# PRESERVING WHAT WAS SAID: EVIDENCE-GUIDED AUDIO-VISUAL SPEECH ENHANCEMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generative audio-visual speech enhancement can improve acoustic quality while increasing speech-recognition errors. Under severe noise, we observe a *cleaner-but-wrong* regime in which a frozen audio-visual diffusion enhancer improves acoustic quality yet increases word error rate (WER) from $0.335$ for the noisy input to $0.504$ after enhancement. To mitigate this failure without retraining the enhancer, we propose AVEC (audio-visual evidence consistency), an inference-time framework that evaluates multiple enhancement candidates against complementary evidence from the original observation. AVEC compares each candidate transcript with visible articulation and with the noisy-input transcript, and converts these evidence scores into input-dependent mixture coefficients for waveform aggregation. On the primary evaluation, AVEC reduces WER from $0.504$ to $0.313$ while keeping acoustic quality close to that of the original enhancer. The higher-budget AVEC$^+$ further reduces WER to $0.279$, below the noisy input, while improving reference-based quality metrics. A human listener study further demonstrates the effectiveness of AVEC and AVEC$^+$.

## 1 INTRODUCTION

Generative audio-visual speech enhancement can substantially improve speech quality under severe noise by leveraging both acoustic and visual information. However, improved acoustic quality does not necessarily imply faithful preservation of the spoken content. Figure 1 illustrates this failure mode. The noisy input is perceptually degraded, but its automatic speech recognition (ASR) transcript remains closer to the original utterance than that of the enhanced output. In contrast, the frozen generative enhancer produces cleaner speech while introducing substantial content errors. We refer to this quality and content mismatch as the *cleaner-but-wrong* phenomenon. This observation highlights a limitation of evaluating generative speech enhancement mainly through acoustic quality. Improving acoustic quality remains a central objective, but it should not alter the content of the original utterance.

To address this problem, we propose **Audio-Visual Evidence Consistency (AVEC)**, an inference-time framework for frozen generative audio-visual speech enhancement (AVSE) models. AVEC combines two key ideas. First, it constructs multiple complementary enhancement candidates using different guidance sources, including visual speech and noisy audio. Second, it evaluates these candidates using external evidence retained in the original observation. Specifically, each candidate is compared with visible articulation and with the transcript of the noisy audio. These evidence scores then set the aggregation weights for each clip. We further introduce **AVEC$^+$**, which uses a larger candidate pool with additional sample-averaged candidates from each generation condition.

Different sources of evidence can be reliable for different inputs. When acoustic evidence is severely corrupted, visual articulation can provide useful constraints. When visual evidence is ambiguous, the noisy audio can still preserve informative linguistic cues. AVEC adapts to this variation by assigning candidate weights independently for each clip, without retraining the enhancer or requiring clean speech or reference transcripts at inference time.

On our primary evaluation with the frozen AV-UDiffSE+ (Ayilo et al., 2025) enhancer, AVEC reduces word error rate (WER) from 0.504 to 0.313, and AVEC$^+$ further reduces it to 0.279, below the noisy input WER of 0.335. AVEC also keeps acoustic quality close to that of the original enhancer, while AVEC$^+$ improves perceptual evaluation of speech quality (PESQ), extended short-time objective intelligibility (ESTOI), and scale-invariant signal-to-distortion ratio (SI-SDR). The



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

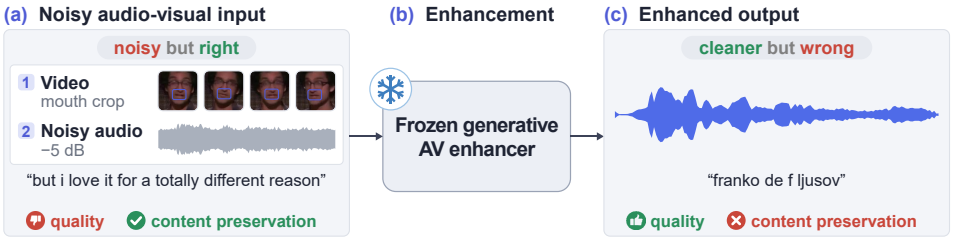



Figure 1: **Cleaner but wrong. (a)** A noisy speech input at $-5$ dB retains enough linguistic information for ASR to recover the original utterance. **(b)** A frozen generative AV enhancer improves the acoustic quality of the input. **(c)** Despite the cleaner output, the recognized content deviates substantially from the original utterance.

content preservation gains are further supported by an independent recognizer and a human listener study.

Our contributions are summarized as follows:

- We characterize a *cleaner-but-wrong* phenomenon in the evaluated generative AVSE settings, where improved acoustic quality can accompany increased recognition errors relative to the noisy input.
- We propose **AVEC**, which combines diverse enhancement candidates with adaptive aggregation based on external evidence at inference time.
- Across the evaluated settings, our approach reduces recognition errors while largely retaining enhancement quality, with condition-dependent trade-offs.

## 2 Related Work

**Audio-visual Speech Enhancement (AVSE).** Visible articulation provides complementary evidence about spoken content that is not directly affected by acoustic noise. Early AVSE approaches (Hou et al., 2018; Gabbay et al., 2018; Afouras et al., 2018a; Ephrat et al., 2018; Michelsanti et al., 2021) incorporate visual information as an input stream, particularly useful when acoustic evidence is weak. Subsequent work has explored self-supervised audio-visual representations (Shi et al., 2022), cross-modal consistency (Gao & Grauman, 2021), and shared evaluation benchmarks (Blanco et al., 2023). Building upon advances in generative speech enhancement (Richter et al., 2023), diffusion and flow-based AVSE models (Ayilo et al., 2025; Jung et al., 2024; Mboungou et al., 2026) incorporate visual information to support speech generation under noise. These approaches primarily use visual information as a condition to estimate or generate speech. We instead use it as evidence for judging which linguistic hypotheses remain compatible with the observation.

**Linguistic Hallucination.** Generative enhancement models often produce plausible speech that differs from the original utterance, particularly when the acoustic evidence is weak. Recent studies (Rong et al., 2026b; Guimarães et al., 2026; Shetu et al., 2026) have identified and investigated this linguistic hallucination phenomenon. Rong et al. (2026a;c); Wang et al. (2026) have explored training objectives, conditioning strategies, and noise-invariant representations to improve content fidelity. Such training-based mitigation is complementary to ours, since AVEC leaves the enhancer frozen and checks its candidates at inference against the original audio-visual observation.

**Guided Sampling and Multiple Restorations.** LipVoicer (Yemini et al., 2024) shows that a lip-reading transcript can serve as a target for an auxiliary recognizer that guides speech generation from silent videos, while SpeechOp (Lovelace et al., 2026) uses the noisy-input ASR transcript for generation-time guidance. Diffusion-based speech enhancement is stochastic, so multiple candidate waveforms can be generated from the same noisy observation (Richter et al., 2023). ArtiFree (Chhaglani et al., 2026) selects among repeated audio-only restorations using speech-embedding consistency. Prior works have explored transcript-guided generation and consistency-based selection, while post-generation validation of heterogeneous candidates using complementary audio-visual source evidence remains less studied.

**Robust Automatic Speech Recognition (ASR).** Prior work has addressed recognition errors after enhancement or transcript correction through signal selection or combination and confidence-based or conservative control (Iwamoto et al., 2022; Sato et al., 2021; 2022; Naderi et al., 2024; Yamauchi et al., 2026; Udagawa et al., 2024). Transcript-level generative error correction (Chen et al., 2023; Hu et al., 2024; Radhakrishnan et al., 2023) instead refines recognition hypotheses with language models, with LipGER (Ghosh et al., 2024) additionally incorporating visual speech evidence. Self-consistency provides another perspective by using agreement among generated candidates for output selection (Wang et al., 2023). Unlike transcript-level correction, our method still produces an enhanced waveform. Unlike candidate-consensus strategies, AVEC evaluates candidates against evidence from the source observation rather than treating their agreement as evidence of correctness.

## 3 Diagnosing Quality-content Mismatch in Generative AVSE

We first examine how acoustic quality and recognition accuracy change after enhancement, and how repeated restorations differ in their transcribed content. We study the frozen AV-UDiffSE+ enhancer (Ayilo et al., 2025) on LRS3 (Afouras et al., 2018b) speech mixed with NTCD (Abdelaziz, 2017) noise. DNSMOS (Reddy et al., 2022) estimates perceptual quality, while Whisper WER against reference transcripts measures recognition error. Figure 2 summarizes the observed quality-content mismatch and the variability of stochastic restorations.

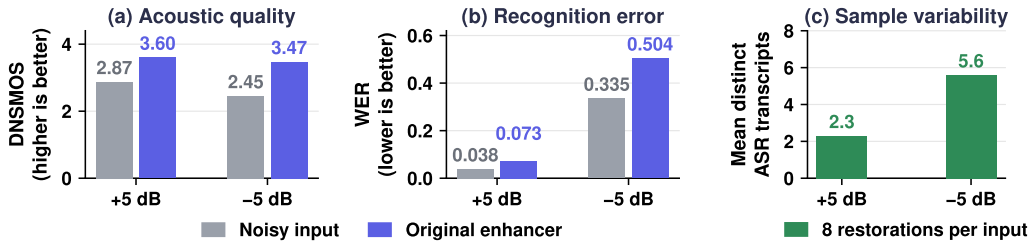



Figure 2: **Acoustic quality, recognition error, and variability across restorations.** **(a)** Enhancement increases DNSMOS at both $+5$ and $-5$ dB. **(b)** WER also increases relative to the noisy input at both evaluated noise levels, with a larger absolute increase at $-5$ dB. **(c)** Eight stochastic restorations yield more distinct ASR transcripts on average at $-5$ than at $+5$ dB.

**Improved quality does not ensure lower recognition error.** As shown in Figure 2(a,b), enhancement improves predicted perceptual quality while increasing recognition error at both noise levels. At $-5$ dB, DNSMOS rises from 2.45 to 3.47, whereas WER increases from 0.335 to 0.504. At $+5$ dB, DNSMOS likewise rises from 2.87 to 3.60, while WER increases from 0.038 to 0.073. The larger absolute WER increase at $-5$ dB makes this mismatch more pronounced under severe noise. These results illustrate the cleaner-but-wrong phenomenon in the evaluated setting. Acoustic quality alone therefore provides an incomplete account of enhancement performance.

**Recognized content varies more across restorations under severe noise.** Figure 2(c) examines the consistency of recognition results across eight stochastic restorations of each input. The mean number of distinct ASR transcripts increases from 2.3 at $+5$ dB to 5.6 at $-5$ dB. Although the underlying utterance and input observation remain fixed across runs, the restored outputs can yield different recognized word sequences. The larger count at $-5$ dB therefore indicates greater variability in ASR-decoded content across stochastic restorations under severe noise.

## 4 AVEC: Evidence-guided Audio-visual Speech Enhancement

AVEC (audio-visual evidence consistency) is an inference-time wrapper around a frozen speech enhancer $f$. Given noisy audio $x_i$ and synchronized mouth video $v_i$, it constructs multiple candidate waveforms, evaluates their linguistic consistency with visual and acoustic evidence retained in the original observation, and combines the candidates using input-adaptive weights. The enhancer itself remains fixed, and AVEC changes only candidate generation and aggregation. The following




Figure 3: **Overview of AVEC**. Two witnesses are transcribed from the inputs, the visual speech recognition (VSR) transcript $t_i^v$ of the mouth video and the ASR transcript $t_i^x$ of the noisy audio. **(1)** A frozen enhancer produces a default candidate and two candidates guided by targets derived from $t_i^v$ and $t_i^x$. **(2)** Each candidate's ASR transcript is scored against both witnesses (viseme distance to $t_i^v$, word error rate against $t_i^x$) and the two scores are combined. Lower is more consistent. **(3)** Scores become clip-specific weights, and the output is the weighted mix of the candidate waveforms.

subsections describe its three stages (Figure 3): candidate construction, source-evidence scoring, and input-adaptive candidate aggregation.

