SITUATED INFERENCE: ADAPTING FROZEN GENERATIVE SPEECH ENHANCEMENT UNDER ACOUSTIC DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow matching models typically enhance speech using fixed inference strategies that may become unsuitable under acoustic distribution shifts in real-world deployment. Our cross-condition analysis shows that the advantage of an inference strategy is not consistently preserved across acoustic conditions, even when the model weights and solver update rule remain fixed. This finding motivates Situated Inference, which specializes the inference strategy of a frozen generative enhancer to its situation, the target acoustic distribution under which it is deployed. Using paired target speech, Situated Inference evaluates a structured inference space that jointly varies observation processing, the reverse starting point, and the solver step count, thereby constructing a target-specific quality and computation profile. From this profile, it selects one strategy either to maximize enhancement quality or to minimize computation subject to the default quality, then fixes that strategy for subsequent inputs. Across four generative backbones and three acoustic conditions, Situated Inference improves perceptual speech quality under quality specialization and reduces the number of function evaluations under efficiency specialization. Deployment on the Ameca humanoid further illustrates the efficiency advantage of Situated Inference in practice, with demonstrations available at `https://anonymous.4open.science/w/situated-inference-demo-0761/`.

1 INTRODUCTION

Diffusion and flow matching speech enhancement (SE) models restore noisy speech through iterative conditional generation, using the noisy mixture as conditioning information to guide the sampling process toward a clean speech estimate (Lee et al., 2025b; Richter et al., 2023; Lee et al., 2025a). These pretrained generative enhancers are typically deployed with a fixed inference strategy across varying acoustic conditions. Changes in talkers, noise, and recording conditions alter the clean speech or its degradation and can induce *acoustic distribution shift*, where the joint distribution of noisy observations and clean speech differs between training and deployment (Liang et al., 2025; Gonzalez et al., 2023; 2024b). Because the noisy observation conditions the generative dynamics and contributes to initialization, changes in its distribution can alter the sampling trajectories, raising the question of whether the same inference strategy remains suitable for a given target acoustic condition.

Existing approaches address acoustic distribution shift by adapting model representations or parameters using target speech (Raichle et al., 2026b;a), which requires updating the deployed model. Beyond model adaptation, generative inference can be modified through the sampler (Karras et al., 2022; Gonzalez et al., 2024a), solver (Shaul et al., 2024), or sampling schedule (Sabour et al., 2024) to improve the quality–cost trade-off, but these designs are fixed for a given model and do not account for the target acoustic condition. Input-adaptive methods further adjust sampling depth or trajectories according to individual observations (Chhaglani et al., 2026; Fabian et al., 2024), making each decision from a single noisy observation without using the acoustic condition shared across inputs. These motivate examining whether the inference strategy of a frozen enhancer should be adapted to the target acoustic condition, and across which inference components.

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.

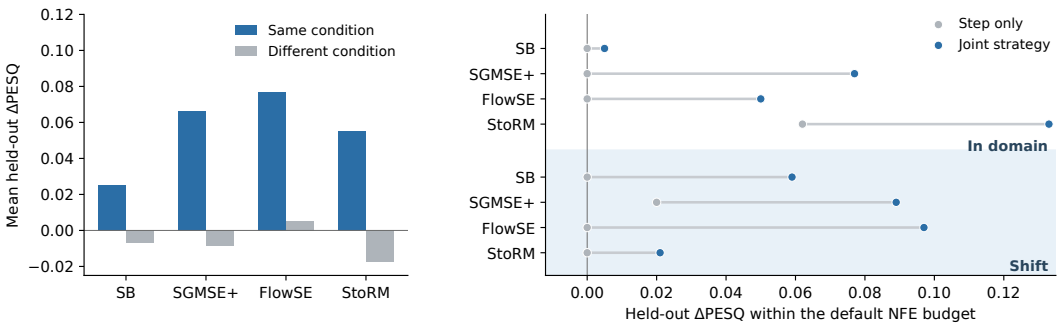

Figure 1: Motivating diagnostics for condition-level inference adaptation, with PESQ gains measured relative to the default inference strategy. **(a) Inference Strategy Transferability Test (ISTT).** Strategies selected under one acoustic condition achieve larger gains when evaluated under the same condition, while their benefits are not consistently preserved under different conditions, showing that strategy effectiveness depends on the acoustic condition and that a fixed strategy across conditions is not reliably suitable. **(b) Joint Adaptation Test.** Under the default number of function evaluations (NFE), jointly adapting observation processing, the reverse starting point, and the solver step count yields larger gains than step count adaptation alone, showing that multiple inference components contribute to target-condition specialization, whereas existing designs fix them for a given model.

To address this question, we introduce the Inference Strategy Transferability Test (ISTT), which selects a strategy on target speech from one acoustic condition and evaluates it on held-out speech from the same and different conditions. As shown in Fig. 1(a), strategies evaluated under the same condition achieve positive average gains in held-out Perceptual Evaluation of Speech Quality (PESQ) over the default strategy across all four backbones, and these gains diminish under different conditions and become negative for three backbones. These results show the limited transferability of inference strategies across acoustic conditions, *motivating situated inference adaptation for generative SE*.

Motivated by this conditional SE insight, we introduce Situated Inference, a training-free framework that adapts the inference strategy of a frozen generative enhancer to its situation, the target acoustic distribution under which it is deployed. We formulate inference adaptation under acoustic distribution shift as a target-specific strategy optimization problem, aiming to specialize the inference behavior of a frozen generative model. We therefore construct a structured inference space to capture the execution choices available at inference time, spanning observation processing, the reverse starting point, and the solver step count, which together determine the acoustic evidence, the integration interval, and the numerical resolution of inference. Our Joint Adaptation Test in Fig. 1(b) shows that adapting these three components jointly yields greater enhancement quality gains than step-only adaptation under the same default number of function evaluations (NFE). Within this structured inference space, we use representative paired target speech to estimate a performance profile that characterizes enhancement quality and inference cost under the target acoustic distribution, capturing how the acoustic condition affects both the quality attained by each strategy and the computation required to reach a given quality. Situated Inference uses this profile to select a strategy under either a quality objective or an efficiency objective subject to a quality requirement. The selected strategy remains fixed for subsequent inputs from the same acoustic condition. Extensive experiments across four generative backbones and three acoustic conditions demonstrate that Situated Inference improves perceptual speech quality and reduces NFE under separate quality and efficiency objectives. We further deploy Situated Inference on the Ameca humanoid robot to illustrate its efficiency advantage in humanlike humanoid communication, where response latency matters.

Our contributions can be summarized as:

- We analyze the transferability of inference strategies across acoustic conditions through the ISTT, showing that a strategy specialized under one acoustic condition does not consistently retain its benefit for another.
- We construct a structured inference space for inference-time adaptation, which turns the execution choices of a frozen generative enhancer into a joint adaptation interface shared across the evaluated generative model families.

- Building on this space, we propose Situated Inference, a training-free framework that specializes the inference strategy of a frozen enhancer to a target acoustic condition under quality or efficiency objectives derived from one performance profile.

2 Speech Enhancement under Acoustic Distribution Shift

Generative SE models produce clean speech estimates by numerically integrating learned conditional dynamics from an initial state constructed using the noisy observation. Specifically, a pretrained enhancer is paired with a fixed inference strategy that determines the observation supplied as conditioning, the reverse starting point, and the solver step count. The suitability of this strategy can change when the acoustic conditions governing these observations differ between training and deployment.

2.1 Conditional generative dynamics and numerical solvers

Diffusion (Richter et al., 2023) and flow matching (Lee et al., 2025b) SE models realize conditional generation through distinct dynamics. Let $\mathbf{x}$ denote clean speech, $\mathbf{y}$ the noisy observation, and $\mathbf{x}_t$ the generative state at continuous time $t \in [0, 1]$. We adopt a common time convention in which $t = 0$ corresponds to the clean speech endpoint and $t = 1$ to an initialization constructed from $\mathbf{y}$, with enhancement proceeding from $t = 1$ toward $t = 0$. Although these model families differ in their path construction and training objectives, each provides a conditional field that governs the evolution of $\mathbf{x}_t$ conditioned on $\mathbf{y}$ (Song et al., 2021; Lipman et al., 2023).

