electronics

MDPI

Article

Multi-Task Flow Matching Speech Enhancement Method with Dual Guidance of Semantic Understanding and Speaker Perception

Bingyu Dai [1], Xia Zou [1,2,*], Yihao Li [1,2], Chong Jia [1,2,*], Meng Sun [1,2], Yongqiang Xie [3] and Xiongwei Zhang [1,2]

[1] College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China; dby44dby103@163.com (B.D.)
[2] State Key Laboratory of Advanced Communication Technology, Nanjing 210007, China
[3] Chinese People's Liberation Army Unit 32092, Beijing 100071, China
* Correspondence: zouxia@aeu.edu.cn (X.Z.); jiachong@aeu.edu.cn (C.J.)

Abstract

Generative speech enhancement models often suffer from the "generative hallucination" issue, where artifacts are introduced or original speech characteristics are distorted during reconstruction due to insufficient prior constraints, severely limiting their practicality. To address this, we propose a multi-task flow-matching speech enhancement network with dual guidance from semantic understanding and speaker perception. The method employs a flow-matching model as the generative backbone, leveraging both semantic and speaker embeddings as auxiliary conditions to guide the denoising process. The semantic embedding, extracted by a pretrained audio foundation model, provides high-level content priors to ensure the intelligibility and clarity of the enhanced speech. The speaker embedding, derived from a speaker verification model, captures timbre representations, which are further regularized by a contrastive InfoNCE loss to enhance intra-class compactness and inter-class separability of speaker characteristics. During optimization, the primary denoising reconstruction task and the auxiliary speaker discrimination task are jointly optimized. The two guidance streams, together with acoustic features, are collaboratively fed into the generation network, effectively suppressing hallucinations from both content correctness and timbre consistency perspectives. Experimental results demonstrate that the proposed multi-task framework substantially mitigates hallucination-related errors in generative scenarios, achieving comprehensive improvements in objective metrics, including both reference-based measures and non-intrusive perceptual metrics, thus validating the effectiveness of the dual-guidance strategy for generative speech enhancement.

Keywords: speech enhancement; flow matching; generative hallucinations; multi-task learning

Check for updates

Academic Editors: Costas Psychalinos, Georgia Tsirimokou, Alexander Barkalov, Esteban Tlelo-Cuautle and Lobna A. Said

Received: 2 September 2026
Revised: 30 September 2026
Accepted: 30 September 2026
Published: 9 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In speech signal processing, discriminative methods perform effectively under stationary noise, yet their deterministic mapping tends to cause over-smoothing in low-SNR or non-stationary conditions, often at the cost of fine speech details. Generative approaches, by contrast, learn the underlying distribution of clean speech as a prior, enabling high-quality and highly natural reconstructions even under severe interference. Early generative enhancement models, such as variational autoencoders (VAEs) [1] and generative adversarial networks (GANs) [2], suffered from posterior estimation errors or training

instability, which limited their fidelity and naturalness. More recently, diffusion models have offered a robust learning paradigm for high-fidelity speech generation, thanks to their well-defined forward diffusion and reverse denoising processes [3]. The step-wise refinement inherent in diffusion models aligns naturally with the time-frequency structure of speech, allowing progressive noise suppression while preserving essential acoustic details. Nevertheless, the iterative sampling required by diffusion models incurs high inference latency, which hinders real-time deployment. To overcome this bottleneck, FlowSE [4] adopts the flow matching framework combined with Diffusion Transformer (DiT) modules. By learning a continuous probability flow ordinary differential equation that maps noise to clean speech, it attains high-fidelity reconstruction with substantially fewer sampling steps than conventional diffusion models, thereby greatly accelerating inference.

Notably, none of the above methods explicitly or sufficiently exploits high-level semantic content or speaker identity cues from the input speech. When the input is heavily distorted, acoustic features alone provide limited guidance in distinguishing "meaningful speech" from erroneous reconstructions. This ambiguity can cause the generative model to introduce spurious components that deviate from the true source distribution, or to warp original voice characteristics—a phenomenon known as generative hallucination. In this work, we define generative hallucination as any content or speaker attribute in the enhanced speech that is not supported by the clean reference. It manifests in two forms: linguistic hallucination, where the model inserts or substitutes words absent from the clean speech, and acoustic hallucination, where the linguistic content is correct but the speaker identity or timbre drifts from the reference. Hallucinations not only degrade intelligibility and naturalness, but can also distort semantic meaning and alter speaker identity, critically undermining the reliability of generative enhancement in practice.

Recent work has explored semantic guidance for flow-matching speech enhancement. However, existing approaches still suffer from certain limitations. For instance, FlowSE [4] relies on external text transcripts as guidance, while allowing this condition to be omitted during inference. Such a loose dependency undermines the stability of guidance and restricts its applicability to scenarios where text annotations are unavailable. SenSE [5], on the other hand, adopts a two-stage training scheme: it first extracts semantic tokens via a language model, and then uses these tokens to guide the flow matching process. The discretization step inevitably incurs a loss of acoustic information, and the optimization objectives of the two stages may become misaligned, potentially introducing bias into the final reconstruction.

From a complementary perspective, G-MaP-SE [6] tackles the problem from a speaker information perspective. It constructs a prior distribution over clean-speech speaker embeddings using a Gaussian mixture model (GMM), and then matches the noisy speaker embeddings towards this prior, thereby providing stable speaker identity guidance for the enhancement network. Nevertheless, the reliance on a fixed GMM prior fundamentally limits its generalization ability. In real-world scenarios, the distribution of speaker embeddings is highly complex and constantly evolving; a pre-fitted static prior can hardly cover unseen speakers, new accents, or novel acoustic conditions. When test data deviates from the training prior, the matching process may fail or even introduce additional bias, thus compromising the robustness and practical reliability of the method. To address the above limitations, we propose a multi-task flow matching speech enhancement method with dual guidance from semantic understanding and speaker perception. The main contributions of this work are threefold:

1. Building a unified multi-task learning framework in which semantic understanding and speaker perception serve as consistently active auxiliary tasks. The semantic

branch provides a stable content prior, while the speaker branch supplies timbre identity guidance; together, they jointly constrain the generation process.

2. Leveraging a frozen pretrained large audio model (Qwen2-Audio-7B) to extract continuous high-level semantic embeddings, thus circumventing discrete quantization of semantic information. This yields richer and more direct content priors for the generation process.
3. Devising a speaker branch to explicitly model speaker identity. Instead of the absolute GMM matching used in G-MaP-SE [6], we impose a relative discriminative constraint via InfoNCE loss on the embedding space, which preserves greater representational freedom for unseen speakers and effectively avoids timbre distortion.

The two streams of guidance, together with acoustic features, are fed into the generation network, suppressing linguistic and acoustic hallucinations from content correctness and timbre consistency, respectively. Experiments verify that the proposed multi-task framework significantly mitigates hallucination phenomena in generative scenarios and effectively improves the overall performance of generative speech enhancement.

We name our method SS-FMSE (Semantic and Speaker-aware Flow Matching for Speech Enhancement). The rest of this paper is organized as follows: Section 2 reviews related work; Section 3 details the proposed SS-FMSE; Section 4 presents the experimental setup and results; and Section 5 concludes the paper.

2. Related Work

2.1. Flow Matching-Based Generative Speech Enhancement Methods

Flow matching [7] is a generative modeling framework that learns a continuous-time probability flow to smoothly transform a simple noise distribution into a complex target data distribution. It directly learns a deterministic, time-dependent velocity field $v_t(x_t)$, where t denotes the time step along the flow matching trajectory and x_t is the data distribution at that position. This velocity field describes how a noise point x_0 is smoothly pushed to a data point x_1 over the interval $t \in [0, 1]$. The underlying assumption is that for any pair of samples $x_0 \sim p_0$ (from a simple distribution, e.g., Gaussian noise) and $x_1 \sim p_1$ (from the target data distribution), there exists a smooth path connecting them. Taking linear interpolation as an example,

$$x_t = (1 - t)x_0 + tx_1 \tag{1}$$

the true velocity is given by

$$\dot{x}_t = x_1 - x_0. \tag{2}$$

The neural network is trained to approximate this velocity field by minimizing the mean-square error (MSE) between predicted and true velocities. Generation then only requires solving an ordinary differential equation (ODE), which can be completed in a few time steps, bypassing the redundant iterative denoising steps.

