INPUT-ADAPTIVE ACOUSTIC CONDITION MODULATION ACROSS SPEECH-TO-LLM BRIDGES FOR ROBUST SPEECH RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech large language models (Speech LLMs) depend on a trainable bridge that compresses acoustic representations into LLM input embeddings. Even with strong frozen speech and language backbones, a bridge learned on one corpus can map unfamiliar acoustic conditions to poorly supported regions of its representation space. We introduce Input-Adaptive Acoustic Condition Modulation (IAACM), a bridge-local calibration method for robust speech recognition. IAACM routes an utterance through a bank of acoustic condition vectors, predicts token-level deviation from a frozen source-bridge manifold, and applies a risk-gated residual FiLM correction. The speech encoder and LLM remain frozen, and inference is self-conditioned on the target utterance without external reference audio. We evaluate matched conditioned and unconditioned systems across eight bridge architectures and six datasets. IAACM lowers the six-domain macro word error rate (WER) for every evaluated bridge and improves 35 of 48 bridge–domain pairs. Averaged uniformly over all pairs, macro WER decreases from 12.22% to 11.65%; the relative reduction grows from 4.68% overall to 17.15% on the highest-error domain of each bridge. Controlled noise and reverberation tests further support robustness. These results demonstrate that bridge-level acoustic calibration improves speech recognition robustness across diverse Speech-to-LLM bridges.

1 INTRODUCTION

Speech-aware large language models commonly join a pretrained speech encoder to a text LLM through a small trainable connector, or *bridge* (Tang et al., 2024; Yu et al., 2024). The bridge must solve two problems simultaneously: it compresses a long acoustic sequence into a tractable number of embeddings that the LLM can interpret. This interface has enabled strong automatic speech recognition (ASR) while keeping most foundation-model parameters frozen. It is also a fragile point. A bridge trained on clean read speech may encounter unfamiliar speakers, channels, speaking styles, and recording environments at test time, even though its frozen speech encoder has seen far broader variation during pretraining.

Most work on Speech LLMs connectivity asks *which* bridge to use. The bridge architectures considered here—Q-Former (Li et al., 2023; Yu et al., 2024), Window Q-Former (Tang et al., 2024), Linear Projector (Shu et al., 2023; Ma et al., 2024), Pooled MLP Projector (Cui et al., 2026), Non-Causal Conv (Wang et al., 2023), Causal Conv (Oh & Kim, 2026), CIF (Dong & Xu, 2020; Wang et al., 2024), and CTC (Gaido et al., 2021; Wu et al., 2023)— make different assumptions about temporal structure and speech–text alignment (Yu et al., 2024; Verdini et al., 2025). Their behavior varies substantially across speech encoders and LLMs (Verdini et al., 2025). Yet a bridge is normally a fixed mapping after training: the same transformation is applied to every utterance, irrespective of how far its acoustic condition lies from the bridge's training distribution.

We study a complementary question: *can the bridge adapt its local representation to each utterance without updating either foundation backbone?* To address this question, we propose Input-Adaptive Acoustic Condition Modulation (IAACM). IAACM combines a learned bank of acoustic condition vectors with token-level risk estimates to produce input-dependent corrections. These corrections are bounded relative to the bridge representations, allowing local adaptation while limiting the magnitude

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of the update. Both foundation backbones remain frozen, and inference relies only on the target utterance, requiring neither external reference audio nor a second speech-encoder pass.

Our main empirical contribution is a controlled study across eight bridge architectures: Q-Former, Window Q-Former, Linear Projector, Pooled MLP Projector, Non-Causal Conv, Causal Conv, CIF, and CTC. Every comparison uses Whisper large-v2 and Qwen3-8B as frozen backbones, the same LibriSpeech training partition, a fixed three-epoch protocol, and identical decoding. We evaluate on LibriSpeech test-clean/test-other and four shifted test sets: TED-LIUM, Common Voice 23, GigaSpeech, and VCTK.

Our contributions are:

- We introduce a centered condition bank and bounded, risk-gated residual FiLM that is self-conditioned at inference and leaves both foundation backbones frozen.
- We provide a paired evaluation over eight bridge families and six test sets. IAACM improves all eight bridge-level macro averages and 35/48 individual bridge–domain pairs, with larger gains as baseline domain error increases.
- We evaluate all eight bridges under 28 controlled noise/reverberation conditions, with improvements in 156/224 bridge–condition pairs, and complement these results with a representation-distance analysis.

2 Related Work

2.1 Speech-aware large language models

Speech LLMs extend a text LLM with continuous or discrete acoustic inputs. SpeechGPT (Zhang et al., 2023), AudioPaLM (Rubenstein et al., 2023), SALMONN (Tang et al., 2024), and Qwen2-Audio (Chu et al., 2024) demonstrate that an LLM can support speech recognition and broader spoken/audio understanding after modality alignment. Recent work also targets sample-efficient speech adaptation and preservation of text capabilities (Cuervo et al., 2026). We focus on hidden-representation ASR: a speech encoder emits continuous features, a trainable bridge converts them into audio tokens, and a frozen autoregressive LLM predicts the transcript.

2.2 Speech-to-LLM Bridges

Speech-to-LLM bridges differ in their feature mappings and temporal aggregation mechanisms. The Q-Former was introduced in BLIP-2 (Li et al., 2023) for vision–language alignment and was subsequently explored as a speech-to-LLM connector by Yu et al. (2024), who also introduced segment-level processing. SALMONN adopts a Window Q-Former for local temporal aggregation (Tang et al., 2024).

Projection-based bridges provide a lightweight and simple interface for mapping speech representations into the LLM input space. They range from simple linear mappings to multi-layer projections, sometimes combined with temporal aggregation or downsampling. Such projection-based interfaces have been widely used in speech-to-LLM systems (Ma et al., 2024; Shu et al., 2023; Cui et al., 2026).

Convolutional bridges incorporate temporal modeling and downsampling into the modality interface. Both non-causal and causal convolutional connectors have been explored for speech–language integration (Wang et al., 2023; Oh & Kim, 2026).

A separate family performs content-dependent compression using explicit speech–text alignment signals. Continuous Integrate-and-Fire (CIF) was introduced as a mechanism for converting variable-length acoustic sequences into token-level representations (Dong & Xu, 2020); BLSP-KD subsequently uses CIF for fine-grained speech–text alignment in language–speech pretraining (Wang et al., 2024). CTC-based compression groups acoustic frames according to CTC predictions (Gaido et al., 2021) and has been explored in decoder-only speech–LLM integration (Wu et al., 2023).

These studies motivate our bridge-diverse protocol. Rather than proposing one more replacement connector, we ask whether the same calibration principle transfers across connector families.

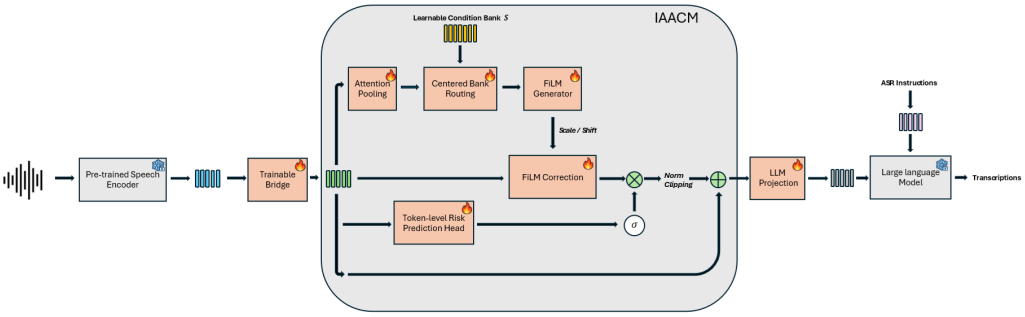


Figure 1: Overview of IAACM. The acoustic condition is inferred in the trainable bridge space. A frozen copy of the Stage-A bridge supplies source-manifold targets only during training. At inference, target bridge tokens are reused for self-conditioning, so the method requires neither external reference audio nor a second speech-encoder pass.

2.3 Conditional modulation and acoustic condition banks

FiLM applies feature-wise affine transformations predicted from conditioning information (Perez et al., 2018). Global style tokens learn a bank of latent vectors and attend to their mixtures for unsupervised variation modeling in speech synthesis (Wang et al., 2018). We borrow the useful engineering primitives—a trainable bank and feature-wise modulation—to calibrate bridge representations according to the acoustic conditions of each input utterance. In IAACM, the bank represents an input-dependent acoustic correction state, and its effect is localized by a bridge-manifold risk estimator.

3 Input-Adaptive Acoustic Condition Modulation

3.1 Speech-to-LLM formulation

Let x be an input waveform, $y = (y_1, \ldots, y_N)$ its transcript, E a frozen speech encoder, G_θ a trainable bridge, and L a frozen autoregressive LLM. The encoder produces $H = E(x) \in \mathbb{R}^{T \times d_e}$ and the bridge emits $B = G_\theta(H) \in \mathbb{R}^{M \times d_b}$. Here T and M are the encoder-frame and bridge-token counts, respectively, and d_e and d_b are their feature dimensions. A learned projection maps B to the LLM embedding dimension, where it is concatenated with a fixed ASR instruction and soft audio boundary embeddings. The unconditioned baseline trains G_θ and the speech-to-LLM projection with the autoregressive ASR objective while keeping E and L frozen.

3.2 Centered acoustic condition bank

IAACM derives its conditioning signal from trainable bridge tokens. Let B^c denote the conditioning sequence and b_t^c its tth token. During training, this sequence follows the reference schedule in Section 4.3; at inference, $B^c = B$, so target bridge tokens are reused without external reference audio or a second bridge pass.