Under this time convention, we express the reverse inference dynamics conditioned on the noisy observation $\mathbf{y}$ in the common form:

$$\mathrm{d}\mathbf{x}_t = F_\theta(\mathbf{x}_t, \mathbf{y}, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\overline{\mathbf{w}}_t, \tag{1}$$

where F_θ denotes the conditional field with pretrained parameters θ, $g(t)$ controls the stochastic component, and $\overline{\mathbf{w}}_t$ is a reverse-time Wiener process, with $\mathrm{d}t < 0$ as inference proceeds toward $t = 0$. For score-based diffusion, Eq. (1) corresponds to the reverse SDE of the forward process, where F_θ combines the forward drift with a conditional score $s_\theta(\mathbf{x}_t, \mathbf{y}, t) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{y})$ (Song et al., 2021). For flow matching, F_θ represents the learned conditional velocity expressed under our time convention, with $g(t) = 0$ (Lipman et al., 2023). Throughout inference adaptation, the conditional dynamics remain fixed, including the model parameters and the functions F_θ and $g(t)$ of Eq. (1).

Numerically integrating these conditional dynamics requires a solver that evaluates F_θ and advances the generative state over discrete time steps. We denote the solver update by Φ:

$$\mathbf{x}_{t_{n-1}} = \Phi(\mathbf{x}_{t_n}, \mathbf{y}, F_\theta(\mathbf{x}_{t_n}, \mathbf{y}, t_n), t_n, t_{n-1}), \qquad t_N = t_\mathrm{s} > \cdots > t_0 \approx 0. \tag{2}$$

Here, Φ advances the generative state from $\mathbf{x}_{t_n}$ to $\mathbf{x}_{t_{n-1}}$ using the conditional field and any stochastic increment or corrector substep required by the underlying solver, while t_n denotes the n-th discrete time point along the reverse path. The upper endpoint t_s specifies the reverse starting point, with the corresponding initialization constructed according to the initialization rule of the model, whereas N determines the number of solver updates from t_s to t_0. Together, t_s and N control the extent and numerical resolution of the reverse inference process.

The inference strategy also determines the acoustic observation supplied to the model. Observation processing modifies the signal used both to condition F_θ and to construct the initial state, without altering the pretrained parameters. Together, the observation processing operation m, the reverse starting point t_s, and the solver step count N define an inference strategy $a = (m, t_\mathrm{s}, N)$.

2.2 Inference adaptation under acoustic distribution shift

Acoustic distribution shift changes the speech distribution on which a given inference strategy operates. We represent each acoustic condition by a joint distribution $\mathcal{P}$ over clean speech $\mathbf{x}$ and noisy observations $\mathbf{y}$, capturing the acoustic variability of the situation in which the enhancer is deployed. Because $\mathbf{y}$ conditions the dynamics and enters the solver update in Eqs. (1) and (2), changes in its distribution can alter the distribution of states encountered along the reverse path. The expected enhancement quality of a given strategy therefore depends on the joint distribution of noisy observations and clean references, and can be written as:

$$Q_\mathcal{P}(a) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[q(G_\theta(\mathbf{y}; a), \mathbf{x})\right], \tag{3}$$

where a denotes the inference strategy, $G_\theta(\mathbf{y}; a)$ the enhanced speech estimate produced under strategy a, and q an enhancement quality measure, with the expectation also taken over sampler randomness for stochastic samplers. Even with θ and a held fixed, $Q_\mathcal{P}(a)$ can vary with the acoustic distribution $\mathcal{P}$, altering which strategy performs best.

In addition to expected enhancement quality, deployment requires computational efficiency, which we characterize by the inference cost $C(a)$. We measure this cost by the NFE, counting the neural network evaluations required by strategy a. Let $\mathcal{A}$ denote the set of inference strategies supported by a frozen backbone, which Sec. 3.1 organizes into a structured inference space. The pairs $\{(C(a), Q_\mathcal{P}(a)) \mid a \in \mathcal{A}\}$ form a performance profile specific to the acoustic distribution. For a fixed strategy, acoustic distribution shift leaves the NFE unchanged but can alter the expected enhancement quality attained at that cost. The situation thus determines the quality $Q_\mathcal{P}(a)$ attained by each strategy and, through the feasible set $\mathcal{A}_0(\mathcal{P}) = \{a \in \mathcal{A} \mid Q_\mathcal{P}(a) \geq Q_\mathcal{P}(a_0)\}$ for the default strategy a_0, the lowest cost that maintains the default quality. This distribution-dependent profile provides a basis for selecting strategies under either a quality objective or an efficiency objective subject to a quality requirement.

Characterizing the target performance profile requires estimating $Q_\mathcal{P}(a)$ from representative paired noisy and clean speech under the target acoustic condition, as q requires clean references. Evaluating strategies in $\mathcal{A}$ on a finite target set yields an empirical estimate of their expected enhancement quality and thus of the target performance profile. We therefore formulate inference adaptation at the level of a situation shared across a sequence of inputs, assuming that $\mathcal{P}$ remains fixed and that a small set of such paired speech is available for strategy selection, whereas subsequent inputs from the same situation require only noisy observations.

3 Situated Inference

We design Situated Inference to adapt the inference strategy of a frozen generative enhancer to its situation $\mathcal{P}$, enhancing each input $\mathbf{y} \sim \mathcal{P}$ as $G_\theta(\mathbf{y}; a_\mathcal{P})$ with its selected strategy $a_\mathcal{P}$. The method comprises three steps: defining a structured inference space, estimating the target performance profile from representative paired speech, and selecting a strategy according to the deployment objective. Fig. 2 illustrates the two resulting stages: target adaptation, which selects the strategy once for a situation, and situated enhancement, which applies it to every subsequent input.

3.1 Structured inference space

Under an inference strategy a, the frozen enhancer maps a noisy observation $\mathbf{y}$ to the enhanced estimate $G_\theta(\mathbf{y}; a)$. Each component of $a = (m, t_\mathrm{s}, N)$ is drawn from its supported set, with $m \in \mathcal{M}$ and with t_s and N following Eq. (2). Applying strategy a first processes the noisy observation through T_m, defined below, and then traverses the reverse path from t_s using N solver steps. The complete enhancement mapping of strategy a is therefore:

$$G_\theta(\mathbf{y}; a) = \widetilde{G}_\theta(T_m(\mathbf{y}); t_\mathrm{s}, N)\,. \tag{4}$$

Here, $\widetilde{G}_\theta(\cdot\,; t_\mathrm{s}, N)$ denotes the reverse integration of Eq. (2) from t_s with N solver steps. For a given backbone, the supported strategies form the structured inference space $\mathcal{A}$, organized by the three components of a, and $a_0 \in \mathcal{A}$ denotes the default inference strategy.

Observation processing defines the acoustic evidence on which the conditional reverse dynamics operate. The available operations in $\mathcal{M}$ retain the original mixture, apply a noise-floor gain derived from the estimated noise floor (Boll, 1979), or apply a modulation-aware gain that augments it with modulation information. In the time–frequency domain, let $\mathbf{Y} = \mathrm{STFT}(\mathbf{y})$ denote the short-time Fourier transform of the noisy observation. The processed signal is $T_m(\mathbf{y}) = \mathrm{iSTFT}(\mathbf{W}_m \odot \mathbf{Y})$, where $\mathbf{W}_m$ denotes the corresponding signal-dependent spectral gain and $\odot$ the elementwise product, with $T_\mathrm{id}(\mathbf{y}) = \mathbf{y}$ for the identity operation $m = \mathrm{id}$. The processed observation is used both to condition F_θ and to construct the initial state at t_s according to the initialization rule of each backbone. The identity operation retains the original observation when spectral processing offers no estimated quality benefit for the target condition.

