Listening Like Humans: Semantics-Guided Noise-Robust Multimodal Speech Recognition

Yan Fang[†], Jun Chen[†], Yian Yao[‡], Shuxin Zhong[†], Min Sun[◇], Kaishun Wu[†*]

[†]Hong Kong University of Science and Technology (Guangzhou),
[‡]Sun Yat-sen University,
[◇]China Mobile Information Technology Co., Ltd

Abstract

Severe acoustic degradation is often caused by overlapping noise, disfluencies, and environmental distortions. This phenomenon results in the dissolution of linguistic structures and the generation of unreliable ASR outputs. Inspired by human speech comprehension, we propose `Speech-MLM`, a novel multimodal framework that reframes ASR as semantics-guided speech reconstruction. This perspective introduces three core challenges: **(C1)** collapse of linguistic structure under acoustic degradation, **(C2)** semantic ambiguity under noise, and **(C3)** misalignment across modalities. To address these issues, we propose `Speech-MLM`, a multimodal ASR framework that integrates speech, spectrogram-derived visual cues, and textual variants to enhance robustness. It consists of: (i) *Cognitive Structure Extractor* that recovers prosodic structure from visualized acoustic features, (ii) *Semantic Weaver* that learns semantic equivalence across varied textual forms, and (iii) *Retrieval-Guided Fusion Learner* that unifies modalities within a shared semantic space. Experiments on multiple real-world noisy datasets demonstrate that `Speech-MLM` achieves an average 38.85% reduction in WER, while also attaining 98.71% BERTScore and 96.7% USE, over advanced baselines, demonstrating substantial gains in semantic robustness and generalization across domains.

1 Introduction

In real-world scenarios, speech rarely occurs under clean or controlled conditions. Instead, it unfolds across diverse and unpredictable noise environments shaped by speaker variation, environment diversity, and complex environment(Mujtaba et al., 2024a). The mismatch between training environments and real-world variability gives rise to systematic biases in ASR and has the capacity to directly influence legal(Wang et al., 2024b), medical(Miner et al., 2020) and academic(Ezema et al., 2025) outcomes. This engenders critical concerns regarding the reliability and fairness of ASR(Koenecke et al., 2024; Mujtaba et al., 2024b). Thus, eliminating bias from structurally degraded or domain-shifted speech is not only a technical concern—it is a matter of social responsibility.

Recent years have seen considerable progress in the field of ASR, due to the development of LLMs(Hu et al., 2024a). In particular, a range of industrial-grade ASR models pretrained on large-scale corpora—such as OpenAI Whisper(Radford et al., 2023), Alibaba Qwen2-Audio(Chu et al., 2024), and ByteDance Seed-ASR(Bai et al., 2024)—have been released to the public, accelerating research and deployment. Pretraining and data augmentation strategies have further improved the robustness of models to speaker variability and linguistic diversity(Wang et al., 2024a). Despite recent progress, current systems remain ill-equipped to address semantic degradation in complex acoustic environment—where noise distorts not only signal clarity but also meaning interpretability.

In stark contrast, human listening is not a transcription task but a cognitively driven reconstruction process. When faced with noisy and ambiguous signals, humans integrate multi-dimensional cues—phonetic expectations, syntactic structures, and real-world context—to reconstruct the intended meaning(Li and Zhang, 2024). Inspired by this capability, we propose to reframe noise-robust ASR from a transcription problem to a task of semantics-guided speech reconstruction. However, acoustic signals alone often provide insufficient information for such inference, particularly in heavy noise conditions. To address this, we incorporate cross-modal contextual signals to facilitate semantic reasoning under uncertainty, enabling the model to "fill in the gaps" just as a human listener would.

However, implementing this reconstruction paradigm poses three key challenges: Firstly, **Col-**

*Corresponding authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16079–16093
July 2-7, 2026

lapse of Linguistic Structure (C1). In noisy environments, the acoustic cues for linguistic structure (e.g. prosodic boundaries) are lost. Secondly, **Semantic Ambiguity under Noise (C2)**. Noise often obscures important lexical items, creating significant semantic ambiguity. Thirdly, **Misalignment of Semantics Across Modality (C3)**. The reconstruction process relies on combining cues from different modalities (e.g. audio patterns, visual spectrograms and textual information). However, these cues are often temporally misaligned or offer conflicting information in the presence of noise.

To address these challenges, we introduce `Speech-MLM`, a novel framework that operationalizes semantics-guided reconstruction through three synergistic components: To Reconstruct Linguistic Structure **(C1)**, the *Cognitive Structure Extractor* treats the spectrogram as a visual landscape. Instead of relying on fragile acoustic cues, it identifies stable, visually-salient patterns corresponding to linguistic structure (e.g., phonemic emphasis). This process constructs a robust structural backbone for the degraded speech, providing the necessary scaffolding for downstream semantic inference. To Resolve Semantic Ambiguity **(C2)**, the *Semantic Weaver* leverages paraphrased textual priors to build a noise-invariant semantic anchor. By learning to abstract the core meaning from multiple phrasings of the same idea, it enables the model to recover the intended message even when specific keywords are acoustically corrupted or lost in the noise. To fuse Misaligned Cross-Modal Cues **(C3)**, the *Retrieval-Guided Fusion Learner* projects the audio, visual, and textual modalities into a unified semantic space. Critically, the model does not rely on perfect alignment across modalities. Instead, it first handles modality-specific failures—such as acoustic dropouts or missing visual cues—by selectively retrieving the most reliable information available. This allows the model to reconcile conflicting information and achieve a coherent, resilient transcription that is robust to modality-specific failures like acoustic dropouts or textual hallucinations.

Our main contributions are listed as follows:

- To the best of our knowledge, this is the first attempt at reframing ASR from direct signal-to-text transcription to cognitively inspired semantic reconstruction. This provides a new lens to understand and tackle the problem of speech recognition under uncertainty.
- We propose `Speech-MLM`, a novel multimodal framework that systematically addresses the core challenges of semantic reconstruction. It uniquely integrates a *Cognitive Structure Extractor* to rebuild linguistic hierarchy from visual spectrogram cues, a *Semantic Weaver* to infer meaning from abstract textual priors, and a *Retrieval-Guided Fusion Learner* to unify misaligned cross-modal information.
- Experiments demonstrate that our model achieves a new state-of-the-art on diverse noisy speech benchmarks. `Speech-MLM` significantly improves recognition accuracy (e.g., reducing WER by up to 38.85% over strong baselines like Whisper), validating the superiority of the semantic reconstruction paradigm.

2 Related Work

2.1 Audio LLMs

Audio LLMs predominantly encompass two distinct paradigms. Discrete approaches convert raw audio into symbolic sequences. Models such as HuBERT(Hsu et al., 2021) and w2v-BERT(Chung et al., 2021) are employed to transform speech waveforms into discrete semantic units. Representative systems include AudioPaLM(Rubenstein et al., 2023) and SpeechGPT(Zhang et al., 2023). Continuous approaches utilize pre-trained audio encoders to extract continuous embeddings from audio waveforms. This design retains fine-grained acoustic and prosodic information while allowing flexible fusion with textual inputs. Notable examples include Qwen2-Audio(Chu et al., 2024) and Seed-ASR(Bai et al., 2024), which demonstrate strong performance in tasks involving speech-text reasoning and generation.