### 4.1 Candidate Construction

For the primary diffusion enhancer, let $\mathcal{C} = \{\mathrm{def}, \mathrm{lip}, \mathrm{aud}\}$ denote default, lip-guided, and audio-guided generation. For condition $c \in \mathcal{C}$, let $\tilde{y}_{i,k}^{(c,K)}$ denote the $k$-th trajectory, initialized with independent noise, used to construct a $K$-trajectory candidate:

$$a_i^{(c,K)} = \frac{1}{K} \sum_{k=1}^{K} \tilde{y}_{i,k}^{(c,K)}. \tag{1}$$

Each standard AVEC candidate is a $K = 4$ batch mean under its corresponding generation condition, and the default candidate $a_i^{(\mathrm{def},4)}$ coincides with the enhancer's deployed output. The candidate pools are

$$\mathcal{A}_i^{\mathrm{AVEC}} = \{a_i^{(c,4)} : c \in \mathcal{C}\}, \qquad \mathcal{A}_i^{\mathrm{AVEC}^+} = \mathcal{A}_i^{\mathrm{AVEC}} \cup \{a_i^{(c,8)} : c \in \mathcal{C}\}. \tag{2}$$

Thus, AVEC contains three candidates and AVEC$^+$ contains six. Let $M_i = |\mathcal{A}_i|$ denote the candidate count. The corresponding sampling budgets are 12 and 36 trajectories per non-fallback input, respectively; further sampling details are given in Section 5.1 and Appendix B.

**Guided candidates.** The lip-guided condition uses phonemes derived from the VSR transcript of $v_i$ as the target of an auxiliary CTC loss (Graves et al., 2006). The corresponding recognition-derived guidance steers sampling toward that target. The audio-guided condition uses the same mechanism with phonemes derived from the noisy-input ASR transcript. The two conditions thus differ only in the target sequence. Implementation details and the fixed guidance strength are given in Appendix C.1. These paths offer alternative restorations when visual or acoustic evidence retains useful content. Neither target is assumed correct on every clip.

**Fallback.** If the normalized noisy-input transcript is empty, the acoustic witness and audio-guidance target are undefined. In this case, we return the deployed default output without candidate aggregation. The frequency and full-set handling of this case are reported in Section 5.1 and Appendix B.

For enhancement backbones that do not support iterative or differentiable guided sampling, we construct candidate pools using operations supported by the backbone while retaining the same evidence-scoring and aggregation rules. The backbone-specific candidate pools are described in the corresponding experiments.

### 4.2 SOURCE-EVIDENCE SCORING

Let $t_i^v$ be the VSR transcript of $v_i$, $t_i^x$ the ASR transcript of $x_i$, and $t_i^a$ the ASR transcript of candidate $a$. The visual witness is therefore transcript-level evidence derived from visible articulation rather than a direct frame-level waveform-video correspondence.

Let $V(t)$ convert text into phonemes, map them to visemes, and collapse consecutive repeated visemes, and let $D(u, w)$ be unit-cost Levenshtein distance between sequences. For transcripts $r, h$, define

$$d_{\mathrm{vis}}(r, h) = \frac{D(V(r), V(h))}{\max\{1, |V(r)|, |V(h)|\}}, \qquad s_i^{\mathrm{lip}}(a) = d_{\mathrm{vis}}(t_i^v, t_i^a). \tag{3}$$

This symmetric visual inconsistency score lies in $[0, 1]$; normalization and edge-case conventions are detailed in Appendix C.2.

For non-empty word reference $r$ and hypothesis transcript $h$, define the clipped word-error score

$$\mathrm{WER}(r, h) = \frac{D(r, h)}{|r|}, \qquad \mathrm{WER}_2(r, h) = \min\{2, \mathrm{WER}(r, h)\}. \tag{4}$$

The acoustic inconsistency is

$$s_i^{\mathrm{aud}}(a) = \mathrm{WER}_2(t_i^x, t_i^a). \tag{5}$$

The noisy-input transcript is an imperfect reference: agreement can preserve recoverable content but can also preserve its recognition errors. Unlike the visual score, this reference-normalized word distance is generally asymmetric.

The combined inconsistency is

$$s_i(a) = s_i^{\mathrm{lip}}(a) + \lambda_{\mathrm{aud}} s_i^{\mathrm{aud}}(a), \qquad \lambda_{\mathrm{aud}} \geq 0. \tag{6}$$

The global coefficient calibrates the relative influence of the two scores $s_i^{\mathrm{lip}}(a), s_i^{\mathrm{aud}}(a)$, with ranges $[0, 1]$ and $[0, 2]$ respectively. Lower scores indicate closer agreement with the witnesses, not calibrated probabilities of correct content. The witnesses need not be independent or accurate. They can share errors, and guided candidates can agree with an erroneous target precisely because generation was steered toward it. Viseme agreement additionally cannot distinguish all phonetic or lexical alternatives. The benefit of combining the witnesses is therefore evaluated empirically rather than assumed.

### 4.3 INPUT-ADAPTIVE CANDIDATE AGGREGATION

For each input, scores determine Gibbs weights and a deterministic waveform mixture:

$$\omega_i(a) = \frac{\exp(-s_i(a)/\tau)}{\sum_{b \in \mathcal{A}_i} \exp(-s_i(b)/\tau)}, \qquad y_i = \sum_{a \in \mathcal{A}_i} \omega_i(a) a, \quad \tau > 0. \tag{7}$$

The global parameters $\lambda_{\mathrm{aud}}$ and $\tau$ are selected on the separate train-validation corpus and stay fixed for evaluation (Section 5.1), so input adaptivity comes from each clip's scores, not from per-clip tuning. Up to peak rescaling, the returned waveform is a convex combination of the candidates on a common time grid (Appendix C.2). The noisy waveform serves only as evidence and is never pooled or mixed in, and no clean speech or reference transcript is needed at inference.

**What the weights guarantee.** Let $z_i$ be the reference transcript, unavailable at inference, and $L_i(a) = \mathrm{WER}_2(z_i, t_i^a)$ the loss of candidate $a$. Write the signed score error $\varepsilon_i(a) = s_i(a) - L_i(a)$, its range over the pool $R_i^{\varepsilon} = \max_a \varepsilon_i(a) - \min_a \varepsilon_i(a)$, a best candidate $a_i^{\star} \in \arg\min_a L_i(a)$, a selected candidate $\hat{a}_i \in \arg\min_a s_i(a)$, and the entropy $H(\omega_i)$ of the weights, in nats.

**Proposition 1** (Selection regret is bounded by candidate-relative score error)**.** *For any input with a finite pool $\mathcal{A}_i$ of size $M_i$, any finite scores, and the weights $\omega_i$ of Equation* (7) *at $\tau > 0$,*

$$L_i(\hat{a}_i) - L_i(a_i^\star) \le \varepsilon_i(a_i^\star) - \varepsilon_i(\hat{a}_i) \le R_i^\varepsilon, \tag{8}$$

$$\mathbb{E}_{A\sim\omega_i}\big[L_i(A)\big] - L_i(a_i^\star) \le R_i^\varepsilon + \tau H(\omega_i) \le R_i^\varepsilon + \tau \log M_i. \tag{9}$$

*Both bounds are unchanged by a score error common to all candidates. Since the enhancer's own output $a_i^{(\mathrm{def},4)}$ is in the pool, in particular $L_i(\hat{a}_i) \le L_i(a_i^{(\mathrm{def},4)}) + R_i^\varepsilon$.*

Nothing is assumed about the witnesses. The bound charges only score error that differs across candidates, such as a guided candidate agreeing with the erroneous target it was steered toward (Section 4.2), the analog of reward hacking in best-of-$N$ selection (Jinnai et al., 2025; Huang et al., 2025). Because it holds for any score, the support for these two witnesses comes from the mismatched-witness and consensus controls (Tables 2 and IX) and the score error by candidate type in Appendix A.6. On the test set, the lowest-WER candidate of AVEC$^+$ averages $0.219$, hard selection $0.267$, the frozen mixture $0.279$, and the uniform average $0.351$, so hard selection is $0.048$ above the best candidate (Appendix A.6). The entropy term $\tau \log 6 = 0.179$ bounds the weighted candidate loss, not the mixed waveform's transcript, so the further $0.012$ to the mixture lies outside the bound. Proofs are in Appendix A.

**What the temperature trades.** The weighting tends to selection among the minimum-score candidates as $\tau \to 0^+$ and to uniform averaging as $\tau \to \infty$ (Appendix A.3). The expected score only rises with $\tau$, at rate $\mathrm{Var}_{\omega_\tau}(s_i)/\tau^2$, while the mixture's squared distortion to any reference sits below the weighted candidate distortion by exactly the pool's spread around it, the ambiguity decomposition of Krogh & Vedelsby (1995) (Proposition 3). Selection favors candidates with low evidence scores, while averaging reduces squared distortion relative to the weighted mean candidate distortion. The frozen $\tau$ is one point on this curve.

**Stability.** With the pool fixed, the returned waveform moves by at most $D_i \min\{1, R_i/(4\tau)\}$ when the relative scores move by a range $R_i$, where $D_i$ is the largest distance between two candidates (Proposition 4, a temperature-scaled Lipschitz property of the softmax (Gao & Pavel, 2017)), while hard selection has no such bound near a tie. These guarantees are relative to the best candidate and the enhancer's own output, never to the input, which is out of the pool by design. Decoding is nonconvex in the waveform, so the mixture's own WER carries no bound.

# 5 Experiments

## 5.1 Experimental setup and evaluation protocol

**Systems and data.** Our primary evaluation uses the frozen AV-UDiffSE+ posterior sampler (Ayilo et al., 2025) on LRS3-NTCD at $-5$ dB ($n = 657$) (Afouras et al., 2018b; Abdelaziz, 2017). To assess transfer beyond this primary condition, we additionally evaluate other noise conditions, speech corpora, SNRs, and enhancement backbones (Section 5.4). All primary-set results use the official $657$ clips. For the one clip whose noisy-input ASR transcript is empty, the pre-specified fallback returns the deployed enhancer output, which is scored in place of the wrapper output (Appendix B). The noisy input and the enhancer's deployed output (*always-enhance*) serve as reference anchors for pre-enhancement content and enhancement quality, respectively.

**Evaluation metrics.** We use Whisper WER against the reference transcript as the primary ASR-based content metric (Radford et al., 2023), and evaluate acoustic quality using DNSMOS (Reddy et al., 2022), PESQ, ESTOI, and SI-SDR. Because Whisper also supplies the acoustic evidence and audio-guidance target, we additionally re-score the final policies with independent speech recognizers (Appendix D.5). To complement mean WER with a per-clip measure of degradation, let $z_i$ denote the reference transcript and $\ell(y, z_i)$ the clipped WER of waveform $y$ against $z_i$. For noisy input $x_i$, returned waveform $y_i$, and tolerance $\delta \ge 0$, we define the harm indicator as

$$\mathsf{h}_i(y_i; \delta) = \mathbf{1}[\ell(y_i, z_i) - \ell(x_i, z_i) > \delta]\,. \tag{10}$$

Unless otherwise stated, we report the fraction of clips with $\mathsf{h}_i = 1$ at $\delta = 0$, which we term *input-relative content harm*. This metric measures how often processing increases recognition error relative to the same input. It does not by itself establish audible fabrication, and returning the noisy input gives zero additional harm by definition. Severity analyses are reported in Appendix B.

**Implementation details.** Each of the three AVEC candidates is the mean of $K{=}4$ independently initialized trajectories under its generation condition, requiring 12 trajectories per non-fallback input. AVEC$^+$ adds one $K{=}8$ sample-mean candidate per condition, for 36 trajectories in total. The lip- and audio-guided candidates use phoneme-CTC guidance with $\lambda_{\text{guide}} = 0.05$. Candidate and noisy-input transcripts come from Whisper `medium.en`, and visual transcripts from a pretrained VSR. The scoring functions and waveform aggregation follow Sections 4.2 to 4.3. Full sampling, recognition, transcript-processing, and guidance details are provided in Appendices C.1 and C.2.