In addition to observation processing, an inference strategy specifies the reverse starting point and solver step count. The reverse starting point t_s denotes the time at which reverse sampling is initialized

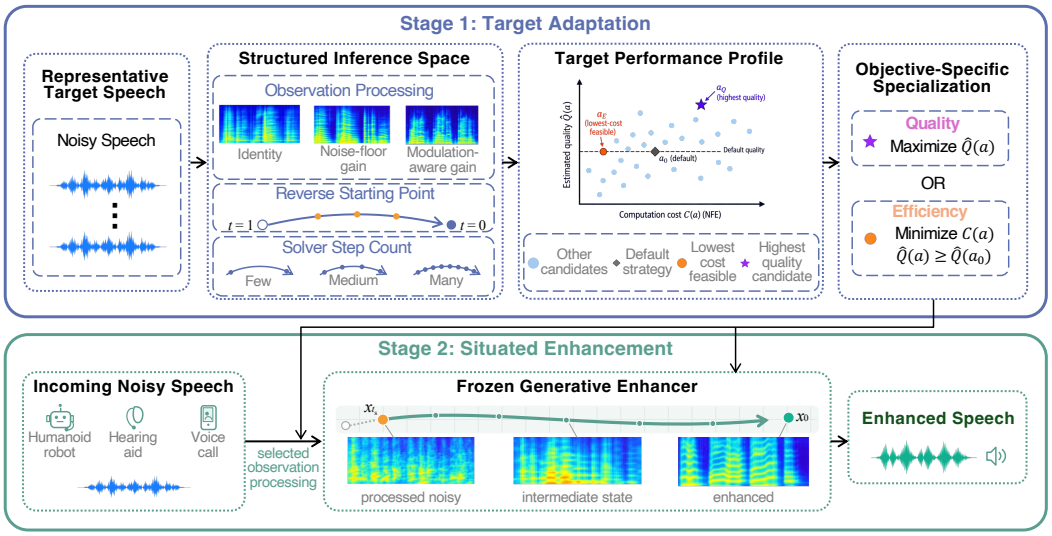

Figure 2: Overview of Situated Inference. **Stage 1: Target adaptation.** Representative paired target speech is used to evaluate strategies in the structured inference space, forming a target performance profile of estimated enhancement quality versus NFE. Objective-specific specialization selects either the highest-quality strategy or the lowest-cost strategy whose estimated quality meets or exceeds that of the default strategy. **Stage 2: Situated enhancement.** The selected observation processing is applied to incoming noisy speech from the same acoustic condition, and the frozen enhancer integrates from the selected reverse starting point with the selected solver step count.

in Eq. (2), with each backbone constructing the corresponding state $\mathbf{x}_{t_s}$ according to its existing initialization rule, which approximates from the observation a marginal that also depends on the unknown clean speech. Intermediate starting points have been studied in diffusion sampling (Meng et al., 2022; Zheng et al., 2023) and generative SE (Lay et al., 2023; 2024). From this state, the solver step count N specifies the number of updates used to reach the clean speech endpoint t_0, while $C(a)$ measures the corresponding inference cost in NFE. For a fixed N, changing t_s changes both the initial state and the reverse sampling interval spanned by these updates. These two controls therefore allow the integration interval and solver step count to be adapted jointly. Because observation processing enters both conditioning and initialization, and t_s and N jointly set the integration interval and step size, the three components interact and are therefore selected jointly within $\mathcal{A}$.

3.2 Constructing the Target Performance Profile

To construct the target performance profile over $\mathcal{A}$, which characterizes how the situation affects quality and efficiency, we evaluate the available strategies in terms of enhancement quality and inference cost. The cost $C(a)$ is determined by the inference configuration, while quality is estimated using representative paired speech from the target acoustic condition. We denote these K noisy and clean speech pairs by $\mathcal{D}_{\text{tar}} = \{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^{K}$. Speech used for final evaluation is kept separate from $\mathcal{D}_{\text{tar}}$ and is used only after strategy selection. Clean references are required for target quality estimation, and situated enhancement uses only noisy observations.

Using $\mathcal{D}_{\text{tar}}$, we estimate the population quality $Q_{\mathcal{P}}(a)$ defined in Eq. (3) with the empirical mean:

$$\widehat{Q}(a) = \frac{1}{K} \sum_{i=1}^{K} q(G_\theta(\mathbf{y}_i; a), \mathbf{x}_i)\,. \tag{5}$$

Each strategy is evaluated on the same noisy observations against the corresponding clean references, so differences in $\widehat{Q}(a)$ reflect strategy performance on a common sample of the target distribution. Together with the inference costs, these estimates form the empirical target performance profile $\{(C(a), \widehat{Q}(a)) \mid a \in \mathcal{A}\}$. This profile supports comparison of candidate strategies under the quality and efficiency objectives, while remaining a finite-sample estimate of their population performance.

The target performance profile separates the estimation of strategy performance from the choice of deployment objective. The default strategy a_0 provides a reference for assessing quality gains and identifying candidates that reduce computation while maintaining its estimated quality.

3.3 Inference specialization for quality and efficiency

The quality objective uses the target performance profile to select the inference strategy with the highest estimated enhancement quality on the target set:

$$a_{\mathrm{Q}} = \arg\max_{a \in \mathcal{A}} \widehat{Q}(a). \tag{6}$$

The maximization compares all supported strategies in $\mathcal{A}$, jointly considering all three components of a through their estimated quality on the target set. Inference cost does not constrain this maximization, but serves as a secondary criterion when multiple strategies attain the same maximum estimated quality. Among these strategies, Situated Inference selects the one with the lowest $C(a)$. A fixed ordering of strategy identifiers resolves any remaining tie, ensuring a uniquely determined strategy.

The efficiency objective uses the same profile to minimize inference cost while maintaining at least the estimated default quality. It therefore restricts selection to the empirical feasible set $\widehat{\mathcal{A}}_0 = \{a \in \mathcal{A} \mid \widehat{Q}(a) \geq \widehat{Q}(a_0)\}$, which estimates $\mathcal{A}_0(\mathcal{P})$, and selects its lowest-cost strategy:

$$a_{\mathrm{E}} = \arg\min_{a \in \widehat{\mathcal{A}}_0} C(a). \tag{7}$$

Because $a_0 \in \widehat{\mathcal{A}}_0$, the feasible set is nonempty, and the selected strategy satisfies $C(a_{\mathrm{E}}) \leq C(a_0)$. Among strategies with the same minimum cost, Situated Inference selects the one with the highest estimated quality $\widehat{Q}(a)$. Any remaining tie is resolved using the same fixed ordering of strategy identifiers as for the quality objective. The quality constraint is evaluated on $\mathcal{D}_{\mathrm{tar}}$ and therefore preserves the estimated baseline quality on the target set without guaranteeing it on unseen speech.

Although the quality and efficiency objectives use different selection criteria, both yield a single inference strategy for the target acoustic condition. This shared formulation separates target adaptation from situated enhancement: representative paired speech is used to select either a_{Q} or a_{E} as the situated strategy $a_{\mathcal{P}}$, which is then applied across inputs from the same situation. For each new noisy observation, the enhancer applies the selected strategy with F_θ and Φ unchanged. Clean references are needed only for target adaptation, and the estimated profile and selected strategy remain fixed during situated enhancement. The adaptation cost of $|\mathcal{A}|K$ enhancements per run is thus incurred once for the target condition, while each subsequent input incurs the inference cost $C(a_{\mathrm{Q}})$ or $C(a_{\mathrm{E}})$.

4 Experiments

We evaluate Situated Inference across four generative backbones and three acoustic conditions, assessing perceptual quality gains under the Quality objective and NFE reductions together with the resulting enhancement quality under the Efficiency objective. Further analyses examine the benefits of jointly adapting inference components beyond the solver step count alone and the influence of target speech quantity on strategy selection and enhancement quality. Deployment on the Ameca further illustrates the applicability of Situated Inference in humanoid communication.

4.1 Experimental setup

We evaluate Situated Inference on four pretrained generative SE backbones: Schrödinger bridge (SB, M7) (Richter et al., 2025), SGMSE+ (Richter et al., 2023), FlowSE (Lee et al., 2025b), and StoRM (Lemercier et al., 2023). These backbones cover different conditional generative dynamics and sampling procedures, with their pretrained weights and solver update rules fixed throughout evaluation. We consider three acoustic conditions, each defining a distinct situation: VoiceBank–DEMAND (VBD) (Valentini-Botinhao, 2017), the synthetic non-reverberant subset of Deep Noise Suppression (DNS) (Reddy et al., 2020), and LibriSpeech–MUSAN (LSM) mixtures (Panayotov et al., 2015; Snyder et al., 2015). These conditions differ in their speech, noise, and mixture distributions, with VBD serving as the in-domain condition and DNS and LSM as dataset-level shifted conditions

Table 1: Comparison of enhancement quality and inference cost across three acoustic conditions and four generative SE backbones. The best quality scores and lowest NFE within each backbone and condition are shown in **bold**, and the second best are underlined.