Flow matching has shown promising results in speech enhancement and separation. FlowSep [8] combines flow matching with a VAE and introduces language conditioning to achieve semantically guided audio source separation. FlowAVSE [9] adopts a U-Net backbone for flow matching, conditioned on both audio and visual features, and achieves competitive performance. FlowSE [4] integrates flow matching with the Transformer-based diffusion framework DiT, laying a foundation for applying flow matching to speech enhancement. However, FlowSE relies on optional external text transcripts rather than directly perceiving semantic content from the speech waveform, which renders semantic guidance unstable and restricts its applicability in real-world scenarios where text annotations are often unavailable.

2.2. Semantic-Guided Generative Methods

Leveraging high-level semantic information to guide the generation process has become an important trend in modern generative models, and its effectiveness has been well demonstrated in image generation. Zhang et al. [10] exploit the reasoning capabilities of multimodal large language models to guide the image generation process of diffusion models, using a semantic-to-detail connector that aligns fine-grained visual features with semantic space in advance, thereby enabling joint guidance from both global semantics and local details. Chang et al. [11] directly optimize semantic embeddings to guide text-to-image diffusion models for editing tasks, without relying on text prompts or model training. The core idea behind these works is highly consistent: they extract high-level semantic embeddings using pretrained models and inject them as conditions into the diffusion model to constrain the semantic direction of the generation process.

Similar approaches have also been widely adopted in audio and speech generation. Ren et al. [12] propose STA-V2A, a video-to-audio generation model that extracts global semantic and local temporal features from video, and combines them with text as cross-modal guidance, producing audio that closely matches the video content. Li et al. [13] present PL-TTS, which integrates prompt embeddings extracted by a large language model with a diffusion-based TTS system, effectively improving both synthesis quality and style controllability. These studies demonstrate that using semantic embeddings from pretrained models as conditional guidance for diffusion or flow matching models has become a well-established and effective technical approach in the generative AI community.

In the domain of speech enhancement, researchers have also explored the introduction of semantic information. SenSE [5] adopts a two-stage generation framework. It first extracts semantic tokens from distorted speech using a semantic-aware speech language model to capture semantic priors, and then feeds these tokens together with the mel-spectrogram of the distorted speech as conditions to guide the flow matching-based enhancement process. Experimental results confirm that the introduction of semantic information effectively alleviates semantic ambiguity in generation. Nevertheless, the two-stage decoupled design of SenSE has inherent limitations. In the first stage, the abstraction of continuous speech content into discrete tokens inevitably leads to the loss of some fine-grained acoustic information.

Moreover, the optimization objectives of the two stages are independent, which may cause misalignment between them and restrict the model's ability to fully exploit rich contextual information. More importantly, none of the existing methods simultaneously leverages semantic content and speaker identity as conditional guidance for speech enhancement (see Table 1 for a systematic comparison). Relying solely on content-level prior constraints is insufficient to suppress acoustic hallucination, i.e., identity feature drift during the generation process.

Table 1. Comparison of representative guidance-based speech enhancement methods. "Semantic content" and "Speaker identity" indicate whether the corresponding information is explicitly used as a condition. "Training scheme" indicates whether the method is trained in one stage or multiple stages. "Prior type" indicates the form of the guidance prior.

Method	Semantic Content	Speaker Identity	Training Scheme	Prior Type
SenSE [5]	✓	×	Two-stage	Discrete semantic tokens
FlowSE [4]	✓	×	One-stage	Text embedding
G-MaP-SE [6]	×	✓	One-stage	Static GMM prior

3. Proposed Method

3.1. Model Overview

We propose a multi-task flow matching speech enhancement method with dual guidance from semantic understanding and speaker perception. The model adopts flow matching as the generative backbone, and its overall architecture is illustrated in Figure 1. To address the generative hallucination problem caused by the lack of high-level priors, the model leverages both semantic content and speaker identity as conditional information to guide the entire flow-field prediction process. Specifically, the semantic branch provides content-correctness constraints to suppress linguistic hallucination, while the speaker branch imposes identity-consistency constraints to suppress acoustic hallucination.

The model treats clean-speech mel-spectrograms as samples from the target data distribution, which are mixed with randomly sampled standard Gaussian noise along a probability path. Let M^T denote the intermediate state at time step T, which serves as the core input to the DiT network. The mel-spectrogram of the distorted speech, denoted as M^N, is incorporated as a conditional embedding. For guidance, the semantic branch employs a pretrained audio large language model to extract semantic content embeddings E^{Se}; the speaker branch adopts a frozen CAM++ model to extract speaker embeddings E^{Sp}, which are then fine-tuned via a lightweight linear adaptation layer, with a discriminative constraint imposed on the mapped embedding space through the InfoNCE contrastive loss. The three streams of conditional information—namely, the distorted speech mel-spectrogram M^N, semantic embeddings E^{Se}, and speaker embeddings E^{Sp}—are concatenated with M^T along the channel dimension and fed into the DiT network, which stacks N transformer modules to predict the flow field. The flow loss serves as the primary optimization objective for joint multi-task training. During inference, the model integrates the predicted flow field using an ODE solver, iteratively sampling from pure noise to clean data, and the resulting mel-spectrogram is reconstructed into a time-domain enhanced waveform via a vocoder.

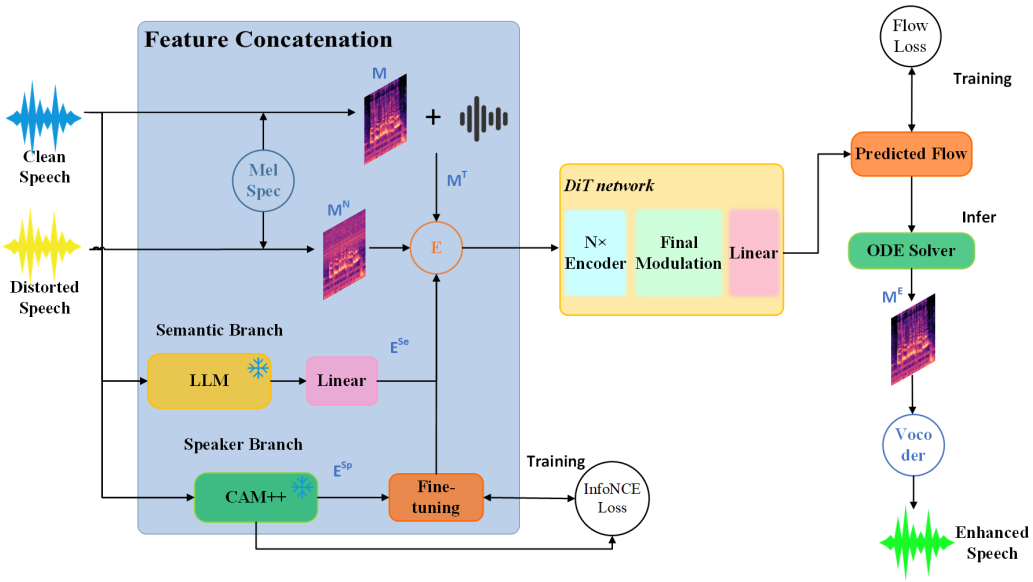


Figure 1. Overview of the SS-FMSE architecture. Three conditioning streams—distorted-speech mel-spectrogram M^N, semantic embeddings E^{Se}, and speaker embeddings E^{Sp}—are concatenated with M^T and fed into the DiT backbone. The training path jointly optimizes the flow loss and the InfoNCE contrastive loss on the speaker branch, while the inference path integrates the flow field via an ODE solver and reconstructs the waveform with a vocoder.

3.2. Semantic and Speaker Embedding

For semantic guidance, we adopt Qwen2-Audio-7B, a pretrained audio large language model, as the semantic feature extractor. Benefiting from pretraining on massive and diverse audio data, this model has demonstrated a strong ability to abstract speech content under the evaluated conditions. When the input waveform is distorted by environmental noise or the local spectral envelope is altered, it can still typically disentangle interference components from the time-frequency aliased acoustic signal and recover useful semantic information. Traditional acoustic features, such as MFCCs or STFT magnitude spectra, tend to degrade rapidly as noise power increases, struggling to retain valid phonetic content cues. In contrast, the high-level semantic embeddings produced by Qwen2-Audio-7B tend to be more anchored to the linguistic content of speech, providing relatively stable content-level priors for the subsequent backbone generation network. By injecting these embeddings as conditional information into the DiT network, the flow matching model can focus its limited generative capacity on frequency regions that are semantically plausible when solving the probability transport path, thereby reducing erroneous speech reconstruction caused by noise-induced ambiguity.