For valid-token indicators $m_t \in \{0, 1\}$, masked attention pooling gives

$$v_t = \tanh(W_p b_t^c + a_p), \quad \omega_t = \frac{m_t \exp(w_s^\top v_t)}{\sum_j m_j \exp(w_s^\top v_j)}, \qquad q = \mathrm{LN}\left(\sum_t \omega_t v_t\right). \tag{1}$$

Let d_c denote the dimension of the conditioning space. The learned projection $W_p \in \mathbb{R}^{d_c \times d_b}$ and bias $a_p \in \mathbb{R}^{d_c}$ map bridge tokens into the conditioning space; $w_s \in \mathbb{R}^{d_c}$ is a learned pooling-score vector. The resulting $q \in \mathbb{R}^{d_c}$ is an utterance-level routing query.

It attends to a trainable condition bank $S = [s_1; \ldots; s_K] \in \mathbb{R}^{K \times d_c}$, where $K = 64$ and $d_c = 512$. With normalized bank values $\widehat{s}_k = s_k / \|s_k\|_2$, routing is

$$\begin{aligned} \overline{q} &= \frac{W_q q}{\|W_q q\|_2}, & \overline{s}_k &= \frac{W_s s_k}{\|W_s s_k\|_2}, \\ \alpha_k &= \mathrm{softmax}_k\left(\frac{\overline{q}^\top \overline{s}_k}{\tau_c}\right), & z &= \sum_{k=1}^{K} \left(\alpha_k - \frac{1}{K}\right) \widehat{s}_k. \end{aligned} \tag{2}$$

The learned matrices $W_q, W_s \in \mathbb{R}^{d_c \times d_c}$ project the query and bank vectors into a shared routing space. The routing temperature is fixed at $\tau_c = 0.5$. The centered weights make uniform routing an exact no-op: $\alpha_k = 1/K$ for all k implies $z = 0$.

3.3 The bridge manifold and token risk

Stage A first aligns the unconditioned bridge with the frozen LLM. We snapshot this bridge as G_0 and freeze it for subsequent stages. It defines the *source bridge manifold*: the token distribution learned from clean source-corpus inputs before adaptive modulation is introduced.

During Stages B and C, a weak perturbation $\mathcal{A}$ is applied to valid speech representations, using random gain and Gaussian noise. The current bridge and frozen anchor produce

$$\begin{aligned} B &= G_\theta(\mathcal{A}(H)), \\ B^0 &= G_0(H). \end{aligned} \tag{3}$$

For valid aligned token positions, the local off-manifold target is

$$\rho_t^* = \mathrm{clip}\big(1 - \cos\big(\mathrm{sg}(b_t), \mathrm{sg}(b_t^0)\big),\, 0,\, 2\big), \tag{4}$$

where sg stops gradients. A trainable token head $f_\rho : \mathbb{R}^{d_b} \to \mathbb{R}$ predicts nonnegative risk, $\widehat{\rho}_t = \mathrm{softplus}(f_\rho(b_t))$.

3.4 Risk-gated bounded residual FiLM

The trainable map $f_c : \mathbb{R}^{d_c} \to \mathbb{R}^{2d_b}$ transforms the condition state z into feature-wise scale and bias vectors, $[\Delta\gamma(z); \beta(z)] = f_c(z) - f_c(0)$. Subtracting $f_c(0)$ ensures that the bridge tokens remain unchanged when $z = 0$. The token-level gate and raw correction are

$$\begin{aligned} g_t &= \sigma(\widehat{\rho}_t / \tau_\rho + b_\rho), \\ d_t &= g_t[\Delta\gamma(z) \odot \mathrm{LN}(b_t) + \beta(z)]. \end{aligned} \tag{5}$$

The gate temperature $\tau_\rho = 1$ controls sensitivity to predicted risk, while the bias $b_\rho = -2$ sets its baseline activation; both are fixed hyperparameters. To prevent an uncertain condition estimate from overwhelming the bridge, we cap the residual norm:

$$\begin{aligned} c_t &= \min\left(1,\, \frac{\epsilon \|b_t\|_2}{\max(\|d_t\|_2, \delta)}\right), \\ \widetilde{b}_t &= b_t + c_t d_t, \end{aligned} \tag{6}$$

Here $\epsilon = 0.05$ is the maximum residual-to-token norm ratio, and $\delta = 10^{-12}$ is a numerical floor on the denominator. Thus modulation is input-dependent in two ways: z controls the direction and feature-wise form of the correction, while g_t controls where it is applied along the bridge-token sequence.

3.5 Training objective

The recognition loss is teacher-forced autoregressive cross-entropy. For a minibatch $\mathcal{B}$, let y_i denote the target token sequence including EOS and N_i its length. Let P_i be the conditioning prefix

containing the fixed instruction, projected acoustic tokens, and audio boundary embeddings. Then

$$\mathcal{L}_{\mathrm{ASR}} = -\frac{1}{\sum_{i\in\mathcal{B}} N_i} \sum_{i\in\mathcal{B}} \sum_{n=1}^{N_i} \log p_L(y_{i,n} \mid P_i, y_{i,<n}) \,. \tag{7}$$

Instruction, audio-prefix, and padding positions are excluded from the prediction targets. Gradients propagate through the frozen LLM to the trainable acoustic interface.

The risk head regresses the detached deviation target in Equation 4:

$$\begin{aligned} \mathcal{L}_{\mathrm{risk}} &= \frac{1}{|\mathcal{V}|} \sum_{t\in\mathcal{V}} \ell_{\mathrm{smooth}}(\widehat{\rho}_t - \rho_t^*) \,, \\ \ell_{\mathrm{smooth}}(e) &= \begin{cases} \frac{1}{2}e^2, & |e| < 1, \\ |e| - \frac{1}{2}, & |e| \geq 1, \end{cases} \end{aligned} \tag{8}$$

where $\mathcal{V}$ indexes valid aligned bridge tokens across the minibatch. This objective supervises the token-wise deviation estimate used by the correction gate; the frozen anchor and the target ρ_t^* receive no gradients.

The bank is regularized at the usage and geometry levels. Let $\widehat{S} = [\widehat{s}_1; \ldots; \widehat{s}_K] \in \mathbb{R}^{K\times d_c}$ collect the normalized bank vectors. For recent routing assignments, $\bar{\alpha}$ denotes mean condition-vector usage and $U_k = 1/K$:

$$\begin{aligned} \mathcal{L}_{\mathrm{bal}} &= D_{\mathrm{KL}}(\bar{\alpha} \| U) \,, \\ \mathcal{L}_{\mathrm{div}} &= \frac{\|\widehat{S}\widehat{S}^\top - I\|_F^2}{K(K-1)}. \end{aligned} \tag{9}$$

A queue of 256 detached routing vectors stabilizes $\bar{\alpha}$ for micro-batches of two without turning the regularizer into a per-example uniformity constraint.

The restoration objective gives larger weight to positions with greater measured deviation:

$$\begin{aligned} w_t &= \frac{\rho_t^* + 10^{-3}}{\frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}}(\rho_j^* + 10^{-3})}, \\ \mathcal{L}_{\mathrm{restore}} &= \frac{1}{|\mathcal{V}|} \sum_{t\in\mathcal{V}} \frac{w_t}{d_b} \left\| \widetilde{b}_t - \mathrm{sg}(b_t^0) \right\|_2^2, \end{aligned} \tag{10}$$

The full objective is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{ASR}} &+ \lambda_b \mathcal{L}_{\mathrm{bal}} + \lambda_d \mathcal{L}_{\mathrm{div}} \\ &+ \lambda_\rho \mathcal{L}_{\mathrm{risk}} + \lambda_m \mathcal{L}_{\mathrm{restore}} + \mathcal{L}_{\mathrm{bridge}}. \end{aligned} \tag{11}$$

$\mathcal{L}_{\mathrm{bridge}}$ supplies architecture-specific alignment supervision for the CIF and CTC bridges and is zero for the other six bridges. The CIF bridge uses CTC and quantity losses, whereas the CTC bridge uses only CTC loss. These auxiliary losses are shared by each baseline–IAACM pair and are defined in Appendix A.1. We use $\lambda_b = 0.001$, $\lambda_d = 0.1$, and $\lambda_\rho = \lambda_m = 0.1$ in the adaptive stages.

4 Experimental Setup

4.1 Backbones and training data

All systems use frozen Whisper large-v2 (Radford et al., 2023) and frozen Qwen3-8B (Yang et al., 2025). The LLM receives the instruction "Transcribe the speech. Return only the transcript." in non-thinking mode. We train the speech-to-LLM interface and, for IAACM, the acoustic conditioning and calibration modules, while keeping both the speech encoder and the LLM frozen. Training uses the LibriSpeech 960-hour corpus (Panayotov et al., 2015). A deterministic speaker-disjoint holdout produces 276,157 training utterances and 5,084 internal validation utterances.

4.2 Bridge families

We cover fixed-rate, query-based, convolutional, and content-aligned bridges. The set includes: (1) a Q-Former with 32 learned queries; (2) a Window Q-Former; (3) a Linear Projector; (4) a Pooled

Table 1: Evaluation corpora. Training uses only LibriSpeech 960h; every listed partition is held out.

Dataset	Utterances	Regime	Condition
In-domain			
LibriSpeech test-clean	2,620	ID	Clean read speech
LibriSpeech test-other	2,939	ID	Challenging read speech
Out-of-domain			
TED-LIUM test	1,155	OOD	Prepared talks
Common Voice 23 test	16,397	OOD	Crowdsourced speech
GigaSpeech test	19,870	OOD	Multi-domain audio
VCTK test partition	29,380	OOD	Multi-speaker read speech

MLP Projector with valid-length temporal pooling; (5) a Non-Causal Conv bridge; (6) a Causal Conv bridge; (7) a CIF bridge; and (8) a CTC bridge. The Window Q-Former, Linear Projector, Non-Causal Conv, CIF, and CTC follow bridge families studied in prior speech-to-LLM work (Yu et al., 2024; Verdini et al., 2025; Wang et al., 2024; Wu et al., 2023; Ma et al., 2024); full implementations and compression details are in Appendix A.