	VBD				DNS				LSM			
Method	PESQ↑	ESTOI↑	SI-SDR↑	NFE↓	PESQ↑	ESTOI↑	SI-SDR↑	NFE↓	PESQ↑	ESTOI↑	SI-SDR↑	NFE↓
Noisy	1.932	0.784	8.389	–	1.576	0.822	9.157	–	1.699	0.806	9.848	–
Spectral subtraction	2.009	0.787	9.075	–	1.615	0.826	9.575	–	1.730	0.810	10.228	–
SB	3.559	0.872	**13.106**	50	2.792	**0.908**	11.608	50	2.299	0.856	10.586	50
+ ArtiFree	3.555	0.869	13.062	51	2.783	0.907	11.571	51	2.293	0.855	10.548	51
+ Situated (Efficiency)	3.561	0.870	12.912	**18**	2.794	**0.908**	11.626	**18**	2.308	0.857	10.541	**6**
+ Situated (Quality)	**3.564**	**0.873**	12.961	50	**2.803**	**0.908**	**11.652**	50	**2.358**	**0.862**	**10.755**	18
SGMSE+	2.857	0.861	17.062	60	2.460	0.908	15.492	60	2.088	0.861	13.859	60
+ ArtiFree	2.621	0.852	16.312	60	2.273	0.898	14.587	60	1.934	0.850	12.906	60
+ Situated (Efficiency)	2.907	0.862	16.963	**40**	2.484	0.906	15.372	**40**	2.103	0.860	13.712	**10**
+ Situated (Quality)	**2.934**	**0.863**	**17.071**	60	**2.493**	**0.909**	**15.511**	60	**2.177**	**0.864**	**13.942**	20
FlowSE	3.038	0.872	**18.890**	5	2.417	**0.908**	15.956	5	1.972	0.854	13.970	5
+ ArtiFree	2.958	0.868	18.302	6	2.362	0.904	15.402	6	1.941	0.852	13.561	6
+ Situated (Efficiency)	3.052	0.869	18.402	**3**	2.425	0.905	15.611	**3**	1.990	0.857	13.806	**1**
+ Situated (Quality)	**3.092**	**0.873**	18.716	10	**2.491**	0.907	**16.012**	20	**2.074**	**0.859**	**14.038**	20
StoRM	2.862	0.868	**18.463**	101	2.409	**0.908**	15.943	101	2.048	0.860	14.191	101
+ ArtiFree	2.836	0.867	18.301	99	2.386	0.906	15.804	99	2.031	0.859	14.102	99
+ Situated (Efficiency)	2.943	0.868	18.238	**21**	2.414	0.907	15.902	**21**	2.055	0.857	14.087	**11**
+ Situated (Quality)	**2.995**	**0.869**	18.317	41	**2.421**	**0.908**	**15.994**	61	**2.069**	**0.861**	**14.226**	21

without reverberation. Within each condition, paired speech is partitioned into a target set for target adaptation and a disjoint set for situated enhancement, using identical utterance-level assignments across all four backbones and all compared methods.

For comparison, we include unprocessed noisy speech, spectral subtraction (Boll, 1979), the default inference strategy of each backbone, and the input-adaptive sampling-depth selection of ArtiFree (Chhaglani et al., 2026), based on the estimated signal-to-noise ratio (SNR) category of each utterance. We evaluate Situated Inference under the Quality and Efficiency objectives defined above. Quality maximizes estimated target quality, whereas Efficiency minimizes NFE while maintaining at least the estimated default quality. Both objectives use the same frozen backbones, target sets, and evaluation inputs, and all compared methods keep the model weights fixed.

We evaluate enhancement quality and inference cost using metrics aligned with the two specialization objectives. PESQ (Rix et al., 2001) guides strategy specialization and serves as the primary quality metric for evaluation, while extended short-time objective intelligibility (ESTOI) (Jensen & Taal, 2016) and scale-invariant signal-to-distortion ratio (SI-SDR) (Le Roux et al., 2019) provide complementary measures of intelligibility and signal fidelity that are not used for selection. NFE counts the neural network evaluations of each backbone. All scores are averaged over three runs.

4.2 Quality and efficiency across situations

As shown in Table 1, adapting inference to each situation improves enhancement performance on held-out utterances. Under the Quality objective, Situated Inference achieves the highest PESQ among the compared methods in all 12 combinations, with statistically significant improvements over the default strategies ($p < 0.05$, paired t-test over held-out utterances). The magnitude of these gains varies across backbones and conditions, with the largest improvements observed for StoRM on VBD and FlowSE on LSM, at 0.133 and 0.102 PESQ, respectively. Gains also appear in the in-domain VBD condition, and all four backbones benefit on LSM, where each attains its lowest default PESQ. Situated Inference also consistently exceeds ArtiFree in PESQ, suggesting the benefit of selecting a shared strategy for each situation over the evaluated input-adaptive variant.

The Efficiency objective tests whether the same target performance profile supports NFE reductions while maintaining perceptual quality on held-out utterances. Situated Inference reduces NFE in all 12 combinations, with PESQ remaining above the corresponding default values in every case. The ratio of default to selected NFE ranges from $1.5\times$ to $9.2\times$ across the full evaluation and from $5.0\times$ to $9.2\times$ on

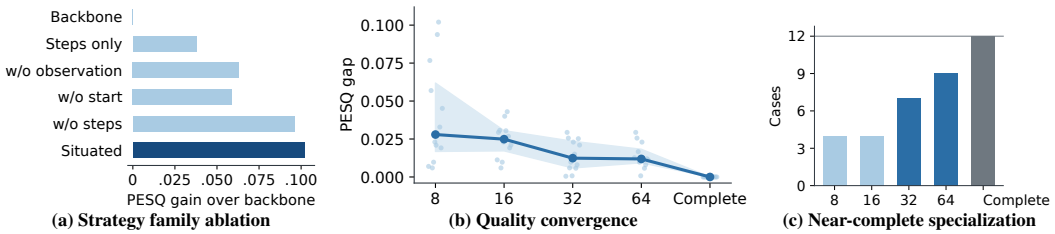

Figure 3: Analysis of inference components and target speech requirements. **(a)** PESQ gains over the default strategy for Situated Inference and restricted variants on FlowSE under LSM. **(b)** Absolute PESQ differences between strategies selected using smaller target sets and those selected using the complete target set across 12 combinations of backbone and acoustic condition. The line and shaded band indicate the median and interquartile range across combinations, and points represent individual combinations. **(c)** Number of combinations with an absolute PESQ difference of at most 0.02 from the result obtained using the complete target set.

LSM. These results show that substantial NFE reductions are possible while preserving or improving the observed PESQ. The selected NFE is also lowest on LSM for every backbone, for example 6 rather than 18 for SB, suggesting that the situation changes not only the attainable quality but also the lowest cost that maintains the default quality. Compared with ArtiFree, the situated strategies achieve higher PESQ with fewer NFE in all 12 combinations. The consistency of these benefits across the four backbones supports its applicability to different conditional generative formulations. Together, the two objectives demonstrate that a shared target performance profile supports either higher enhancement quality or lower inference cost through objective-specific selection.

4.3 Ablation of the structured inference space

To examine the contributions of individual components of the structured inference space to target specialization, we compare Situated Inference with restricted strategy variants in Fig. 3(a). We conduct this analysis on FlowSE, which has the largest evaluated strategy grid, under LSM, the condition with the lowest default PESQ. Each variant varies a subset of the three components while fixing the others at their default values. The step-only variant yields the smallest PESQ gain, indicating that step count adjustment captures only a limited portion of the available improvement. The w/o observation variant adapts the reverse starting point and solver step count, while the w/o start variant adapts observation processing and solver step count. Both outperform the step-only variant, showing that either observation processing or reverse starting point adaptation provides additional gains beyond step count adjustment. The w/o steps variant adapts observation processing and the reverse starting point while retaining the default step count, recovering most of the joint adaptation gain. These results indicate that observation processing and reverse starting point adaptation account for much of the improvement in this setting, while Situated Inference achieves the highest evaluation PESQ by jointly adapting all three components.