For speaker identity representation, we employ a pretrained speaker recognition model, CAM++, to extract embeddings related to individual attributes such as timbre and prosody. Although CAM++ has demonstrated a certain degree of generalization and noise robustness in public benchmarks, the noise types and reverberation conditions in our experimental environment still exhibit non-negligible domain shifts from the pretraining corpus. If the model parameters are directly frozen, the output speaker features cannot be perfectly aligned with the true identity distribution under the current noisy conditions. To address this, we cascade a lightweight trainable adaptation module, consisting of a single linear layer, after the CAM++ output, and impose a discriminative constraint on the mapped embedding space via the InfoNCE loss. This constraint forces the projected vectors of utterances from the same speaker to cluster tightly while pushing those from different speakers apart (as illustrated in Figure 2). This design leverages the small parameter size and rapid convergence of the linear layer, helping the speaker features to better align with the current experimental environment. With only a minimal increase in computational overhead, this ensures that the speaker guidance can accurately inject the target speaker's identity attributes into the flow-field prediction process of the DiT network.

Building on the above motivations, we detail the implementation as follows: For the semantic branch, we use the hidden states from the last transformer layer of Qwen2-Audio-7B, which are 4096-dimensional. Since semantic content evolves over time, we preserve the temporal dimension and apply linear interpolation along the time axis to align the hidden states with the mel-spectrogram frame rate; the interpolated features are then projected from 4096 to 512 dimensions, yielding $E^{Se} \in \mathbb{R}^{B\times T\times 512}$. For the speaker branch, CAM++ produces a 192-dimensional utterance-level embedding, which is projected to 256 dimensions by a single trainable linear layer. Because speaker identity is stable within an utterance, we broadcast the projected embedding along the time axis to obtain $E^{Sp} \in \mathbb{R}^{B\times T\times 256}$. Finally, M^T, M^N, E^{Se}, and E^{Sp} are concatenated along the channel dimension, resulting in a tensor of size $B \times T \times (2F + 512 + 256)$, which is projected to the DiT hidden dimension before entering the backbone.

3.3. Loss Function

The optimization objective of our method consists of two components: the primary flow matching loss and the auxiliary contrastive learning loss. The total loss is defined as

$$\mathcal{L} = \lambda \mathcal{L}_{\text{flow}} + (1 - \lambda)\mathcal{L}_{\text{InfoNCE}}, \tag{3}$$

where λ is set to 0.9 to balance the optimization weights between the primary and auxiliary tasks.

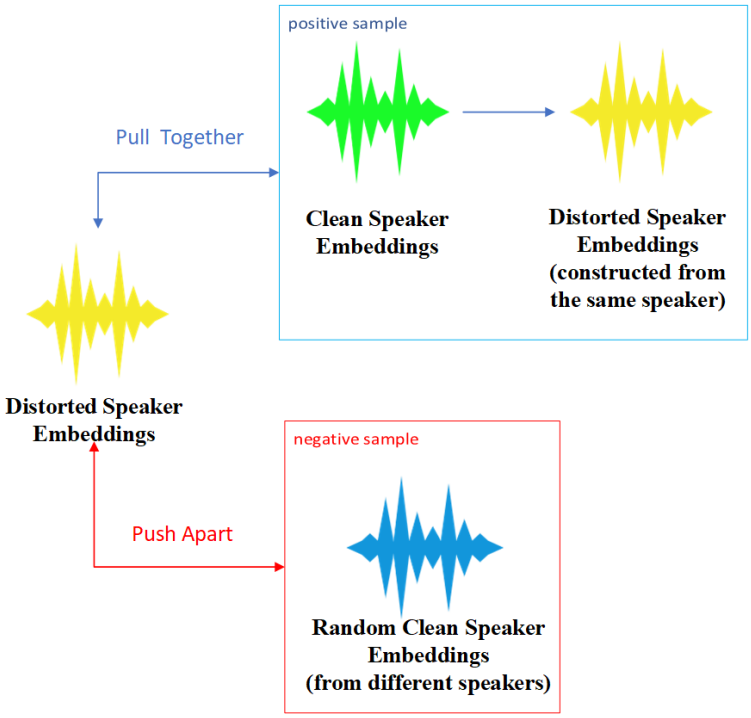


Figure 2. Contrastive loss on speaker embeddings.

The flow matching loss $\mathcal{L}_{\text{flow}}$ serves as the main optimization objective for the generative backbone. We adopt a conditional straight-line probability path. Let $x_0 \sim p_0 = \mathcal{N}(0, \mathbf{I})$ denote a Gaussian noise sample and $x_1 \sim p_1 = p_{\text{data}}$ denote a clean mel-spectrogram sample. Let $\mathbf{c} = (M^N, E^{Se}, E^{Sp})$ denote the conditioning variable, where M^N is the distorted-speech mel-spectrogram, and E^{Se} and E^{Sp} are the semantic and speaker embeddings, respectively. For $t \in [0, 1]$, the intermediate state x_t is defined by linear interpolation between x_0 and x_1, as defined in Equation (1). The target velocity field along this path is

$$u_t(x_t \mid \mathbf{c}) = \frac{dx_t}{dt} = x_1 - x_0. \tag{4}$$

The DiT network predicts the velocity field $v_t(x_t, \mathbf{c})$ to approximate $u_t(x_t \mid \mathbf{c})$. The conditional flow matching objective is then defined as

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{x_0 \sim p_0, x_1 \sim p_1}\left[\|v_t(x_t, \mathbf{c}) - (x_1 - x_0)\|_2^2\right]. \tag{5}$$

During inference, we solve the ordinary differential equation

$$\frac{dx_t}{dt} = v_t(x_t, \mathbf{c}), \tag{6}$$

starting from $x_0 \sim \mathcal{N}(0, \mathbf{I})$ at $t = 0$ and integrating to $t = 1$ with the Euler method over 32 steps to obtain the clean estimate x_1, which is finally reconstructed into a waveform by the vocoder.

The contrastive loss $\mathcal{L}_{\text{InfoNCE}}$ serves as a discriminative constraint for the speaker branch, applied to the feature space mapped by the lightweight adapter after the CAM++ output. Given an anchor sample $\mathbf{z}_a$, its positive sample set is denoted as $\mathbf{z}_p \in \mathcal{P}$ (different distorted speaker embeddings synthesized from the same utterance), and its negative sample set is denoted as $\mathbf{z}_n \in \mathcal{N}$ (random clean speaker embeddings from different speak-

ers). Let $\mathrm{sim}(\cdot,\cdot)$ denote the cosine similarity function and τ the temperature coefficient. The InfoNCE loss is defined as

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\sum_{\mathbf{z}_p \in \mathcal{P}} \exp\left(\mathrm{sim}(\mathbf{z}_a, \mathbf{z}_p)/\tau\right)}{\sum_{\mathbf{z}_p \in \mathcal{P}} \exp\left(\mathrm{sim}(\mathbf{z}_a, \mathbf{z}_p)/\tau\right) + \sum_{\mathbf{z}_n \in \mathcal{N}} \exp\left(\mathrm{sim}(\mathbf{z}_a, \mathbf{z}_n)/\tau\right)}\right]. \quad (7)$$

This loss maximizes the mutual information lower bound between the anchor and positive samples while minimizing the similarity with negative samples, forcing mapped embeddings of utterances from the same speaker to cluster tightly in the feature space and those from different speakers to be pushed apart. This provides a discriminative identity prior for the flow matching process, as illustrated in Figure 2.

In the total loss, the flow matching loss dominates, ensuring that generation quality remains the primary priority. The InfoNCE loss acts as an auxiliary constraint, injecting explicit timbre consistency guidance into the model without interfering with the optimization of the primary task. Through joint optimization of the two losses, the model not only pursues high-fidelity reconstruction but also effectively preserves the target speaker's identity characteristics, jointly suppressing hallucinations from both the generation path and the training objective.

4. Experimental Setup and Results

4.1. Datasets

To construct a diverse training dataset, we integrate multiple public data sources. For speech data, we adopt the VoiceBank corpus and the DNS Challenge 2021 dataset, utilizing both the English and Chinese portions for multilingual training. To simulate complex noise environments, we incorporate environmental noise from the DEMAND dataset and background sounds from the WHAM! dataset, enhancing model robustness under noisy conditions. In addition, we employ open-source room impulse response (RIR) data (OpenSLR26 and OpenSLR28) to generate reverberant speech samples.

In our experiments, we adopt two data construction strategies: (1) directly using the provided pairs of distorted and clean speech in the datasets, and (2) programmatically adding noise and reverberation to clean speech by controlling the signal-to-noise ratio (SNR) and reverberation time to generate distorted samples. Combining these two strategies effectively improves the diversity of the training set.