Each bridge has a separately trained Baseline with the condition path disabled. The paired IAACM model uses the same bridge architecture, frozen backbones, training corpus, and total number of epochs.

4.3 Training protocol

Training proceeds in three stages, each lasting one epoch. Stage A trains bridge alignment with modulation disabled. Stage B activates IAACM and samples conditioning evidence from same-speaker utterances (50%), a target prefix (25%), or the target-derived fallback (25%). Stage C uses same-speaker (60%), same-session (15%), target prefix (10%), fallback (10%), and deliberately mismatched speaker (5%) references. References need not share transcripts with the target. At inference, the target utterance always supplies the condition.

We use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4}, weight decay 0.01, bfloat16 arithmetic, batch size 2, and four-step gradient accumulation. Each model is trained on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. Weak manifold perturbation uses gain sampled from $[0.95, 1.05]$ and Gaussian noise with standard deviation 0.05 on valid bridge inputs. We evaluate all systems using the final checkpoint after three training epochs. Decoding is greedy with a maximum of 128 new tokens and no repetition or logit heuristics.

4.4 Evaluation

Table 1 summarizes the six test sets. LibriSpeech test-clean and test-other are source-domain evaluations. TED-LIUM (Hernandez et al., 2018), Common Voice 23 English (Ardila et al., 2020), GigaSpeech (Chen et al., 2021), and VCTK (Veaux et al., 2017) probe progressively different speaking and recording conditions. Common Voice and GigaSpeech use their official test partitions. Because VCTK does not provide a universal ASR test split, we use a fixed utterance-level hash partition of mic-1 recordings (seed 2027), assigning one third to development and two thirds to the reported test partition. All main and appendix WERs use the Qwen2-Audio English normalization procedure (Chu et al., 2024), applied identically to reference and hypothesis. We compute corpus-level WER as the total number of edits divided by the total number of reference words, and define relative word error rate reduction (WERR) as

$$\mathrm{WERR}\ (\%) = 100\left(1 - \frac{\mathrm{WER}_{\mathrm{IAACM}}}{\mathrm{WER}_{\mathrm{Base}}}\right), \tag{12}$$

where positive values indicate improvement. For aggregate results, WERR is computed from the corresponding mean WERs.

Table 2: WER (%, lower is better) after Stage C. Macro is the unweighted mean across all six test sets. Bold values denote the better result within each bridge pair, computed from unrounded scores.

Bridge	System	LS clean	LS other	TED	CV23	Giga	VCTK	Macro
Q-Former	Base	9.53	12.12	10.70	25.09	17.04	4.28	13.13
	IAACM	**8.92**	**11.95**	**10.26**	**23.78**	**16.58**	**4.20**	**12.61**
Window Q-Former	Base	3.12	7.19	6.60	23.71	**15.18**	**4.31**	10.02
	IAACM	**3.06**	**6.53**	**6.42**	**23.69**	15.85	4.44	**10.00**
Linear Projector	Base	2.46	**5.67**	**4.91**	18.83	14.48	**3.45**	8.30
	IAACM	**2.39**	5.87	5.27	**18.52**	**13.24**	3.67	**8.16**
Pooled MLP Projector	Base	3.91	7.62	6.78	22.66	13.50	**3.88**	9.73
	IAACM	**3.76**	**7.49**	**6.61**	**19.86**	**13.29**	3.99	**9.17**
Non-Causal Conv	Base	2.95	7.12	6.50	26.18	16.18	4.52	10.58
	IAACM	**2.67**	**6.48**	**5.53**	**22.56**	**15.77**	**4.01**	**9.50**
Causal Conv	Base	4.08	7.67	6.66	24.46	14.87	4.35	10.35
	IAACM	**4.02**	**7.52**	**6.18**	**20.69**	**14.49**	**4.20**	**9.52**
CIF	Base	**3.96**	**10.84**	**10.63**	52.92	31.20	**8.73**	19.71
	IAACM	8.86	11.04	19.36	**31.76**	**23.84**	17.77	**18.77**
CTC	Base	**3.97**	**8.63**	**8.92**	42.23	26.01	5.96	15.95
	IAACM	4.78	10.00	13.85	**34.73**	**24.12**	**5.21**	**15.45**

5 RESULTS

5.1 MAIN BRIDGE-FAMILY COMPARISON

Table 2 reports matched results. IAACM reduces the six-domain macro WER for all eight bridges. Across the 48 individual bridge–domain comparisons, 35 improve. The equally weighted average of the eight row macros decreases from 12.22% to 11.65%, a 4.68% relative reduction.

The largest bridge-level macro reductions occur for the Non-Causal Conv (10.16% relative), Causal Conv (8.04%), and Pooled MLP Projector (5.73%). The result is not merely a consequence of starting from a weak bridge: the Linear Projector has the best baseline macro WER and still improves, while the stronger source-domain performance of Non-Causal Conv is preserved and extended across all six test sets. Window Q-Former is the weakest positive case: its macro changes by only 0.02 absolute despite improvements on the two LibriSpeech sets, TED-LIUM, and Common Voice. We therefore claim transfer across diverse bridge families at the aggregate level, not universal improvement on every domain.

5.2 PERFORMANCE BY BRIDGE ALIGNMENT STRATEGY

We refer to the first six bridges in Table 2 as *Direct* bridges: they construct LLM input embeddings without an explicit speech–text token-alignment step. They improve in 30/36 bridge–domain pairs. Their aggregate macro WER decreases from 10.35% to 9.83%; both the LibriSpeech mean (6.12% to 5.89%) and the four-set OOD mean (12.46% to 11.80%) improve. These results show that bridge-local modulation improves recognition across these architectures while preserving source-domain performance.

The *Alignment-driven* group comprises CIF and CTC, which use explicit speech–text alignment to guide temporal compression and show a different performance pattern. IAACM improves their six-set macro WERs, primarily through reductions on Common Voice and GigaSpeech, while WER increases on LibriSpeech and TED-LIUM. Their combined LibriSpeech mean rises from 6.85% to 8.67%, while their four-set OOD mean falls from 23.32% to 21.33%. The shared direction of change on five of six datasets suggests that acoustic conditioning interacts with alignment-driven compression, with benefits concentrated on domains exhibiting large baseline errors.

Table 3: Error-tail analysis. "Hardest k" denotes the k datasets with the highest baseline WER among the six evaluation datasets, selected separately for each bridge. Baseline and IAACM WERs are averaged uniformly over the same selected bridge–dataset pairs. WERR is computed from these mean WERs. The three highest-WER datasets for each bridge are listed in Table 8.

Domains / bridge	Base	IAACM	WERR (%) ↑
All six	12.22	11.65	4.68
Hardest 3	18.82	16.81	10.66
Hardest 2	24.03	20.80	13.46
Hardest 1	29.51	24.45	17.15

Table 4: SRB noise/reverberation subset: mean WER (%) over all 28 corrupted conditions, excluding clean. Positive WERR denotes relative improvement in WER over the separately trained baseline; wins count conditions with lower WER. The lower WER in each row is bold. Table 9 gives the 27-condition sensitivity analysis.

Bridge	Base	IAACM	WERR (%) ↑	Wins ↑
Q-Former	22.34	**20.45**	+8.47	27/28
Window Q-Former	**14.72**	15.93	-8.21	18/28
Linear Projector	15.03	**11.76**	+21.72	18/28
Pooled MLP Projector	13.08	**11.20**	+14.40	22/28
Non-Causal Conv	13.46	**10.09**	+25.03	28/28
Causal Conv	14.83	**12.11**	+18.31	25/28
CIF	**12.92**	13.06	-1.11	6/28
CTC	15.81	**11.66**	+26.24	12/28

5.3 Benefit grows with baseline difficulty

To describe how gains vary with baseline difficulty, for each bridge, we rank the six datasets only by Baseline WER, then evaluate IAACM on the corresponding top-k hardest domains. Table 3 shows a monotonic pattern. Relative WER reduction increases from 4.68% over all domains to 10.66% on the three hardest, 13.46% on the two hardest, and 17.15% on the single hardest domain per bridge.

Common Voice is the highest-error domain for all eight bridges; the bridge-specific top-three rankings are provided in Table 8.

5.4 Trainable-parameter cost

Both foundation backbones remain frozen in every experiment. For bridges with 1,024-dimensional internal tokens, IAACM adds approximately 3.84M trainable parameters. For the Linear Projector, calibration operates directly in the 4,096-dimensional LLM embedding space, compared with the 1,024-dimensional intermediate space of the other bridges. Because the FiLM generator uses a hidden width of d_b and an output width of $2d_b$, its parameter count grows quadratically with d_b. This larger calibration dimension brings the total IAACM parameter overhead to 38.86M, the largest among the evaluated bridges. Total trainable parameters range from 10.42M for the conditioned Pooled MLP Projector to 154.81M for conditioned CIF, dominated in the latter case by the content-alignment bridge itself. Table 6 reports parameter counts.

5.5 Robustness to noise and reverberation

Cross-corpus differences combine acoustic and linguistic shifts. To isolate acoustic changes while holding utterances and transcripts fixed, we evaluate all eight bridges on a noise-and-reverberation subset of Speech Robust Bench (SRB) (Shah et al., 2025), using its released LibriSpeech test-clean recordings. Seven selected perturbation types at four severity levels yield 28 corrupted versions of each of 2,620 utterances. Each comparison uses the separately trained baseline and full IAACM model from the main evaluation, with matched input handling and decoding. Table 4 reports the mean WER across the 28 noise and reverberation conditions, excluding clean speech.

IAACM lowers mean corrupted-condition WER for six of eight bridges and improves 156/224 bridge–condition pairs. Non-Causal Conv improves all 28 conditions, reducing mean WER from 13.46% to 10.09% (25.03% relative). Causal Conv improves 25/28 conditions with an 18.31% relative reduction, while Q-Former improves 27/28. Pooled MLP Projector, Linear Projector, and CTC also improve their mean WER, whereas Window Q-Former and CIF show increases in mean WER. These results extend the cross-corpus evidence to controlled acoustic perturbations across multiple bridge designs. Appendix C provides detailed results by corruption family and individual condition.