**Parameter selection and evaluation protocol.** The aggregation parameters $\tau$ and $\lambda_{\text{aud}}$ are selected on a separate train-validation corpus reconstructed from LRS3 `trainval`. The pre-specified rule selects the configuration with the lowest train-validation WER among those whose PESQ, ESTOI, and SI-SDR point estimates exceed *always-enhance* and whose WER is below the noisy input. The guidance strength $\lambda_{\text{guide}} = 0.05$ is also validated on train-validation data. All parameters are then frozen for the reported evaluations. The evaluation sets were used in the exploratory diagnostic analyses of Section 3 and are therefore not untouched confirmatory holdouts, but no aggregation parameter is selected on these sets. Detailed train-validation construction, parameter grids, evaluation history, and uncertainty analyses are provided in Appendices B and D.1.

### 5.2 Main results

We first evaluate two diffusion backbones that support the same guided candidate construction: AV-UDiffSE+ and AV-CA-DiffUSE (Mboungou et al., 2026). Backbones requiring alternative candidate construction are considered in Section 5.3. Table 1 shows the cleaner-but-wrong trade-off on both systems. The deployed enhancers improve acoustic quality but substantially increase recognition error relative to the noisy input. On AV-UDiffSE+, WER rises from $0.335$ to $0.504$ while DNSMOS increases from $2.45$ to $3.47$.

Across both backbones, AVEC reduces WER and input-relative harm while largely retaining enhancement quality, and AVEC$^+$ further improves the content-quality trade-off. On AV-UDiffSE+, AVEC$^+$ reduces WER to $0.279$ and harm from $44.6\%$ to $18.1\%$, while improving PESQ, ESTOI, and SI-SDR over *always-enhance*. Its DNSMOS remains slightly lower. Its WER also falls below the noisy input, with the same ordering under Parakeet-TDT (NVIDIA NeMo Team, 2025) ($0.307$ vs. $0.339$; Appendix D.5). On AV-CA-DiffUSE, AVEC$^+$ lowers WER from $0.580$ to $0.381$ and harm from $51.6\%$ to $27.4\%$ with small reference-quality gains, although WER remains above the noisy input. Thus, recovery is backbone-dependent, but enhancer-induced content degradation decreases on both diffusion systems.

Table 1: Main results on LRS3-NTCD at $-5$ dB. Harm is the percentage of clips with higher WER than the noisy input; $\text{WER}_{\text{Pk}}$ denotes Parakeet-TDT WER. Bold marks the best value per enhancer.

| Enhancer | Policy | WER ↓ | Harm (%) ↓ | DNSMOS ↑ | PESQ ↑ | ESTOI ↑ | SI-SDR (dB) ↑ | $\text{WER}_{\text{Pk}}$ ↓ |
|---|---|---|---|---|---|---|---|---|
| *Noisy input (reference)* | | 0.335 | 0.0 | 2.45 | 1.108 | 0.456 | −4.99 | 0.339 |
| AV-UDiffSE+ | Always enhance | 0.504 | 44.6 | **3.47** | 1.470 | 0.580 | +3.24 | 0.481 |
| | AVEC (ours) | 0.313 | 22.4 | 3.41 | 1.469 | 0.579 | +3.33 | 0.334 |
| | **AVEC$^+$ (ours)** | **0.279** | **18.1** | 3.41 | **1.485** | **0.586** | **+3.46** | **0.307** |
| AV-CA-DiffUSE | Always enhance | 0.580 | 51.6 | **3.41** | 1.426 | 0.533 | +2.04 | 0.567 |
| | AVEC (ours) | 0.401 | 32.1 | 3.35 | 1.437 | 0.537 | +2.08 | 0.432 |
| | **AVEC$^+$ (ours)** | **0.381** | **27.4** | 3.36 | **1.452** | **0.543** | **+2.17** | **0.410** |

### 5.3 Contributions of candidates and evidence

**Candidate construction.** Table 2 compares lip- and audio-guided candidates generated from VSR- and noisy-input ASR-derived phoneme guidance, respectively. Each reduces recognition errors but sacrifices some acoustic quality. An unguided mean using the same 36-trajectory budget as AVEC$^+$ improves quality but leaves WER above the noisy input. Combining default and guided candidates with evidence-based weights achieves a better content-quality balance, suggesting that increased sampling alone does not explain the gains.

**Evidence and aggregation.** We fix the six AVEC$^+$ candidates and vary their aggregation. Candidate consensus scores each candidate by mean pairwise WER against the other candidates, relying on

Table 2: Candidate-construction and aggregation ablations on LRS3-NTCD at $-5$ dB. Upper: candidate construction; lower: aggregation on a fixed six-candidate pool. $\mathrm{WER}_{\mathrm{Pk}}$ denotes Parakeet-TDT WER. Bold marks the best value within each block.

| Component | Method | WER ↓ | Harm (%) ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR (dB) ↑ | $\mathrm{WER}_{\mathrm{Pk}}$ ↓ |
|---|---|---|---|---|---|---|---|
| Candidate construction | Always enhance | 0.504 | 44.6 | 1.470 | 0.580 | +3.24 | 0.481 |
| | Lip-guided only | 0.366 | 33.0 | 1.416 | 0.564 | +3.10 | 0.374 |
| | Audio-guided only | 0.401 | 31.5 | 1.423 | 0.564 | +3.09 | 0.402 |
| | Unguided mean (36 traj.) | 0.450 | 40.2 | **1.501** | **0.591** | **+3.58** | 0.447 |
| | AVEC (3 cand., 12 traj.) | 0.313 | 22.4 | 1.469 | 0.579 | +3.33 | 0.334 |
| | **AVEC$^+$ (6 cand., 36 traj.)** | **0.279** | **18.1** | 1.485 | 0.586 | +3.46 | **0.307** |
| Evidence and aggregation | Uniform average | 0.351 | 29.4 | **1.489** | **0.589** | **+3.54** | 0.353 |
| | Candidate consensus | 0.328 | 25.7 | 1.487 | 0.587 | +3.50 | 0.346 |
| | Acoustic evidence only | 0.325 | 21.8 | 1.487 | 0.586 | +3.49 | 0.333 |
| | Visual evidence only ($\lambda_{\mathrm{aud}}$=0) | 0.295 | 22.5 | 1.486 | 0.588 | +3.49 | 0.312 |
| | Fixed global weights | 0.337 | 28.0 | 1.485 | 0.588 | +3.53 | 0.348 |
| | **AVEC$^+$ weighted average** | **0.279** | **18.1** | 1.485 | 0.586 | +3.46 | **0.307** |

Table 3: Transfer to non-diffusion backbones on LRS3-NTCD at $-5$ dB with primary-system parameters fixed. ‡Two candidates: deployed output and the mean of eight outputs from the original input and seven perturbed copies. §SI-SDR is omitted for PASE because vocoder resynthesis limits waveform-based interpretation.

| Backbone | Policy | WER ↓ | Harm (%) ↓ | DNSMOS ↑ | PESQ ↑ | ESTOI ↑ | SI-SDR (dB) ↑ |
|---|---|---|---|---|---|---|---|
| FlowAVSE‡ | Always enhance | 0.256 | 15.4 | **3.11** | **1.49** | **0.65** | **+5.0** |
| | Perturbed mean | 0.276 | 19.3 | 3.05 | 1.41 | 0.64 | +4.2 |
| | **AVEC (2 cand.)** | **0.242** | **14.0** | 3.10 | 1.47 | **0.65** | +4.6 |
| PASE‡§ | Always enhance | 0.294 | 16.4 | **3.70** | **1.61** | **0.69** | – |
| | Perturbed mean | 0.302 | 19.0 | 3.59 | 1.53 | 0.65 | – |
| | **AVEC (2 cand.)** | **0.278** | **14.0** | 3.66 | 1.59 | 0.68 | – |

agreement within the generated pool rather than evidence from the input. Single-source scores use only visual or noisy-input evidence. These rules yield similar acoustic quality, but joint visual and acoustic evidence gives lower WER under both Whisper and Parakeet, supporting source evidence over candidate agreement alone. It also outperforms fixed global weights, which favors input-adaptive over corpus-level mixing. Additional controls appear in Appendices D.1 and D.4.

**Across non-diffusion enhancers.** We next consider FlowAVSE (Jung et al., 2024) and PASE (Rong et al., 2026b), neither of which supports the guided sampling above. For each backbone, we use two candidates: the deployed output and the mean of eight input-perturbed outputs (Table 3). AVEC lowers Whisper WER from $0.256$ to $0.242$ on FlowAVSE and from $0.294$ to $0.278$ on PASE, while reducing harm to $14.0\%$ on both, with modest quality losses. The perturbed mean alone is worse than the deployed output, again indicating that the gain comes from evidence-guided use of the pool rather than averaging alone. For PASE, the wrapper also adds video evidence to this audio-only backbone. Further details and ablations are provided in Appendix F.2.

### 5.4 Generalization across conditions and additional enhancers

We evaluate transfer across input conditions and enhancement backbones, keeping $\tau$ and $\lambda_{\mathrm{aud}}$ fixed at the values selected on the primary trainval corpus.

**Across conditions.** Figure 4 evaluates LRS3-NTCD, LRS3-DEMAND (Thiemann et al., 2013), TCD-DEMAND (Harte & Gillen, 2015; Thiemann et al., 2013), and TCD-NTCD (Harte & Gillen, 2015; Abdelaziz, 2017), where the latter remixes TCD speech with NTCD noise, at input SNRs of $-5$, $0$, and $+5$ dB. Both AVEC and AVEC$^+$ reduce WER relative to *always-enhance* in all twelve conditions while largely preserving PESQ. The WER advantage of AVEC$^+$ over AVEC is most pronounced at $-5$ dB and becomes smaller at $+5$ dB, indicating that the expanded candidate pool is particularly useful under severe noise. In contrast, PESQ gains over always-enhance are less consistent, with small decreases in some milder-noise settings. Overall, the results show transfer across speech corpora and noise conditions, mainly as lower recognition error with enhancement quality retained. Detailed results, the official TCD-NTCD mixtures, and independent-recognizer evaluations are provided in Tables XV and XVI.

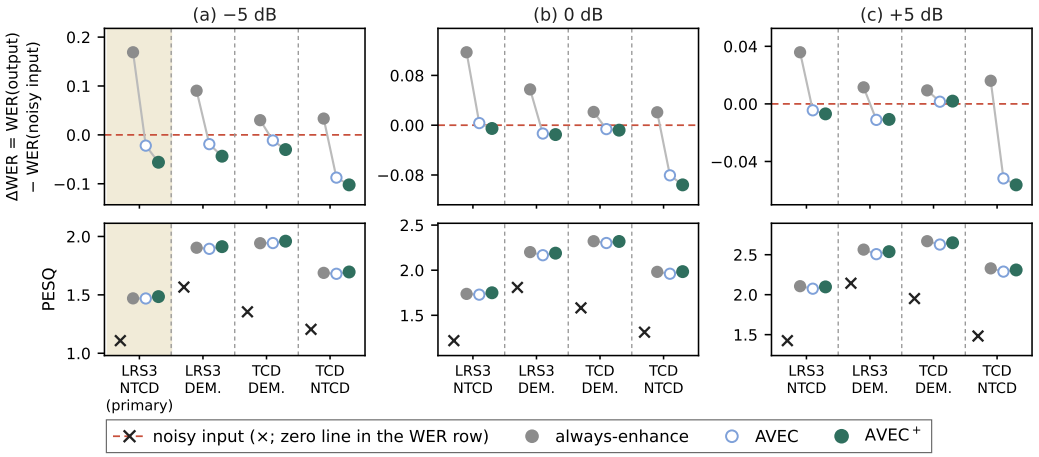



Figure 4: Evaluation across speech corpora, noise types, and SNRs. Columns (a)-(c) show $-5$, $0$, and $+5$ dB. The top row shows WER change relative to the noisy input, and the bottom row reports PESQ, with crosses marking the noisy input. Shading marks the primary condition. AVEC and AVEC$^+$ reduce WER relative to always-enhance while largely preserving PESQ across conditions.