For evaluation, we use the English and Chinese subsets of the VoiceBank and DNS Challenge 2021 test sets. Since dedicated Chinese and long-utterance (duration > 30 s) test sets are scarce, we construct an additional held-out test set (hereafter, the supplementary test set). To evaluate generalization to unseen speakers and utterances, we perform the data split at the source-data level before synthesis, ensuring that no speaker or utterance in the supplementary test set appears in the training set. The supplementary test set is synthesized using the same noise and reverberation pipeline as the training set; therefore, the training and test data share the same noise and RIR distributions. This supplementary test set accounts for approximately 50% of the total test set and is never used for training or model selection. Table 2 summarizes the dataset configurations.

Table 2. Summary of dataset configurations. Ratio denotes the proportion of each subset within the total training or test duration.

Purpose	Dataset	Noise Source	Duration (Hours)	SNR (dB)	Ratio
Training	VoiceBank + DEMAND Training set	Provided by the official source	23.03	0 to 15	3%

Table 2. *Cont.*

Purpose	Dataset	Noise Source	Duration (Hours)	SNR (dB)	Ratio
Training	DNS Challenge 2021 Training set	DEMAND + WHAM! + RIRs	725.13	−5 to 20	97%
Test	VoiceBank + DEMAND Test set	Provided by the official source	0.41	2.5 to 17.5	10%
Test	DNS Challenge 2021 Test set	Provided by the official source	1.5	−5 to 20	40%
Test	Supplementary	DEMAND + WHAM! + RIRs	1.93	−5 to 20	50%

4.2. Evaluation Metrics

We employ a comprehensive set of metrics to evaluate the proposed model, covering perceptual quality, semantic similarity, signal-level fidelity, and generation efficiency.

DNSMOS (Deep Noise Suppression Mean Opinion Score) is a non-intrusive metric proposed by Microsoft for the DNS Challenge. It predicts the mean opinion score without requiring a clean reference, outputting three scores: SIG (signal quality), BAK (background noise suppression), and OVRL (overall quality). Each score ranges from 1 to 5, with higher values indicating better performance.

PESQ (Perceptual Evaluation of Speech Quality) is a standardized intrusive metric established by ITU-T (P.862). It compares the enhanced speech against a clean reference to predict the subjective MOS, with scores ranging from −0.5 to 4.5, where higher scores indicate better perceptual quality.

STOI (Short-Time Objective Intelligibility) is an objective intelligibility measure ranging from 0 to 1, with higher values indicating better intelligibility. It is derived by computing the correlation between the enhanced and clean speech in the short-time Fourier transform domain.

SECS (Speaker Embedding Cosine Similarity) serves as a direct metric for acoustic hallucination, defined as speaker-identity or timbre drift in the enhanced speech relative to the clean reference. It computes the cosine similarity between speaker embeddings extracted from the enhanced and clean speech using a pretrained ECAPA-TDNN model. The score ranges from −1 to 1, with higher values indicating less acoustic hallucination and better speaker preservation.

WER (Word Error Rate) is a metric for evaluating the preservation of semantic information. It employs the Whisper large-v2 ASR model to compare the transcriptions of enhanced speech against those of the clean reference. Lower WER values indicate better semantic retention. We further decompose WER into substitution (S), deletion (D), and insertion (I) rates. Since linguistic hallucination can manifest as both content insertion and content substitution, we use the combined S+I rate as a proxy for linguistic hallucination event frequency, denoted as the **Hallucination Event Rate (HER)**. HER is an ASR-output-based operational proxy for hallucination, not a human-annotated hallucination rate; its results should therefore be interpreted as relative indicators of hallucination-like errors rather than actual hallucination rates.

Since WER additionally includes deletions, which are omissions rather than hallucinations, WER serves as a broader content-fidelity metric, while HER provides a hallucination-focused refinement of it. We therefore report WER in the main comparison and ablation study, and report HER only on the supplementary synthetic test set, which provides clean references and controllable SNRs. The official VoiceBank + DEMAND test set has a fixed SNR range, and the DNS Challenge 2021 test set is reference-free; therefore, HER is not

reported on these datasets. For the supplementary test set, we report HER at SNRs of -5, 0, and ≥ 5 dB to provide an SNR-stratified analysis of hallucination behavior. This stratification is motivated by the observation that generative models are more prone to hallucination under low-SNR conditions, where degraded inputs provide weaker semantic cues; thus, the -5 dB condition is particularly diagnostic, while 0 and ≥ 5 dB serve as intermediate and relatively clean references.

We quantify computational efficiency using four metrics: FLOPs, total parameters (Params), peak GPU memory (GPU Mem), and real-time factor (RTF). FLOPs are measured for one complete end-to-end inference pass under identical input conditions for all compared methods. Params denote the total number of model parameters. GPU memory is the peak inference memory measured on an NVIDIA A100 GPU under BF16 precision. RTF is defined as inference time divided by audio duration; lower values indicate better efficiency, and RTF < 1 means faster than real time.

4.3. Experimental Configuration

We use 100-dimensional log mel-filterbank features extracted from 16 kHz audio as the input representation, with a window length of 1024, an FFT size of 1024, and a frame shift of 256. Let B and T denote the batch size and the number of mel frames, and let $F = 100$ denote the mel dimension. The notation follows Section 3: M^T and M^N denote the intermediate state and the noisy mel-spectrogram, E^{Se} and E^{Sp} denote the semantic and speaker embeddings, and x_1 and x_0 denote the clean target and the Gaussian noise, respectively.

Semantic branch. We adopt Qwen2-Audio-7B as the semantic feature extractor, which is frozen during training. The hidden states from its last transformer layer have a dimension of 4096. Since the semantic content evolves over time, we preserve the temporal dimension and apply linear interpolation along the time axis to align the hidden states with the mel frame rate. The interpolated features are then projected from 4096 to 512 dimensions via a fully connected layer, yielding $E^{Se} \in \mathbb{R}^{B\times T\times 512}$.

Speaker branch. We employ the CAM++ model to extract speaker embeddings, which is also kept frozen. CAM++ produces a 192-dimensional utterance-level embedding for each utterance. A lightweight adaptation module consisting of a single linear layer projects it to 256 dimensions. Because speaker identity is stable within an utterance, the projected embedding is broadcast along the time axis to obtain $E^{Sp} \in \mathbb{R}^{B\times T\times 256}$. The adaptation module is trainable and is optimized by an InfoNCE loss with temperature $\tau = 0.1$. For each anchor (the projected embedding of a noisy training utterance), we construct the positive set from projected embeddings of different distorted versions of the same utterance, and the negative set from projected embeddings randomly drawn from clean utterances of different speakers.

Conditioning and fusion. The four features $M^T \in \mathbb{R}^{B\times T\times F}$, $M^N \in \mathbb{R}^{B\times T\times F}$, E^{Se}, and E^{Sp} are concatenated along the channel dimension, yielding a tensor of size $B \times T \times (2F + 512 + 256)$. A linear layer projects it to the DiT hidden dimension, followed by a convolutional positional embedding with a residual connection.

DiT backbone. The above features are fed into the DiT network to estimate the velocity field in the flow matching model. The network comprises 16 encoder layers, each containing 16 attention heads, with a hidden dimension of 1024 and a feed-forward network dimension of 2048. A dropout rate of 0.1 is applied to both the attention layers and the feed-forward layers.

Optimization. We train all models with AdamW using a peak learning rate of 7.5×10^{-5}, a linear warm-up over the first 40,000 steps, and a subsequent linear decay. Training is conducted for 20 epochs with a batch size of 2048 on a single NVIDIA A100

GPU. To ensure statistical reliability, each model is trained with 5 different random seeds, and all baselines are trained independently under the same seed set. The test set is fixed across all runs, and we report all main results as mean ± standard deviation over the 5 runs. This protocol ensures a fair comparison across all evaluated methods.

Inference. During inference, we adopt the Euler method as the ordinary differential equation solver, performing 32 integration steps to sample from the noise distribution to the clean data distribution.

The overall training and inference procedure of SS-FMSE is summarized in Algorithm 1.

Algorithm 1: Training and inference of SS-FMSE

Input : Noisy mel M^N, state M^T, clean mel x_1, noise x_0
Output: Enhanced speech waveform
// Training
for *each mini-batch* **do**
  Extract semantic features from frozen Qwen2-Audio-7B;
  Interpolate along time axis and project to 512 dimensions;
  Extract speaker embedding from frozen CAM++ and project to 256 dimensions;
  Broadcast the speaker embedding along the time axis;
  Let $H(M^T) = [M^T, M^N, E^{Se}, E^{Sp}]$ denote the concatenated conditioning features;
  Predict flow field with DiT and compute $\mathcal{L}_{\text{flow}}$;
  Compute $\mathcal{L}_{\text{spk}}$ with InfoNCE;
  Update all trainable parameters;
// Inference
Initialize $x_0 \sim \mathcal{N}(0, I)$;
for $T = 0, \ldots, 1$ *with 32 Euler steps* **do**
  Predict flow field $\hat{v} = \text{DiT}(\text{Proj}(H(M^T)))$;
  Update $x \leftarrow x + \Delta t \cdot \hat{v}$;
Reconstruct waveform via the vocoder;

4.4. Baseline Methods

To facilitate the presentation of the experimental results, we briefly introduce several mainstream speech enhancement models included in our comparison.