5.6 Representation movement

We examine representation movement across all eight bridges on the six evaluation datasets and the selected SRB noise and reverberation conditions. For each bridge, the separately trained baseline Stage-C model encodes 276,157 LibriSpeech training utterances to form a fixed source reference. We compare paired IAACM Stage-C representations before and after correction, mean-pool valid LLM-input embeddings, project them to 512 dimensions, and standardize using source statistics. We measure corpus-centroid distance and mean distance to the ten nearest source neighbors (10-NN); Appendix D gives the protocol and detailed distance tables.

On the six evaluation datasets, Q-Former, Linear Projector, Non-Causal Conv, Causal Conv, CIF, and CTC reduce both centroid and 10-NN distances on every dataset (Table 14). Their mean 10-NN distance reductions range from 1.14% to 3.69%, with Causal Conv and Non-Causal Conv reaching 3.69% and 3.21%, respectively. The same six bridges reduce 10-NN distance under every one of the 28 selected SRB corruption conditions, with mean reductions of 0.94%–3.68% (Table 15). The 27-condition sensitivity analysis excluding Real RIR 4 retains this pattern (Table 16). In contrast, Window Q-Former shows almost no movement in either evaluation, while Pooled MLP Projector increases its mean 10-NN distance by 3.26% across the six datasets and 3.41% across SRB.

6 Conclusion

We introduced IAACM, an input-adaptive calibration method operating in the trainable bridge space of a speech-to-LLM system. A centered acoustic condition bank determines the correction direction, while a token-level bridge-manifold risk controls a bounded residual FiLM update. Across eight bridge architectures and six test sets, IAACM improves every bridge's macro WER and 35/48 individual pairs, with increasingly large gains on each bridge's highest-error domains. Controlled SRB corruptions extend these gains to six of eight bridge-level means and 156/224 bridge–condition pairs. Representation-distance analysis shows movement toward the baseline Stage-C source for six bridge architectures across both evaluation suites. The results establish bridge-local acoustic calibration as an effective approach to improving speech recognition robustness across diverse bridge architectures.

AI use statement

Generative AI tools assisted with literature search, implementation and debugging, result analysis and presentation, and manuscript editing. The authors are responsible for verifying AI-assisted content and for the final manuscript, code, results, and claims.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proceedings of Interspeech*, pp. 3670–3674, 2021. doi: 10.21437/Interspeech.2021-1965. URL https://www.isca-archive.org/interspeech_2021/chen21o_interspeech.html.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. Closing the gap between text and speech understanding in LLMs. In *Proceedings of International Conference on Learning Representations*, volume 2026, pp. 146313–146335, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ec4c05921d70202023898568a2a8c002-Paper-Conference.pdf.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026. URL https://arxiv.org/abs/2604.27393.

Linhao Dong and Bo Xu. CIF: Continuous integrate-and-fire for end-to-end speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6079–6083, 2020. doi: 10.1109/ICASSP40776.2020.9054250. URL https://doi.org/10.1109/ICASSP40776.2020.9054250.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 690–696, 2021. doi: 10.18653/v1/2021.eacl-main.57. URL https://aclanthology.org/2021.eacl-main.57/.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, volume 11096 of *Lecture Notes in Computer Science*, pp. 198–208. Springer International Publishing, 2018. doi: 10.1007/978-3-319-99579-3_21. URL https://doi.org/10.1007/978-3-319-99579-3_21.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. An embarrassingly simple approach for LLM with strong ASR capacity. *arXiv preprint arXiv:2402.08846*, 2024. URL https://arxiv.org/abs/2402.08846.

Junseok Oh and Ji-Hwan Kim. Adapter-only bridging of frozen speech encoder and frozen LLM for ASR. In *Advances in Knowledge Discovery and Data Mining*, volume 16600 of *Lecture Notes in Computer Science*, pp. 251–262. Springer Nature Singapore, 2026. doi: 10.1007/978-981-92-1468-6_15. URL https://doi.org/10.1007/978-981-92-1468-6_15.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28492–28518, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

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, et al. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023. URL https://arxiv.org/abs/2306.12925.

Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkilä, Bhiksha Raj, and Nicolas Kourtellis. Speech Robust Bench: A robustness benchmark for speech recognition. In *Proceedings of International Conference on Learning Representations*, volume 2025, pp. 38625–38651, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/605e02ae04cba1ebf6a08206299e76b9-Paper-Conference.pdf.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. LLaSM: Large language and speech model. *arXiv preprint arXiv:2308.15930*, 2023. URL https://arxiv.org/abs/2308.15930.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *Proceedings of International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/476ab8f369e489c04187ba84f68cfa68-Paper-Conference.pdf.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017. URL https://doi.org/10.7488/ds/1994.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sébastien Bratières, et al. How to connect speech foundation models and large language models? What matters and what does not. In *Proceedings of Interspeech*, pp. 1813–1817, 2025. doi: 10.21437/Interspeech.2025-2245. URL https://www.isca-archive.org/interspeech_2025/verdini25_interspeech.html.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. BLSP: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*, 2023. URL https://arxiv.org/abs/2309.00916.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. BLSP-KD: Bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*, 2024. URL https://arxiv.org/abs/2405.19041.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 5180–5189, 2018. URL https://proceedings.mlr.press/v80/wang18h.html.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. On decoder-only architecture for speech-to-text and large language model integration. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8, 2023. doi: 10.1109/ASRU57964.2023.10389705. URL https://doi.org/10.1109/ASRU57964.2023.10389705.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Connecting speech encoder and large language model for ASR. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12637–12641, 2024. doi: 10.1109/ICASSP48485.2024.10445874. URL https://doi.org/10.1109/ICASSP48485.2024.10445874.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, 2023. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

Table 5: Bridge architectures in the paired study. Rows group Q-Former, projection, convolutional, and alignment-based designs; the second column records the exact variant implemented.

Bridge	Architecture details
Q-Former	32 learned queries with four cross-attention decoder layers, 1,024 hidden units, and eight heads.
Window Q-Former	Non-overlapping 16-frame windows, one shared learned query per window, and four decoder layers.
Linear Projector	A single framewise linear map from the Whisper hidden dimension to the Qwen embedding dimension.
Pooled MLP Projector	A framewise two-layer MLP followed by length-aware adaptive pooling to 32 tokens over each utterance's valid frames.
Non-Causal Conv	Two Transformer layers, two symmetric stride-2 temporal convolutions, and two post-convolution Transformer layers; hidden size 768, FFN size 3,072, and 12 heads.
Causal Conv	One left-padded temporal convolution ($k = 5$, stride 4), followed by a two-layer MLP; outputs are capped at 128 tokens.
CIF	Two Transformer layers, Continuous Integrate-and-Fire emission, and two post-compression Transformer layers, with CTC and quantity losses.
CTC	Two Transformer layers, averaging of consecutive frames assigned the same nonblank CTC label, and two post-compression Transformer layers.

A Bridge Implementations

Table 5 records the exact bridge variants used in the main comparison.

A.1 Bridge-specific auxiliary objectives

The alignment-driven bridges receive transcript supervision before temporal compression. For utterance i, let u_i be its sequence of transcript tokens excluding EOS and padding, $U_i = |u_i|$, and T_i the number of valid pre-compression frames. A framewise classification head predicts $q_{i,t}$ over the transcript vocabulary augmented with a distinct CTC blank. Let $\mathcal{B}_+ = \{i \in \mathcal{B} : T_i > 0\}$. If $\mathcal{C}$ merges consecutive repeated labels and then removes blanks, and $\pi = (\pi_1, \ldots, \pi_{T_i})$ is a frame-level label path that may include blanks, the auxiliary objective is

$$\mathcal{L}_{\mathrm{CTC}} = -\frac{1}{|\mathcal{B}_+|} \sum_{i \in \mathcal{B}_+} \frac{1}{\max(U_i, 1)} \log \left[\sum_{\pi:\, \mathcal{C}(\pi) = u_i} \prod_{t=1}^{T_i} q_{i,t}(\pi_t) \right]. \tag{13}$$

The reduction first normalizes each utterance's loss by its target length and then averages over the minibatch. Utterances without valid frames are omitted; infinite losses from infeasible alignments contribute zero.

For the CIF bridge, let $a_{i,t} \in (0, 1)$ be the predicted integration weight at valid frame t. The quantity objective matches the total unnormalized integration mass to the number of transcript tokens:

$$\mathcal{L}_{\mathrm{qty}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left| \sum_{t=1}^{T_i} a_{i,t} - \max(U_i, 1) \right|. \tag{14}$$

This loss is evaluated before the training-time rescaling of integration weights used to control the number of emitted tokens.

The bridge term in Equation 11 is

$$\mathcal{L}_{\mathrm{bridge}} = \begin{cases} 0.1\mathcal{L}_{\mathrm{CTC}} + 0.1\mathcal{L}_{\mathrm{qty}}, & \text{CIF bridge,} \\ 0.1\mathcal{L}_{\mathrm{CTC}}, & \text{CTC bridge,} \\ 0, & \text{all other bridges.} \end{cases} \tag{15}$$

Table 6: Trainable parameter counts in millions (M), rounded to two decimal places, for each matched pair. The final column is the additional capacity introduced by IAACM.

Bridge	Base (M)	IAACM (M)	Added (M)
Q-Former	72.74	76.58	3.84
Window Q-Former	72.71	76.55	3.84
Linear Projector	5.26	44.12	38.86
Pooled MLP Projector	6.57	10.42	3.84
Non-Causal Conv	37.87	41.72	3.84
Causal Conv	12.86	16.71	3.84
CIF	150.97	154.81	3.84
CTC	150.97	154.81	3.84

These auxiliary objectives and weights are identical for the separately trained baseline and IAACM model of each architecture and are retained throughout all three training epochs.