2.2 Noise-robust ASR

Noise-robust ASR systems have made initial progress, primarily through the mapping of noisy speech features to a 'cleaner' representation space prior to recognition(Mehrish et al., 2023). Speech enhancement methods, for instance, improve speech quality by reducing background noise through front-end modules(Fu et al., 2021); domain adversarial training is dedicated to learning speech representations that are insensitive to noise variations(Prasad et al., 2021); and some of the most recent ASR foundation model rely on large-scale data to enhance their adaptability to real speech scenes(Koenecke et al., 2024). Alternatively, ASR

candidate hypotheses are denoised at the semantic level in the hidden space, thereby guiding the model to identify and rectify potential errors in the expression recovery process(Hu et al., 2024b). However, their robustness in complex noisy environments still faces significant challenges.

2.3 Multimodal Learning in Audio

Multimodal learning has emerged as a promising direction for improving ASR robustness. Previous studies have investigated the visual underpinnings of speech through the use of lip reading or facial dynamics(Wang et al., 2024c). However, these methodologies predominantly depend on video input, restricting their applicability to audio-only scenarios. Recent studies use spectrogram-based visual features to improve speech representation learning, treating the time-frequency structure as an image processed by a visual encoder(Gong et al., 2022). Concurrently, cross-modal learning has achieved progress(Zhang et al., 2024; Lin et al., 2025) through retrieval augmented architectures and shared semantic spaces, to align spoken and written modalities, yet without addressing issues of expression variability or structural degradation.

3 Methodology

3.1 Problem Definition

We formulate ASR in noisy environments as a cognitively inspired process of structure and semantics-guided speech reconstruction, grounded in the integration of linguistic priors extracted from cross-modal cues under uncertainty. Let $\mathbf{X}_{\text{clean}} \in \mathbb{R}^T$ denote the clean speech waveform, and $\mathbf{Y}_{\text{true}} \in \mathbb{R}^L$ the corresponding ground-truth transcription, where T and L represent the lengths of the input and output sequences, respectively. In real-world scenarios, what we observe is a noisy speech signal $\tilde{\mathbf{X}} = \mathcal{N}(\mathbf{X}_{clean})$, where $\mathcal{N}(\cdot)$ represents an unknown real-world corruption noise (e.g., additive background noise, clipping, speaker accents). Our objective is to train an ASR model F_θ that predicts a robust transcription $\hat{\mathcal{Y}}$ given the noisy input $\tilde{\mathcal{X}}$, leveraging multimodal semantic priors:

$$\hat{\mathbf{Y}} = F_\theta(\tilde{\mathbf{X}}; \theta_{(E_{\text{audio}}, E_{\text{textual}}, E_{\text{visual}})}), \quad (1)$$

where θ represents the trainable parameters, and E_{audio}, E_{textual}, E_{visual} are modality-specific embedding that simulate human perceptual processing by extracting complementary features from textual, audio, and visual views of the degraded input.

3.2 Overall Framework

`Speech-MLM` provides a cognitively inspired framework for robust speech recognition under diverse and unpredictable noisy conditions. The architecture consists of three synergistic modules, as illustrated in Figure 1:

(i) Cognitive Structure Extractor: Extracts basic audio scaffolds from visual representations of speech, simulating human attention to discourse boundaries, and prosodic contours.

(ii) Semantic Weaver: Generates and encodes semantically consistent paraphrases to model the diversity of linguistic expressions, enabling robust semantic inference under ambiguity.

(iii) Retrieval-Guided Fusion Learner: Projects audio, visual, and textual representations into a unified latent space, enabling effective modality fusion and enhancing transcription robustness through multimodal alignment.

3.3 Cognitive Structure Extractor

In noisy environments, the literal content of speech (e.g. the specific representation of phonemes) may be obscured, but the contextual semantic relationships still exist. *Cognitive Structural Extractor* adaptively converts the input speech sequence $\tilde{\mathbf{X}}$ into an image representation and extracts underlying structural features, such as phoneme composition and resonance peak orientation, through a JEPA-based visual encoder(Assran et al., 2025). JEPA uses contextual embedding to predict the abstract semantic vectors of the current block. This makes the model robust to semantically consistent but superficially different spectral segments in noisy environments. The process is in Two steps:

- **Basic Linguistic Scaffolds Encoding**: The short-time Fourier transform (STFT) extracts frequency components from raw input $\tilde{\mathbf{X}} \in \mathbb{R}^T$, the log mel spectrogram is computed as:

$$\mathbf{Mel} = \log\left(\mathbf{S} \cdot |\text{STFT}(\tilde{\mathbf{X}})|^2 + \epsilon\right), \quad (2)$$

  where $\mathbf{Mel} \in \mathbb{R}^{H \times W}$, $\mathbf{S}$ is the Mel filter bank matrix, and ϵ is a small constant for numerical stability. To capture prosodic patterns at multiple linguistic timescales, two parallel dilated convolutional branches are applied to extract phoneme-level and syllable-level features respectively. Let $\text{Conv}_{\text{ph}}(\cdot)$ and $\text{Conv}_{\text{sy}}(\cdot)$ denote dilated convolutional layers with kernel sizes k_{ph}

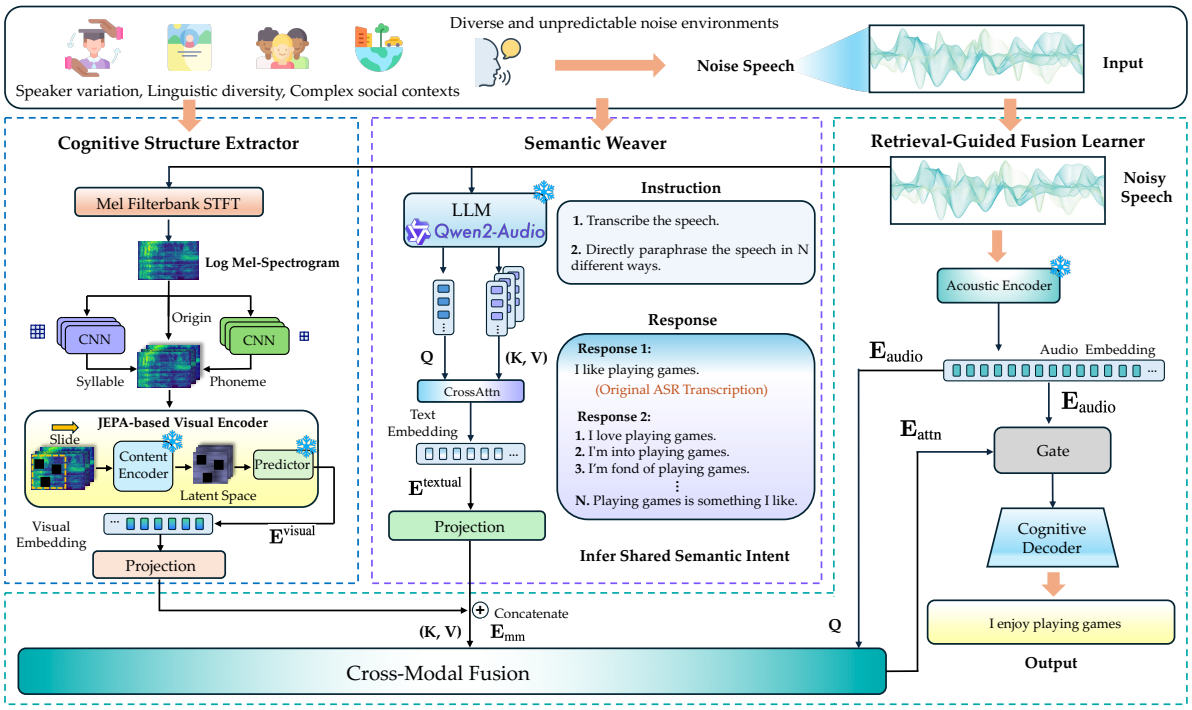

Figure 1: The Framework of Speech-MLM.