4.4 Target evidence for strategy specialization

To assess the target evidence required for strategy specialization, we select strategies using progressively larger nested target sets and compare their enhancement quality with that obtained using the complete target set in Fig. 3(b,c). Across the 12 combinations, the median absolute PESQ gap decreases from 0.028 with eight target pairs to 0.012 with 32 pairs, indicating that additional target evidence brings the selected strategies closer to the quality achieved using the complete set. The median gap remains at 0.012 with 64 pairs, but the number of combinations within 0.02 PESQ of the complete-set result rises from four with eight pairs to seven with 32 pairs and nine with 64 pairs. The two summaries suggest that the benefit of additional target speech varies across settings, with some combinations still improving after the median gap has stabilized.

4.5 Deployment on the Ameca humanoid robot

We deploy Situated Inference as the SE front end of the Ameca humanoid to illustrate its efficiency advantage in practice. For each evaluated situation, target adaptation selects a situated strategy

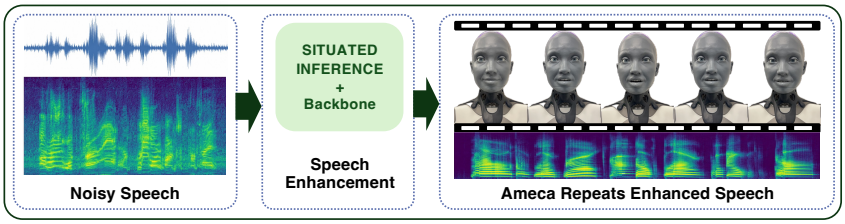

Figure 4: Deployment of Situated Inference as the SE front end of the Ameca humanoid robot.

for the frozen enhancer from representative paired recordings, and situated enhancement applies it during subsequent interactions. As illustrated in Fig. 4, the selected observation processing is applied to microphone recordings, and the enhanced speech is passed to the robot speech interface. The deployment follows the benchmark procedure, with the model and solver unchanged.

5 Related Work

Diffusion (Richter et al., 2023), stochastic regeneration (Lemercier et al., 2023), Schrödinger bridge (Jukić et al., 2024), and flow matching models (Lee et al., 2025b) formulate SE as conditional generation. Their inference procedures expose choices that affect enhancement quality and computation. Prior studies investigated reverse initialization and sampling steps (Lay et al., 2024; Gonzalez et al., 2024a) and evaluated a sampler under matched and mismatched acoustic conditions (Gonzalez et al., 2024b), while the broader diffusion literature has examined intermediate initialization (Meng et al., 2022) and configurable and optimized sampling schedules (Karras et al., 2022; Sabour et al., 2024; Li et al., 2023). These works show that individual inference choices can depend on the data and on acoustic mismatch, but none selects observation processing, reverse starting point, and step count jointly for a target acoustic condition or tests whether it transfers.

Adaptation under acoustic distribution shift commonly modifies either the learned model or the inference behavior of individual inputs. Test-time adaptation (Liang et al., 2025) updates model representations (Raichle et al., 2026b; Zhang et al., 2024) or parameters (Wang et al., 2021; Raichle et al., 2026a) using unlabeled target data, and adaptive generative methods adjust sampling depth for each input (Chhaglani et al., 2026; Fabian et al., 2024) or steer individual sampling runs (Guo et al., 2025; Ren et al., 2026). Frozen generative speech priors also adapt to unseen noise or reverberation per input without clean references (Nortier et al., 2024; Moliner et al., 2024), label-free methods update SE models on unlabeled mixtures (Tzinis et al., 2022; Behera et al., 2025), and validation-based step selection has been used in diffusion SE (Tai et al., 2023). Test-time adaptation thus changes what the model has learned, and input-adaptive methods change how each input is sampled. Situated Inference instead adapts the inference strategy at the level of a situation shared across inputs while leaving the model and solver fixed, which we term inference-space adaptation. Unlike most test-time adaptation, it uses a paired target set with clean references, and Fig. 3(b,c) examines how much target speech this requires. The formulation applies empirical algorithm configuration (Balcan et al., 2020; 2024) to coupled inference choices and quality and computation requirements in generative SE.

6 Conclusion

We studied the relationship between acoustic distribution shift and inference in frozen generative SE models. Our findings show that a strategy effective under one acoustic condition can become less effective under another, and that the lowest cost that maintains quality also varies across conditions, indicating that inference choices depend on the situation in which the enhancer is deployed. Building on this observation, Situated Inference uses representative paired target speech to jointly adapt inference components under quality or efficiency objectives. Experiments across four generative backbones and three acoustic conditions demonstrate the benefits of this adaptation, including gains beyond step-only adaptation, and deployment on the Ameca humanoid illustrates its efficiency in practice. More broadly, this work offers a perspective on inference-space adaptation, where a frozen generative model adapts only its inference strategy to its situation and deployment objective.

AI USE STATEMENT

In this work, we used generative AI tools for grammar polishing. We have not used generative AI tools for methodology design, implementation, experiments, or result interpretation, and the other tasks requiring disclosure are not applicable to this work. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

REFERENCES

Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Refined Bounds for Algorithm Configuration: The Knife-Edge of Dual Class Approximability. In *International Conference on Machine Learning*, 2020.

Maria-Florina Balcan, Christopher Seiler, and Dravyansh Sharma. Accelerating ERM for Data-Driven Algorithm Design Using Output-Sensitive Techniques. In *Advances in Neural Information Processing Systems*, 2024.

Avishkar Behera, Riya Ann Easow, Venkatesh Parvathala, and K. Sri Rama Murty. Test-Time Training for Speech Enhancement. In *Annual Conference of the International Speech Communication Association*, 2025.

Steven Boll. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1979.

Bhawana Chhaglani, Yang Gao, Julius Richter, Xilin Li, Syavosh Zadissa, Tarun Pruthi, and Andrew Lovitt. ArtiFree: Detecting and Reducing Generative Artifacts in Diffusion-based Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.

Zalan Fabian, Berk Tınaz, and Mahdi Soltanolkotabi. Adapt and Diffuse: Sample-Adaptive Reconstruction via Latent Diffusion Models. In *International Conference on Machine Learning*, 2024.

Philippe Gonzalez, Tommy Sonne Alstrøm, and Tobias May. Assessing the Generalization Gap of Learning-Based Speech Enhancement Systems in Noisy and Reverberant Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Philippe Gonzalez, Zheng-Hua Tan, Jan Østergaard, Jesper Jensen, Tommy Sonne Alstrøm, and Tobias May. Investigating the Design Space of Diffusion Models for Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024a.

Philippe Gonzalez, Zheng-Hua Tan, Jan Østergaard, Jesper Jensen, Tommy Sonne Alstrøm, and Tobias May. Diffusion-Based Speech Enhancement in Matched and Mismatched Conditions Using a Heun-based Sampler. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024b.

Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. Training-Free Guidance Beyond Differentiability: Scalable Path Steering with Tree Search in Diffusion and Flow Models. In *Advances in Neural Information Processing Systems*, 2025.

Jesper Jensen and Cees H. Taal. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

Ante Jukić, Roman Korostik, Jagadeesh Balam, and Boris Ginsburg. Schrödinger Bridge for Generative Speech Enhancement. In *Annual Conference of the International Speech Communication Association*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, 2022.

Bunlong Lay, Simon Welker, Julius Richter, and Timo Gerkmann. Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-based Speech Enhancement. In *Annual Conference of the International Speech Communication Association*, 2023.

Bunlong Lay, Jean-Marie Lemercier, Julius Richter, and Timo Gerkmann. Single and Few-Step Diffusion for Generative Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – Half-Baked or Well Done? In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

Ching-Hua Lee, Chouchang Yang, Jaejin Cho, Yashas Malur Saidutta, Rakshith Sharma Srinivasa, Yilin Shen, and Hongxia Jin. RestoreGrad: Signal Restoration Using Conditional Denoising Diffusion Models with Jointly Learned Prior. In *International Conference on Machine Learning*, 2025a.