B Additional Aggregate Views

Table 7: Aggregate WER by bridge regime. Each value is a uniform mean over the corresponding bridge–dataset cells.

Regime	Base	IAACM	WERR (%) ↑
All bridges, all sets	12.22	11.65	4.68
Direct, all sets	10.35	9.83	5.05
Direct, ID	6.12	5.89	3.78
Direct, OOD	12.46	11.80	5.36
Alignment-driven, ID	6.85	8.67	-26.59
Alignment-driven, OOD	23.32	21.33	8.55

Table 7 groups bridges by how they construct LLM input tokens. 'Direct' contains the first six bridges in Table 2, which map and aggregate speech features without an explicit speech–text token-alignment step. 'Alignment-driven' contains CIF and CTC, whose token construction and temporal compression are guided by auxiliary speech–text alignment supervision.

Table 8: The three highest-WER datasets for each bridge, ranked by baseline WER in descending order. Hardest 1 in Table 3 uses Rank 1; Hardest 2 uses Ranks 1 and 2; Hardest 3 uses Ranks 1, 2, and 3.

Bridge	Rank 1	Rank 2	Rank 3
Q-Former	Common Voice 23	GigaSpeech	LibriSpeech test-other
Window Q-Former	Common Voice 23	GigaSpeech	LibriSpeech test-other
Linear Projector	Common Voice 23	GigaSpeech	LibriSpeech test-other
Pooled MLP Projector	Common Voice 23	GigaSpeech	LibriSpeech test-other
Non-Causal Conv	Common Voice 23	GigaSpeech	LibriSpeech test-other
Causal Conv	Common Voice 23	GigaSpeech	LibriSpeech test-other
CIF	Common Voice 23	GigaSpeech	LibriSpeech test-other
CTC	Common Voice 23	GigaSpeech	TED-LIUM

C Noise and Reverberation Evaluation

Corruptions and input protocol. We evaluate all eight bridges on a noise-and-reverberation subset of the released SRB LibriSpeech test-clean data (Shah et al., 2025), using the provided waveforms without resynthesizing corruptions. The selected perturbations are Gaussian Noise, Environmental Noise from MS-SNSD, ESC-50, MUSAN, and WHAM!, and simulated and real room impulse

responses, labeled RIR and Real RIR. The condition tables identify each environmental-noise source in parentheses. Each of these seven perturbation types has four released severity settings. Together with clean, this yields 29 conditions and 75,980 inputs per model. Each condition contains the same 2,620 utterances and 53,029 reference words after Qwen2-Audio English normalization.

The baseline and full IAACM systems use the separately trained Stage-C checkpoints from Table 2, with identical input handling and decoding. Recordings longer than 30 seconds are divided into non-overlapping chunks. Each chunk is decoded greedily with a maximum of 128 new tokens; transcripts are concatenated before scoring. This input handling differs from the main cross-corpus evaluation, so the SRB clean result is reported separately. We compute corpus WER for each condition and take the unweighted mean over the 28 corrupted conditions in Table 4.

Sensitivity to the highest real-reverberation severity. Table 9 excludes the real-room reverberation condition at SRB severity level 4, labeled Real RIR 4 in the condition-level tables. Clean remains excluded, and the full 28-condition result remains the primary evaluation. On the remaining 27 conditions, Q-Former, both convolutional bridges, both projectors, and Window Q-Former retain or achieve mean improvements, ranging from 2.18% to 11.34% relative. Non-Causal Conv improves all 27 conditions, followed by Q-Former with 26/27 and Causal Conv with 24/27. Thus, the gains for these architectures extend across the broader corruption set.

Table 9: SRB sensitivity analysis: mean WER (%) over 27 corrupted conditions, excluding only Real RIR 4 from Table 4. Models, decoding, normalization, and metric definitions are identical to the primary result.

Bridge	Base	IAACM	WERR (%) ↑	Wins ↑
Q-Former	12.67	**12.17**	+3.94	26/27
Window Q-Former	6.65	**6.50**	+2.18	18/27
Linear Projector	5.15	**4.94**	+4.15	17/27
Pooled MLP Projector	7.51	**7.30**	+2.91	21/27
Non-Causal Conv	6.98	**6.19**	+11.34	27/27
Causal Conv	7.95	**7.22**	+9.26	24/27
CIF	**9.02**	11.12	-23.30	5/27
CTC	**9.20**	9.21	-0.17	11/27

The condition-level results distinguish different severity responses. CTC benefits especially from Real RIR 4, where WER falls from 194.44% to 77.83%; its remaining-condition mean is nearly unchanged (9.20% to 9.21%). CIF also improves on this condition (118.13% to 65.39%), while its remaining-condition mean rises from 9.02% to 11.12%. Window Q-Former shows the reverse pattern: its 27-condition mean improves from 6.65% to 6.50%, whereas WER on Real RIR 4 rises from 232.61% to 270.37%. WER can exceed 100% because insertions are unbounded.

Condition-level uncertainty. For each bridge and condition, we compute the IAACM minus baseline difference in corpus WER and a percentile 95% interval from 1,000 paired utterance bootstrap resamples (seed 2027). Tables 10–13 report all 29 selected conditions for all eight bridges. Intervals describe evaluation-sample uncertainty for fixed checkpoints and are marginal, without multiple-comparison adjustment.

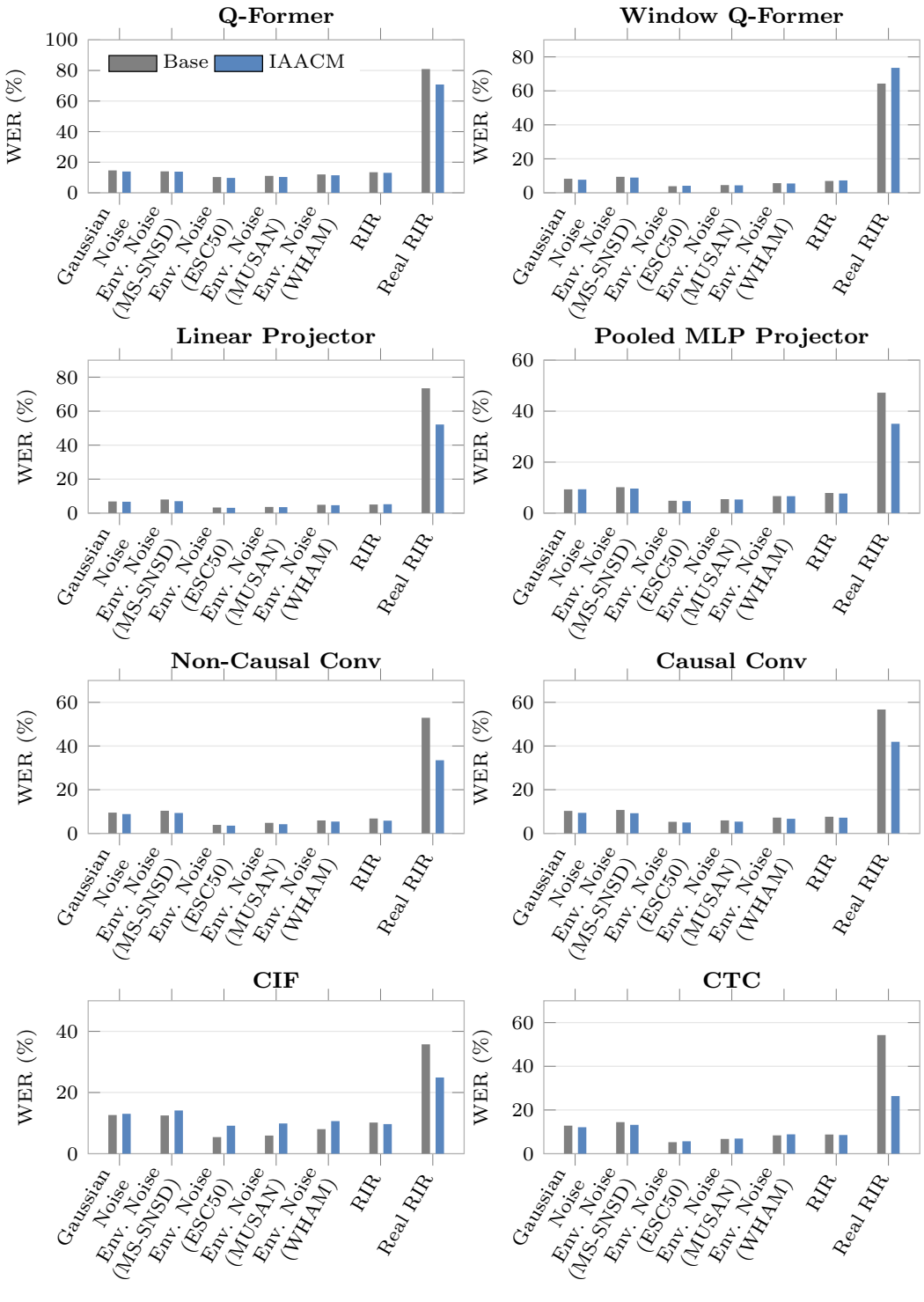


Figure 2: WER by perturbation type in the SRB noise-and-reverberation subset for all eight bridges, averaged over each type's four severity settings. Gray: separately trained baseline; blue: full IAACM. All 28 corruption conditions, including Real RIR 4, contribute to these means. Axis ranges vary by bridge. Table 9 separately reports the 27-condition sensitivity analysis.

Table 10: Q-Former and Window Q-Former on all 29 selected SRB conditions. WER is in percent; $\Delta = \mathrm{IAACM} - \mathrm{Base}$ and its paired bootstrap 95% interval are in percentage points. Negative differences favor IAACM; the lower WER in each pair is bold. Suffixes 1–4 denote released severity settings.