and k_{sy} corresponding to the average temporal span of phonemes and syllables ($k_{\text{ph}} < k_{\text{sy}}$):

$$\mathbf{Mel}_{\text{ph}} = \text{Conv}_{\text{ph}}(\mathbf{Mel}), \tag{3}$$

$$\mathbf{Mel}_{\text{sy}} = \text{Conv}_{\text{sy}}(\mathbf{Mel}). \tag{4}$$

To retain the original spectral structure and facilitate information fusion, the original $\mathbf{Mel}$ is also preserved as a base channel. The three spectrograms—original, phoneme-level, and syllable-level—are concatenated along the channel dimension to form a 3-channel visual representation:

$$\mathbf{Mel}_{+} = \text{Concat}(\mathbf{Mel}, \mathbf{Mel}_{\text{ph}}, \mathbf{Mel}_{\text{sy}}). \tag{5}$$

This multi-resolution spectrographic image ($\mathbf{Mel}_{+} \in \mathbb{R}^{3\times H'\times W'}$) serves as the input to downstream structural modeling modules.

- **JEPA-based Visual Encoder**: To extract structural cues from the spectral representation, we segment the 3-channel spectrogram $\mathbf{Mel}_{+}$ along the temporal dimension into n non-overlapping frames, yielding a sequence of spectro-temporal patches:

$$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}, \quad \mathbf{v}_i \in \mathbb{R}^{3\times H'\times w}, \tag{6}$$

where w is the window width in time. Each frame $\mathbf{v}_i$ is fed into a pretrained visual encoder f_{JEPA}, which is designed to learn hierarchical representations from phonetic to semantic levels by modeling the predictive relationships across local contexts.

$$\mathbf{E}^{i}_{\text{visual}} = f_{\text{JEPA}}(\mathbf{v}_i) \in \mathbb{R}^{d}, \tag{7}$$

The visual encoder extracts representations $\{\mathbf{E}^{i}_{\text{visual}}, \ldots, \mathbf{E}^{n}_{\text{visual}}\}$ that reflect not only phoneme- and syllable-level structrue, but also abstract semantic content relevant to speech recovery. Unlike contrastive objectives or autoencoding, our encoder employs predictive coding to estimate latent states of masked patches $(i_1, \ldots, i_M)$ conditioned on visible ones $(j_1, \ldots, j_N)$. The context encoder E_θ processes the visible patches to obtain latent representations:

$$z_j = E_\theta(v_j) \tag{8}$$

The target encoder $E_{\bar{\theta}}$, whose parameters are updated via an exponential moving average (EMA) of E_θ, produces the latent representations of the masked patches:

$$z_i = E_{\bar{\theta}}(v_i) \tag{9}$$

The predictor P_ϕ takes the visible representations z_j together with positional encoding Δp_{ij} to predict the latent state of each masked patch $\hat{z}_i$. The training objective minimizes the squared

ℓ_2 distance between predicted and target latent representations:

$$L_{\text{JEPA}} = \frac{1}{M} \sum_{k=1}^{M} \| P_\phi(z_j, \Delta p_{i_k,j}) - z_{i_k} \|_2^2 \quad (10)$$

This reduces sensitivity to noise perturbations and provides a robust representation of semantically consistent, yet superficially different, spectral fragments. This process yields basic linguistic scaffolds embeddings that serve as input to downstream semantic fusion and decoding modules.

3.4 Semantic Weaver

To enable the system to infer shared semantic intent across diverse surface expressions, we introduce the *Semantic Weaver*, a language-centric module that models the symbolic, sequential, and syntactic regularities underlying paraphrastic variation. Given a noisy speech segment $\tilde{\mathcal{X}}$, we first employ a pre-trained instruction-tuned audio-language model $\mathcal{A}_{\text{gen}}$ (Qwen2-Audio-7B-Instruct) to generate a set of semantically aligned paraphrases:

$$\{\mathbf{Y}_{trans}^{(1)}, \mathbf{Y}_{trans}^{(2)}, \ldots, \mathbf{Y}_{trans}^{(N)}\} = \mathcal{A}_{\text{gen}}(\tilde{\mathcal{X}}) \,. \quad (11)$$

Each $\mathbf{Y}_{trans}^{(i)}$ is a plausible transcription that preserves the core intent of $\tilde{\mathcal{X}}$, but may differ in lexical choices, syntactic structure, or pragmatic framing. The module then extract a shared semantic embedding by applying a token-level cross-attention mechanism between the original ASR transcription $\hat{\mathbf{Y}}_{trans}$ of $\tilde{\mathcal{X}}$ and the generated paraphrase set:

$$\mathbf{E}_{\text{textual}} = \text{CrossAttn}(\hat{\mathbf{Y}}_{trans}, \{\mathbf{Y}_{trans}^{(i)}\}_{i=1}^{N}), \quad (12)$$

where $\mathbf{E}_{\text{textual}} \in \mathbb{R}^{T \times d}$ denotes the token-level fused embedding, capturing semantic consensus across diverse expressions. By distilling this cross-variant signal, the Semantic Weaver equips the system with the ability to tolerate surface-level ambiguity, allowing it to focus on intent-invariant semantics during downstream fusion and decoding.

3.5 Retrieval-Guided Fusion Learner

To achieve effective cross-modal integration, the Retrieval-Guided Fusion Learner uses audio as the anchoring modality, to project the audio, textual, and visual modalities into a shared semantic space. This enables rich interactions among symbolic, prosodic, and structural cues, while preserving fidelity to the original speech by retrieving acoustically consistent representations based on similarity to stored audio embeddings.

Cross-Modal Fusion. Let the audio input $\tilde{\mathcal{X}}$ be encoded by a pre-trained audio encoder $\mathcal{E}_{\text{audio}}$ (Radford et al., 2023), resulting in temporal embeddings:

$$\mathbf{E}_{\text{audio}} = \mathcal{E}_{\text{audio}}(\tilde{\mathbf{X}}). \quad (13)$$

The textual and visual modalities are respectively encoded and projected via learned transformations. The projected embeddings are concatenated to obtain the multimodal representation:

$$\mathbf{E}_{\text{mm}} = [\mathcal{E}_{\text{text}}(\mathbf{E}_{\text{textual}}),\ \mathcal{E}_{\text{vis}}(\mathbf{E}_{\text{visual}})]. \quad (14)$$

Audio and multimodal features are aligned through a Cross-Modal Attention (CM-Attn) mechanism, where audio embeddings act as queries and multimodal embeddings as keys and values:

$$\mathbf{E}_{\text{attn}} = \text{LayerNorm}(\mathbf{E}_{\text{audio}} + \text{CM-Attn}(Q, K, V)). \quad (15)$$

$$\text{CM-Attn}(Q, K, V) = \mathbf{W} \cdot \text{Cat}(\text{head}_1, \ldots, \text{head}_h), \quad (16)$$

Here, $Q = \mathbf{E}_{\text{audio}}$, $K = V = \mathbf{E}_{\text{mm}}$ and W is a learnable projection matrix. The residual connection with layer normalization stabilizes the output.