MP-SEnet [14] is a time-frequency domain single-channel speech enhancement model whose core innovation lies in the parallel explicit estimation of magnitude and phase spectra. The model adopts a Transformer-based encoder–decoder architecture, achieving state-of-the-art performance on various speech enhancement tasks.

G-MaP-SE [6] is a guided enhancement framework that constructs a speaker embedding prior from clean speech using a Gaussian mixture model (GMM), and refines distorted conditioning embeddings by matching them to this prior. The matched embeddings are then injected into the enhancement backbone via a lightweight gated fusion module without requiring enrollment audio at inference time.

AnyEnhance [15] is a general-purpose generative speech enhancement model based on masked generative models. It introduces a prompt guidance mechanism for in-context learning with reference timbre cues and a self-critique mechanism that iteratively refines the generated output, achieving excellent performance on both objective metrics and subjective listening tests.

FlowSE [4] is a generative speech enhancement method based on flow matching. By learning a continuous transformation between the distributions of distorted and clean speech, it achieves high-fidelity reconstruction while significantly reducing inference latency.

SenSE [5] is a two-stage universal generative speech enhancement framework that leverages a semantic-aware speech language model to extract semantic tokens from distorted speech, and then integrates these semantic priors into a flow-matching-based enhancement process via a semantic guidance mechanism to effectively mitigate semantic ambiguity.

These methods represent state-of-the-art approaches across different speech enhancement paradigms and provide comprehensive baselines for our study.

4.5. Experimental Results

4.5.1. Overall Comparison and Qualitative Analysis

To comprehensively evaluate the performance of SS-FMSE, we compare our method against several mainstream speech enhancement models on a unified test set. To ensure comparability with public benchmarks, we conduct experiments separately on the public test set and the synthetic test set, as shown in Tables 3 and 4. For fair comparison, all baseline methods are retrained using their official code, with the same training data as our method, and evaluated under consistent input conditions and experimental settings. The evaluation metrics include the no-reference speech quality metric DNSMOS (SIG, BAK, OVRL), the reference-based quality metrics STOI and PESQ, the downstream word error rate (WER), and the speaker embedding cosine similarity (SECS). Since the DNS Challenge 2021 test set provides only distorted speech without clean references, reference-based metrics cannot be computed; therefore, we report only no-reference metrics for this test set. Unless otherwise stated, all main results are reported as mean ± standard deviation across 5 random seeds.

Among the baseline methods, MP-SEnet [14], with its PESQ-optimized loss function and explicit phase modeling, achieves the best performance on the objective metrics PESQ and STOI. Nevertheless, its improvement on the BAK (background noise suppression) dimension is relatively limited, indicating insufficient capability in suppressing background noise. G-MaP-SE [6], built upon MP-SEnet, introduces a GMM-based speaker embedding prior matching mechanism. By refining distorted speaker embeddings towards a clean prior, it achieves consistent improvements over MP-SEnet across most metrics, particularly in BAK and OVRL, while also obtaining a higher SECS, demonstrating the effectiveness of speaker identity guidance in enhancing perceptual quality and preserving speaker consistency.

In contrast, generative methods generally suffer from phase mismatch issues, which degrade the alignment between enhanced outputs and reference signals, resulting in a notable gap from MP-SEnet on PESQ and STOI. However, the DNSMOS metrics (SIG, BAK, OVRL) do not strictly depend on phase alignment and better reflect human auditory perception. On these three dimensions, generative methods consistently outperform MP-SEnet. AnyEnhance [15], through its prompt learning and self-critique mechanisms, achieves excellent performance on OVRL and WER. FlowSE [4] adopts flow matching framework as an alternative to iterative diffusion sampling, striking a balance between inference efficiency and generation fidelity. Its performance on most metrics is comparable to that of AnyEnhance, validating the advantages of flow matching in perceptual quality. SenSE [5], building upon FlowSE, further incorporates semantic tokens extracted by a semantic-aware speech language model as guidance. This semantic-aware design effectively mitigates semantic ambiguity and acoustic discontinuities, achieving improvements over FlowSE on both DNSMOS and WER.

Compared with the above methods, our proposed SS-FMSE achieves the best results on SIG, BAK, OVRL, WER, and SECS, validating the effectiveness of the dual guidance mechanism—semantic understanding and speaker perception—in improving perceptual

quality and content intelligibility. The SECS results further confirm that the speaker branch effectively preserves the target timbre, complementing the semantic branch. Notably, SS-FMSE outperforms FlowSE and SenSE on all metrics, demonstrating that the joint guidance from semantic embeddings and speaker identity information mitigates generative hallucinations, significantly enhancing both semantic fidelity and timbre consistency while maintaining generation quality.

Table 3. Performance comparison of different models on the public test sets.

Model/Method	DNS Challenge 2021			VoiceBank + DEMAND						
	SIG	BAK	OVRL	SIG	BAK	OVRL	STOI	PESQ	WER	SECS
Noisy	2.45	2.235	1.968	3.244	3.125	2.695	0.687	1.877	0.095	**0.972**
MP-SEnet [14]	$3.443_{\pm 0.018}$	$3.923_{\pm 0.016}$	$2.969_{\pm 0.021}$	$3.343_{\pm 0.019}$	$3.832_{\pm 0.017}$	$3.179_{\pm 0.023}$	**0.824** $_{\pm 0.007}$	**2.581**$_{\pm 0.031}$	$0.074_{\pm 0.003}$	$0.915_{\pm 0.009}$
G-MaP-SE [6]	$3.424_{\pm 0.016}$	$3.979_{\pm 0.014}$	$3.028_{\pm 0.019}$	$3.419_{\pm 0.017}$	$4.042_{\pm 0.015}$	$3.228_{\pm 0.020}$	$0.795_{\pm 0.006}$	$2.329_{\pm 0.028}$	$0.072_{\pm 0.003}$	$0.949_{\pm 0.008}$
AnyEnhance [15]	$3.542_{\pm 0.015}$	$3.934_{\pm 0.017}$	$3.121_{\pm 0.018}$	$3.426_{\pm 0.016}$	$4.114_{\pm 0.014}$	$3.261_{\pm 0.019}$	$0.768_{\pm 0.008}$	$1.997_{\pm 0.026}$	$0.062_{\pm 0.003}$	$0.931_{\pm 0.009}$
FlowSE [4]	$3.486_{\pm 0.014}$	$4.028_{\pm 0.013}$	$3.212_{\pm 0.016}$	$3.401_{\pm 0.017}$	$4.132_{\pm 0.013}$	$3.229_{\pm 0.020}$	$0.744_{\pm 0.007}$	$2.033_{\pm 0.025}$	$0.072_{\pm 0.003}$	$0.925_{\pm 0.008}$
SenSE [5]	$3.523_{\pm 0.012}$	$4.086_{\pm 0.011}$	$3.283_{\pm 0.014}$	$3.496_{\pm 0.014}$	$4.144_{\pm 0.011}$	$3.312_{\pm 0.014}$	$0.743_{\pm 0.006}$	$2.112_{\pm 0.019}$	$0.054_{\pm 0.002}$	$0.933_{\pm 0.007}$
SS-FMSE	**3.594**$_{\pm 0.008}$	**4.118**$_{\pm 0.007}$	**3.317**$_{\pm 0.010}$	**3.534**$_{\pm 0.009}$	**4.185**$_{\pm 0.007}$	**3.345**$_{\pm 0.011}$	$0.763_{\pm 0.004}$	$2.144_{\pm 0.015}$	**0.049**$_{\pm 0.001}$	$0.962_{\pm 0.005}$

Note: Bold values indicate the best result in each column.

Table 4. Comparison of different models on the synthetic test set.

