0.5, 0.15)$, which compresses the dynamic range while preserving phase. The DPS loop in Section III-C returns a clean compressed estimate $\hat{\tilde{s}}_{nft}$, from which the raw PSD is recovered as $\hat{\lambda}_{nft} = (|\hat{\tilde{s}}_{nft}|/\beta)^{2/\alpha}$ and calibrated against the observed mixture

$$\lambda_{nft} = u_{nf}\,\hat{\lambda}_{nft}\,v_{nt}, \tag{6}$$

where $u_{nf}, v_{nt} \geq 0$ absorb the scale ambiguity of the diffusion-derived PSD and let the spatial model refine it without retraining the score network.

C. LGM-based diffusion posterior sampling

Under the LGM, with $\mathbf{Q}_f$ and $\mathbf{g}_n$ fixed, the negative log-likelihood reduces to an Itakura–Saito divergence

$$\mathcal{L}_{\text{IS}} = \sum_{f,t,m}\left[\frac{|z_{ftm}|^2}{Y_{ftm}} + \log Y_{ftm}\right], \tag{7}$$

where $z_{ftm} = \{\mathbf{Q}_f\mathbf{x}_{ft}\}_m$ is the demixed signal and $Y_{ftm} = \sum_{n=1}^{N}\lambda_{nft}\,g_{nm}$ with $g_{nm} \triangleq \{\mathbf{g}_n\}_m$ denoting the m-th element of $\mathbf{g}_n$. The posterior score driving the reverse SDE combines the prior score $\mathbf{s}_\theta$ with this likelihood gradient

$$\begin{aligned}\mathbf{s}_{\text{post}} &= \mathbf{s}_\theta - \zeta(\tau)\,\nabla\mathcal{L}_{\text{IS}},\\ \zeta(\tau) &= \frac{\xi}{\|\nabla\mathcal{L}_{\text{IS}}\|_2/\sqrt{N_{\text{el}}}\,\sigma(\tau) + \varepsilon},\end{aligned} \tag{8}$$

where $\xi > 0$ is a constant guidance strength, N_{el} is the number of elements in $\nabla\mathcal{L}_{\text{IS}}$, $\sigma(\tau)$ is the noise scale of the OUVE diffusion process, and $\varepsilon > 0$ is a numerical safeguard, with the adaptive weight $\zeta(\tau)$ keeping the prior and likelihood gradients on comparable scales across the reverse trajectory, in the spirit of DPS step-size normalization for inverse problems [4].

The likelihood gradient back-propagates through inverse compression and calibration to the diffusion state, so each reverse-SDE step is guided by the current spatial and noise model. At each reverse step, $\nabla\mathcal{L}_{\text{IS}}$ is evaluated with $\mathbf{Q}_f$, $\mathbf{g}_n$, w_{nfk}, h_{nkt} at their current values and u_{nf}, v_{nt} detached, and the gradient flows through $\hat{\lambda}_{nft}$ to $\tilde{\mathbf{s}}_{n,\tau_i}$, so the diffusion update sees the spatial and noise model through the likelihood. The diffusion state is then updated by the Euler–Maruyama discretization of the reverse SDE

$$\begin{aligned}\tilde{\mathbf{s}}_{n,\tau_i-\Delta\tau} = \tilde{\mathbf{s}}_{n,\tau_i} &- \left[\mathbf{f}(\tilde{\mathbf{s}}_{n,\tau_i},\tau_i) - \sigma(\tau_i)^2\,\mathbf{s}_{\text{post}}\right]\Delta\tau\\ &+ \sigma(\tau_i)\sqrt{\Delta\tau}\,\mathbf{z},\end{aligned} \tag{9}$$

where $\mathbf{f} = -\theta\,\tilde{\mathbf{s}}_{n,\tau_i}$ is the OUVE drift with drift coefficient $\theta > 0$, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

D. Inference procedure

Inference jointly estimates the spatial parameters $\mathbf{Q} \triangleq \{\mathbf{Q}_f\}$, $\mathbf{G} \triangleq \{\mathbf{g}_n\}$, the noise NMF parameters $\mathbf{W}_{\text{n}}, \mathbf{H}_{\text{n}}$, the speech calibration parameters $\mathbf{U}_{\text{s}}, \mathbf{V}_{\text{s}}$, and the diffusion-derived speech PSDs $\hat{\mathbf{\Lambda}}_{\text{s}}$. The full procedure is outlined in Algorithm 1.