Gated Fusion. To balance the contributions of each modality, we introduce a gating mechanism:

$$\mathbf{Gate} = \sigma(\mathbf{W}_g[\mathbf{E}_{\text{audio}}; \mathbf{E}_{\text{mm}}] + \mathbf{b}_g), \quad (17)$$

$$\mathbf{E}_{\text{fused}} = \mathbf{Gate} \odot \mathbf{E}_{\text{attn}} + (1 - \mathbf{Gate}) \odot \mathbf{E}_{\text{audio}} \,, \quad (18)$$

where $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable parameters, and $\sigma(\cdot)$ is the sigmoid function. This fusion adaptively emphasizes robust signals for semantic reasoning.

3.6 Cognitive Decoder

The fused representation is fed into Cognitive Decoder to generate the final ASR response(Radford et al., 2023). By leveraging the complementary strengths of various modalities, the decoder captures both fine-grained patterns and high-level semantic context. The optimization objective is a token-level cross-entropy loss:

$$\mathcal{L} = -\frac{1}{T} \sum_{t=1}^{T} \log P(\mathbf{Y}_{true} \mid \tilde{\mathbf{X}}; \theta_{(E_{\text{textual}}, E_{\text{audio}}, E_{\text{visual}})}), \quad (19)$$

where $\mathbf{Y}_{true}$ is the ground-truth tokens at timestep t, and $P(Y_{true} \mid \tilde{\mathbf{X}}; \theta)$ is the predicted probability for the correct token conditioned on the noisy input speech. The model parameters θ include modality encoders and fusion layers, while the backbone language model remains frozen during fine-tuning to reduce overfitting and ensure robustness.

4 Evaluation

4.1 Evaluation Setting

4.1.1 Datasets

We evaluate our model on four benchmarks covering diverse real-world and synthetic noise conditions, including vocalizations, natural soundscapes, domestic settings, and complex urban environments. **CHiME-4**(Vincent et al., 2016) focuses on noisy, far-field speech with real and simulated multi-environmental recordings. **VB-DEMAND**(Valentini-Botinhao et al., 2016) blends clean utterances featuring various accents with a variety of background noises from the DEMAND corpus. **Noizeus**(Hu and Loizou, 2006) introduces synthetic distortions across multiple Signal-to-Noise Ratios (SNR) levels, allowing fine-grained control of noise severity. **GigaSpeech-ESC50**(Chen et al., 2021; Piczak, 2015) features speech with natural disfluency and ambient variations of male and female speakers, combined with the diverse environmental and artificial noises from ESC-50. Please refer to the Appendix for more detailed information on the processing of these datasets (Sec. A).

4.1.2 Metrics

We assess transcription quality using **Word Error Rate (WER)**(Morris et al., 2004) as the primary metric. To evaluate semantic preservation beyond surface form, we further report **BERTScore (BTS)**(Zhang et al., 2019) and **Universal Sentence Encoder (USE)**(Cer et al., 2018) similarity, capturing both lexical overlap and embedding-level meaning alignment.

4.1.3 Baselines

We compare `Speech-MLM` with six competitive baselines:

Whisper(Radford et al., 2023): A strong open-domain ASR pre-trained model.

Whisper (Finetune): Whisper finetuned on the same noisy speech datasets as our model to adapt to domain-specific degradations.

Qwen2-Audio(Chu et al., 2024): A recent large-scale instruction-tuned audio language model.

Qwen2-Audio (Finetune): Qwen2-Audio further finetuned for ASR under noisy conditions.

STAR(Hu et al., 2024a): A noise-robust ASR system that leverages unlabeled noise data to improve performance.

4.1.4 Implementation Details

We adopt ViT-L/16 (vjepa2-vitl-fpc64-256) as the default visual backbone for all methods. Optimization is performed using the AdamW optimizer with a learning rate linearly warmed up from 3×10^{-7} to 3×10^{-5} over the first 10% of total training steps. After warmup, the learning rate is scheduled with CosineAnnealing, decaying to 3×10^{-6} at the end of training. During paraphrase generation, we adopt the default setting (Temperature = 0.7). All models are trained for up to 3 epochs, using a batch size of 16. Experiments are conducted on 4 NVIDIA A800 GPUs (80GB), with distributed training enabled across all devices.

4.2 Overall Performance

Table 1 compares `Speech-MLM` with representative baselines on four diverse datasets.

• **Robustness in Highly Challenging audio Environments**. Our model's superiority is most evident in challenging noisy scenarios. When evaluated on the CHiME-4 and VB-DEMAND—which emulate challenging real-world noise scenarios—`Speech-MLM` reduces WER to 13.09% and 1.77%, achieving reductions of 21.66% and 55.53% over the best baseline. The performance gap is even more dramatic on the Noizeus under extreme low-SNR conditions. At 0dB SNR, `Speech-MLM` achieves a WER of just 4.00%, marking a 69.4% relative reduction compared to the best baseline. As noise decreases, our model's performance approaches perfection, reaching a 0% WER at 10dB.

• **Superiority in Semantic Fidelity and Understanding**. Beyond mere word-level accuracy, `Speech-MLM` demonstrates a superior ability to preserve semantic meaning, as measured by BTS and USE similarity. This is critical for downstream applications where understanding is paramount. For instance, on Noizeus at 5dB SNR, while other models struggle to maintain meaning, `Speech-MLM` achieves near-perfect scores of 99.88 in BTS and 99.74 in USE. This trend continues across all datasets, including GigaSpeech-ESC50, which emphasizes semantic fluency. At a low 10dB SNR, `Speech-MLM` scores 98.53 in BTS and 95.91 in USE, consistently leading all other models.