Model/Method	SIG	BAK	OVRL	STOI	PESQ	WER	SECS
Noisy	2.231	1.868	1.787	0.637	1.432	0.426	**0.921**
MP-SEnet [14]	$3.341_{\pm 0.018}$	$3.437_{\pm 0.016}$	$2.743_{\pm 0.021}$	$0.714_{\pm 0.007}$	$2.182_{\pm 0.031}$	$0.291_{\pm 0.008}$	$0.789_{\pm 0.009}$
G-MaP-SE [6]	$3.384_{\pm 0.016}$	$3.646_{\pm 0.014}$	$2.943_{\pm 0.019}$	**0.718**$_{\pm 0.006}$	**2.192**$_{\pm 0.028}$	$0.251_{\pm 0.007}$	$0.819_{\pm 0.008}$
AnyEnhance [15]	$3.327_{\pm 0.015}$	$3.843_{\pm 0.017}$	$3.081_{\pm 0.018}$	$0.688_{\pm 0.008}$	$1.732_{\pm 0.026}$	$0.263_{\pm 0.006}$	$0.811_{\pm 0.009}$
FlowSE [4]	$3.356_{\pm 0.014}$	$3.834_{\pm 0.013}$	$3.062_{\pm 0.016}$	$0.637_{\pm 0.007}$	$1.643_{\pm 0.025}$	$0.241_{\pm 0.008}$	$0.803_{\pm 0.008}$
SenSE [5]	$3.441_{\pm 0.012}$	$3.932_{\pm 0.011}$	$3.143_{\pm 0.014}$	$0.702_{\pm 0.006}$	$1.882_{\pm 0.022}$	$0.213_{\pm 0.005}$	$0.814_{\pm 0.007}$
SS-FMSE	**3.483**$_{\pm 0.008}$	**4.095**$_{\pm 0.007}$	**3.221**$_{\pm 0.010}$	$0.703_{\pm 0.004}$	$1.963_{\pm 0.016}$	**0.174**$_{\pm 0.003}$	$0.868_{\pm 0.005}$

Note: Bold values indicate the best result in each column.

Figure 3 presents example 100-dimensional mel-spectrograms of different models on the synthetic test set. Panel (a) shows the clean spectrogram, panel (b) shows the noisy spectrogram, and panels (c)–(f) display the outputs of different baseline models and our SS-FMSE. Although MP-SEnet achieves relatively good performance on STOI and PESQ, its spectrograms still contain considerable residual noise, resulting in less satisfactory subjective listening quality compared to other methods. AnyEnhance removes noise effectively, yet suffers from energy loss in certain frequency bands. This is evident in panel (d), where the region marked by the blue box corresponds to the same time-frequency region marked by the blue box in the clean reference panel (a). FlowSE also recovers the audio well, but introduces spurious energy in some frequency regions. This can be seen in panel (e), where the region marked by the green box corresponds to the same time-frequency region marked by the green box in the clean reference panel (a). In contrast, the spectrograms reconstructed by our proposed SS-FMSE most closely resemble the clean reference: background noise is effectively suppressed, the harmonic structure across the full frequency band is restored with greater completeness and continuity, and no obvious spectral artifacts or spurious harmonics are observed.

4.5.2. SNR-Stratified Hallucination Analysis

To examine how generative hallucination behaves under different noise conditions, we stratify the supplementary synthetic test set into three SNR groups: -5, 0, and ≥ 5 dB. This stratification is motivated by the fact that lower SNRs provide weaker semantic cues and thus make generative models more prone to hallucination, manifesting as content insertion or substitution. As shown in Table 5, all models improve as SNR increases, but the gap between SS-FMSE and the baselines is most pronounced at the lowest SNR. In the -5 dB

condition, SS-FMSE achieves the lowest HER and WER, reducing the HER to 16.2%. This suggests that dual guidance is particularly beneficial for reducing hallucination-related errors under severe degradation. At higher SNRs, the performance gap decreases, yet SS-FMSE still achieves the best results.

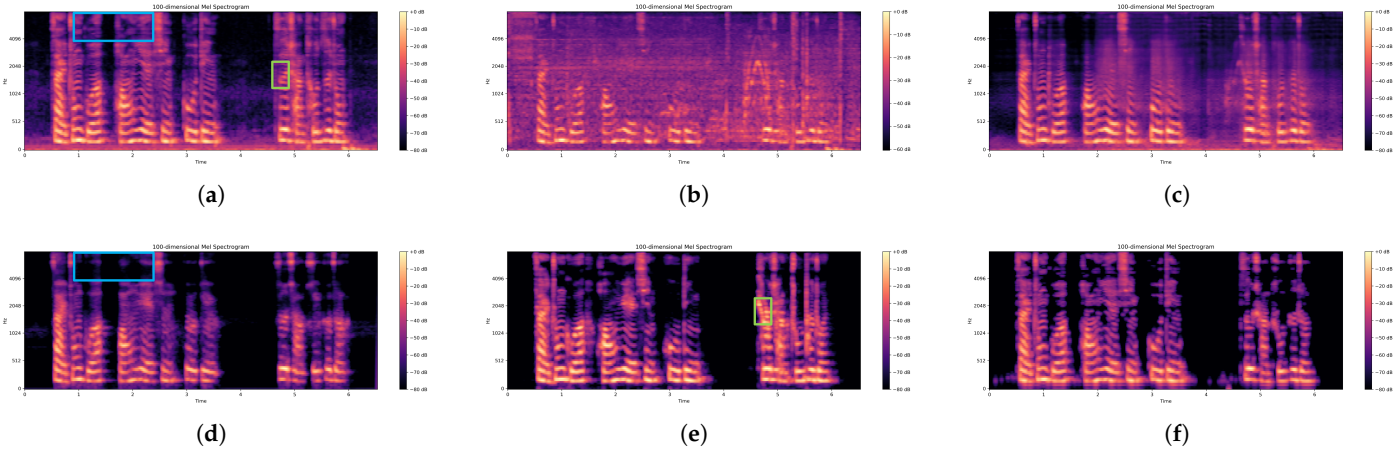


Figure 3. Comparison of mel-spectrograms reconstructed by different generative models on the synthetic test set. (**a**) Clean, (**b**) Distorted, (**c**) MP-SEnet, (**d**) AnyEnhance, (**e**) FlowSE, and (**f**) SS-FMSE (Proposed). Colored boxes mark the same time-frequency regions across panels; no magnification is applied.

Table 5. SNR-stratified performance comparison on the supplementary synthetic test set.

Model/Method	−5 dB			0 dB			≥5 dB		
	SECS	WER	HER	SECS	WER	HER	SECS	WER	HER
Noisy	**0.823**	0.566	0.368	**0.908**	0.352	0.231	**0.917**	0.145	0.085
MP-SEnet [14]	$0.701_{\pm 0.012}$	$0.352_{\pm 0.012}$	$0.254_{\pm 0.011}$	$0.792_{\pm 0.011}$	$0.268_{\pm 0.011}$	$0.192_{\pm 0.010}$	$0.831_{\pm 0.010}$	$0.113_{\pm 0.010}$	$0.076_{\pm 0.009}$
G-MaP-SE [6]	$0.744_{\pm 0.010}$	$0.315_{\pm 0.010}$	$0.231_{\pm 0.009}$	$0.835_{\pm 0.009}$	$0.224_{\pm 0.009}$	$0.164_{\pm 0.008}$	$0.874_{\pm 0.009}$	$0.089_{\pm 0.008}$	$0.062_{\pm 0.007}$
AnyEnhance [15]	$0.726_{\pm 0.011}$	$0.328_{\pm 0.011}$	$0.242_{\pm 0.010}$	$0.823_{\pm 0.010}$	$0.235_{\pm 0.010}$	$0.172_{\pm 0.009}$	$0.871_{\pm 0.009}$	$0.095_{\pm 0.009}$	$0.066_{\pm 0.008}$
FlowSE [4]	$0.712_{\pm 0.009}$	$0.302_{\pm 0.009}$	$0.225_{\pm 0.008}$	$0.801_{\pm 0.008}$	$0.216_{\pm 0.008}$	$0.159_{\pm 0.007}$	$0.872_{\pm 0.008}$	$0.084_{\pm 0.007}$	$0.059_{\pm 0.006}$
SenSE [5]	$0.741_{\pm 0.008}$	$0.276_{\pm 0.008}$	$0.208_{\pm 0.007}$	$0.832_{\pm 0.007}$	$0.189_{\pm 0.007}$	$0.141_{\pm 0.006}$	$0.886_{\pm 0.007}$	$0.072_{\pm 0.006}$	$0.051_{\pm 0.005}$
SS-FMSE	$0.794_{\pm 0.005}$	$\mathbf{0.238}_{\pm 0.005}$	$\mathbf{0.162}_{\pm 0.004}$	$0.881_{\pm 0.004}$	$\mathbf{0.152}_{\pm 0.004}$	$\mathbf{0.104}_{\pm 0.003}$	$0.909_{\pm 0.004}$	$\mathbf{0.058}_{\pm 0.003}$	$\mathbf{0.038}_{\pm 0.002}$

Note: Bold values indicate the best result in each column.