The warm-up runs FastMNMF2 for I_{init} MM iterations on the observed mixture to obtain a spatially informed initialization of $\mathbf{Q}$, $\mathbf{G}$, $\mathbf{W}_{\text{n}}$, $\mathbf{H}_{\text{n}}$, and the warm-up speech images $\hat{\mathbf{s}}_{nft}$ via MWF (2). The calibration parameters are initialized to $u_{nf} = 1/F$ and $v_{nt} = 1$. For each speech source, the reference-channel signal is taken from $\hat{\mathbf{s}}_{nft}$ and compressed via (5) to form the clean compressed array $\tilde{\mathbf{s}}_{n,0}$, which is then noised by the SDEdit procedure [22] as

$$\tilde{\mathbf{s}}_{n,\tau_{\max}} = \tilde{\mathbf{s}}_{n,0} + \sigma(\tau_{\max})\,\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

so that the reverse diffusion starts from a spatially separated estimate rather than pure Gaussian noise.

The DPS refinement loop performs T_{DPS} reverse-SDE steps. To prevent the noisy early diffusion state from disrupting the warm-up spatial structure, Phase A holds $\mathbf{Q}$, $\mathbf{G}$, $\mathbf{W}_{\text{n}}$, and $\mathbf{H}_{\text{n}}$ at their warm-up values for the first D reverse steps, with $\mathbf{U}_{\text{s}}$ and $\mathbf{V}_{\text{s}}$ kept at their normalization initialization, while the likelihood gradient still guides the reverse SDE. Phase B then runs a small number of FastMNMF2 MM iterations before each reverse-SDE update to update $\mathbf{Q}$, $\mathbf{G}$, $\mathbf{U}_{\text{s}}$, $\mathbf{V}_{\text{s}}$, $\mathbf{W}_{\text{n}}$, and $\mathbf{H}_{\text{n}}$ so that the spatial and noise models track the evolving speech PSD. At step i with diffusion time τ_i, Tweedie's formula for the OUVE process gives a clean compressed estimate

$$\hat{\tilde{\mathbf{s}}}_{n,0} = \frac{\tilde{\mathbf{s}}_{n,\tau_i} + \sigma(\tau_i)^2\,\mathbf{s}_\theta(\tilde{\mathbf{s}}_{n,\tau_i}, \tau_i)}{e^{-\theta\tau_i}}. \tag{11}$$

The raw PSD is recovered by inverting (5) and calibrated via (6). When $i \geq D$, the Phase B MM iterations refresh the calibrated speech PSDs and the noise PSDs before the likelihood gradient is evaluated as in Section III-C. The diffusion state is then updated by (9).

A final refit aligns the spatial and noise model with the converged diffusion estimate and produces the separated sources. Let $\hat{\lambda}_{\text{final}}$ denote the speech PSD derived from the last Tweedie estimate via the inverse of (5). FastMNMF2 then runs for I_{refit} additional MM iterations to align the spatial and noise parameters with this converged PSD, after which the MWF in (2) produces the separated speech images $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{N_{\text{s}}}$.