Table 4: Listener evaluation (30 randomly sampled test clips, 53 listeners). Harm is the share of clips with higher WER than their noisy input. Whisper is scored on the same clips.

| | Listener | | Whisper | | |
|---|---|---|---|---|---|
| Condition | WER ↓ | Harm (%) ↓ | WER ↓ | Harm (%) ↓ | Quality ↑ |
| Clean source | 0.147 | 6.7 | 0.012 | 3.3 | 4.12 |
| Noisy input | 0.608 | – | 0.362 | – | 2.28 |
| Always enhance | 0.586 | 36.7 | 0.515 | 43.3 | 2.83 |
| Always lip-guided | 0.465 | 23.3 | 0.373 | 36.7 | 2.92 |
| Always audio-guided | 0.468 | 23.3 | 0.407 | 36.7 | 2.97 |
| AVEC (3 cand.) | 0.455 | 26.7 | 0.300 | 23.3 | 2.84 |
| **AVEC$^+$ (6 cand.)** | **0.433** | **13.3** | **0.261** | **16.7** | **3.03** |

### 5.5 Listener corroboration

As shown in Table 4, we evaluate AVEC, AVEC$^+$, always-enhance, and the individually guided conditions with 53 listeners on 30 randomly sampled test clips (details in Appendices E.1 and E.2). Relative to always-enhance, AVEC reduces listener WER from $0.586$ to $0.455$ while maintaining perceived quality ($2.84$ vs. $2.83$). AVEC$^+$ further reduces listener WER to $0.433$ and harm to $13.3\%$, while improving perceived quality to $3.03$, the highest among the enhancement policies. These results show that evidence-guided aggregation improves content preservation without sacrificing perceived quality. Listener and Whisper WER are also strongly correlated across the $180$ outputs (Pearson $r$=$0.76$), suggesting that the automatic metric broadly reflects human transcription performance.

## 6 Conclusion

Generative AVSE can improve acoustic quality while increasing recognition error relative to the noisy input, revealing a cleaner-but-wrong regime. AVEC combines frozen enhancer outputs using source-video and noisy-input evidence, while AVEC$^+$ expands the candidate pool. On the primary system, AVEC$^+$ substantially reduces content harm while retaining enhancement quality, with gains corroborated by an independent recognizer and human listeners. The method also transfers across noise conditions, speech corpora, and enhancer backbones without retuning, although the content-quality trade-off remains condition- and backbone-dependent. These findings support source-evidence-guided aggregation as a practical post-hoc approach to reducing content degradation in fixed generative enhancers.

## ETHICS STATEMENT

**Human participants.** The listening study described in Appendix E.1 was conducted in accordance with institutional review procedures (protocol no. 64610). It was hosted on Amazon Mechanical Turk, with recruitment capped at 200 submissions. Participation was voluntary, and listeners provided consent before audio playback and could stop at any time without penalty. Participants who completed the approximately 18-minute session received USD $5.00$. A headphone screening preceded the study items; listeners who did not pass the screening ended the session without compensation, as stated in the recruitment text and consent screen. Submissions were reviewed by the researchers, and those that did not meet the stated attention-check or good-faith criteria were not approved for payment. The study did not collect voice recordings from participants, and all responses were stored using random identifiers.

**Speech data and foreseeable harm.** All audio and video derive from LRS3, NTCD-TIMIT, DEMAND, and TCD-TIMIT under their research licenses. We do not infer speaker attributes or redistribute source media beyond license-compatible short demonstrations. Fabricated speech is especially consequential for assistive, clinical, captioning, and evidentiary uses. A reduction in measured content harm must not be presented as a per-utterance safety certificate. High-consequence deployment requires representative in-domain validation, monitoring, and appropriate human oversight.

## AI USE STATEMENT

AI assistants were used under author direction: Claude Code supported coding, and ChatGPT assisted with manuscript revision. Claude Code wrote analysis and evaluation scripts and ran experiments. Literature-search outputs were checked against primary sources.

## REPRODUCIBILITY STATEMENT

All experiments use frozen checkpoints, public corpora, fixed seeds, and clip-level manifests: the enhancers, Whisper, Parakeet, and the VSR are public checkpoints, and the one model trained by us, the auxiliary phoneme CTC recognizer that scores guidance targets, is trained once on clean LRS3 `trainval` audio (Appendix C.1); its checkpoint, configuration, and training script are released with the code. The release includes per-clip metrics, the averaging and selection code, deterministic split construction, figure scripts, the guidance recipe, and blinded human-study materials. The combination rule has no learned parameters. The official LRS3–NTCD $-5\,\mathrm{dB}$ set contains $657$ clips; all primary-set tables use the full $657$ clips, applying the documented fallback to the one clip with an empty input transcript. No clip is excluded based on its outcome. Code, per-clip metrics, human-study materials, and the audio demonstration are provided as supplementary material.

## REFERENCES

Ahmed Hussen Abdelaziz. NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition. In *Proc. Interspeech*, 2017.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proc. Interspeech*, pp. 3244–3248, 2018a.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018b.

Gholamali Aminian, Idan Shenfeld, Amir R. Asadi, Ahmad Beirami, and Youssef Mroueh. Best-of-N through the smoothing lens: KL divergence and regret analysis. In *ICML 2025 Workshop on Efficient Systems for Foundation Models*, 2025. arXiv:2507.05913.

Jean-Eudes Ayilo, Mostafa Sadeghi, Romain Serizel, and Xavier Alameda-Pineda. Diffusion-based unsupervised audio-visual speech enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Andrea Lorena Aldana Blanco, Cassia Valentini-Botinhao, Ondřej Klejch, Mandar Gogate, Kia Dashtipour, Amir Hussain, and Peter Bell. AVSE challenge: Audio-visual speech enhancement challenge. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 465–471, 2023.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng Siong Chng. HyPoradise: An open baseline for generative speech recognition with large language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

Bhawana Chhaglani, Yang Gao, Julius Richter, Xilin Li, Syavosh Zadissa, Tarun Pruthi, and Andrew Lovitt. ArtiFree: Detecting and reducing generative artifacts in diffusion-based speech enhancement. In *Proc. ICASSP*, 2026.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):112:1–112:11, 2018.

Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. In *Proc. Interspeech*, pp. 1170–1174, 2018.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Ruohan Gao and Kristen Grauman. VisualVoice: Audio-visual speech separation with cross-modal consistency. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15495–15505, 2021.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Purva Chiniya, Utkarsh Tyagi, Ramani Duraiswami, and Dinesh Manocha. LipGER: Visually-conditioned generative error correction for robust automatic speech recognition. In *Proc. Interspeech*, 2024.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Heitor R. Guimarães, Jiaqi Su, Rithesh Kumar, Tiago H. Falk, and Zeyu Jin. DiTSE: High-fidelity generative speech enhancement via latent diffusion transformers. In *Proc. ICASSP*, 2026. arXiv:2504.09381.

Naomi Harte and Eoin Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015.

Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

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

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-N the best of them? coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126, 2025.

Kazuma Iwamoto, Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Hiroshi Sato, Shoko Araki, and Shigeru Katagiri. How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR. In *Proc. Interspeech*, 2022.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-N sampling with minimum Bayes risk objective for language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 9321–9347, 2025. doi: 10.18653/v1/2025.naacl-long.472.

Chaeyoung Jung, Suyeon Lee, Ji-Hoon Kim, and Joon Son Chung. FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching. In *Proc. Interspeech*, 2024.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, pp. 231–238. MIT Press, 1995.

Justin Lovelace, Rithesh Kumar, Jiaqi Su, Ke Chen, Kilian Q. Weinberger, and Zeyu Jin. SpeechOp: Inference-time task composition for generative speech processing. In *International Conference on Learning Representations*, 2026.

Colombe Mboungou, Mostafa Sadeghi, Jean-Eudes Ayilo, and Romain Serizel. Audio-visual contrastive alignment for diffusion-based visual-conditioned speech enhancement. In *Proc. Interspeech*, 2026. arXiv:2606.23712.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021.

Maryam Naderi, Enno Hermann, Alexandre Nanchen, Sevada Hovsepyan, and Mathew Magimai-Doss. Towards interfacing large language models with ASR systems using confidence measures and prompting. In *Proc. Interspeech*, 2024.

Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Hervé Glotin, Dimitra Vergyri, June Sison, Azad Mashari, and Jie Zhou. Audio-visual speech recognition. Final workshop 2000 report, Center for Language and Speech Processing, Johns Hopkins University, 2000.

NVIDIA NeMo Team. Parakeet-TDT 0.6b v2: FastConformer transducer for English speech recognition. `https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2`, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2023.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegnér. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In *Proc. EMNLP*, 2023.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*, pp. 886–890, 2022.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

Xiaobin Rong, Jun Gao, Zheng Wang, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. StuPASE: Towards low-hallucination studio-quality generative speech enhancement. In *Proc. Interspeech*, 2026a. arXiv:2603.09234.

Xiaobin Rong, Qinwen Hu, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. PASE: Leveraging the phonological prior of WavLM for low-hallucination generative speech enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32826–32834, 2026b.

Xiaobin Rong, Zheng Wang, Yushi Wang, Jun Gao, and Jing Lu. UniPASE: A generative model for universal speech enhancement with high fidelity and low hallucinations. *IEEE Transactions on Audio, Speech and Language Processing*, 34:3901–3915, 2026c. arXiv:2604.14606.

Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Takafumi Moriya, and Naoyuki Kamo. Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition. In *Proc. Interspeech*, 2021.

Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Naoyuki Kamo, and Takafumi Moriya. Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition. In *Proc. ICASSP*, 2022.

Shrishti Saha Shetu, Emanuël A. P. Habets, and Andreas Brendel. A comparison of generative and discriminative methods for speech enhancement: Robustness, complexity, and hallucination. *arXiv preprint arXiv:2606.02913*, 2026.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. In *Proc. Meetings on Acoustics*, 2013.

Takuma Udagawa, Masayuki Suzuki, Masayasu Muraoka, and Gakuto Kurata. Robust ASR error correction with conservative data filtering. In *Proc. EMNLP (Industry Track)*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Zheng Wang, Xiaobin Rong, Hang Su, Tianyi Tan, Junnan Wu, Lichun Fan, Zhenbo Luo, Jian Luan, and Jing Lu. Reducing linguistic hallucination in LM-based speech enhancement via noise-invariant acoustic-semantic distillation. *arXiv preprint arXiv:2605.08608*, 2026.

Kevin J. P. Woods, David H. Siegel, James Traer, and Josh H. McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79(7): 2064–2072, 2017.

Kazuki Yamauchi, Masato Murata, and Shogo Seki. Confidence-based filtering for speech dataset curation with generative speech enhancement using discrete tokens. In *Proc. ICASSP*, Barcelona, Spain, 2026. arXiv:2601.12254.

Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. Lipvoicer: Generating speech from silent videos guided by lip reading. In *International Conference on Learning Representations*, 2024.

# APPENDIX CONTENTS

**A Mathematical interpretation of evidence weighting** **15**

A.1 Variational characterization . . . 15
A.2 Proof of Proposition 1 . . . 15
A.3 Temperature limits and ties . . . 16
A.4 What the temperature trades . . . 16
A.5 Waveform stability . . . 17
A.6 Measured quantities . . . 19

**B Experimental protocol and setup** **19**

B.1 Datasets, systems, and evaluation conventions . . . 19
B.2 Harm severity and noise-type breakdown . . . 20

**C Implementation details** **21**

C.1 Candidate generation . . . 21
C.2 Transcripts, evidence scores, and aggregation . . . 22

**D Complete results on the primary condition** **23**

D.1 Parameter selection: freezing rule, grid, and robustness . . . 23
D.2 Full ablation table and further controls . . . 23
D.3 Compute budget and candidate count . . . 25
D.4 Does the evidence carry the content? Substitutions, degradation, and a clean-input control . . . 26
D.5 The final policies under an independent recognizer . . . 27

**E Independent validation beyond the primary evaluator** **29**

E.1 Listener corroboration protocol . . . 29
E.2 Results . . . 29
E.3 Qualitative examples . . . 29

**F Generalization: full results** **31**

F.1 Condition transfer . . . 31
F.2 Non-diffusion backbones . . . 33

**G Diagnostics behind the analysis of Section 3** **35**

G.1 Counterfactual-video probe: construction and full tables . . . 35
G.2 Speech-corpus by noise-family analysis . . . 35
G.3 Evaluator sensitivity . . . 36

**H Limitations and scope** **36**

# A MATHEMATICAL INTERPRETATION OF EVIDENCE WEIGHTING

The results below concern a fixed input with a non-empty, finite indexed candidate pool and finite scores. They are standard finite-dimensional properties of Gibbs weighting, included to specify the rule's interpretation and limitations, rather than as separate theoretical contributions. All logarithms are natural. Write $M_i = |\mathcal{A}_i|$ for the candidate count; $K$ elsewhere denotes trajectories averaged into a candidate.