Seonggyu Lee, Sein Cheong, Sangwook Han, and Jong Won Shin. FlowSE: Flow Matching-based Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025b.

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*, 2023.

Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, and Rongrong Ji. AutoDiffusion: Training-Free Optimization of Time Steps and Architectures for Automated Diffusion Model Acceleration. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts. *International Journal of Computer Vision*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

Rainer Martin. Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. *IEEE Transactions on Speech and Audio Processing*, 2001.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*, 2022.

Eloi Moliner, Jean-Marie Lemercier, Simon Welker, Timo Gerkmann, and Vesa Välimäki. BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models. In *International Workshop on Acoustic Signal Enhancement*, 2024.

Berné Nortier, Mostafa Sadeghi, and Romain Serizel. Unsupervised Speech Enhancement with Diffusion-based Generative Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. Single-Channel Speech Enhancement Using Spectral Subtraction in the Short-Time Modulation Domain. *Speech Communication*, 2010.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR Corpus based on Public Domain Audio Books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

Tobias Raichle, Erfan Amini, and Bin Yang. Test-Time Adaptation for Speech Enhancement via Mask Polarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026a.

Tobias Raichle, Niels Edinger, and Bin Yang. Test-Time Adaptation for Speech Enhancement via Domain Invariant Embedding Transformation. *IEEE Open Journal of Signal Processing*, 2026b.

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 *Annual Conference of the International Speech Communication Association*, 2020.

Yinuo Ren, Wenhao Gao, Lexing Ying, Grant Rotskoff, and Jiequn Han. DriftLite: Lightweight Drift Control for Inference-Time Scaling of Diffusion Models. In *International Conference on Learning Representations*, 2026.

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*, 2023.

Julius Richter, Danilo de Oliveira, and Timo Gerkmann. Investigating Training Objectives for Generative Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual Evaluation of Speech Quality (PESQ): A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align Your Steps: Optimizing Sampling Schedules in Diffusion Models. In *International Conference on Machine Learning*, 2024.

Neta Shaul, Juan Perez, Ricky T. Q. Chen, Ali Thabet, Albert Pumarola, and Yaron Lipman. Bespoke Solvers for Generative Flow Models. In *International Conference on Learning Representations*, 2024.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Volker Stahl, Alexander Fischer, and Rolf Bippus. Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000.

Wenxin Tai, Yue Lei, Fan Zhou, Goce Trajcevski, and Ting Zhong. DOSE: Diffusion Dropout with Adaptive Prior for Speech Enhancement. In *Advances in Neural Information Processing Systems*, 2023.

Efthymios Tzinis, Yossi Adi, Vamsi Krishna Ithapu, Buye Xu, Paris Smaragdis, and Anurag Kumar. RemixIT: Continual Self-Training of Speech Enhancement Models via Bootstrapped Remixing. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

Cassia Valentini-Botinhao. Noisy Speech Database for Training Speech Enhancement Algorithms and TTS Models. University of Edinburgh DataShare, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*, 2021.

Zhen-Yu Zhang, Zhiyu Xie, Huaxiu Yao, and Masashi Sugiyama. Test-Time Adaptation in Non-Stationary Environments via Adaptive Representation Alignment. In *Advances in Neural Information Processing Systems*, 2024.

Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated Diffusion Probabilistic Models and Diffusion-based Adversarial Auto-Encoders. In *International Conference on Learning Representations*, 2023.

APPENDIX CONTENTS

A Backbone Realizations of Conditional Inference **14**

A.1 Backbone Dynamics and Initialization 14

A.2 Solver Steps and Function Evaluations 15

B Implementation of Situated Inference **15**

B.1 Observation Processing 15

B.2 Structured Inference Spaces 15

B.3 Target Profile and Strategy Specialization 16

B.4 Target Adaptation and Situated Enhancement 16

C Experimental Protocol **16**

C.1 Acoustic Conditions and Data Construction 16

C.2 Target and Held-Out Assignment 17

C.3 Backbones and Comparison Methods 17

C.4 Quality and Computation Measures 17

D Detailed Support for the Main Analyses **18**

D.1 Inference Strategy Transferability Test 18

D.2 Joint Adaptation Test 18

D.3 Target Speech Sensitivity 19

E Ameca Deployment Details **19**

F Reproducibility Record **19**

A Backbone Realizations of Conditional Inference

This section summarizes how the conditional dynamics and numerical update introduced in the main text are realized by the four evaluated backbones. We retain the common time convention of Eqs. (1) and (2): the observation-anchored endpoint lies at $t = 1$, the clean endpoint lies at $t = 0$, and reverse inference proceeds from t_{s} toward 0. Integrating the common dynamics over two adjacent solver times yields the increment:

$$\mathbf{x}_{t_{n-1}} - \mathbf{x}_{t_n} = \int_{t_n}^{t_{n-1}} F_\theta(\mathbf{x}_t, \mathbf{y}, t)\,\mathrm{d}t + \int_{t_n}^{t_{n-1}} g(t)\,\mathrm{d}\overline{\mathbf{w}}_t. \tag{8}$$

The solver update rule Φ in Eq. (2) numerically approximates these deterministic and stochastic terms. Its backbone-specific implementation contains the required stochastic increment, predictor-corrector substeps, and initialization rule, where the corrector substeps refine each state rather than discretize Eq. (8). Situated Inference preserves this implementation and changes only the observation processing, reverse starting point, and solver step count supplied to it.

A.1 Backbone Dynamics and Initialization

Table 2 relates the common formulation to the prediction and initialization used by each backbone. The table distinguishes the quantity predicted by the neural network from the conditional field F_θ constructed from that prediction.

Table 2: Realization of the common conditional inference formulation by the evaluated backbones.

Backbone	Learned prediction	Realization of F_θ	Initialization at t_{s}	Solver
SB	Clean speech estimate	Bridge update coefficients	Observation-anchored bridge state	Bridge ODE
SGMSE+	Conditional score	Known drift and score	Perturbation around the observation	Predictor-corrector
FlowSE	Conditional velocity	Time-reversed velocity	Perturbation around the observation	Euler ODE
StoRM	Initial estimate and score	Conditional reverse drift	Perturbation around the initial estimate	Predictor-corrector

For SGMSE+, the network estimates the conditional score appearing in the reverse drift. Substituting this estimate into the released stochastic process gives the reverse field:

$$F_\theta(\mathbf{x}_t, \mathbf{y}, t) = \mu(\mathbf{x}_t, \mathbf{y}, t) - g(t)^2 s_\theta(\mathbf{x}_t, \mathbf{y}, t). \tag{9}$$

The drift μ and the diffusion coefficient g are fixed by the released stochastic process. This expression shows that Situated Inference does not replace the learned score or its stochastic process. It changes the observation and the numerical interval through which the same reverse field is evaluated.

FlowSE learns a conditional velocity and uses a deterministic ordinary differential equation. Because its published time coordinate runs from the noisy endpoint to the clean endpoint, reversing that coordinate to match the convention of the main text gives the field:

$$F_\theta(\mathbf{x}_t, \mathbf{y}, t) = -v_\theta(\mathbf{x}_t, \mathbf{y}, 1 - t), \qquad g(t) = 0. \tag{10}$$

The sign and transformed time argument preserve the direction of the released FlowSE trajectory. Its Euler solver then realizes this field over the interval selected by t_{s} using the specified solver step count.

The evaluated SB model predicts a clean speech estimate $\widehat{\mathbf{x}}_\theta$ instead of a score or velocity. Its released bridge sampler converts that estimate into the next reverse state by the update:

$$\mathbf{x}_{t_{n-1}} = \alpha_n \mathbf{x}_{t_n} + \beta_n \widehat{\mathbf{x}}_\theta + \gamma_n \mathbf{y}. \tag{11}$$

The fixed bridge coefficients α_n, β_n, and γ_n are determined by adjacent solver times. They provide the backbone-specific mapping represented abstractly by F_θ and Φ in the main text, and Situated Inference retains this mapping intact.