Algorithm 1 MNMF-DPS inference

Require: Mixture $\mathbf{X}$, score model $\mathbf{s}_\theta$, steps T_{DPS}, freeze length D, guidance strength ξ
Ensure: $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{N_{\text{s}}}$

Warm-up
Run FastMNMF2 for I_{init} iterations on $\mathbf{X}$ and apply MWF, obtaining $\mathbf{Q}$, $\mathbf{G}$, $\mathbf{W}_{\text{n}}$, $\mathbf{H}_{\text{n}}$, and warm-up speech images $\hat{\mathbf{s}}_{nft}$, $n \leq N_{\text{s}}$
Initialize $u_{nf} = 1/F$ and $v_{nt} = 1$
$\tilde{\mathbf{s}}_{n,0} \leftarrow$ reference channel of $\hat{\mathbf{s}}_{nft}$ compressed via (5)
$\tilde{\mathbf{s}}_{n,\tau_{\max}} \leftarrow$ SDEdit noising of $\tilde{\mathbf{s}}_{n,0}$ via (10)
DPS refinement loop
for $i = 0$ **to** $T_{\text{DPS}} - 1$ with decreasing diffusion times τ_i from $\tau_{\max}$ to ϵ **do**
  $\hat{\tilde{\mathbf{s}}}_{n,0} \leftarrow$ Tweedie (11)
  $\hat{\lambda}_{nft} \leftarrow$ inverse compression of $\hat{\tilde{\mathbf{s}}}_{n,0}$ via (5), then λ_{nft} from (6)
  if $i \geq D$ **then**
    Update $\mathbf{Q}$, $\mathbf{G}$, $\mathbf{U}_{\text{s}}$, $\mathbf{V}_{\text{s}}$, $\mathbf{W}_{\text{n}}$, $\mathbf{H}_{\text{n}}$ via FastMNMF2 MM iterations
  end if
  Compute $\nabla\mathcal{L}_{\text{IS}}$ and update $\tilde{\mathbf{s}}_{n,\tau_i}$ via (9) with $\mathbf{s}_{\text{post}}$ from (8)
end for
Final refit and beamforming
Refit FastMNMF2 on $\hat{\lambda}_{\text{final}}$ for I_{refit} MM iterations
Apply MWF (2) to obtain $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{N_{\text{s}}}$

IV. Evaluation

We evaluate MNMF-DPS on single-speaker enhancement ($N_{\text{s}} = N_{\text{n}} = 1$) and noisy two-speaker separation ($N_{\text{s}} = 2$, $N_{\text{n}} = 1$), which share the same model and inference procedure. The FastMNMF2 warm-up serves as a baseline that swaps the diffusion prior for an NMF speech model under the same LGM, and thereby isolates the contribution of the diffusion prior. ArrayDPS [7] serves as a second baseline that swaps the LGM for a convolutive mixture under the same class of diffusion prior, and thereby isolates the contribution of the observation model.

A. Experimental conditions

All recordings are six-channel at 16 kHz with matched preprocessing and metrics across tasks. For enhancement, we use 100 utterances from CHiME-3 [23], 25 per noise type (BUS, CAF, PED, STR), each with a single speaker in real diffuse noise. For separation, we use 900 spatialized WSJ0-2mix mixtures [24] mixed with CHiME-3 diffuse noise at 10 dB SNR, partitioned by RT60 into low $[0.2, 0.4)$ s, mid $[0.4, 0.6)$ s, and high $[0.6, 0.8)$ s (300 per regime).

Performance is measured by scale-invariant signal-to-distortion ratio (SI-SDR) [25] against early reference signals (direct path plus reflections up to 50 ms), computed in the linear time domain after inverse compression and the MWF and evaluated per source with utterance-level permutation-invariant assignment. We additionally report PESQ [26] for enhancement.

For enhancement, we add an oracle MVDR with clean-reference SCMs as an upper bound and a supervised single-channel SGMSE+ [5] trained on CHiME-3 noisy–clean pairs as an absolute reference. Classical IVA-family methods [9], [11] are omitted because they do not explicitly model full-rank diffuse noise, and supervised multichannel methods (e.g., neural beamformers [27], multichannel diffusion models [28]) are out of scope as they require paired training data.