## A.1 VARIATIONAL CHARACTERIZATION

**Proposition 2** (Entropy-regularized evidence weighting)**.** *For $\tau > 0$, let*

$$\Delta_i = \left\{ q : \mathcal{A}_i \to [0,1] \;\middle|\; \sum_a q(a) = 1 \right\}. \tag{11}$$

*The coefficients in Equation* (7) *uniquely minimize $F_i(q) = \sum_a q(a)s_i(a) + \tau \sum_a q(a) \log q(a)$ over $\Delta_i$, with $0 \log 0 = 0$.*

*Proof.* Set $Z_i = \sum_a \exp(-s_i(a)/\tau)$ and $\omega_i(a) = \exp(-s_i(a)/\tau)/Z_i$. Every entry of $\omega_i$ is strictly positive. For every $q \in \Delta_i$, including boundary points,

$$\begin{aligned} \tau \,\mathrm{KL}(q \| \omega_i) &= \tau \sum_a q(a) \log q(a) + \sum_a q(a) s_i(a) + \tau \log Z_i, \\ F_i(q) &= \tau \,\mathrm{KL}(q \| \omega_i) - \tau \log Z_i. \end{aligned} \tag{12}$$

KL divergence is nonnegative and is zero if and only if $q = \omega_i$. Hence $\omega_i$ is the unique minimizer. Subtracting a common score offset multiplies numerator and denominator by the same positive factor and leaves these coefficients unchanged. □

This proof includes the boundary of the simplex explicitly; a Lagrange-multiplier proof would additionally need to justify an interior optimum before differentiating $q \log q$. This is the Gibbs variational principle (Jaynes, 1957), and the same form appears in soft best-of-$N$ selection (Aminian et al., 2025).

## A.2 PROOF OF PROPOSITION 1

The candidate-level loss $L_i(a)$ of Proposition 1 matches the clipped word error $\ell(a, z_i)$ of Equation (10) evaluated on candidate $a$.

*Proof.* Suppress the input index.

*Hard selection.* Because $\hat{a}$ minimizes the score,

$$\begin{aligned} L(\hat{a}) &= s(\hat{a}) - \varepsilon(\hat{a}) \\ &\le s(a^\star) - \varepsilon(\hat{a}) \\ &= L(a^\star) + \varepsilon(a^\star) - \varepsilon(\hat{a}) \\ &\le L(a^\star) + R^\varepsilon, \end{aligned}$$

where the last inequality follows because the difference between any two values of $\varepsilon$ is at most its range.

*Soft weighting.* Write

$$\mathbb{E}_\omega[L] = \mathbb{E}_\omega[s] - \mathbb{E}_\omega[\varepsilon].$$

By Proposition 2, $F(\omega) \le F(\delta_{a^\star})$ for the point mass $\delta_{a^\star}$, and therefore

$$\begin{aligned} \mathbb{E}_\omega[s] &\le s(a^\star) + \tau H(\omega) \\ &= L(a^\star) + \varepsilon(a^\star) + \tau H(\omega). \end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}_\omega[L] - L(a^\star) &\le \varepsilon(a^\star) - \mathbb{E}_\omega[\varepsilon] + \tau H(\omega)\\ &\le R^\varepsilon + \tau H(\omega)\\ &\le R^\varepsilon + \tau \log M.\end{aligned}$$

Adding a constant to $\varepsilon$ on every candidate changes neither $R^\varepsilon$ nor any difference of its values, so both bounds are invariant to a common error. The last claim follows from $L(a^\star) \le L(a^{(\mathrm{def},4)})$. □

The smoothing cost is small when the score margin is large. With $g(a) = s(a) - s(\hat{a})$,

$$\mathbb{E}_\omega[s] - s(\hat{a}) = \frac{\sum_{a\neq\hat{a}} g(a)e^{-g(a)/\tau}}{1 + \sum_{a\neq\hat{a}} e^{-g(a)/\tau}} \le \frac{(M-1)\tau}{e},$$

where the bound uses $ue^{-u/\tau} \le \tau/e$. For example, a candidate whose score exceeds the best by $0.3$ receives relative weight $e^{-3} \approx 0.05$ at $\tau$=0.10.

**An imperfect acoustic reference.** The noisy-input transcript contains recognition errors. Such an error cancels in Proposition 1 only if it raises every candidate's acoustic score by the same amount, and nothing guarantees this. When the candidate transcripts differ, the same reference error adds different amounts to different candidates and enters $R^\varepsilon$ as a cost (Appendix A.6). The choice does not rest on the acoustic reference alone. Replacing the acoustic witness with another speaker's transcript still leaves WER at $0.294$, below the input, because the visual term carries the choice (Table IX), and on clips whose input transcript is wrong the lip term decides instead (Table XIV, second and third cases).

**Remark 1** (Scope)**.** $\mathbb{E}_{A\sim\omega_i}[L_i(A)]$ *is the loss of a hypothetical randomized selection procedure. It is generally unrelated by Jensen's inequality to* $\mathrm{WER}_2(z_i, \mathrm{ASR}(\sum_a \omega_i(a)a))$*, since waveform-to-transcript decoding and transcript error are not convex in the waveform. A WER guarantee for the mixture would require additional assumptions; these candidate-selection inequalities and the waveform-stability result provide neither an input-relative nor a deployed-enhancer-relative WER guarantee.*

### A.3 Temperature limits and ties

Let $B_i = \arg\min_a s_i(a)$. Subtracting the minimum score gives a numerator of one for each entry in $B_i$ and a numerator tending to zero for every other entry as $\tau \to 0^+$. Consequently,

$$\lim_{\tau\to 0^+} \omega_i(a) = \begin{cases} |B_i|^{-1}, & a \in B_i,\\ 0, & a \notin B_i.\end{cases}$$

If $|B_i| = 1$, the mixture converges to that candidate. Otherwise it converges to the average over $B_i$, which is not generally a tie-broken hard selection. For $\tau \to \infty$, every exponential tends to one, so $\omega_i(a) \to 1/M_i$.

### A.4 What the temperature trades

**Proposition 3** (What the temperature trades)**.** *Fix a finite pool of equal-length candidate waveforms, scores $s$, and Gibbs weights $\omega_\tau$.*

*(i) The expected score is nondecreasing in $\tau$, with*

$$\frac{d}{d\tau}\mathbb{E}_{\omega_\tau}[s] = \frac{\mathrm{Var}_{\omega_\tau}(s)}{\tau^2} \ge 0,$$

*and ranges from* $\min_a s(a)$ *as $\tau \to 0^+$ to the pool mean as $\tau \to \infty$. Likewise,*

$$\frac{d}{d\tau}H(\omega_\tau) = \frac{\mathrm{Var}_{\omega_\tau}(s)}{\tau^3} \ge 0.$$

*(ii) For the mixture $y = \sum_a \omega(a)a$, any reference waveform $r$, and any squared seminorm induced by a positive semidefinite inner product,*

$$\|y - r\|^2 = \sum_a \omega(a)\|a - r\|^2 - \sum_a \omega(a)\|a - y\|^2. \tag{13}$$

*Thus, the squared distortion of the mixture lies below the weighted mean candidate distortion by exactly the weighted spread of the pool around the mixture. The spread vanishes under hard selection and is at most $D^2/2$, where $D = \max_{a,b} \|a - b\|$.*

*Proof.* (i) Differentiating

$$\log \omega_\tau(a) = -\frac{s(a)}{\tau} - \log Z_\tau$$

gives

$$\dot{\omega}_\tau(a) = \frac{\omega_\tau(a)}{\tau^2} \left(s(a) - \mathbb{E}_{\omega_\tau}[s]\right).$$

Therefore,

$$\begin{aligned}\frac{d}{d\tau}\mathbb{E}_{\omega_\tau}[s] &= \sum_a \dot{\omega}_\tau(a) s(a) \\ &= \frac{\mathrm{Var}_{\omega_\tau}(s)}{\tau^2}.\end{aligned}$$

The entropy satisfies

$$H(\omega_\tau) = \frac{\mathbb{E}_{\omega_\tau}[s]}{\tau} + \log Z_\tau,$$

and differentiating yields

$$\frac{d}{d\tau} H(\omega_\tau) = \frac{\mathrm{Var}_{\omega_\tau}(s)}{\tau^3}.$$

The limits follow from Appendix A.3.

(ii) Expand

$$\|a - r\|^2 = \|a - y\|^2 + \|y - r\|^2 + 2\langle a - y, y - r\rangle$$

and take the $\omega$-weighted sum. The cross term vanishes because

$$\sum_a \omega(a)(a - y) = 0.$$

For the spread,

$$\sum_a \omega(a)\|a - y\|^2 = \frac{1}{2}\sum_{a,b} \omega(a)\omega(b)\|a - b\|^2 \le \frac{D^2}{2}.$$

□

The identity in Equation (13) is the ambiguity decomposition of Krogh & Vedelsby (1995), and (i) is the finite-pool form of a nonnegative heat capacity. The contribution is placing both on the same $\tau$ as the measurements. WER is not exactly monotone in $\tau$. It is $0.264$ at $\tau$=$0.05$ against $0.267$ under hard selection, while the expected score is monotone (Table II).

### A.5 Waveform stability

The witnesses are imperfect, so small score changes should not cause an abrupt switch between substantially different waveforms. Positive-temperature weighting provides a quantitative stability property for the actual waveform output, without assuming that the scores approximate true WER.

**Proposition 4** (Stability of evidence-weighted waveform aggregation)**.** *Fix a finite indexed pool $\mathcal{A}_i$ of candidate waveforms of equal length and a temperature $\tau > 0$. Let $s_i, s_i'$ be two finite score vectors on this same pool, and let $(\omega_i, y_i)$ and $(\omega_i', y_i')$ be their weights and outputs under Equation* (7)*. For any norm on the waveform space, define*

$$\delta_i(a) = s_i'(a) - s_i(a), \quad R_i = \max_a \delta_i(a) - \min_a \delta_i(a), \quad D_i = \max_{a,b \in \mathcal{A}_i} \|a - b\|. \tag{14}$$

*Then*

$$\|\omega_i' - \omega_i\|_1 \le \min\left\{2, \frac{R_i}{2\tau}\right\}, \qquad \|y_i' - y_i\| \le D_i \min\left\{1, \frac{R_i}{4\tau}\right\}. \tag{15}$$

*In particular, if $\max_a |s_i'(a) - s_i(a)| \le \eta_i$, then $\|y_i' - y_i\| \le D_i \min\{1, \eta_i/(2\tau)\}$.*

The proof follows below. The bound depends only on changes in relative candidate scores: a common offset leaves the weights and output unchanged. Thus, $\tau$ controls sensitivity to score perturbations as well as the selection–averaging trade-off. In contrast, hard selection can switch between different candidates under arbitrarily small perturbations near a tie. Proposition 4 therefore provides a useful robustness perspective for smooth weighting: with a fixed candidate pool, modest changes in the relative evidence scores lead to bounded changes in the returned waveform. This property complements, rather than replaces, the empirical evaluation of content fidelity. Because ASR decoding and WER are nonlinear functions of the waveform, we assess their behavior empirically in Section 5. The result concerns the aggregation stage with fixed candidate waveforms and does not address changes in the candidate pool induced by perturbations to the input or guidance targets.