StoRM combines an initial speech estimate $D_\theta(\mathbf{y})$, produced by its predictive network, with a conditional score model. Before reverse sampling, its released initialization constructs the state at the selected starting point as follows:

$$\mathbf{x}_{t_{\mathrm{s}}} = D_\theta(\mathbf{y}) + \sigma(t_{\mathrm{s}})\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{12}$$

The noise scale $\sigma(t_s)$ is that of the released stochastic process at t_s. The score model subsequently conditions on both the noisy observation and this initial estimate. Across all four backbones, $\mathbf{y}$ in these expressions is replaced consistently by the processed observation $T_m(\mathbf{y})$, while the learned prediction, its mapping to reverse dynamics, and the solver update remain unchanged.

A.2 Solver Steps and Function Evaluations

The solver step count N and the number of function evaluations measure different quantities when one numerical update invokes a neural field more than once. Table 3 gives the accounting used throughout the paper. The additional StoRM evaluation produces the initial estimate used by its generative stage. Reporting NFE counts neural computation under the original numerical solver of each backbone and excludes the STFT-domain observation processing T_m.

Table 3: Mapping from solver steps to the number of function evaluations.

Backbone	Solver update	Sampling NFE	Additional evaluation	NFE
SB	Bridge ODE	N	0	N
SGMSE+	Predictor-corrector	$2N$	0	$2N$
FlowSE	Euler ODE	N	0	N
StoRM	Predictor-corrector	$2N$	1	$2N+1$

B Implementation of Situated Inference

This section details the observation processing in Eq. (4), the structured inference spaces, the target profile and strategy specialization of Eqs. (5) to (7), and the separation between target adaptation and situated enhancement.

B.1 Observation Processing

The operation T_m instantiates the observation component of the inference strategy in Eq. (4). The implementation first represents the noisy speech with a centered 510-sample Hann short-time Fourier transform and a 128-sample hop at 16 kHz. Writing this representation as $\mathbf{Y} = \mathrm{STFT}(\mathbf{y})$, each available operation applies a spectral gain $\mathbf{W}_m$ before reconstructing the waveform:

$$T_m(\mathbf{y}) = \mathrm{iSTFT}(\mathbf{W}_m \odot \mathbf{Y}), \qquad T_{\mathrm{id}}(\mathbf{y}) = \mathbf{y}. \tag{13}$$

The identity operation retains the original mixture. The noise-floor gain estimates the noise level from the tenth percentile of the spectral power in each frequency bin, smooths the resulting gain over a 3×5 time–frequency neighborhood, and applies a retention floor. The modulation-aware gain further measures the relative modulation energy between 2 and 8 Hz within a broader modulation range and uses this evidence to apply bounded attenuation before the same smoothing and retention floor. The percentile noise floor follows quantile-based noise estimation (Stahl et al., 2000) related to minimum statistics (Martin, 2001), and the modulation gain relates to modulation-domain enhancement (Paliwal et al., 2010).

These are the same identity operation, noise-floor gain, and modulation-aware gain introduced in the main text. In every case, the resulting $T_m(\mathbf{y})$ supplies both the conditioning observation for F_θ and the observation anchor used to construct $\mathbf{x}_{t_s}$. The identity operation allows target speech to determine whether spectral processing belongs in the specialized inference strategy instead of imposing a front end on every acoustic condition.

B.2 Structured Inference Spaces

Table 4 reports the discrete strategy spaces evaluated for each backbone. Each row lists the values represented in the recorded grid, and only combinations present in that grid enter specialization. The common support is the intersection used for the ISTT and the restricted families in Fig. 3(a). The LSM Efficiency selections for SGMSE+, FlowSE, and StoRM use the smallest evaluated N,

Table 4: Structured inference spaces used by Situated Inference.

Backbone and condition	Observation processing	t_s	N	Default a_0	Strategies
SB: VBD/DNS	All three	0.06, 0.12, 0.6, 1	1,2,4,6,12,18,25,30,50,75	identity, 1, 50	42
SB: LSM	All three	0.03, 0.06, 0.12, 0.6, 1	1,2,4,6,12,18,25,30,50,75	identity, 1, 50	51
SGMSE+: VBD/DNS	All three	0.4, 0.6, 1	5,10,12,15,20,30,40,50	identity, 1, 30	39
SGMSE+: LSM	All three	0.4, 0.6, 1	5,10,12,15,20,30	identity, 1, 30	27
FlowSE: all	All three	0.4, 0.6, 1	1,2,3,4,5,7,10,15,20	identity, 1, 5	63
StoRM: all	All three	0.4, 0.6, 1	5,10,15,20,30,50	identity, 1, 50	28
Common support	Backbone-specific intersection across VBD, DNS, and LSM				42/27/63/28

so the grid bounds their NFE ratios. The listed value sets describe the support of each grid instead of a Cartesian product: the strategy count records the combinations actually evaluated. Each grid is defined for one backbone and acoustic condition, and its exact tuples (m, t_s, N) belong to the strategy inventory in Table 10. This preserves each backbone's supported execution while keeping the learned dynamics and solver update rule fixed.

B.3 Target Profile and Strategy Specialization

The structured inference space specifies the available executions, while paired target speech supplies the information needed to compare them under the target acoustic condition. Following Fig. 2, every strategy $a \in \mathcal{A}$ enhances the same target speech pairs according to Eq. (4), the resulting signals determine $\widehat{Q}(a)$ through Eq. (5), and the numerical execution of each strategy determines $C(a)$. The quality objective applies Eq. (6) to this profile to obtain a_Q, and the efficiency objective applies Eq. (7) to its feasible set $\widehat{\mathcal{A}}_0$ to obtain a_E. Both objectives therefore use one empirical profile and differ only in the selection criterion. Exact quality ties are resolved by lower $C(a)$ and then by the fixed ordering of strategy identifiers. Equal-cost efficiency candidates are resolved by higher $\widehat{Q}(a)$ and then by the same ordering. Because $a_0 \in \widehat{\mathcal{A}}_0$, the default inference strategy always satisfies the efficiency quality requirement. The selected strategy is stored and used for subsequent speech from the same acoustic condition.

B.4 Target Adaptation and Situated Enhancement

Table 5 makes the information boundary between target adaptation and situated enhancement explicit. Target adaptation evaluates the structured inference space once for an acoustic condition. Subsequent

Table 5: Information and operations used during target adaptation and situated enhancement.

Stage	Noisy speech	Clean reference	Strategy evaluation	Weight update
Target adaptation	Yes	Yes	Yes	No
Situated enhancement	Yes	No	No	No

inputs execute the stored observation processing, reverse starting point, and solver step count directly. The target evidence is therefore shared across subsequent speech without reconstructing the target performance profile or changing the learned field.

C Experimental Protocol

This section details the data, comparison methods, and measurement rules underlying the experiments in Sec. 4.

C.1 Acoustic Conditions and Data Construction

Table 6 describes the sources of the three evaluated acoustic conditions and the target and held-out assignments used in the paper. VBD retains the speech, noise, and mixture protocol associated with

Table 6: Speech and noise sources for the evaluated acoustic conditions.

Condition	Clean speech	Noise	Mixture construction	Role	Target	Held-out
VBD	VoiceBank	DEMAND	Standard VoiceBank–DEMAND test mixtures	In-domain	413	411
DNS	DNS clean speech	DNS noise recordings	Synthetic non-reverberant test set	Shifted	74	76
LSM	LibriSpeech test-clean	MUSAN	2.5, 7.5, 12.5, 17.5 dB mixtures	Shifted	982	1018

the development benchmark of the evaluated enhancers. DNS changes the recorded speech and noise conditions. LSM combines LibriSpeech utterances with MUSAN noise at SNRs assigned in round-robin order, changing both speech material and noise family relative to VBD. These differences instantiate the acoustic distribution shifts considered in the main text without changing the enhancement task.

C.2 Target and Held-Out Assignment

The stored utterance identifier determines a reproducible assignment that is shared by all four backbones. For an identifier u without its file extension, target adaptation uses identifiers whose SHA256 digest, read as an integer $h(u)$, satisfies $h(u) \bmod 2 = 0$. The remaining identifiers form the held-out set. The assignment does not use enhancement scores. All strategies for a backbone and condition are evaluated on identical target speech, and the specialized strategy is subsequently measured on the corresponding held-out speech.