MNMF-DPS uses $T_{\text{DPS}} = 400$ reverse-SDE steps, $I_{\text{init}} = 100$ warm-up iterations, freeze length $D = 170$, guidance strength $\xi = 0.5$, 3 MM iterations per DPS step during Phase B, 3 refit MM iterations, and $K_{\text{n}} = 16$ noise NMF basis vectors. The FastMNMF2 warm-up additionally uses $K_{\text{s}} = 32$ speech NMF basis vectors, which the diffusion prior then replaces during the DPS loop. The UDiffSE score model is trained on WSJ0 (`si_tr_s`) at 16 kHz. For fair comparison, the ArrayDPS score model is retrained from scratch on the same WSJ0 set at 16 kHz with the publicly released code, since only an

TABLE I
ENHANCEMENT RESULTS ON CHIME-3 ($n=100$). †ORACLE MVDR USES CLEAN-REFERENCE SCMS. ‡SUPERVISED SINGLE-CHANNEL REFERENCE.

Method	SI-SDR (dB) BUS	CAF	PED	STR	Avg.	PESQ
Unprocessed	3.35	4.27	0.71	6.37	3.67	—
Oracle MVDR†	14.09	12.30	10.24	12.73	12.34	—
SGMSE+‡	13.07	9.95	6.46	13.91	10.85	—
FastMNMF2 warm-up	6.99	5.61	3.24	9.18	6.26	—
ArrayDPS	3.40	5.07	1.50	7.50	4.37	1.38
MNMF-DPS	9.72	10.41	5.84	7.24	**8.30**	**1.84**
MNMF-DPS, $N_n=0$	2.80	4.27	0.71	6.37	3.55	—

8 kHz checkpoint is publicly available. Both DPS methods use the same $T_{\text{DPS}} = 400$ reverse-SDE steps and matched WPE preprocessing $(1024/256/1)$. ArrayDPS does not include WPE in its public pipeline; we add the same WPE for fair comparison and verified on a held-out subset that it also benefits ArrayDPS.

B. Enhancement results

MNMF-DPS achieves 8.30 dB SI-SDR and 1.84 PESQ on CHiME-3, outperforming ArrayDPS by +3.93 dB SI-SDR (paired t-test $p < 0.001$, 95% CI $[2.31, 5.54]$) and +0.46 PESQ ($p < 0.001$, 95% CI $[0.38, 0.55]$), and the FastMNMF2 warm-up by +2.04 dB SI-SDR ($p = 0.018$). The two gains support the LGM-based design at both axes of comparison, as detailed below and broken down in Table I. SGMSE+ is supervised and single-channel, hence not directly comparable to the blind systems; it serves only as an absolute reference.

The +2.04 dB gain over the FastMNMF2 warm-up isolates the diffusion speech prior, because the spatial and noise models are identical between the two methods. The gain therefore reflects the richer speech distribution that the diffusion model captures over an NMF parameterization.

The +3.93 dB gain over ArrayDPS isolates the LGM-based observation model, because both methods use a clean-speech score trained on the same WSJ0 data. ArrayDPS improves SI-SDR by only +0.70 dB over the unprocessed mixture, while MNMF-DPS improves by +4.63 dB. Under heavy diffuse noise, the convolutive-mixture likelihood routes residual noise energy into the speech estimate through the diffusion prior, whereas the LGM allocates that energy to an explicit noise source. The gap is largest in BUS and CAF (+6.32 and +5.34 dB over ArrayDPS), where non-stationary noise benefits most from the noise-aware likelihood.

Removing the noise source from MNMF-DPS ($N_n = 1 \to 0$, other settings unchanged) collapses SI-SDR from 8.30 dB to 3.55 dB (mean of 99 finite estimates, with one sample that diverged), which matches the unprocessed level of 3.67 dB. The diffusion prior alone cannot disentangle speech from diffuse noise. Without an explicit noise source in the likelihood, the model has no place to allocate the noise energy, confirming that the LGM-based noise modeling, not the diffusion prior, performs the noise decoupling.