*Proof of Proposition 4.* Suppress the input index. Interpolate the scores as $s_t(a) = s(a) + t\delta(a)$ for $t \in [0, 1]$, and let $w_t$ be their Gibbs weights at temperature $\tau$. Differentiating the normalized exponential gives

$$\frac{\mathrm{d}w_t(a)}{\mathrm{d}t} = -\frac{w_t(a)}{\tau}\left(\delta(a) - \sum_b w_t(b)\delta(b)\right). \tag{16}$$

Set $\ell = \min_a \delta(a)$ and $u = \max_a \delta(a)$. If a random variable $X$ takes values in $[\ell, u]$ with mean $\mu$, then $(X - \ell)(u - X) \geq 0$ implies

$$\mathrm{Var}(X) \leq (u - \mu)(\mu - \ell) \leq \frac{(u - \ell)^2}{4}.$$

By Cauchy–Schwarz, $\mathbb{E}|X - \mu| \leq \sqrt{\mathrm{Var}(X)} \leq (u - \ell)/2$. Apply this to $X = \delta(A)$ with $A \sim w_t$. Equation (16) yields

$$\left\| \frac{\mathrm{d}w_t}{\mathrm{d}t} \right\|_1 = \frac{1}{\tau}\mathbb{E}_{A\sim w_t} |\delta(A) - \mathbb{E}_{B\sim w_t}\delta(B)| \leq \frac{R}{2\tau}.$$

Integrating from 0 to 1 gives $\|w_1 - w_0\|_1 \leq R/(2\tau)$; the distance between two probability vectors is also at most 2.

For the waveform bound, put $d(a) = w_1(a) - w_0(a)$ and $T = \frac{1}{2}\sum_a |d(a)|$. The positive and negative parts $d_+, d_-$ each sum to $T$, since $\sum_a d(a) = 0$. If $T = 0$, the outputs coincide. Otherwise, $p(a) = d_+(a)/T$ and $n(a) = d_-(a)/T$ are probability vectors, and

$$\begin{aligned}
\|y_1 - y_0\| &= T \left\| \sum_a p(a)a - \sum_b n(b)b \right\| \\
&= T \left\| \sum_{a,b} p(a)n(b)(a - b) \right\| \\
&\leq T \sum_{a,b} p(a)n(b)\|a - b\| \leq TD.
\end{aligned}$$

Using $T \leq \min\{1, R/(4\tau)\}$ proves the claimed waveform inequality. Finally, $|\delta(a)| \leq \eta$ implies $R \leq 2\eta$, giving the stated special case. □

The $1/\tau$ sensitivity is not merely an artifact of this proof. Consider two distinct candidates $a, b$ with scores $(0, 0)$ and perturbed scores $(h, -h)$. At $h = 0$ the output is $(a + b)/2$; differentiating at zero gives

$$\left\| \left. \frac{\mathrm{d}y(h)}{\mathrm{d}h} \right|_{h=0} \right\| = \frac{\|a - b\|}{2\tau}.$$

Thus the local constant in the sup-norm score bound can be attained to first order. Hard selection instead chooses $a$ for $h < 0$ and $b$ for $h > 0$, exhibiting a discontinuity at the tie. Neither observation asserts continuity of ASR decoding or of WER.

### A.6 MEASURED QUANTITIES

On the 656 test clips with a defined pool (the remaining clip returns the enhancer's output), we computed every quantity of Propositions 1 and 3 for AVEC$^+$ from the stored scores, transcripts, and candidate waveforms. Neither bound of Proposition 1 is violated on any clip. Hard selection picks a lowest-loss candidate on $84.5\%$ of clips (554 of 656) and is at least as good as the enhancer's output on $95.1\%$. The mean realized regret $L_i(\hat{a}_i) - L_i(a_i^\star)$ is $0.044$. Here $\hat{a}_i$ is the first minimizer in pool order and $L_i$ is read from its own transcript. When scores tie, the returned waveform can differ from $\hat{a}_i$, and transcribing the returned waveforms gives $0.266$ on these clips and $0.267$ with the fallback clip, $0.048$ above the lowest-WER candidate.

The score error is positive on average ($+0.174$), but it is not a common offset. Within a clip, its range $R_i^\xi$ across the six candidates has mean $0.257$ and median $0.185$, larger than the mean error itself. Table I separates score and loss by candidate type. Each guided candidate lowers the score of the witness it was steered toward, but its loss also falls. For the lip-guided candidate the visual score falls by $0.083$ and the loss by $0.139$, so the lower score is not evidence of a scoring bias. The audio-guided candidate reduces both acoustic inconsistency and reference loss. Its mean acoustic-score residual decreases from 0.083 to 0.031; this difference alone does not establish guidance-induced scoring bias. By type, the mean combined error lies between $0.152$ and $0.190$, a spread much smaller than the mean within-clip range.

When the score margin between the two best candidates is at least $0.30$ (38 clips), the effective pool size $1/\sum_a \omega_i(a)^2$ is $1.07$ and the mixture's transcript equals the top candidate's on $92.1\%$ of them. Below $0.05$ (432 clips) the effective size is $4.74$.

The identity in Equation (13) holds on every clip to within $3.9 \times 10^{-13}$. The mixture's SI-SDR exceeds the weighted mean of its candidates' by $0.235$ dB on average, and the gain grows with the pool spread (Pearson $r$=0.38). Along $\tau$, the expected score rises monotonically from $0.414$ (hard selection) through $0.422$ ($\tau$=0.05) and $0.439$ ($\tau$=0.10) to $0.584$ (uniform), and the pool spread rises from 0 to 7.46. At the frozen $\tau$, the measured smoothing margin $0.025$ lies below its exponential bound $0.055$, and $\tau H(\omega) = 0.133 \le \tau \log 6 = 0.179$.

Table I: Score and loss by candidate type for AVEC$^+$ on the test set (656 clips with a defined pool, $\lambda_{\text{aud}}$=0.5). $L$ is the candidate's clipped WER against the reference; $s^{\text{lip}}$ and $s^{\text{aud}}$ are the visual and acoustic witness scores, and $s = s^{\text{lip}} + \lambda_{\text{aud}} s^{\text{aud}}$. Residual columns report the corresponding score minus $L$. All entries are means over clips.

| | | Visual witness | | Acoustic witness | | Combined | |
|---|---|---|---|---|---|---|---|
| Candidate | $L$ | $s^{\text{lip}}$ | $s^{\text{lip}} - L$ | $s^{\text{aud}}$ | $s^{\text{aud}} - L$ | $s$ | $s - L$ |
| Default ($K$=4, deployed) | 0.504 | 0.382 | −0.122 | 0.587 | +0.083 | 0.675 | +0.172 |
| Lip-guided ($K$=4) | 0.365 | 0.299 | −0.066 | 0.509 | +0.144 | 0.554 | +0.189 |
| Audio-guided ($K$=4) | 0.400 | 0.342 | −0.059 | 0.431 | +0.031 | 0.557 | +0.157 |
| Default mean ($K$=8) | 0.459 | 0.365 | −0.094 | 0.556 | +0.098 | 0.643 | +0.184 |
| Lip-guided mean ($K$=8) | 0.349 | 0.290 | −0.059 | 0.498 | +0.149 | 0.539 | +0.190 |
| Audio-guided mean ($K$=8) | 0.382 | 0.325 | −0.057 | 0.418 | +0.036 | 0.534 | +0.152 |

**The temperature on test.** Table II sweeps $\tau$ on the test set with $\lambda_{\text{aud}}$=0.5 and the six candidates fixed. The expected evidence score rises monotonically, as Proposition 3 requires, and WER follows it except at the hard end. Every quality margin grows with $\tau$, from hard selection ($\tau \to 0$) through the frozen point to the uniform average ($\tau \to \infty$). The frozen $\tau$=0.10 is an interior point of that band, not its optimum on either axis, and $0.05$ is better on content at a smaller quality margin.

## B EXPERIMENTAL PROTOCOL AND SETUP

### B.1 DATASETS, SYSTEMS, AND EVALUATION CONVENTIONS

**Datasets.** The primary condition is the complete LRS3–NTCD test mixture at $-5$ dB with $n$=657, using LRS3 test speech and NTCD noise mixed according to the corpus protocol. Condition transfer covers combinations of LRS3 and TCD-TIMIT speech, NTCD and DEMAND noise, and input SNRs

Table II: Temperature sweep of AVEC$^+$ on the test set ($n$=657, $\lambda_{\text{aud}}$=0.5, six candidates). $\mathbb{E}_\omega[s]$ is the mean expected evidence score over the 656 clips with a defined candidate pool. Markers indicate whether the paired 95% interval against the noisy input (WER 0.335) or always-enhance (PESQ 1.470, ESTOI 0.580, SI-SDR +3.24 dB) excludes zero favorably ($^*$), unfavorably ($^\dagger$), or includes zero ($^\circ$). The shaded row denotes the frozen temperature selected on trainval.

| | | Content | | Quality | | |
|---|---|---|---|---|---|---|
| $\tau$ | $\mathbb{E}_\omega[s]\uparrow$ | WER ↓ | Harm ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR (dB) ↑ |
| → 0 (hard selection) | 0.414 | 0.267$^*$ | 0.167 | 1.477$^*$ | 0.580$^\circ$ | +3.35$^*$ |
| 0.05 | 0.422 | 0.264$^*$ | 0.170 | 1.482$^*$ | 0.584$^*$ | +3.42$^*$ |
| 0.10 (frozen) | 0.439 | 0.279$^*$ | 0.181 | 1.485$^*$ | 0.586$^*$ | +3.46$^*$ |
| 0.15 | 0.457 | 0.291$^*$ | 0.202 | 1.486$^*$ | 0.587$^*$ | +3.49$^*$ |
| 0.20 | 0.472 | 0.298$^*$ | 0.213 | 1.487$^*$ | 0.588$^*$ | +3.51$^*$ |
| 0.30 | 0.497 | 0.309$^*$ | 0.227 | 1.488$^*$ | 0.589$^*$ | +3.52$^*$ |
| → ∞ (uniform) | 0.584 | 0.351$^\circ$ | 0.294 | 1.489$^*$ | 0.589$^*$ | +3.54$^*$ |

of $-5$, 0, and $+5$ dB. We include both official mixtures and remixed conditions where needed to complete the comparison across speech corpora, noise families, and SNRs. Full set definitions and sample counts are reported in Appendix F.1. Backbone transfer evaluates AV-CA-DiffUSE (Table 1), FlowAVSE, and PASE (Table 3) on the primary condition. No clip is filtered based on its outcome. The one primary clip with an empty input transcript receives the pre-specified fallback described in Appendix C.2.

**Evaluation conventions.** Automatic content evaluation uses Whisper WER against the corpus transcript. Harm is always computed relative to the corresponding input of the same clip following Equation (10). PESQ, ESTOI, and SI-SDR are computed against clean speech, while DNSMOS (OVRL) is computed without a reference. Unless stated otherwise, uncertainty intervals are obtained using paired clip bootstrap with 10,000 resamples at the 95% level. The TCD-DEMAND sets instead use a cluster bootstrap over the 90 utterances because each utterance appears under five noise conditions. The four criteria in Section 5.1 are pre-declared and reported jointly for each policy and set without a multiplicity correction. A policy is called "strict" only when all four intervals exclude zero in the favorable direction. Every transfer set is reported in full, including unfavorable results. Speaker-clustered intervals are provided in Appendix D.1, and the independent Parakeet re-scoring is reported in Appendix D.5. The listener study repeats the paired comparison using free-transcription WER on a frozen subset as described in Appendix E.1.