	Q-Former				Window Q-Former			
Condition	Base	IAACM	$\Delta \downarrow$	95% CI	Base	IAACM	$\Delta \downarrow$	95% CI
Clean	9.52	**8.78**	-0.74	[-1.20, -0.34]	3.08	**2.97**	-0.11	[-0.27, 0.06]
Gaussian Noise 1	9.79	**9.02**	-0.77	[-1.31, -0.22]	3.18	**3.02**	-0.15	[-0.31, 0.01]
Gaussian Noise 2	10.10	**9.62**	-0.48	[-1.09, 0.23]	3.49	**3.18**	-0.31	[-0.47, -0.15]
Gaussian Noise 3	11.57	**10.96**	-0.61	[-1.04, -0.23]	5.02	**4.97**	-0.05	[-0.39, 0.47]
Gaussian Noise 4	27.13	**26.11**	-1.02	[-2.20, 0.23]	21.37	**19.61**	-1.75	[-3.13, -0.36]
Env. Noise (MS-SNSD) 1	9.35	**9.30**	-0.05	[-0.46, 0.43]	**3.29**	3.47	+0.18	[-0.19, 0.73]
Env. Noise (MS-SNSD) 2	9.76	**9.60**	-0.17	[-0.62, 0.29]	**4.12**	4.63	+0.51	[-0.07, 1.20]
Env. Noise (MS-SNSD) 3	**11.65**	11.68	+0.03	[-0.53, 0.65]	7.02	**6.26**	-0.77	[-1.57, -0.12]
Env. Noise (MS-SNSD) 4	25.28	**24.71**	-0.58	[-1.60, 0.44]	23.22	**21.48**	-1.74	[-3.93, 0.44]
Env. Noise (ESC50) 1	9.70	**8.81**	-0.89	[-1.35, -0.48]	**3.12**	3.45	+0.34	[-0.20, 1.04]
Env. Noise (ESC50) 2	9.63	**9.05**	-0.58	[-1.11, -0.07]	**3.34**	3.36	+0.02	[-0.21, 0.35]
Env. Noise (ESC50) 3	9.64	**9.34**	-0.30	[-0.74, 0.12]	**3.44**	3.90	+0.47	[-0.22, 1.37]
Env. Noise (ESC50) 4	12.21	**11.70**	-0.50	[-0.92, -0.09]	**5.62**	5.81	+0.19	[-0.37, 0.91]
Env. Noise (MUSAN) 1	9.67	**9.19**	-0.48	[-1.09, 0.13]	3.11	**2.95**	-0.16	[-0.31, -0.00]
Env. Noise (MUSAN) 2	9.85	**9.27**	-0.58	[-1.20, -0.01]	3.25	**3.02**	-0.23	[-0.38, -0.06]
Env. Noise (MUSAN) 3	10.38	**9.81**	-0.57	[-1.26, 0.15]	3.66	**3.48**	-0.17	[-0.33, 0.00]
Env. Noise (MUSAN) 4	14.38	**13.07**	-1.31	[-1.96, -0.75]	8.15	**8.11**	-0.05	[-0.98, 0.92]
Env. Noise (WHAM) 1	9.69	**8.82**	-0.86	[-1.40, -0.41]	3.09	**2.98**	-0.11	[-0.27, 0.04]
Env. Noise (WHAM) 2	9.20	**8.99**	-0.21	[-0.60, 0.15]	3.17	**3.04**	-0.14	[-0.29, 0.04]
Env. Noise (WHAM) 3	10.19	**9.85**	-0.35	[-0.71, 0.03]	3.79	**3.62**	-0.17	[-0.42, 0.14]
Env. Noise (WHAM) 4	18.94	**18.19**	-0.75	[-1.37, -0.14]	12.66	**12.30**	-0.36	[-1.14, 0.43]
RIR 1	10.61	**10.42**	-0.19	[-0.81, 0.56]	4.24	**4.22**	-0.02	[-0.78, 0.82]
RIR 2	12.28	**11.64**	-0.64	[-1.18, -0.09]	**5.50**	5.81	+0.31	[-0.43, 1.11]
RIR 3	15.05	**14.82**	-0.24	[-1.45, 0.84]	**9.10**	9.72	+0.63	[-0.68, 1.99]
RIR 4	15.83	**15.35**	-0.48	[-1.28, 0.24]	**9.02**	9.39	+0.36	[-0.80, 1.40]
Real RIR 1	9.68	**9.24**	-0.45	[-0.88, 0.02]	3.33	**3.09**	-0.24	[-0.38, -0.08]
Real RIR 2	15.41	**15.38**	-0.03	[-0.87, 0.73]	11.33	**11.20**	-0.13	[-0.36, 0.10]
Real RIR 3	14.97	**14.57**	-0.41	[-1.20, 0.32]	9.88	**9.51**	-0.37	[-1.28, 0.57]
Real RIR 4	283.63	**244.11**	-39.52	[-46.59, -32.29]	**232.61**	270.37	+37.76	[28.02, 46.73]

Table 11: Linear Projector and Pooled MLP Projector (with valid-length masking) on all 29 selected SRB conditions. Definitions and paired bootstrap intervals follow Table 10.

	Linear Projector				Pooled MLP Projector			
Condition	Base	IAACM	$\Delta \downarrow$	95% CI	Base	IAACM	$\Delta \downarrow$	95% CI
Clean	**2.42**	2.59	+0.17	[-0.28, 0.77]	3.87	**3.83**	-0.04	[-0.56, 0.42]
Gaussian Noise 1	2.46	**2.40**	-0.05	[-0.32, 0.14]	3.83	**3.68**	-0.14	[-0.46, 0.11]
Gaussian Noise 2	2.75	**2.62**	-0.13	[-0.44, 0.08]	4.26	**4.09**	-0.17	[-0.60, 0.14]
Gaussian Noise 3	4.00	**3.97**	-0.04	[-0.29, 0.17]	**5.89**	6.14	+0.25	[-0.10, 0.67]
Gaussian Noise 4	18.21	**17.76**	-0.44	[-1.69, 0.79]	**23.25**	23.48	+0.23	[-1.15, 1.62]
Env. Noise (MS-SNSD) 1	**2.52**	2.70	+0.17	[-0.21, 0.58]	4.19	**3.91**	-0.28	[-0.80, 0.15]
Env. Noise (MS-SNSD) 2	**3.00**	3.16	+0.16	[-0.49, 0.70]	5.13	**4.46**	-0.67	[-1.34, -0.11]
Env. Noise (MS-SNSD) 3	**4.64**	4.87	+0.22	[-0.16, 0.55]	7.55	**7.49**	-0.06	[-0.62, 0.56]
Env. Noise (MS-SNSD) 4	21.87	**17.33**	-4.54	[-6.72, -2.55]	23.76	**22.65**	-1.11	[-2.89, 0.55]
Env. Noise (ESC50) 1	2.84	**2.61**	-0.23	[-1.08, 0.51]	3.99	**3.83**	-0.16	[-0.65, 0.24]
Env. Noise (ESC50) 2	2.67	**2.42**	-0.25	[-0.77, 0.05]	4.21	**4.04**	-0.17	[-0.59, 0.21]
Env. Noise (ESC50) 3	2.96	**2.89**	-0.07	[-0.74, 0.66]	4.29	**4.15**	-0.14	[-0.35, 0.05]
Env. Noise (ESC50) 4	4.72	**4.49**	-0.23	[-0.71, 0.12]	6.87	**6.82**	-0.04	[-0.73, 0.52]
Env. Noise (MUSAN) 1	2.38	**2.35**	-0.03	[-0.33, 0.17]	4.11	**3.87**	-0.24	[-0.91, 0.32]
Env. Noise (MUSAN) 2	2.54	**2.46**	-0.07	[-0.40, 0.15]	**4.05**	4.08	+0.02	[-0.15, 0.20]
Env. Noise (MUSAN) 3	**2.80**	2.81	+0.01	[-0.11, 0.12]	4.84	**4.42**	-0.42	[-1.18, 0.12]
Env. Noise (MUSAN) 4	6.78	**6.42**	-0.36	[-1.12, 0.24]	**8.88**	8.98	+0.10	[-0.64, 0.81]
Env. Noise (WHAM) 1	2.49	**2.40**	-0.10	[-0.40, 0.11]	4.13	**3.78**	-0.35	[-1.00, 0.17]
Env. Noise (WHAM) 2	2.53	**2.42**	-0.11	[-0.40, 0.10]	4.16	**3.98**	-0.18	[-0.65, 0.26]
Env. Noise (WHAM) 3	3.20	**2.95**	-0.25	[-0.86, 0.13]	4.81	**4.48**	-0.32	[-0.83, 0.04]
Env. Noise (WHAM) 4	11.35	**10.83**	-0.53	[-1.64, 0.41]	**13.55**	14.29	+0.74	[-0.07, 1.68]
RIR 1	**3.07**	3.14	+0.07	[-0.33, 0.65]	5.26	**4.72**	-0.54	[-1.27, -0.04]
RIR 2	**4.10**	4.31	+0.21	[-0.28, 0.77]	6.67	**6.58**	-0.10	[-0.79, 0.65]
RIR 3	**6.10**	6.40	+0.30	[-0.32, 1.01]	**8.78**	9.19	+0.41	[-0.29, 1.07]
RIR 4	**6.92**	7.05	+0.13	[-0.52, 0.73]	10.88	**10.24**	-0.63	[-1.81, 0.40]
Real RIR 1	**2.51**	2.55	+0.04	[-0.08, 0.14]	4.19	**3.85**	-0.34	[-0.80, 0.02]
Real RIR 2	**3.12**	3.58	+0.46	[0.19, 0.80]	11.05	**9.55**	-1.50	[-2.74, -0.29]
Real RIR 3	6.53	**6.41**	-0.12	[-0.86, 0.66]	10.31	**10.20**	-0.11	[-1.09, 0.81]
Real RIR 4	281.64	**196.04**	-85.60	[-96.45, -74.94]	163.44	**116.60**	-46.84	[-56.02, -37.85]

Table 12: Non-Causal Conv and Causal Conv on all 29 selected SRB conditions. Definitions and paired bootstrap intervals follow Table 10.