TABLE II
NOISY TWO-SPEAKER SEPARATION SI-SDR (DB) ON SPATIALIZED WSJ0-2MIX WITH CHIME-3 NOISE AT 10 DB SNR, $n=300$ PER RT60 REGIME, MATCHED WPE. EACH CELL IS MEAN / STD / MEDIAN. ♭ABLATION ROW.

Method	Low $[0.2, 0.4)$ s	Mid $[0.4, 0.6)$ s	High $[0.6, 0.8)$ s
FastMNMF2 warm-up	7.27/ − /8.48	4.49/ − /5.66	2.96/ − /3.80
ArrayDPS	7.51/2.42/7.88	5.48/2.43/5.85	3.60/2.52/3.91
MNMF-DPS	7.95/4.67/**8.91**	5.01/4.70/**6.32**	3.59/4.12/**4.62**
MNMF-DPS, $N_n=0$♭	3.21/ − /5.16	1.09/ − /3.33	0.79/ − /1.64

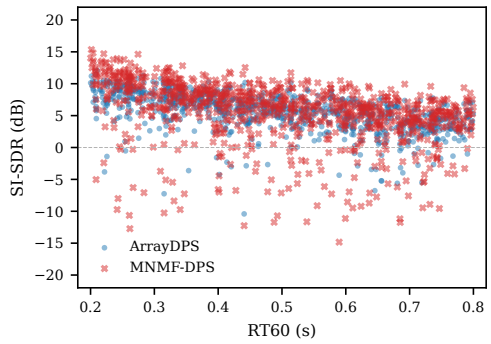


Fig. 3. Per-sample SI-SDR against the sampled RT60 for the 900 noisy two-speaker mixtures. MNMF-DPS estimates cluster above ArrayDPS for most mixtures but exhibit a heavier left tail.

C. Noisy two-speaker separation results

MNMF-DPS and ArrayDPS achieve essentially equal mean SI-SDR (5.52 vs 5.53 dB) on the 900 separation samples, yet MNMF-DPS dominates ArrayDPS on a per-sample basis. It wins in $627/900$ (69.7%) pairs with a pooled median paired improvement of +0.93 dB (paired Wilcoxon signed-rank $p < 10^{-13}$). MNMF-DPS also achieves the higher median SI-SDR in every RT60 regime by +0.5 to +1.0 dB, with per-regime Wilcoxon tests confirming dominance in each (low $p = 5\times10^{-9}$, mid $p = 3\times10^{-3}$, high $p = 1.5\times10^{-6}$). Table II summarizes per-regime statistics and Fig. 3 plots the per-sample SI-SDR against RT60.

These separation results use the same implementation and hyperparameters as for enhancement, with only the source counts changed from $(1, 1)$ to $(2, 1)$. ArrayDPS, by contrast, gains only +0.70 dB over unprocessed on enhancement, consistent with the absence of an explicit noise source in its likelihood, and therefore cannot match MNMF-DPS in the unified setting. The unification is a direct consequence of the LGM, where speech and noise sources enter the likelihood symmetrically and the source counts (N_s, N_n) become the only task-specific parameters.

The mean tie reflects a heavier left tail rather than a worse central tendency, with MNMF-DPS exhibiting standard deviation 4.5 dB against ArrayDPS's 2.5 dB and 12% of estimates below 0 dB SI-SDR against 4%. This pattern follows from the design difference between the two methods. ArrayDPS updates relative RIRs as point estimates decoupled from the diffusion state, which yields locally conservative trajectories, whereas MNMF-DPS treats the spatial and noise parameters as free like-

lihood variables, which produces larger gains when the warm-up is informative but harder recovery when it is not.