**Trainval corpus.** Combination-parameter selection uses a separate trainval corpus reconstructed from LRS3 `trainval`. It contains 587 clips with one clip per speaker and is duration-decile matched to the test set. Both sets have a median duration of 1.92 s. The clips are mixed at $-5$ dB using the corpus' own protocol, while the noise is drawn from NTCD volunteers disjoint from the test speakers. As a result, no source utterance, speaker, or noise recording appears in both trainval and test data. Mouth crops follow the geometry of the released crops. Validation against the official crops on held-out clips shows statistically indistinguishable generation WER and lip scores, with a per-clip disagreement correlation of 0.95. The two aggregation parameters of the proposed method are fixed using the pre-declared rule in Appendix D.1. The reconstructed corpus contains 587 clips, while the freezing grid is evaluated on the 585 usable clips reported there. The selection of $\tau$ and $\lambda_{\text{aud}}$ is fully reproducible from the trainval data.

### B.2 Harm severity and noise-type breakdown

**Severity of input-relative harm.** The harm rate in Equation (10) treats every harmed clip equally regardless of the magnitude of the increase in WER. To characterize the severity of these changes, Table III reports the distribution of the paired difference $\Delta\text{WER} = \text{WER}(y_i) - \text{WER}(x_i)$ over all 657 clips with the fallback applied. Always-enhance worsens more clips than it improves and produces relatively large increases in WER among the harmed clips, with a median $\Delta$WER of 0.40. Both AVEC and AVEC$^+$ reverse this balance, improving more clips than they worsen while also reducing the degradation among clips that remain harmed. The harmed subsets differ across policies,

so the conditional mean and median values should be interpreted as severity summaries rather than paired improvements.

Table III: Distribution of input-relative $\Delta$WER on the primary LRS3–NTCD $-5$ dB test set. "Improved," "tied," and "worse" are defined relative to the same clip's noisy input with tolerance $\delta$=0. The "worse by" bins partition harmed clips according to the magnitude of $\Delta$WER. The final two columns report the mean and median $\Delta$WER among harmed clips.

| | Outcome (%) | | | Worse by $\Delta$WER (%) | | | Harmed clips | |
|---|---|---|---|---|---|---|---|---|
| Policy | Improved | Tied | Worse | $(0, .25]$ | $(.25, .5]$ | $> .5$ | Mean | Median |
| always enhance | 15.7% | 39.7% | 44.6% | 14.9% | 14.0% | 15.7% | 0.50 | 0.40 |
| AVEC, weighted average | 26.8% | 50.8% | 22.4% | 12.3% | 6.8% | 3.2% | 0.34 | 0.25 |
| **AVEC$^+$, weighted average** | 28.3% | 53.6% | 18.1% | 10.4% | 5.0% | 2.7% | 0.31 | 0.25 |

**By noise type and tolerance.** Table IV breaks down performance by NTCD noise type and examines the sensitivity of the harm rate to the tolerance $\delta$ in Equation (10). Always-enhance produces the highest harm rates for Babble, Cafe, and White noise, where the noisy-input WER is also relatively high. In these conditions, AVEC$^+$ substantially reduces input-relative harm and lowers WER relative to always-enhance. Its WER is also below the noisy input for all noise types except Cafe. In the near-clean Car condition, all policies have low WER and low harm rates. Increasing $\delta$ reduces the measured harm rate for every policy while preserving the overall pattern. At $\delta$=0.25, the harm rate decreases from $29.7\%$ for always-enhance to $7.8\%$ for AVEC$^+$.

Table IV: Performance by NTCD noise type on the primary LRS3–NTCD $-5$ dB test set. For each noise type, policy entries report WER and input-relative harm at $\delta$=0, while the noisy-input WER is reported separately. The bottom rows show harm rates over all 657 clips for increasing values of the tolerance $\delta$ in Equation (10).

| | | WER / harm by policy | | | |
|---|---|---|---|---|---|
| Noise type ($n$) | Input WER | Always enhance | AVEC | AVEC$^+$ | Uniform (6) |
| Babble (130) | 0.620 | 0.776 / 0.508 | 0.554 / 0.292 | 0.472 / 0.192 | 0.572 / 0.338 |
| Cafe (129) | 0.472 | 0.750 / 0.651 | 0.531 / 0.380 | 0.488 / 0.349 | 0.558 / 0.442 |
| Car (121) | 0.019 | 0.016 / 0.025 | 0.013 / 0.017 | 0.010 / 0.017 | 0.016 / 0.033 |
| LR (141) | 0.231 | 0.320 / 0.326 | 0.233 / 0.191 | 0.209 / 0.149 | 0.251 / 0.255 |
| White (136) | 0.323 | 0.639 / 0.691 | 0.225 / 0.228 | 0.210 / 0.191 | 0.347 / 0.382 |
| *Harm sensitivity to tolerance $\delta$ over all clips* | | | | | |
| All (657), $\delta$=0.1 | – | 0.432 | 0.209 | 0.155 | 0.271 |
| All (657), $\delta$=0.25 | – | 0.297 | 0.100 | 0.078 | 0.140 |
| All (657), $\delta$=0.5 | – | 0.157 | 0.032 | 0.027 | 0.064 |

# C Implementation details

This section details the framework in Section 4, covering candidate generation and guidance, transcript processing, evidence scoring, and waveform aggregation.

## C.1 Candidate generation

**Sampling.** Candidates for the primary system are constructed by averaging reverse-diffusion trajectories of the frozen AV-UDiffSE+ sampler, each initialized with independent noise. We use the released sampling configuration with 30 reverse steps and a pseudo-likelihood weight of 1.75. The deployed, lip-guided, and audio-guided candidates each average $K$=4 trajectories. AVEC$^+$ additionally includes a $K$=8 candidate for each generation condition. This results in 12 trajectories per non-fallback input for AVEC and 36 for AVEC$^+$. All trajectories use fixed released seeds assigned by clip and sampling round. The resulting candidates are 16 kHz waveforms generated from the same input with identical STFT framing.

**Primary candidate-pool construction.** The $K{=}4$ and additional $K{=}8$ candidates are generated in separate sampling runs without trajectory reuse. Reusing trajectories from a cached run is considered only in the later budget ablations and is not used for the primary candidate pools.

**Lip-guided candidate behavior and limits.** The lip-guided candidate steers the frozen sampler toward the VSR-derived phoneme sequence using a test-time CTC objective (cf. Yemini et al., 2024). The guidance procedure is detailed in Appendix C.1. On the primary set, lip guidance reduces WER from $0.504$ to $0.366$ and LIP from $0.383$ to $0.300$, with a small decrease in DNSMOS of $0.09$. Despite these improvements, substantial variation remains across stochastic samples. Moreover, selecting unguided samples based on lip agreement does not recover the same WER improvement as guided generation. These results support the use of lip guidance to construct complementary candidates in the evaluated setting, but do not imply that guidance is necessary for every candidate pool or sampling budget.

**Guidance implementation.** The base model remains frozen throughout guidance. At each reverse-diffusion step, guidance applies a gradient update to the current clean-signal estimate $\hat{S}_0$. We first convert $\hat{S}_0$ to an 80-band log-mel spectrogram using a 16 kHz sampling rate, $n_{\text{fft}}{=}1024$, a window length of $1024$, a hop length of $256$, the Slaney scale, and a log floor of $10^{-5}$. The spectrogram is standardized using the training mean and variance and passed to a frozen audio CTC phoneme recognizer with a Conformer-style encoder, $d_{\text{model}}{=}384$, and six heads. Guidance then follows the gradient of the CTC loss with respect to the target phoneme sequence

$$S_t \leftarrow S_t - \underbrace{\lambda_{\text{guide}} \frac{\|S_t\|_1}{\|\nabla\|_1 + 10^{-8}}}_{\text{scale}} \nabla_{\hat{S}_0} \mathcal{L}_{\text{CTC}}\big(\text{CTC}(\text{mel}(\hat{S}_0)), p^{\text{VSR}}\big).$$

The magnitude normalization makes $\lambda_{\text{guide}}$ insensitive to the scale of the update across clips and noise conditions. We use $\lambda_{\text{guide}}{=}0.05$ throughout. This value is selected using trainval-DEV, where we additionally evaluate $\lambda_{\text{guide}} \in \{0.02, 0.1\}$. Neither alternative shows a clear WER improvement over $0.05$, so we keep this value.

For lip guidance, $p^{\text{VSR}}$ is obtained by applying grapheme-to-phoneme (G2P) conversion to the VSR transcript. The CTC recognizer uses blank index $0$ with `zero_infinity`. Guidance is applied at every reverse-diffusion step after the likelihood update without modifying the diffusion hyperparameters of the base model. The audio-guided candidate uses the same procedure with $p^{\text{ASR}}$, obtained from the Whisper transcript of the noisy input using the same G2P conversion. Both guidance conditions use $\lambda_{\text{guide}}{=}0.05$ and the same number of reverse steps. For each batch, the gradient is computed from the batch-mean clean estimate and applied to all four trajectories, whose time-domain mean forms the resulting candidate.

**Access requirements and candidates for other backbones.** Guidance needs (i) a differentiable, iterative sampler exposing its clean-signal estimate at each reverse step, and (ii) the auxiliary recognizer above. It therefore applies to open diffusion checkpoints, but not to closed APIs or to single-step samplers. FlowAVSE's one-step flow leaves no reverse trajectory to steer, which is why its candidate set omits the guided arm. For enhancers that admit neither guidance nor posterior samples (FlowAVSE, PASE) the second candidate used here is the *mean of input-perturbed outputs*: the mixture and seven copies of it with seeded Gaussian noise at $-20$ dB relative to the mixture RMS are each enhanced, and the eight outputs are averaged in the time domain, the same construction as the $K{=}8$ sample mean with the sampler's randomness replaced by input randomness. For the $K$-sample variant each posterior sample is nudged independently rather than through the batch mean, so the gradient acts on individual trajectories.

### C.2 TRANSCRIPTS, EVIDENCE SCORES, AND AGGREGATION

**Recognizers and text processing.** The ASR used for every transcript in the score (candidates and noisy input) and for the WER estimand is Whisper `medium.en` (Hugging Face `openai/whisper-medium.en`), run with its default English transcription prompt and greedy decoding (`generate`, at most 200 new tokens). Parakeet-TDT is used for independent re-scoring and, in one control, as the acoustic evidence source. Text normalization lower-cases, keeps letters and apostrophes, and collapses whitespace; the reference transcripts receive the same normalization. Word error is computed with `jiwer` with the reference as the first argument and clipped at 2, matching $\text{WER}_2$

in Equation (4). The VSR is the public auto-avsr checkpoint `vsr_trlrs2lrs3vox2avsp_base` on the $96\times96$ mouth ROI of its own preprocessing, with greedy decoding. Grapheme-to-phoneme conversion uses `g2p_en` with stress digits removed (39 ARPAbet phones plus silence); phonemes map to 14 viseme classes adapted from the 13-class set of Neti et al. (2000), repeated visemes are collapsed, and $d_{\mathrm{vis}}$ is the Levenshtein distance divided by the longer sequence length. Consistent with Equation (3), two empty viseme sequences have distance 0, while exactly one empty sequence gives distance 1. The same pipeline produces the CTC guidance targets. The mapping table and the normalization code are released.

**Waveform aggregation implementation.** Candidate waveforms are sampled at 16 kHz from the same input using the same STFT framing. Before aggregation, candidates are truncated to the shortest candidate length and combined sample by sample using Equation (7), without RMS normalization. If the mixture's peak exceeds 0.99, it is rescaled to 0.99. The common time grid makes waveform addition well-defined, although it does not imply exact phase or fine phonetic alignment across candidates.

For numerical stability, the implementation subtracts the minimum candidate score before exponentiation. Specifically, with

$$m_i = \min_{b\in\mathcal{A}_i} s_i(b),$$

we compute

$$\omega_i(a) = \frac{\exp(-(s_i(a)-m_i)/\tau)}{\sum_{b\in\mathcal{A}_i}\exp(-(s_i(b)-m_i)/\tau)}.$$

This is exactly equivalent to the unshifted Gibbs weighting in Equation (7), since the common factor $\exp(m_i/\tau)$ cancels during normalization.