The main quality and efficiency results, the ISTT, and restricted-family comparisons use this fixed assignment. Target-size analysis orders the target identifiers by the SHA256 digest of `size-v1:`u and uses progressively enlarged prefixes. This produces nested target sets while leaving held-out speech unchanged.

C.3 Backbones and Comparison Methods

Table 7: Backbone checkpoints and inference interfaces used in the evaluation.

Backbone	Evaluated model	Model weights	Inference interface
SB	M7 SB-VE data-prediction checkpoint	Fixed	Bridge ODE
SGMSE+	Released conditional score checkpoint	Fixed	Predictor-corrector
FlowSE	Released conditional velocity checkpoint	Fixed	Euler ODE
StoRM	Released predictor and score checkpoints	Fixed	Predictor-corrector

Table 8 separates the information and intervention available to each comparison in Table 1. It describes the executed comparison protocol, in which ArtiFree contributes its SNR-dependent sampling-depth selection and only Situated Inference uses paired target speech.

Table 8: Information and inference choices used by the comparison methods.

Comparison	Adaptation information	Changed object	Decision scope	Reported computation
Noisy	None	None	None	N/A
Spectral subtraction	No target adaptation	Signal processing	Each execution	Preprocessing
Backbone	None	None	Default inference strategy	NFE
ArtiFree	Estimated SNR of each input	Sampling depth	Individual input	NFE
Situated Inference	Paired target speech	Complete inference strategy	Acoustic condition	NFE

C.4 Quality and Computation Measures

PESQ at 16 kHz provides the quality measure q used to construct $\widehat{Q}(a)$. ESTOI and SI-SDR are computed on held-out speech as complementary measures of intelligibility and signal fidelity. Signals are aligned at their common minimum length before scoring, and VBD entries are means over the

held-out half of the test set, so they are not directly comparable with published full-test-set scores. All reported method differences use the same held-out utterances. Neural computation is reported as NFE according to Table 3, and solver steps are not compared directly across different numerical solvers.

D Detailed Support for the Main Analyses

This section specifies how the analyses in Figs. 1 and 3 are constructed from the experimental protocol. Numerical outcomes remain in the main figures and Table 1, and the solver step count of each selected strategy follows from its NFE in Table 1 through Table 3.

D.1 Inference Strategy Transferability Test

The Inference Strategy Transferability Test (ISTT) examines whether an inference strategy specialized under one acoustic condition retains its benefit under another. It uses the three acoustic conditions in Table 6, the four backbones in Table 7, and the fixed target and held-out assignment of App. C. The learned dynamics, solver update rule, and held-out speech remain unchanged throughout the test.

Candidate strategies. For each backbone, candidates are restricted to the common support in Table 4, which is the intersection of the strategy spaces evaluated under VBD, DNS, and LSM. Every candidate is therefore executable and scored under all three conditions, so a strategy selected under one condition can be evaluated under any other. The common support contains 42, 27, 63, and 28 strategies for SB, SGMSE+, FlowSE, and StoRM, respectively.

Specialization. For a source condition c, the quality objective of Eq. (6) is applied to the target speech of c, yielding the strategy $a_{\mathrm{Q}}^{(c)}$ with the tie-breaking rules of App. B. The default strategy a_0 of each backbone in Table 4 serves as the reference.

Evaluation. The fixed strategy $a_{\mathrm{Q}}^{(c)}$ and the default strategy a_0 are evaluated on held-out speech from every evaluation condition c'. The resulting gain is computed as follows:

$$\Delta_{c\to c'} = Q_{c'}^{\mathrm{ho}}\left(a_{\mathrm{Q}}^{(c)}\right) - Q_{c'}^{\mathrm{ho}}(a_0). \tag{14}$$

The term $Q_{c'}^{\mathrm{ho}}(a)$ denotes the mean PESQ of strategy a on the held-out speech of condition c'.

Aggregation. Taking $c, c' \in \{\mathrm{VBD}, \mathrm{DNS}, \mathrm{LSM}\}$ produces one 3×3 source-by-evaluation matrix of $\Delta_{c\to c'}$ for each backbone. Diagonal entries ($c = c'$) represent target-matched specialization, and off-diagonal entries ($c \neq c'$) represent reuse under another acoustic distribution. The same-condition bar in Fig. 1(a) averages the three diagonal entries, and the different-condition bar averages the six off-diagonal entries. Both bars are computed on identical held-out speech for the same backbone, so their difference reflects the condition under which the strategy was selected.

D.2 Joint Adaptation Test

The Joint Adaptation Test in Fig. 1(b) restricts every candidate to the default NFE budget of its backbone. The step-only strategy varies N while retaining the identity operation and the default reverse starting point, so within the budget it can only keep or reduce the default N. The joint strategy instead considers every strategy in the structured inference space that satisfies the same budget. The test therefore examines whether the observation and reverse starting point provide useful choices beyond reallocating solver steps, without granting the joint strategy additional neural evaluations.

The ablation in Fig. 3(a) uses the strategy families defined in Table 9. Each restricted family preserves the same target speech, quality measure, and specialization rule while holding the remaining inference components at their default values. Unlike the Joint Adaptation Test, these families are selected without an NFE budget. The full method and every restricted family therefore differ only in the inference components available during specialization.

Table 9: Strategy families used to ablate the structured inference space.

Strategy family	Observation processing	Reverse starting point	Solver step count
Backbone	Fixed	Fixed	Fixed
Step only	Fixed	Fixed	Specialized
w/o observation	Fixed	Specialized	Specialized
w/o start	Specialized	Fixed	Specialized
w/o steps	Specialized	Specialized	Fixed
Situated Inference	Specialized	Specialized	Specialized

D.3 Target Speech Sensitivity

Target speech sensitivity uses progressively enlarged subsets of the same ordered target pool and preserves one fixed held-out set. At each subset size, Situated Inference reconstructs the target performance profile, specializes a quality strategy, and evaluates that strategy on the unchanged held-out speech. The complete target set supplies the reference result for each backbone and acoustic condition. Fig. 3(b) summarizes the absolute PESQ difference from this reference across all evaluated combinations, while Fig. 3(c) counts combinations that fall within the tolerance defined in its caption. This construction measures convergence toward complete-set specialization without mixing target speech into held-out evaluation.

E Ameca Deployment Details

The Ameca evaluation applies the same separation between target adaptation and situated enhancement used in the benchmark experiments. Paired speech recorded for each evaluated acoustic condition constructs the target performance profile, after which the specialized inference strategy remains fixed during subsequent interaction. Each microphone recording is processed by the selected operation T_m and enhanced by the frozen generative model, and the enhanced speech is passed to the robot speech interface. The deployment protocol records the microphone and room arrangement, source placement, target speech construction, the selected strategy (m, t_s, N) with its NFE, the measured processing time, and the audio examples corresponding to Fig. 4. These details instantiate the method defined in Eqs. (4), (6) and (7) without adding a robot-specific learning or selection component.

F Reproducibility Record

Table 10 summarizes the records connecting the method definition, executed strategy grids, and reported figures and tables. The recorded analysis comprises 505,528 distinct utterance-strategy score records, each averaged over the three runs. Each input grid is rectangular in utterances and strategies, contains finite quality scores, and contains no duplicate utterance-strategy keys. The default and Situated Inference rows of Table 1 and the analyses in Figs. 1 and 3 are reconstructed from these score records using the target and held-out assignments specified in App. C.

Table 10: Reproducibility records for the reported evaluation.

Record	Content
Source revision	Enhancement and scoring implementation identified by repository revision.
Backbone identity	Checkpoint, released solver, default inference strategy, and conditioning interface for each backbone.
Strategy inventory	Exact identifiers in each structured inference space and its common support.
Data assignment	Deterministic target and held-out identifier lists shared across backbones.
Execution control	Deterministic utterance-strategy random-state mapping for recorded stochastic executions.
Score inventory	PESQ, ESTOI, SI-SDR, and NFE for every recorded utterance-strategy pair.
Result mapping	Source tables for Table 1 and Figs. 1 and 3.