In terms of SECS, SS-FMSE consistently outperforms all enhancement baselines across the three SNR groups, confirming that the speaker branch suppresses acoustic hallucination, i.e., timbre drift, even under strong noise. The SECS gap is also largest at −5 dB, indicating that speaker guidance is particularly effective when the input is severely degraded. Note that the distorted input itself yields a high SECS, since the degradation process mainly affects the acoustic environment rather than the speaker identity; SECS therefore specifically reflects whether the enhancement process introduces speaker-level distortion, i.e., acoustic hallucination.

Overall, the SNR-stratified results show that SS-FMSE maintains consistent improvements across noise levels and specifically reduces hallucination-related errors in low-SNR scenarios.

4.5.3. Ablation Study

To investigate the impact of each module on various speech quality evaluation metrics and speaker consistency, we conduct ablation experiments on the synthetic test set. As shown in Table 6, Method A (flow matching only) performs the weakest across all metrics, with a particularly high WER of 24.1%, indicating that generative models without high-level guidance suffer from significant deficiencies in content intelligibility.

For the speaker branch, we compare two configurations: B1 adds a speaker embedding with a lightweight adapter but without InfoNCE, while B2 further applies the InfoNCE

regularization to form the full speaker branch. As shown in Table 6, B1 improves over A on all DNSMOS dimensions and SECS, suggesting that speaker identity constraints guide the model toward more consistent target timbre. However, its WER reduction is limited (from 24.1% to 23.3%), implying that timbre constraints alone are insufficient to fully address linguistic hallucinations. In contrast, B2 achieves consistent improvements over B1 across all metrics, especially on SECS (from 0.829 to 0.849) and WER (from 23.3% to 22.1%). This confirms that InfoNCE enhances intra-class compactness and inter-class separability of speaker characteristics and contributes independently to hallucination suppression. Note that the adapter is required for dimension alignment before feature concatenation and is therefore included in both B1 and B2; the comparison between them isolates the effect of InfoNCE.

Table 6. Ablation study results. "Full Speaker Branch" denotes the complete speaker branch with adapter and InfoNCE, and "Full Model" denotes the full dual-guidance model combining the semantic branch and the complete speaker branch.

Model/Method	Embedding Info	InfoNCE	SIG	BAK	OVRL	STOI	PESQ	WER	SECS
A. Flow Matching	/	/	$3.356_{\pm 0.014}$	$3.834_{\pm 0.013}$	$3.062_{\pm 0.016}$	$0.637_{\pm 0.007}$	$1.643_{\pm 0.025}$	$0.241_{\pm 0.008}$	$0.803_{\pm 0.008}$
B1. + Speaker Branch (only Adapter)	Speaker	/	$3.372_{\pm 0.018}$	$3.856_{\pm 0.016}$	$3.094_{\pm 0.021}$	$0.652_{\pm 0.007}$	$1.663_{\pm 0.028}$	$0.233_{\pm 0.005}$	$0.829_{\pm 0.009}$
B2. + Full Speaker Branch	Speaker	✓	$3.441_{\pm 0.015}$	$3.937_{\pm 0.013}$	$3.192_{\pm 0.018}$	$0.694_{\pm 0.006}$	$1.742_{\pm 0.025}$	$0.221_{\pm 0.004}$	$0.849_{\pm 0.008}$
C. + Semantic Branch	Semantic	/	$3.397_{\pm 0.017}$	$3.943_{\pm 0.015}$	$3.143_{\pm 0.020}$	$0.696_{\pm 0.006}$	$1.835_{\pm 0.027}$	$0.185_{\pm 0.004}$	$0.813_{\pm 0.009}$
D. Full Model	Semantic, Speaker	✓	$\mathbf{3.483}_{\pm 0.009}$	$\mathbf{4.095}_{\pm 0.007}$	$\mathbf{3.221}_{\pm 0.011}$	$\mathbf{0.703}_{\pm 0.004}$	$\mathbf{1.963}_{\pm 0.016}$	$\mathbf{0.174}_{\pm 0.002}$	$\mathbf{0.868}_{\pm 0.005}$

Note: Bold values indicate the best result in each column.

Method C introduces semantic embeddings and substantially reduces WER to 18.5%, confirming the core role of the semantic branch in enhancing content intelligibility. Steady improvements are also observed across all DNSMOS dimensions. Nevertheless, its SECS is only 0.813, lower than that of B2, indicating that semantic priors alone cannot maintain timbre consistency. Method D, the full model with dual guidance, achieves the best results across all metrics, attaining the highest PESQ and STOI scores, leading all methods on the three DNSMOS metrics, and further reducing WER to 17.4% while achieving the highest SECS of 0.868. These results demonstrate the synergistic effect of dual semantic and speaker guidance: the semantic branch ensures the plausibility of generated content, while the complete speaker branch with InfoNCE maintains timbre consistency. The two branches complement each other, jointly suppressing generative hallucinations from both content and identity perspectives.

4.5.4. Sensitivity Analysis of InfoNCE Loss Weight

The InfoNCE loss weight λ_{InfoNCE} controls the strength of the contrastive regularization on the speaker branch. To examine its impact, we vary λ_{InfoNCE} over $\{0.01, 0.05, 0.1, 0.2, 0.5, 0.9\}$ while keeping all other settings identical to the full model. We report SECS and WER to measure speaker consistency and semantic fidelity, respectively, and OVRL as an overall perceptual quality indicator. The results are shown in Table 7.

A very small weight (e.g., 0.01) barely activates the contrastive regularization, yielding the lowest SECS and relatively high WER. As λ_{InfoNCE} increases to 0.1, SECS improves steadily and both WER and OVRL reach their best values, indicating that a moderate contrastive strength effectively compacts the speaker embedding space and improves overall training stability, which also benefits content preservation. Interestingly, SECS reaches its maximum at $\lambda_{\text{InfoNCE}} = 0.2$, slightly higher than at 0.1, but this comes at the cost of a higher WER and a lower OVRL. This suggests a mild trade-off between speaker consistency and semantic fidelity: once the contrastive objective becomes dominant, the model tends to over-emphasize timbre consistency at the expense of content preservation.

When λ_{InfoNCE} is further increased to 0.5 and 0.9, the performance degrades sharply, with WER rising to 0.418 and OVRL dropping to 2.934 at $\lambda = 0.9$. This suggests that the speaker branch should serve as an auxiliary constraint rather than dominate the training objective, since an overly strong contrastive signal interferes with the optimization of the primary denoising objective, ultimately degrading generation fidelity. We therefore adopt $\lambda_{\text{InfoNCE}} = 0.1$ in all experiments, as it achieves the best balance among speaker consistency, semantic fidelity, and overall perceptual quality.

Table 7. Sensitivity analysis of the InfoNCE loss weight λ_{InfoNCE}.

λ_{InfoNCE}	0.01	0.05	0.1	0.2	0.5	0.9
OVRL	$3.135_{\pm 0.021}$	$3.172_{\pm 0.018}$	$\mathbf{3.221}_{\pm 0.010}$	$3.122_{\pm 0.015}$	$3.079_{\pm 0.023}$	$2.934_{\pm 0.028}$
WER	$0.189_{\pm 0.006}$	$0.182_{\pm 0.005}$	$\mathbf{0.174}_{\pm 0.003}$	$0.189_{\pm 0.004}$	$0.295_{\pm 0.009}$	$0.418_{\pm 0.013}$
SECS	$0.816_{\pm 0.012}$	$0.849_{\pm 0.009}$	$0.868_{\pm 0.005}$	$\mathbf{0.876}_{\pm 0.004}$	$0.861_{\pm 0.006}$	$0.826_{\pm 0.011}$

Note: Bold values indicate the best result in each row.

4.5.5. Computational Efficiency and Quality Trade-Off

Table 8 reports the quality and computational efficiency of SS-FMSE and the baselines. All efficiency metrics are measured end-to-end on the complete model under identical input and hardware conditions. SS-FMSE provides a controllable quality–efficiency trade-off through the number of ODE steps: more steps consistently improve enhancement quality, while FLOPs and RTF also increase. The total parameter count and peak GPU memory remain unchanged across step counts. The inference cost consists of a fixed part from the frozen condition encoders, i.e., the Qwen2-Audio-7B semantic encoder and the CAM++ speaker encoder together with their projection layers, which run once per utterance and are independent of the ODE step count, and a variable part from repeated DiT evaluations. The total parameter count of about 7.4B is dominated by the frozen Qwen2-Audio-7B encoder, while the trainable parameters (DiT backbone, projection layers, and speaker adapter) amount to approximately 400 M. As the ODE steps increase, RTF rises from 0.54 to 1.17. Compared with FlowSE, SS-FMSE achieves better quality at sufficient steps at the cost of a larger model. Compared with SenSE, SS-FMSE achieves better quality at 32 steps with a slightly higher RTF, but at 16 steps it already attains comparable quality while being faster, and the additional overhead mainly stems from the frozen encoder that does not scale with the number of ODE steps. Overall, SS-FMSE is not limited to a single operating point and can be adapted to different latency and quality requirements.