	Non-Causal Conv				Causal Conv			
Condition	Base	IAACM	$\Delta \downarrow$	95% CI	Base	IAACM	$\Delta \downarrow$	95% CI
Clean	2.91	**2.83**	-0.08	[-0.39, 0.42]	**4.12**	4.21	+0.08	[-0.53, 0.94]
Gaussian Noise 1	3.21	**2.89**	-0.32	[-0.95, 0.26]	4.52	**4.15**	-0.37	[-1.17, 0.31]
Gaussian Noise 2	3.33	**3.09**	-0.24	[-0.57, 0.26]	4.94	**4.36**	-0.58	[-1.37, 0.06]
Gaussian Noise 3	5.48	**4.97**	-0.51	[-1.38, 0.25]	6.58	**6.05**	-0.53	[-1.19, 0.14]
Gaussian Noise 4	26.01	**24.42**	-1.60	[-4.14, 0.83]	25.12	**23.06**	-2.06	[-3.80, -0.35]
Env. Noise (MS-SNSD) 1	2.98	**2.82**	-0.16	[-0.29, -0.02]	4.54	**4.19**	-0.35	[-0.88, 0.18]
Env. Noise (MS-SNSD) 2	3.77	**3.48**	-0.28	[-0.72, 0.00]	5.30	**5.11**	-0.19	[-0.76, 0.46]
Env. Noise (MS-SNSD) 3	6.62	**5.92**	-0.70	[-1.64, 0.18]	7.50	**6.79**	-0.71	[-1.29, -0.01]
Env. Noise (MS-SNSD) 4	28.08	**25.16**	-2.92	[-5.89, -0.23]	25.54	**20.83**	-4.71	[-6.78, -2.66]
Env. Noise (ESC50) 1	3.00	**2.86**	-0.14	[-0.45, 0.39]	**4.09**	4.11	+0.02	[-0.59, 0.79]
Env. Noise (ESC50) 2	3.27	**2.91**	-0.36	[-1.01, 0.21]	**4.27**	4.38	+0.12	[-0.53, 0.88]
Env. Noise (ESC50) 3	3.25	**2.98**	-0.27	[-0.43, -0.12]	4.92	**4.47**	-0.44	[-1.07, 0.01]
Env. Noise (ESC50) 4	6.04	**5.41**	-0.63	[-1.21, -0.11]	7.92	**6.99**	-0.93	[-1.89, -0.04]
Env. Noise (MUSAN) 1	2.95	**2.58**	-0.38	[-0.52, -0.23]	4.34	**4.12**	-0.22	[-0.67, 0.22]
Env. Noise (MUSAN) 2	3.31	**2.75**	-0.56	[-1.09, -0.23]	4.34	**4.20**	-0.14	[-0.54, 0.31]
Env. Noise (MUSAN) 3	3.77	**3.55**	-0.22	[-0.91, 0.40]	4.75	**4.73**	-0.02	[-0.47, 0.43]
Env. Noise (MUSAN) 4	9.25	**7.94**	-1.30	[-2.62, -0.02]	10.34	**8.47**	-1.87	[-2.97, -0.92]
Env. Noise (WHAM) 1	2.95	**2.67**	-0.28	[-0.42, -0.14]	4.25	**4.01**	-0.24	[-0.69, 0.28]
Env. Noise (WHAM) 2	3.08	**2.72**	-0.36	[-0.50, -0.21]	**4.21**	4.22	+0.02	[-0.53, 0.62]
Env. Noise (WHAM) 3	3.68	**3.47**	-0.21	[-0.57, 0.35]	4.87	**4.74**	-0.13	[-0.59, 0.37]
Env. Noise (WHAM) 4	13.99	**12.78**	-1.21	[-2.39, -0.07]	15.51	**13.74**	-1.77	[-2.90, -0.66]
RIR 1	3.73	**3.62**	-0.11	[-0.46, 0.47]	4.96	**4.67**	-0.29	[-0.90, 0.35]
RIR 2	5.72	**5.15**	-0.57	[-1.57, 0.37]	6.48	**6.02**	-0.46	[-0.93, -0.01]
RIR 3	8.30	**6.87**	-1.43	[-2.44, -0.61]	8.82	**8.57**	-0.25	[-0.83, 0.43]
RIR 4	9.52	**7.65**	-1.88	[-3.02, -0.94]	10.22	**9.46**	-0.76	[-1.62, -0.02]
Real RIR 1	3.21	**2.86**	-0.35	[-0.52, -0.20]	4.42	**4.32**	-0.11	[-0.51, 0.26]
Real RIR 2	9.80	**7.30**	-2.50	[-3.32, -1.74]	11.52	**9.26**	-2.26	[-2.95, -1.61]
Real RIR 3	10.07	**8.21**	-1.87	[-2.69, -1.18]	10.40	**9.78**	-0.62	[-1.34, 0.15]
Real RIR 4	188.59	**115.59**	-72.99	[-82.40, -63.33]	200.50	**144.36**	-56.14	[-65.33, -46.43]

Table 13: CIF and CTC on all 29 selected SRB conditions. Definitions and paired bootstrap intervals follow Table 10.

	CIF				CTC			
Condition	Base	IAACM	Δ ↓	95% CI	Base	IAACM	Δ ↓	95% CI
Clean	**3.77**	8.87	+5.11	[3.85, 6.45]	**3.91**	4.73	+0.83	[-0.00, 1.68]
Gaussian Noise 1	**3.62**	8.43	+4.81	[3.57, 6.17]	**3.81**	4.80	+0.99	[0.14, 1.97]
Gaussian Noise 2	**3.73**	8.01	+4.28	[3.23, 5.47]	4.45	**4.45**	-0.00	[-1.03, 0.92]
Gaussian Noise 3	**7.28**	9.62	+2.33	[0.60, 4.05]	**7.22**	7.88	+0.67	[-0.80, 2.04]
Gaussian Noise 4	35.91	**26.09**	-9.82	[-13.43, -6.40]	35.83	**31.27**	-4.56	[-8.27, -0.84]
Env. Noise (MS-SNSD) 1	**3.97**	8.81	+4.85	[3.65, 6.23]	**3.89**	4.54	+0.65	[-0.43, 1.74]
Env. Noise (MS-SNSD) 2	**4.48**	9.81	+5.32	[3.94, 6.74]	6.71	**5.99**	-0.72	[-2.13, 0.58]
Env. Noise (MS-SNSD) 3	**8.72**	12.16	+3.44	[1.61, 5.55]	11.97	**10.84**	-1.13	[-3.34, 1.10]
Env. Noise (MS-SNSD) 4	32.81	**25.73**	-7.07	[-10.69, -3.52]	35.15	**31.40**	-3.75	[-7.58, -0.14]
Env. Noise (ESC50) 1	**4.32**	8.68	+4.36	[3.08, 5.85]	**3.92**	4.71	+0.79	[-0.14, 1.72]
Env. Noise (ESC50) 2	**3.93**	8.30	+4.36	[3.19, 5.59]	**3.90**	4.79	+0.89	[-0.02, 1.82]
Env. Noise (ESC50) 3	**5.54**	8.99	+3.45	[1.82, 5.13]	**4.34**	4.84	+0.49	[-0.47, 1.45]
Env. Noise (ESC50) 4	**7.87**	10.61	+2.74	[1.19, 4.28]	8.85	**8.34**	-0.51	[-2.10, 0.99]
Env. Noise (MUSAN) 1	**3.94**	8.81	+4.87	[3.52, 6.30]	**4.09**	4.63	+0.54	[-0.49, 1.58]
Env. Noise (MUSAN) 2	**3.74**	8.95	+5.22	[4.10, 6.47]	4.60	**4.26**	-0.34	[-1.44, 0.63]
Env. Noise (MUSAN) 3	**4.78**	8.47	+3.69	[2.31, 5.05]	**4.61**	5.80	+1.19	[0.14, 2.30]
Env. Noise (MUSAN) 4	**11.24**	13.29	+2.05	[0.05, 3.96]	13.68	**12.95**	-0.73	[-2.91, 1.53]
Env. Noise (WHAM) 1	**3.47**	8.68	+5.21	[4.04, 6.46]	**3.72**	4.77	+1.05	[0.12, 2.05]
Env. Noise (WHAM) 2	**3.92**	8.53	+4.62	[3.32, 5.90]	**4.08**	4.84	+0.76	[-0.17, 1.77]
Env. Noise (WHAM) 3	**4.84**	7.89	+3.05	[1.76, 4.26]	**4.82**	5.69	+0.88	[-0.27, 1.96]
Env. Noise (WHAM) 4	19.84	**17.50**	-2.33	[-4.74, 0.07]	20.87	**20.09**	-0.77	[-3.71, 1.72]
RIR 1	**5.56**	7.82	+2.26	[0.79, 3.69]	**4.83**	5.82	+0.99	[-0.17, 2.13]
RIR 2	**7.89**	8.91	+1.02	[-0.61, 2.70]	7.99	**6.45**	-1.54	[-3.19, 0.02]
RIR 3	12.17	**10.92**	-1.24	[-3.19, 0.57]	**9.53**	10.27	+0.74	[-0.69, 2.14]
RIR 4	15.08	**11.04**	-4.04	[-6.72, -1.82]	12.73	**11.64**	-1.09	[-3.15, 0.80]
Real RIR 1	**4.44**	7.80	+3.36	[2.10, 4.77]	**3.94**	5.03	+1.08	[-0.06, 2.13]
Real RIR 2	**9.89**	12.92	+3.03	[1.64, 4.49]	**8.77**	10.60	+1.83	[0.17, 3.64]
Real RIR 3	**10.57**	13.52	+2.95	[1.18, 4.71]	**10.02**	12.02	+2.00	[0.49, 3.44]
Real RIR 4	118.13	**65.39**	-52.73	[-60.91, -44.76]	194.44	**77.83**	-116.61	[-128.89, -104.93]

D REPRESENTATION MOVEMENT RELATIVE TO BASELINE STAGE C

Source reference and paired representations. For each of the eight bridges, the fixed source bank contains the same 276,157 LibriSpeech training utterances encoded by its separately trained baseline Stage-C checkpoint. The corresponding IAACM Stage-C checkpoint supplies paired representations before and after feature correction for the same evaluation utterances. We analyze all 72,361 utterances across the six evaluation datasets and the selected SRB noise and reverberation conditions described in Appendix C. Each SRB condition contains 2,620 utterances.