Decomposing the failures, 99 of the 105 MNMF-DPS samples below 0 dB (94.3%) already had a negative warm-up SI-SDR, so the DPS loop inherits failures rather than producing them. Across all samples, the DPS loop improves the warm-up estimate by $+0.61$ dB on average and by $+0.15$ dB on the failure subset, so the loop is not the source of the left tail. Improving the warm-up initialization, e.g., through a pretrained NMF speech basis or a neural initializer, is therefore the most direct route to closing the mean gap and is orthogonal to the LGM-based DPS design evaluated here.

V. Conclusion

We presented MNMF-DPS, an LGM-based framework combining FastMNMF2 spatial updates, an NMF noise PSD, and a diffusion speech prior in one inference procedure that yields explicit noise modeling, likelihood-based spatial-parameter estimation, and gradient coupling between the diffusion state and the spatial–noise model. On CHiME-3, MNMF-DPS gains $+3.93$ dB SI-SDR over ArrayDPS (isolating the LGM observation model) and $+2.04$ dB over FastMNMF2 warm-up (isolating the diffusion prior); an $N_{\mathrm{n}}=0$ ablation falls to the unprocessed level, confirming explicit noise modeling as the source of noise decoupling. On 900 noisy two-speaker mixtures, changing only the source counts gives paired dominance over ArrayDPS in 69.7% of samples and higher median SI-SDR in every RT60 regime. The heavier left tail is tied to FastMNMF2 warm-up failures rather than the DPS loop, making better warm-up initializers, joint dereverberation, and extension beyond two speakers the direct next steps.

Acknowledgment

This work was partially supported by JST FOREST No. JPMJFR2270 and JSPS KAKENHI Nos. 24H00742, 25H01142, 25K22841, and 26K14896.

References

[1] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[2] H. Sawada et al., "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, e12, 2019.

[3] Y. Song et al., "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[4] H. Chung et al., "Diffusion posterior sampling for general noisy inverse problems," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[5] J. Richter et al., "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[6] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised speech enhancement with diffusion-based generative models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 12 481–12 485.

[7] Z. Xu et al., "ArrayDPS: Unsupervised blind speech separation with a diffusion prior," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2025, pp. 69 160–69 188.

[8] K. Sekiguchi et al., "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2610–2625, 2020.

[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2011, pp. 189–192.

[10] H. Sawada et al., "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[11] D. Kitamura et al., "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[12] K. Sekiguchi et al., "Autoregressive moving average jointly-diagonalizable spatial covariance analysis for joint source separation and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2368–2382, 2022.

[13] H. Kameoka et al., "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, 2019.

[14] L. Li, H. Kameoka, and S. Makino, "FastMVAE2: On improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 96–110, 2023.

[15] K. Sekiguchi et al., "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2197–2212, 2019.

[16] A. A. Nugraha et al., "Flow-based fast multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 501–505.

[17] Y. Bando et al., "Neural full-rank spatial covariance analysis for blind source separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1670–1674, 2021.

[18] J.-M. Lemercier et al., "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2724–2737, 2023.

[19] J.-E. Ayilo et al., "Diffusion-based unsupervised audio-visual speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2025, pp. 1–5.

[20] Z. Xu et al., "ArrayDPS-Refine: Generative refinement of discriminative multi-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2026.

[21] Z. Xu et al., *Unified diffusion refinement for multi-channel speech enhancement and separation*, arXiv:2603.24810, 2026.

[22] C. Meng et al., "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[23] J. Barker et al., "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Underst. (ASRU)*, 2015, pp. 504–511.

[24] L. Drude et al., *SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition*, arXiv:1910.13934, 2019.

[25] J. Le Roux et al., "SDR – half-baked or well done?" In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 626–630.

[26] A. W. Rix et al., "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 2001, pp. 749–752.

[27] Y. Fujita et al., "Run-time adaptation of neural beamforming for robust speech dereverberation and denoising," in *Proc. APSIPA Annu. Summit Conf. (APSIPA ASC)*, 2024, pp. 1–6.

[28] R. Kimura et al., "Diffusion model-based MIMO speech denoising and dereverberation," in *Proc. HSCMA*, 2024, pp. 455–459.