Metric	Model	CHiME-4				VB-DEMAND		Noizeus			GigaSpeech-ESC50			
		test_real	test_simu	dev_real	dev_simu	clean	noisy	0	5	10	0	10	20	30
WER ↓	Whisper (Frozen)	40.14	42.25	37.71	39.28	5.80	8.63	37.97	14.11	4.77	29.01	22.75	21.90	21.82
	Whisper (Finetune)	17.82	20.96	15.19	16.88	2.15	2.53	27.20	9.13	2.07	18.54	12.35	11.55	11.86
	Qwen2-Audio (Frozen)	38.55	36.37	33.78	31.26	12.50	31.84	47.10	14.11	8.92	64.04	26.01	20.60	19.37
	Qwen2-Audio (Finetune)	16.71	18.72	14.01	14.83	2.34	3.98	13.07	5.39	1.45	19.69	12.27	10.87	12.18
	STAR	20.76	23.66	25.17	18.67	3.08	4.07	27.39	10.17	3.73	23.24	18.52	14.86	13.79
	Ours	**13.09**	**15.66**	**11.98**	**13.66**	**1.14**	**1.77**	**4.00**	**0.41**	**0.00**	**15.84**	**8.09**	**8.01**	**8.70**
BTS ↑	Whisper (Frozen)	94.77	94.35	95.27	94.91	99.31	98.80	93.69	97.49	99.14	95.92	97.23	97.46	97.51
	Whisper (Finetune)	97.77	96.98	98.10	97.55	99.75	99.75	95.25	98.58	99.43	96.44	98.04	98.32	98.27
	Qwen2-Audio (Frozen)	93.53	90.24	94.80	95.31	94.57	90.21	92.92	97.93	98.73	85.52	90.84	89.99	87.42
	Qwen2-Audio (Finetune)	97.46	97.07	97.99	97.61	94.99	94.84	98.10	99.08	99.58	95.88	97.03	97.30	97.35
	STAR	95.40	95.03	95.59	96.14	99.68	99.53	94.82	98.15	99.45	95.85	97.39	98.35	97.34
	Ours	**97.88**	**97.27**	**98.17**	**97.69**	**99.93**	**99.87**	**99.03**	**99.88**	**100.00**	**97.62**	**98.53**	**98.84**	**98.57**
USE ↑	Whisper (Frozen)	85.95	83.58	89.32	87.52	98.65	95.15	62.34	90.39	97.12	87.40	95.21	96.23	96.57
	Whisper (Finetune)	96.06	91.92	97.30	95.49	98.74	99.15	77.41	95.18	99.33	85.34	95.61	96.45	96.48
	Qwen2-Audio (Frozen)	80.00	55.32	85.25	87.44	74.38	62.25	62.17	91.68	96.59	41.45	62.79	73.29	74.87
	Qwen2-Audio (Finetune)	94.21	91.51	96.49	94.99	97.95	96.67	90.34	95.49	98.39	84.46	92.05	93.40	93.65
	STAR	83.82	81.26	84.14	86.72	98.93	98.04	75.66	93.69	97.90	83.95	88.16	92.67	93.35
	Ours	**96.18**	**93.33**	**97.51**	**96.00**	**99.72**	**99.18**	**96.99**	**99.74**	**100.00**	**87.84**	**95.91**	**98.05**	**97.18**

Table 1: Performance Comparison across Multiple Datasets (% for BTS and USE).
Note: CHiME-4 *datasets include different test sets: test_real, test_simu, dev_real, and dev_simu.* ***VB-DEMAND*** *dataset is divided into clean and noisy subsets.* ***Noizeus*** *and* ***GigaSpeech-ESC50*** *columns report results under different SNRs.*

Metric	Model	Noizeus			GigaSpeech-ESC50			
		SNR=0	SNR=5	SNR=10	SNR=0	SNR=10	SNR=20	SNR=30
WER ↓	Speech-MLM w/o T&I	27.20	9.13	2.07	18.54	12.35	11.55	11.86
	Speech-MLM w/o I	5.60	1.04	0.00	16.71	9.30	8.85	8.77
	Speech-MLM w/o T	16.60	1.64	0.00	16.10	9.34	8.93	8.83
	Speech-MLM	**4.00**	**0.41**	**0.00**	**15.84**	**8.09**	**8.01**	**8.40**
BTS ↑	Speech-MLM w/o T&I	95.25	98.58	99.43	96.44	98.04	98.32	98.27
	Speech-MLM w/o I	98.68	99.77	100.00	96.72	98.32	98.52	98.53
	Speech-MLM w/o T	97.13	99.72	100.00	96.87	98.35	98.52	98.50
	Speech-MLM	**99.03**	**99.88**	**100.00**	**97.62**	**98.53**	**98.84**	**98.57**
USE ↑	Speech-MLM w/o T&I	77.41	95.18	99.33	85.34	95.61	96.45	96.48
	Speech-MLM w/o I	95.47	99.04	100.00	86.52	95.66	96.93	97.13
	Speech-MLM w/o T	84.98	97.57	100.00	87.04	95.73	96.92	96.99
	Speech-MLM	**96.99**	**99.74**	**100.00**	**87.84**	**95.91**	**98.05**	**97.18**

Table 2: Ablation Study Results (% for BTS and USE).

In summary, Speech-MLM enhances robustness to noise and distortion through multimodal integration, effectively mitigating the impact of audio degradation while maintaining statistically significant gains and efficient inference (Sec. E and B).

4.3 Ablation Study

To investigate the contribution of each component, we conduct ablation experiments by removing the *Cognitive Structure Extractor* (w/o I), the *Semantic Weaver* (w/o T), or both visual and textual component (w/o T&I) from the full model, and by replacing the *Semantic Weaver* with other audio LLMs (Team et al., 2024; Kong et al., 2024) as shown in Table 2 and 3.

• **Removing the visual or textual modality** leads to a noticeable drop in WER, suggesting their critical role in disambiguating structural boundaries in noisy environments. **Notably, removing the textual modality** causes a substantial decline in semantic metrics, especially on Noizeus, where expression-level ambiguity is more prevalent. Re-

Metric	Semantic Weaver	SNR=0	SNR=10	SNR=20	SNR=30
WER ↓	Gemma-2-9b-it	17.57	10.41	9.85	9.85
	Audio Flamingo	15.92	10.65	9.21	9.70
	Speech-MLM	**15.84**	**8.09**	**8.01**	**8.40**
BTS ↑	Gemma-2-9b-it	96.89	98.05	98.42	98.26
	Audio Flamingo	96.96	98.00	98.74	**98.64**
	Speech-MLM	**97.62**	**98.53**	**98.84**	98.57
USE ↑	Gemma-2-9b-it	89.18	95.67	97.29	96.37
	Audio Flamingo	**89.64**	95.62	96.78	97.16
	Speech-MLM	87.84	**95.91**	**98.05**	**97.18**

Table 3: Performance When Using Other Audio LLMs on GigaSpeech-ESC50.

placing the Semantic Weaver with other audio LLMs also influences performance, though remain outperform the w/o T configuration.

• **When removing both component (w/o T&I)**, all the metrics drop off sharply, confirming the inherent limitations of systems that rely solely on the audio modality when the noise causes the speech signal to be acoustically indistinguishable. These findings confirm the complementary strengths of each modality in capturing different facets of speech information, and underscore the importance of multimodal integration.

In addition, to disentangle the effect of semantic abstraction from potential text-prior bias, we conducted an additional ablation study on paraphrase sources, with details provided in the Appendix (Sec. C).

4.4 Hyperparameter Study

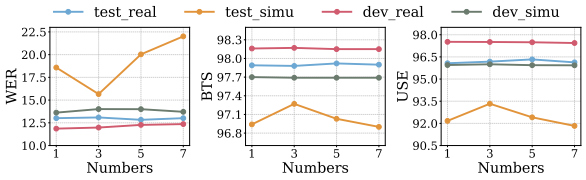

Figure 2: WER, BTS and USE Performance across Different Paraphrasing Quantities on CHiME-4.