We map valid acoustic tokens through the checkpoint's learned projection into the 4,096-dimensional LLM input space, exclude audio delimiters and padding, and mean-pool the tokens within each utterance. For the six evaluation datasets, inputs use the common 30-second truncation limit. For SRB, recordings longer than 30 seconds follow the chunking protocol in Appendix C; chunk representations are pooled with weights proportional to their valid acoustic token counts. A fixed Gaussian projection with seed 2027 reduces the pooled vectors to 512 dimensions. Within each bridge, the same projection and baseline-source population mean and standard deviation are used for both evaluation suites and both correction states.

Distance summaries. Let S be the standardized source bank, H an evaluation set, and μ_S and μ_H their respective centroids. For the six evaluation datasets, we report corpus-centroid distance $D_{\text{centroid}} = \|\mu_H - \mu_S\|_2$. For both evaluation suites, we report the mean individual nearest-neighbor distance

$$D_k(H, S) = \frac{1}{|H|} \sum_{h \in H} \frac{1}{k} \sum_{s \in \mathcal{N}_k(h;S)} \|h - s\|_2, \qquad k \in \{5, 10\}, \tag{16}$$

where $\mathcal{N}_k(h; S)$ contains the k nearest source vectors. All distances are unsquared Euclidean distances in the full standardized 512-dimensional space and use every evaluation utterance. We average dataset-level distances uniformly across the six datasets and condition-level distances uniformly across the selected SRB conditions. Relative distance reduction is $100(1 - \bar{D}_{\text{after}}/\bar{D}_{\text{before}})$; positive values indicate movement toward the source. Distance scales are standardized separately for each bridge, and the reported changes compare paired representations within that bridge.

Table 14: Relative distance reductions (%) averaged over the six evaluation datasets. The reference is the separately trained baseline Stage-C source bank for each bridge. The final column counts datasets with lower 10-NN distance after correction. Very small changes are shown with extra precision.

Bridge	Centroid ↑	5-NN ↑	10-NN ↑	Decreased ↑
Q-Former	+1.02	+1.60	+1.59	6/6
Window Q-Former	+0.0004	+0.0002	+0.0002	5/6
Linear Projector	+0.46	+1.14	+1.14	6/6
Pooled MLP Projector	-3.86	-3.27	-3.26	0/6
Non-Causal Conv	+0.87	+3.23	+3.21	6/6
Causal Conv	+3.44	+3.70	+3.69	6/6
CIF	+1.82	+2.69	+2.69	6/6
CTC	+1.25	+2.79	+2.79	6/6

Q-Former, Linear Projector, Non-Causal Conv, Causal Conv, CIF, and CTC reduce both centroid and nearest-neighbor distances on all six datasets. Causal Conv and Non-Causal Conv have mean 10-NN reductions of 3.69% and 3.21%, respectively. Window Q-Former's mean 10-NN reduction is approximately 0.0002%, while Pooled MLP Projector increases both distance summaries on every dataset. Tables 17 and 18 provide all dataset-level values.

Movement under controlled corruptions. Table 15 reports the primary 28-condition SRB analysis, excluding clean. The same six bridges reduce 10-NN distance under all 28 conditions, with mean reductions from 0.94% for Linear Projector to 3.68% for Causal Conv. Window Q-Former again changes negligibly, and Pooled MLP Projector moves farther from the source under every corruption condition. Table 16 excludes only Real RIR 4, the real-room reverberation condition at severity level 4. All six source-directed bridges retain lower distances in every remaining condition, so this pattern extends across the corruption set. The 5-NN summaries show the same bridge-level directions in both evaluation suites.

Table 15: Representation movement over the 28 selected SRB noise and reverberation conditions, excluding clean. Before and After are mean 10-NN distances to the baseline Stage-C source bank. Reductions are percentages; the final column counts conditions with lower 10-NN distance.

Bridge	Before	After	5-NN red. ↑	10-NN red. ↑	Decreased ↑
Q-Former	43.5294	42.9047	+1.44	+1.44	28/28
Window Q-Former	52.1753	52.1753	+0.0001	+0.0001	27/28
Linear Projector	81.7048	80.9376	+0.95	+0.94	28/28
Pooled MLP Projector	47.3964	49.0147	-3.43	-3.41	0/28
Non-Causal Conv	38.5982	37.5031	+2.85	+2.84	28/28
Causal Conv	249.3687	240.1855	+3.69	+3.68	28/28
CIF	47.4737	46.3254	+2.42	+2.42	28/28
CTC	37.9498	37.0171	+2.46	+2.46	28/28

Table 16: SRB representation-movement sensitivity analysis over 27 conditions, excluding clean and Real RIR 4. Before and After are mean 10-NN distances to the baseline Stage-C source bank. Reductions are percentages; the final column counts conditions with lower 10-NN distance.

Bridge	Before	After	5-NN red. ↑	10-NN red. ↑	Decreased ↑
Q-Former	43.5586	42.9315	+1.45	+1.44	27/27
Window Q-Former	51.9455	51.9455	+0.0001	+0.0001	27/27
Linear Projector	81.5828	80.8228	+0.94	+0.93	27/27
Pooled MLP Projector	47.4886	49.1320	-3.47	-3.46	0/27
Non-Causal Conv	38.5793	37.4858	+2.85	+2.83	27/27
Causal Conv	250.7895	241.5530	+3.69	+3.68	27/27
CIF	47.4814	46.3286	+2.43	+2.43	27/27
CTC	37.9419	37.0121	+2.45	+2.45	27/27

Table 17: Dataset-level distances before and after correction for Q-Former, Window Q-Former, Linear Projector, and Pooled MLP Projector. Each row uses all evaluation utterances and the bridge's fixed baseline Stage-C source bank in 512 dimensions.

Dataset	Centroid: before	After	10-NN: before	After
Q-Former				
LS clean	39.1902	38.8642	43.7454	43.1016
LS other	40.2810	39.9092	44.3769	43.6908
TED-LIUM	39.4828	39.1095	43.6902	43.0378
Common Voice 23	43.4331	42.9839	45.8297	45.0932
GigaSpeech	39.8115	39.4750	44.2976	43.6271
VCTK	45.2956	44.6190	48.2293	47.3200
Window Q-Former				
LS clean	48.3491	48.3491	51.3132	51.3132
LS other	49.1745	49.1744	52.4807	52.4806
TED-LIUM	49.6321	49.6321	53.0784	53.0783
Common Voice 23	56.3933	56.3933	57.0306	57.0306
GigaSpeech	49.5199	49.5197	55.4506	55.4504
VCTK	59.7865	59.7859	60.2877	60.2874
Linear Projector				
LS clean	84.3740	84.1286	81.8598	81.0179
LS other	85.1374	84.8860	81.6849	80.8574
TED-LIUM	87.7082	87.3471	83.3982	82.4992
Common Voice 23	90.1566	89.6715	83.4440	82.5434
GigaSpeech	87.2434	86.9549	84.8508	83.8342
VCTK	92.0216	91.2435	86.3357	85.1186
Pooled MLP Projector				
LS clean	40.3067	42.0707	45.8541	47.4517
LS other	44.1125	46.1320	48.9691	50.7204
TED-LIUM	40.0489	41.5508	45.8011	47.1945
Common Voice 23	53.5923	55.3265	53.3983	55.1429
GigaSpeech	47.0591	49.1233	54.1767	55.8800
VCTK	64.2191	66.2982	61.8062	63.7197

Table 18: Dataset-level distances for Non-Causal Conv, Causal Conv, CIF, and CTC. Each row uses all evaluation utterances and the bridge's fixed baseline Stage-C source bank in 512 dimensions.

Dataset	Centroid: before	After	10-NN: before	After
Non-Causal Conv				
LS clean	60.8878	60.5244	41.7595	40.5332
LS other	61.9213	61.4074	44.6734	43.2584
TED-LIUM	59.2634	59.0233	38.1315	37.0275
Common Voice 23	62.6669	61.9760	46.9238	45.2911
GigaSpeech	58.5519	58.4511	35.8026	34.8344
VCTK	69.6461	68.3122	57.3125	55.1562
Causal Conv				
LS clean	252.9558	244.3079	247.5909	238.4653
LS other	255.7656	247.0124	250.1008	240.8896
TED-LIUM	258.6704	249.7157	253.1675	243.7912
Common Voice 23	254.6821	246.0232	248.0197	238.8779
GigaSpeech	267.1483	257.8481	261.7987	252.1237
VCTK	272.7989	263.3296	266.8968	257.0742
CIF				
LS clean	40.9229	40.2319	47.9239	46.7271
LS other	41.2452	40.5987	49.1582	47.9147
TED-LIUM	41.0085	40.1815	46.9648	45.7605
Common Voice 23	39.5490	39.0579	49.0370	47.7703
GigaSpeech	42.4862	41.6136	51.5518	50.1055
VCTK	43.4835	42.4866	53.1432	51.4885
CTC				
LS clean	27.6990	27.5706	38.7079	37.7486
LS other	28.1192	27.9125	38.7073	37.7050
TED-LIUM	28.9010	28.5865	37.9666	37.0131
Common Voice 23	28.1884	27.9034	39.5993	38.4750
GigaSpeech	28.7605	28.4332	40.1345	39.0309
VCTK	34.9866	34.0465	47.5013	45.8679