Table 8. Quality and computational efficiency comparison.

Model/Method	OVRL	WER	SECS	FLOPs (G)	Params (M)	GPU Mem (GB)	RTF
MP-SEnet [14]	$2.743_{\pm 0.021}$	$0.291_{\pm 0.008}$	$0.749_{\pm 0.009}$	49.98	**4.26**	**6.3**	**0.14**
G-MaP-SE [6]	$2.943_{\pm 0.021}$	$0.251_{\pm 0.007}$	$0.809_{\pm 0.008}$	59.43	5.02	6.6	0.16
AnyEnhance [15]	$3.081_{\pm 0.021}$	$0.263_{\pm 0.006}$	$0.811_{\pm 0.009}$	328.29	122.65	10.4	0.27
FlowSE [4]	$3.062_{\pm 0.016}$	$0.241_{\pm 0.008}$	$0.803_{\pm 0.008}$	384.91	334.44	12.1	0.51
SenSE [5]	$3.143_{\pm 0.014}$	$0.213_{\pm 0.005}$	$0.814_{\pm 0.007}$	814.92	1763.37	18.4	1.03
SS-FMSE (ODE = 4)	$2.847_{\pm 0.013}$	$0.252_{\pm 0.003}$	$0.811_{\pm 0.007}$	1078.34	7423.12	25.5	0.54
SS-FMSE (ODE = 8)	$3.043_{\pm 0.012}$	$0.221_{\pm 0.002}$	$0.835_{\pm 0.004}$	1122.47	7423.12	25.5	0.64
SS-FMSE (ODE = 16)	$3.113_{\pm 0.013}$	$0.209_{\pm 0.003}$	$0.849_{\pm 0.003}$	1241.23	7423.12	25.5	0.82
SS-FMSE (ODE = 32)	$\mathbf{3.221}_{\pm 0.010}$	$\mathbf{0.174}_{\pm 0.003}$	$\mathbf{0.868}_{\pm 0.005}$	1480.31	7423.12	25.5	1.17

Note: Bold values indicate the best result in each column.

5. Conclusions

To address the problem of generative hallucinations in speech enhancement caused by the lack of high-level priors, this paper proposes SS-FMSE, a multi-task flow matching

speech enhancement method with dual guidance from semantic understanding and speaker perception. The method adopts flow matching as the generative backbone and introduces two auxiliary guidance branches—semantic and speaker—on top of acoustic features.

Experimental results demonstrate that SS-FMSE achieves the highest DNSMOS and SECS among all enhancement methods and the lowest WER on both test sets. Within the generative category, it also achieves the best PESQ and the best or comparable STOI. The SNR-stratified analysis using the HER proxy further shows that, at -5 dB, SS-FMSE achieves both the lowest HER (16.2%) and the highest SECS among all enhancement methods. This suggests that SS-FMSE reduces linguistic and acoustic hallucination-related errors under severe noise.

Although SS-FMSE achieves a favorable balance between generation quality and identity fidelity, there remains room for further improvement in computational efficiency, guidance richness, acoustic generalization, and data efficiency. In particular, although separation at both the speaker and utterance levels was enforced between the training and test sets, the noise recordings and RIRs were not split disjointly, so generalization to unseen acoustic conditions has not been evaluated. Future work will address this limitation via disjoint noise/RIR splits and cross-dataset evaluation. In addition, more lightweight semantic extraction schemes, such as distillation or lightweight audio encoders, could be explored to reduce computational overhead; richer guidance dimensions, such as emotion and prosody, could be incorporated to improve controllability and expressiveness; and self-supervised or unsupervised adaptation could be integrated to reduce the reliance on paired training data.

Author Contributions: B.D. and X.Z. (Xia Zou) proposed the idea and designed the framework. B.D. performed the experiments and wrote the draft. Y.L. and C.J. contributed to data preparation and validation. M.S., Y.X. and X.Z. (Xiongwei Zhang) supervised the research and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) under Grant No. 62371469.

Data Availability Statement: The data used in this study are publicly available from the following repositories: VoiceBank corpus (https://datashare.ed.ac.uk/handle/10283/1942), DNS Challenge 2021 (https://github.com/microsoft/DNS-Challenge), DEMAND (https://data-staging.niaid.nih.gov/resources?id=zenodo_1227120), WHAM! (http://wham.whisper.ai), OpenSLR26 (https://www.openslr.org/26/), and OpenSLR28 (https://www.openslr.org/28/) (accessed on 29 September 2026). The supplementary test set was synthesized from held-out source data in the DNS Challenge 2021 corpus using the same noise and reverberation pipeline as the training set. The synthesized supplementary test set is available from the corresponding author upon reasonable request. No new primary data were created.

Conflicts of Interest: The funders had no role in the design of the study; the collection, analysis, or interpretation of data; the writing of the manuscript; or the decision to publish the results.

References

1. Bie, X.; Leglaive, S.; Alameda-Pineda, X.; Girin, L. Unsupervised speech enhancement using dynamical variational autoencoders. *IEEE ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2993–3007. [CrossRef]
2. Pascual, S.; Bonafonte, A.; Serrà, J. SEGAN: Speech Enhancement Generative Adversarial Network. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden, 20–24 August 2017; pp. 3642–3646.
3. Lu, Y.J.; Tsao, Y.; Watanabe, S. A Study on Speech Enhancement Based on Diffusion Probabilistic Model. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 659–666.

4. Wang, Z.; Liu, Z.; Zhu, X.; Zhu, Y.; Liu, M.; Chen, J.; Xiao, L.; Weng, C.; Xie, L. FlowSE: Efficient and High-Quality Speech Enhancement via Flow Matching. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Rotterdam, The Netherlands, 17–21 August 2025; pp. 4858–4862.
5. Li, X.; Xie, H.; Wang, Z.; Zhang, Z.; Xiao, L.; Wang, S.; Xie, L. SenSE: Semantic-Aware High-Fidelity Universal Speech Enhancement. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Bangkok, Thailand, 5–9 July 2026.
6. Zhu, Y.; Wang, Z.; Liu, Z.; Li, X.; Chen, Z.; Xia, X.; Huang, C.; Xie, L. G-MaP-SE: Guided Speech Enhancement via GMM-Based Prior Matching. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Sydney, Australia, 27 September–1 October 2026.
7. Lipman, Y.; Chen, R.T.Q.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow Matching for Generative Modeling. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
8. Yuan, Y.; Liu, X.; Liu, H.; Plumbley, M.D.; Wang, W. FlowSep: Language-Queried Sound Separation with Rectified Flow Matching. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1961–1965.
9. Jung, C.; Lee, S.; Kim, J.H.; Chung, J.S. FlowAVSE: Efficient Audio-Visual Speech Enhancement with Conditional Flow Matching. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Kos, Greece, 1–5 September 2024; pp. 2210–2214.
10. Zhang, P.; He, W.; Liu, M.; Xiao, W.; Zou, S.; Li, Y.; Wang, X.; Zhang, G.; Liu, Y.; Dai, W.; et al. FUSE: Fine-Grained and Semantic-Aware Learning for Unified Image Understanding and Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 28355–28363.
11. Chang, Y.; Yao, Y.; Qin, T.; Wang, M.; Tsang, I.W.; Dai, G. Instructing Text-to-Image Diffusion Models via Classifier-Guided Semantic Optimization. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, Canada, 16–22 August 2025; pp. 747–755.
12. Ren, Y.; Li, C.; Xu, M.; Liang, W.; Gu, Y.; Chen, R.; Yu, D. STA-V2A: Video-to-Audio Generation with Semantic and Temporal Alignment. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
13. Li, S.; Mao, Q.; Shi, J. PL-TTS: A Generalizable Prompt-based Diffusion TTS Augmented by Large Language Model. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Kos, Greece, 1–5 September 2024; pp. 4888–4892.
14. Lu, Y.X.; Ai, Y.; Ling, Z. Explicit Estimation of Magnitude and Phase Spectra in Parallel for High-Quality Speech Enhancement. *Neural Netw.* **2023**, *189*, 107562.
15. Zhang, J.; Yang, J.; Fang, Z.; Wang, Y.; Zhang, Z.; Wang, Z.; Fan, F.; Wu, Z. AnyEnhance: A Unified Generative Model With Prompt-Guidance and Self-Critic for Voice Enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 3085–3098. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.