We analyze how the paraphrase hypothesis number N in Semantic Weaver affects model performance. Results show consistent accuracy across test_real, dev_real, and dev_simu datasets despite varying N, while test_simu exhibits distinct behavior. As shown in Table 1, all models perform worse on test_simu, suggesting its noise patterns are more challenging for speech recognition. When employing a single paraphrase, Semantic Weaver exhibits significant susceptibility to errors. Although increasing the paraphrase number enhances semantic recovery, this improvement is accompanied by the introduction of additional noise, thereby complicating the learning process. However, setting N beyond 3 yields diminishing returns and slightly reduces performance due to increased semantic noise. This suggests that while diverse paraphrase hypotheses enrich semantic representation, excessive variation may introduce inconsistencies. comprehensive experimentation, we establish that employing three paraphrases represents the optimal configuration.

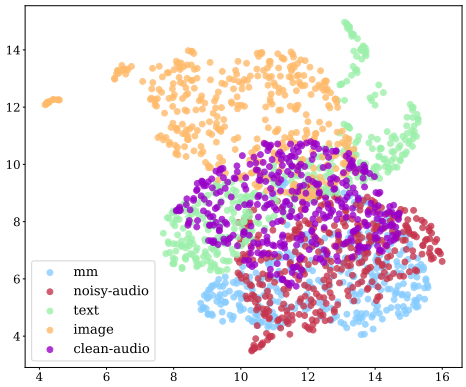

Figure 3: UMAP Visualization of Audio (noisy/clean), Text, and Image Embeddings.

4.5 Multimodal Analysis

To further investigate how multimodal integration enhances semantic representation, we visualize the learned embedding space using UMAP projections, as shown in Figure 3. We compare the embeddings generated from individual modalities—$\mathbf{E}_{\text{audio}}$ (clean vs. noisy), $\mathbf{E}_{\text{textual}}$, and $\mathbf{E}_{\text{visual}}$. The visualization provides direct evidence of how `Speech-MLM` constructs a noise-robust semantic representation, yielding three key insights:

(i) Distinct and Complementary Modalities. The projection reveals that each unimodal embedding space forms a distinct cluster with minimal overlap, indicating that each modality provides complementary, rather than redundant, information.

(ii) Acoustic Degradation Creates a Semantic Gap. While the clean and noisy audio embeddings show some overlap, a significant distributional shift is evident, confirming the severe impact of noise on the audio modality.

(iii) Cross-Modal Semantic Compensation. Notably, the embeddings from the text and visual modalities substantially cover the regions of the clean audio distribution that are lost or vacated by the noisy audio embeddings.

This demonstrates that these modalities enrich the representation with text and visual information, compensating for the degradation in audio quality.

4.6 Case Study

Table 4 showcases a typical case that vividly illustrates the effectiveness of our `Speech-MLM` framework. Here, multiple advanced baselines, including Whisper and Qwen, systematically fail on critical phrases like "Hedge," "may stain," and "green." Even with fine-tuning on noisy data, these models still struggle due to their inability to differen-

Model	Utterance	WER (%) ↓	BTS (%) ↑	USE (%) ↑
Whisper(Frozen)	Hedge apples may sustain your hands for me.	37.50	91.61	66.47
Whisper(Finetune)	Edge apples may stain your hands green.	14.29	99.97	81.57
Qwen2-Audio(Frozen)	Hedge apple made same your hands are green.	50.00	90.46	66.73
Qwen2-Audio(Finetune)	Hedge apples made stain your hands green.	14.29	96.72	91.74
STAR	Edge apples may stain your hands green.	14.29	99.97	81.57
Ours	Hedge apples may stain your hands green.	**0.00**	**100.00**	**100.00**
Ground Truth	Hedge apples may stain your hands green.	-	-	-

Table 4: Case study of Speech-MLM. The test sample is selected from the Noizeus(SNR=10).

tiate acoustically confusable words like "hedge" and "edge" under degraded signal conditions. In contrast, Speech-MLM successfully reconstructs the intended content through multimodal semantic reasoning. Furthermore, to address the limitation of relying solely on proxy metrics (e.g., BTS and USE), we conducted a small-scale human evaluation, with results reported in the Appendix (Sec. D). By leveraging visual phonetic cues and textual priors, our model goes beyond surface-level transcription to infer the most plausible result.

5 Conclusion

This paper presents Speech-MLM, a cognitively inspired ASR framework that reframes speech transcription as the recovery of structured meaning under noise and ambiguity. Through the synergy of three modules: (i) Cognitive Structure Extractor capturing prosodic-syntactic anchors, (ii) Semantic Weaver aligning textual variants, and (iii) Retrieval-Augmented Learner unifying cross-modal signals. Our approach enables robust understanding in scenarios where traditional audio-centric models fail. Empirical results across multiple challenging benchmarks validate its effectiveness, showing marked improvements over both general and noise-specialized baselines—not only in WER, but also in semantic preservation. By emulating the cognitive mechanisms of understanding, our approach contributes to building more robust ASR systems.

Acknowledgments

This work is supported partly by Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007), China NSFC Grant (No.62472366), the Project of DEGP (No.2024GCZX003, 2023KCXTD042), 111 Center (No.D25008), Shenzhen Science and Technology Foundation (ZDSYS20190902092853047)

References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, and 11 others. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *Preprint*, arXiv:2506.09985.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, and 1 others. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and 1 others. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Kelechi Ezema, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran, and Sidney D'Mello. 2025. "it feels like we're not meeting the criteria": Examining and mitigating the cascading effects of bias in automatic speech recognition in spoken language interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. 2021. Metricgan+: An improved version of metricgan for speech enhancement. In *Interspeech 2021*, pages 201–205.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. 2022. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Yi Hu and P.C. Loizou. 2006. Subjective comparison of speech enhancement algorithms. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Yuchen Hu, Chen Chen, Chao-Han Yang, Chengwei Qin, Pin-Yu Chen, Eng-Siong Chng, and Chao Zhang. 2024a. Self-taught recognizer: Toward unsupervised adaptation for speech foundation models. *Advances in Neural Information Processing Systems*, 37:29566–29594.

Yuchen Hu, CHEN CHEN, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EngSiong Chng. 2024b. Large language models are efficient learners of noise-robust speech recognition. In *The Twelfth International Conference on Learning Representations*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: a novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhuoran Li and Dan Zhang. 2024. How does the human brain process noisy speech in real life? insights from the second-person neuroscience perspective. *Cognitive Neurodynamics*, 18(2):371–382.

Yen-Ting Lin, Zhehuai Chen, Piotr Zelasko, Zhen Wan, Xuesong Yang, Zih-Ching Chen, Krishna C Puvvada, Ke Hu, Szu-Wei Fu, Jun Wei Chiu, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Chao-Han Huck Yang. 2025. NeKo: Cross-modality post-recognition error correction with tasks-guided mixture-of-experts language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 222–236, Vienna, Austria. Association for Computational Linguistics.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869.

Adam S Miner, Albert Haque, Jason A Fries, Scott L Fleming, Denise E Wilfley, G Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A Arnow, W Stewart Agras, and 1 others. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ digital medicine*, 3(1):82.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. 2024a. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809, Mexico City, Mexico. Association for Computational Linguistics.

Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. 2024b. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Karol J. Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA. Association for Computing Machinery.

Archiki Prasad, Preethi Jyothi, and Rajbabu Velmurugan. 2021. An investigation of end-to-end models for robust speech recognition. In *ICASSP 2021-2021*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6893–6897. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2016. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 146–152.

Emmanuel Vincent, Shinji Watanabe, Jon Barker, and Ricard Marxer. 2016. The 4th chime speech separation and recognition challenge. *URL: http://spandh. dcs. shef. ac. uk/chime_challenge/(last accessed on 1 August, 2018)*.

Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. 2024a. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12311–12315. IEEE.

Yicheng Wang, Mark Cusick, Mohamed Laila, Kate Puech, Zhengping Ji, Xia Hu, Michael Wilson, Noah Spitzer-Williams, Bryan Wheeler, and Yasser Ibrahim. 2024b. Fairlens: Assessing fairness in law enforcement speech recognition. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3815–3824. IEEE.

Ziyi Wang, Yiming Rong, Deyang Jiang, Haoran Wu, Shiyu Zhou, and Bo Xu. 2024c. Cieasr:contextual image-enhanced automatic speech recognition for improved homophone discrimination. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 915–924, New York, NY, USA. Association for Computing Machinery.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Liang Zhang, Zhen Yang, Biao Fu, Ziyao Lu, Liangying Shao, Shiyu Liu, Fandong Meng, Jie Zhou, Xiaoli Wang, and Jinsong Su. 2024. Multi-level cross-modal alignment for speech relation extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11975–11986, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Detail of Datasets

All datasets were processed using a unified and fully reproducible pipeline. During noise mixing, if a noise clip was shorter than the corresponding speech utterance, it was repeated in a loop until the full speech length was covered; if the noise clip was longer, it was truncated to match the speech duration.

For **GigaSpeech–ESC50**, clean speech segments from the GigaSpeech(Chen et al., 2021) dataset were randomly mixed with environmental and synthetic noise clips from ESC-50(Piczak, 2015). SNRs were uniformly sampled from {0, 10, 20, 30} dB. Each speech utterance was paired with a randomly selected 3–5 s noise clip after amplitude normalization. The resulting dataset was split into training, validation, and test sets with a ratio of 8:1:1, ensuring no speaker overlap across splits.

For **CHiME4**(Vincent et al., 2016), **VB-DEMAND**(Valentini-Botinhao et al., 2016), and **Noizeus**(Hu and Loizou, 2006), we strictly followed the official dataset protocols, including their predefined data splits and noise mixing configurations.

B Real-Time Inference Efficiency

Model	test_real	test_simu
Whisper	0.1563	0.1406
Qwen2-Audio	0.8672	0.4844
Speech-MLM (Ours)	0.9062	0.4531

Table 5: End-to-end inference latency (seconds per second of input audio) on CHiME-4.

To evaluate the computational cost introduced by Speech-MLM, We analyze the inference latency and scalability of Speech-MLM to verify that its multimodal design does not introduce excessive computational overhead or hinder real-time deployment. End-to-end inference latency is evaluated on the CHiME-4 dataset (*test_real* and *test_simu*) using a single NVIDIA A100 GPU. Latency is reported as the processing time per second of input audio, which directly reflects real-time feasibility.

As shown in Table 5, Whisper achieves the lowest latency due to its lightweight, audio-only architecture. Qwen2-Audio exhibits substantially higher latency, reflecting the cost of large-scale audio–language semantic modeling. Speech-MLM shows latency comparable to Qwen2-Audio across both subsets, indicating that the proposed multimodal semantic reconstruction does not introduce a noticeable efficiency drop beyond a strong audio–language backbone. the overall inference complexity remains dominated by the backbone ASR model, while the additional multimodal components contribute only marginal overhead. This design prevents cascading latency growth as input length or modality diversity increases, enabling stable scaling across batch sizes and deployment scenarios.

In practice, Speech-MLM consistently maintains near–real-time performance (less than one second per second of audio), demonstrating that its robustness gains are achieved without sacrificing efficiency. These results confirm that Speech-MLM strikes a practical balance between semantic resilience and computational scalability, making it suitable for real-world applications.

C Controlled Ablation on Paraphrase Sources.

Metric	Model	test_real	test_simu
WER ↓	Speech-MLM w/o Paraphrases	16.59	20.69
	Speech-MLM Human Paraphrases	14.02	17.33
	Speech-MLM (Ours)	**13.09**	**15.66**
BTS ↑	Speech-MLM w/o Paraphrases	97.02	96.73
	Speech-MLM Human Paraphrases	97.72	97.01
	Speech-MLM (Ours)	**97.88**	**97.27**
USE ↑	Speech-MLM w/o Paraphrases	96.18	92.33
	Speech-MLM w Human Paraphrases	94.99	**96.51**
	Speech-MLM (Ours)	**96.18**	93.33

Table 6: Controlled ablation on paraphrase sources.

To disentangle the effect of semantic abstraction from the strength of textual priors, we conduct a controlled ablation study on the source of paraphrases used in the Semantic Weaver. We compare three variants under identical training and inference settings on CHiME-4: (i) Speech-MLM w/o Paraphrases, where semantic weaving is disabled; (ii) Speech-MLM w Human Paraphrases, using manually curated alternative expressions of the same utterances; and (iii) Speech-MLM (Ours), where paraphrases are generated from noisy inputs by an external LLM.

Table 6 reports results on both *test_real* and *test_simu*. Introducing paraphrases yields consistent improvements across all metrics. Compared to the no-paraphrase baseline, both human- and LLM-based paraphrases reduce WER by a substantial margin (e.g., from 16.59% to 14.02% and 13.09% on *test_real*), while simultaneously improving semantic similarity as measured by BTS

Category	Metric / Rater	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Semantic Metrics	BTS	95.10	99.93	98.56	99.87	94.53	99.88	99.25	99.80	99.98	99.88
	USE	91.04	99.72	89.63	99.18	96.10	99.74	100.00	100.00	90.62	100.00
Human Evaluation	Informativeness (R1)	2.9	4.7	3.8	4.6	4.5	4.8	5.0	5.0	4.5	5.0
	Informativeness (R2)	3.2	4.7	4.1	4.5	4.4	4.6	5.0	5.0	4.6	5.0
	Informativeness (R3)	2.8	4.5	3.6	4.7	4.4	4.8	5.0	5.0	4.5	5.0
	Naturalness (R1)	4.3	4.7	4.5	4.6	4.7	4.7	5.0	5.0	4.8	4.9
	Naturalness (R2)	4.6	4.8	4.6	4.7	4.6	4.8	5.0	5.0	4.8	4.8
	Naturalness (R3)	4.2	4.6	4.5	4.7	4.5	4.7	4.9	4.9	4.7	4.8
	Quality (R1)	4.6	4.8	4.3	4.8	4.6	4.7	4.9	4.9	4.5	4.8
	Quality (R2)	4.7	4.8	4.4	4.8	4.6	4.7	4.8	4.9	4.6	4.8
	Quality (R3)	4.5	4.5	4.2	4.6	4.6	4.6	4.9	4.8	4.4	4.8

Table 7: Automatic semantic metrics and human evaluation results on ten sampled utterances. Human scores are reported on a RankME scale (0.0–5.0), with three independent raters (R1–R3).

and USE. Notably, the performance gap between human and LLM paraphrases is small across all metrics. On *test_real*, LLM paraphrases achieve the lowest WER (13.09%) and the highest BTS (97.88%), while human paraphrases yield comparable gains. This trend holds on *test_simu*, where both paraphrase sources consistently outperform the no-paraphrase variant.

These results indicate that the observed robustness gains do not stem from stronger or more accurate textual priors, but from enforcing noise-invariant semantic equivalence during training. In other words, the Semantic Weaver benefits from exposure to multiple meaning-preserving realizations of the same utterance, regardless of whether these variants are authored by humans or generated by an LLM. This confirms that the improvements are driven by semantic abstraction rather than paraphrase source or linguistic fluency.

D Human Assessment

While automatic metrics provide scalable and reproducible evaluation, their reliability in assessing semantic adequacy and acceptability remains limited in safety-critical and socially sensitive scenarios. Human judgment is still more trustworthy for evaluating meaning preservation, informativeness, and linguistic acceptability. To address this limitation, we complement our automatic evaluation with a small-scale human assessment study.

Evaluation Setup. We randomly sampled 10 utterances (**S1-S10**) and asked three human participants (**R1-R3**) to independently evaluate both the ground-truth transcripts and the corresponding system outputs. Each utterance was rated on a 0–5 Likert scale (Joshi et al., 2015) along three dimensions adapted from RankME (Novikova et al., 2018):

- **Informativeness (Adequacy):** Does the utterance preserve all essential information from the ground-truth meaning?
- **Naturalness (Fluency):** Could the utterance plausibly be produced by a native speaker?
- **Quality:** An overall judgment considering grammaticality, fluency, and semantic adequacy.

Results and Analysis. Table 7 summarizes the human ratings alongside automatic semantic metrics (BTS and USE). Across all samples, we observe a strong qualitative alignment between human-rated Informativeness and automatic semantic similarity scores. Utterances with higher BTSand USE consistently receive higher Informativeness ratings from human judges, suggesting that these automatic metrics serve as reasonable proxies for semantic fidelity in noisy ASR settings.

In contrast, Naturalness and Overall Quality scores are consistently high across most samples, even when Informativeness varies. This indicates that while modern LLM-based generation produces fluent and well-formed text, fluency alone does not guarantee correct meaning reconstruction. Importantly, our model demonstrates the ability to transform fragmented or noisy acoustic evidence into coherent, high-quality utterances, rather than merely generating superficially fluent outputs.

Discussion. The consistency between human Informativeness judgments and BTS/USE trends supports the validity of our automatic evaluation protocol. At the same time, the divergence between semantic adequacy and surface fluency highlights the necessity of human assessment when evaluating robustness and fairness in real-world ASR systems.

Metric	CHiME-4				VB-DEMAND		Noizeus			GigaSpeech-ESC50			
	test_real	test_simu	dev_real	dev_simu	clean	noisy	0	5	10	0	10	20	30
WER ↓	13.09±0.32	15.66±0.16	11.98±0.23	13.66±0.86	1.14±0.74	1.77±1.05	4.00±3.83	0.41±0.58	0.00±0.00	15.84±3.50	8.09±0.94	8.01±0.51	8.70±0.72
BTS ↑	97.88±0.38	97.27±0.19	98.17±0.08	97.69±0.30	99.93±0.06	99.87±0.12	99.03±0.81	99.88±0.17	100.00±0.00	97.62±0.32	98.53±0.02	98.84±0.24	98.57±0.20
USE ↑	96.18±0.37	93.33±0.88	97.51±0.30	96.00±1.07	99.72±0.11	99.18±0.46	96.99±2.43	99.74±0.37	100.00±0.00	87.84±2.63	95.91±1.15	98.05±0.55	97.18±0.85

Table 8: Statistical performance of Speech-MLM, aligned with Table 1. BTS and USE are reported in percentage (mean ± std). Statistical significance is assessed using the Wilcoxon signed-rank test (normal approximation), comparing Speech-MLM against the strongest baseline within each dataset group. Overall improvements are statistically significant at $p < 0.05$.

E Statistical Significance Analysis

To evaluate the reliability and reproducibility of the reported improvements, we conducted a thorough statistical analysis of all the datasets and evaluation metrics. Specifically, we repeated all experiments under five different random seeds and report the mean and standard deviation for each configuration. Additionally, we performed a Wilcoxon signed-rank test to compare Speech-MLM with the strongest baseline in each dataset group, in line with standard practice for the non-parametric.

Table 8 summarises the statistical performance of Speech-MLM across all datasets and metrics (WER, BTS and USE). The observed gains are consistent across five random seeds and are statistically significant at the $p < 0.05$ level. This confirms that the improvements reported in the main paper are not driven by a particular initialization or isolated runs, but reflect stable and repeatable behavior. We note that certain settings—most notably Noizeus at high SNR levels—exhibit extremely low WER values (e.g., 0.00% at 10 dB). Importantly, these cases are accompanied by near-zero variance across random seeds, indicating convergence rather than instability. Similar stability trends are also reflected in the corresponding BTS and USE scores, which reach saturation with negligible variance.

Overall, the combination of multi-seed evaluation, variance reporting, and non-parametric significance testing demonstrates that the gains achieved by Speech-MLM are statistically robust and generalize consistently across datasets and noise conditions. These results strengthen the empirical foundation of our claims and enhance the reproducibility and reliability of the proposed approach.

F Limitation

Although Speech-MLM performs well, there remains substantial room for further improvement:

- Although the proposed framework effectively alleviates semantic ambiguity in noisy conditions by using multimodal semantic reconstruction, we have not yet systematically characterised the conditions in which reconstruction errors arise. For instance, it is unclear whether particular noise types, incomplete modality cues or semantically ambiguous inputs invariably result in erroneous reconstructions. Specifically, the gating mechanism is optimized to balance text-derived semantic priors with raw acoustic and visual evidence. When a rare word is clearly articulated, the Cognitive Structure Extractor captures distinct phonetic boundaries that guide the gating mechanism to emphasize acoustic and visual modalities, thereby preserving the authentic pronunciation. Nevertheless, we acknowledge that if the acoustic evidence of an OOV word is severely corrupted or completely masked by noise, textual priors may dominate, potentially resulting in a fluent but normalized substitution. Further refining error patterns and systematically analysing deviations will enhance the reliability and interpretability of this framework, and these remain key areas for future work.

- Although we report on inference latency and demonstrate near-real-time performance across multiple benchmark datasets, practical deployment constraints in real-world settings, such as call centres, healthcare environments, and streaming ASR systems, remain under-explored. A key area for future research is evaluating robustness and scalability under sustained workloads, in resource-constrained conditions.

- Speech-MLM relies on a audio–language model to generate intermediate semantic variants. While this design enables richer semantic abstraction, it inevitably introduces additional computational overhead. In extreme cases, failures or hallucinations during the variant generation stage could lead to cascading errors in downstream semantic reconstruction. Although our empirical results suggest that such risks are relatively well controlled in practice, systematic evaluation of

paraphrase quality, robust fallback mechanisms, and lightweight alternatives remain important avenues for strengthening the framework. Incorporating human-in-the-loop or cognitively informed evaluation mechanisms could also help to improve fairness and reliability.