**Fallback.** If the normalized Whisper transcript of the noisy input is empty, $s^{\mathrm{aud}}$ and the audio-guided target are undefined. Then the wrapper returns the deployed output without aggregation. This occurs on one primary clip (`8Z24LCysq3A/00001`) and on two TCD–NTCD $-5$ dB clips.

**Auxiliary-model provenance (no test overlap).** The CTC recognizer is trained only on clean audio from the LRS3 `trainval` partition (3,964 speakers); the official test partition (412 speakers) is speaker- and utterance-disjoint from it by corpus construction, so no test material enters this auxiliary model. The VSR is the public auto-avsr checkpoint trained on LRS2, the LRS3 `pretrain` and `trainval` partitions, VoxCeleb2, and AVSpeech; the LRS3 test partition is likewise disjoint from all of its training sources. The auxiliary CTC recognizer is trained once for this study and then frozen; the public VSR checkpoint is used without fine-tuning.

# D Complete results on the primary condition

## D.1 Parameter selection: freezing rule, grid, and robustness

**Freezing rule.** On the 585 of 587 reconstructed trainval clips with complete criteria, a pre-declared rule selected the lowest-WER configuration with WER below the input and PESQ, ESTOI, and SI-SDR above always-enhance (point estimates). The grid was $\tau \in \{0.10, 0.15, 0.20, 0.30\}$ and $\lambda_{\mathrm{aud}} \in \{0.5, 1.0\}$, with $\tau$=0.05 additionally tested for AVEC. Both selected $(\tau, \lambda_{\mathrm{aud}}) = (0.10, 0.5)$ (Table V); all AVEC$^+$ points were admissible, whereas AVEC had near-zero PESQ margins and failed PESQ at $\tau$=0.05. The once-evaluated test WERs were 0.313 and 0.279, respectively, with AVEC preserving always-enhance-level quality and AVEC$^+$ improving all three reference metrics (Table 1).

**Speaker clustering.** The 657 test clips come from 312 speakers, so clips are not independent. Resampling speakers instead of clips leaves every headline interval of the frozen configuration essentially unchanged: WER vs. input $[-.081, -.032]$ (clip-level $[-.079, -.033]$), harm vs. always-enhance $[-.302, -.228]$, PESQ $[+.008, +.021]$, ESTOI $[+.004, +.008]$, SI-SDR $[+.18, +.27]$ dB; every interval still excludes zero.

## D.2 Full ablation table and further controls

**Ablations** (Table VI; frozen AVEC$^+$ unless noted). Uniform averaging improves reference quality but does not establish WER below the input (0.351); hard selection lowers WER to 0.267 with smaller

Table V: Freezing grids on the reconstructed trainval set ($n$=585) for AVEC (3 candidates) and AVEC$^+$ (6 candidates). The noisy-input WER is 0.442; always-enhance gives WER 0.551, PESQ 1.446, ESTOI 0.567, and SI-SDR +3.31 dB. "Adm." marks configurations satisfying the pre-declared admissibility rule: WER below the noisy input and PESQ, ESTOI, and SI-SDR above always-enhance on point estimates. For each method, the rule selects the admissible configuration with the lowest WER; both select $\tau$=0.10 and $\lambda_{\text{aud}}$=0.5, highlighted in gray. These parameters are then frozen before test evaluation (Table 1).

| | | AVEC (3 candidates) | | | | | AVEC$^+$ (6 candidates) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | $\lambda_{\text{aud}}$ | WER ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR ↑ | Status | WER ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR ↑ | Status |
| 0.05 | 0.5 | 0.373 | 1.446 | 0.567 | +3.34 | PESQ below | – | – | – | – | – |
| 0.10 | 0.5 | **0.380** | 1.450 | 0.569 | +3.39 | Adm. | **0.360** | 1.463 | 0.576 | +3.50 | Adm. |
| 0.15 | 0.5 | 0.382 | 1.452 | 0.570 | +3.41 | Adm. | 0.370 | 1.465 | 0.577 | +3.52 | Adm. |
| 0.20 | 0.5 | 0.386 | 1.453 | 0.571 | +3.43 | Adm. | 0.372 | 1.467 | 0.577 | +3.53 | Adm. |
| 0.30 | 0.5 | 0.397 | 1.456 | 0.572 | +3.45 | Adm. | 0.387 | 1.468 | 0.578 | +3.55 | Adm. |
| 0.05 | 1.0 | 0.380 | 1.444 | 0.565 | +3.33 | PESQ, ESTOI below | – | – | – | – | – |
| 0.10 | 1.0 | 0.386 | 1.447 | 0.567 | +3.37 | Adm. | 0.369 | 1.461 | 0.574 | +3.47 | Adm. |
| 0.15 | 1.0 | 0.382 | 1.450 | 0.569 | +3.39 | Adm. | 0.372 | 1.463 | 0.575 | +3.50 | Adm. |
| 0.20 | 1.0 | 0.384 | 1.451 | 0.570 | +3.41 | Adm. | 0.377 | 1.464 | 0.576 | +3.51 | Adm. |
| 0.30 | 1.0 | 0.391 | 1.453 | 0.571 | +3.43 | Adm. | 0.383 | 1.467 | 0.577 | +3.53 | Adm. |

quality gains and a nonsignificant ESTOI margin. Candidate-removal variants and lip-only weighting meet the four point-estimate criteria, although one removal variant has a borderline PESQ interval. Thus, the joint rule's lower WER and harm than lip-only weighting do not establish that every candidate or evidence source is individually necessary. Input-only weighting also lacks a significant input-relative WER gain. The tested two-candidate pools fail to improve WER significantly over the input and lose PESQ/ESTOI; their Parakeet WERs (0.368, 0.387) also exceed AVEC's 0.334. The three deployed-form outputs are the smallest tested pool to reach input-level WER.

**Alternative weighting functions.** Uniformly averaging candidates within $\epsilon$ of the best score gives WER 0.267, 0.269, and 0.281 at $\epsilon = 0.05$, 0.10, and 0.20, respectively. Quality margins increase with $\epsilon$; all four paired intervals exclude zero at 0.10 and 0.20. Thus, comparable trade-offs are available without softmax weighting.

**The audio weight has a plateau.** At $\tau$=0.10, all tested $\lambda_{\text{aud}} \in \{0.25, 0.5, 1, 2\}$ satisfy the four criteria with intervals excluding zero. Increasing the audio weight raises WER and reduces quality margins; at $\lambda_{\text{aud}}$=4, the ESTOI interval includes zero. Several joint weights have lower WER point estimates than either single-term score, but lip-only weighting also meets all four criteria.

**A tested per-clip temperature rule.** Switching $\tau$ to 0.05 above the trainval median of DIVERSITY and to 0.30 below (Appendix G) gives test WER/harm of 0.274/19.6%, versus 0.279/18.1% for fixed $\tau$. Quality differences remain within the paired intervals. This tested switch provides no consistent improvement, so DIVERSITY remains a diagnostic.

**The score beats chance and beats quality.** Uniform random selection among six candidates gives WER 0.407 and harm 34.6% over 20 seeds, with worse content than the input and lower PESQ than always-enhance. DNSMOS-argmax over the five scored candidates gives WER 0.456 and harm 39.9%, despite DNSMOS 3.55. Both have higher WER than evidence-based hard selection (0.267).

**Consensus weighting.** Scoring each candidate by mean WER against the other five, at fixed $\tau$, gives WER 0.328 and harm 25.7%. Quality improves over always-enhance, but input-relative WER does not significantly improve ($-0.007$ $[-.032, +.018]$). Joint evidence has lower WER/harm point estimates; this tests one consensus score and temperature, not the entire method family.

**Admitting the input to the pool.** The score is computed against the noisy input's transcript, so the input itself, if admitted to the pool, agrees perfectly with it and takes most of the weight. Doing so with the input RMS-matched to the deployed output lowers WER to 0.270 and harm to 8.5%, and costs 0.11 PESQ, 0.033 ESTOI, and 1.9 dB SI-SDR against always-enhance. Restricting candidates to enhancer outputs avoids this direct route for reintroducing input noise; it does not itself guarantee content fidelity.

Table VI: Ablations of AVEC$^+$ on the primary LRS3–NTCD $-5$ dB test set ($n$=657; $\tau$=0.10 and $\lambda_{\text{aud}}$=0.5 unless noted). The noisy-input WER is 0.335; always-enhance gives WER 0.504, harm 0.446, PESQ 1.470, ESTOI 0.580, and SI-SDR +3.24 dB. PASS denotes WER below the noisy input together with PESQ, ESTOI, and SI-SDR above always-enhance on point estimates; strict additionally requires all four paired 95% intervals to exclude zero in the favorable direction. Rows are grouped by aggregation controls, candidate-pool ablations, evidence/score ablations, and alternative weighting or guidance rules. The DNSMOS-argmax control selects among the five candidates for which DNSMOS was computed.

| Policy | WER | Harm | PESQ | ESTOI | SI-SDR | Criterion |
|---|---|---|---|---|---|---|
| *Primary and aggregation controls* | | | | | | |
| AVEC$^+$, frozen (six candidates) | 0.279 | 0.181 | 1.485 | 0.586 | +3.46 | PASS, strict |
| uniform average, six candidates ($\tau \to \infty$) | 0.351 | 0.294 | 1.489 | 0.589 | +3.54 | fails WER |
| hard selection, same score ($\tau \to 0$) | 0.267 | 0.167 | 1.477 | 0.580 | +3.35 | PASS; ESTOI n.s. |
| AVEC (deployed outputs only) | 0.313 | 0.224 | 1.469 | 0.579 | +3.33 | quality level; WER marginal |
| AVEC, hard selection | 0.304 | 0.215 | 1.463 | 0.575 | +3.23 | fails PESQ, ESTOI |
| AVEC, uniform average | 0.352 | 0.291 | 1.475 | 0.584 | +3.44 | fails WER |
| two cand.: deployed + lip-guided | 0.349 | 0.289 | 1.459 | 0.576 | +3.26 | WER n.s.; fails PESQ, ESTOI |
| same, hard selection | 0.339 | 0.279 | 1.456 | 0.573 | +3.20 | WER n.s.; fails PESQ, ESTOI |
| two cand.: deployed + audio-guided | 0.369 | 0.277 | 1.466 | 0.577 | +3.26 | fails WER, ESTOI |
| same, hard selection | 0.367 | 0.271 | 1.461 | 0.574 | +3.20 | fails WER, PESQ, ESTOI |
| *Candidate-pool ablations* | | | | | | |
| without the audio-guided output | 0.281 | 0.195 | 1.482 | 0.586 | +3.46 | PASS, strict |
| without the deployed output | 0.275 | 0.186 | 1.479 | 0.584 | +3.45 | PASS, strict |
| without the unguided mean | 0.278 | 0.187 | 1.477 | 0.584 | +3.43 | PASS; PESQ marginal |
| the three $K$=8 means only | 0.298 | 0.213 | 1.474 | 0.583 | +3.46 | PASS; PESQ n.s. |
| *Evidence and score ablations* | | | | | | |
| score: lip term only ($\lambda_{\text{aud}}$=0) | 0.295 | 0.225 | 1.486 | 0.588 | +3.49 | PASS, strict |
| score: input term only (lip weight 0) | 0.325 | 0.218 | 1.487 | 0.586 | +3.49 | WER n.s. vs input |
| score: $\lambda_{\text{aud}}$=0.25 | 0.275 | 0.201 | 1.486 | 0.587 | +3.48 | PASS, strict |
| score: $\lambda_{\text{aud}}$=1 | 0.290 | 0.177 | 1.484 | 0.585 | +3.44 | PASS, strict |
| score: $\lambda_{\text{aud}}$=2 | 0.307 | 0.177 | 1.482 | 0.583 | +3.42 | PASS, strict |
| score: $\lambda_{\text{aud}}$=4 | 0.314 | 0.177 | 1.480 | 0.581 | +3.39 | PASS; ESTOI n.s. |
| score: consensus only (no lips, no input) | 0.328 | 0.257 | 1.487 | 0.587 | +3.50 | WER n.s. vs input |
| random selection (20 seeds) | 0.407 | 0.346 | 1.442 | 0.572 | +3.25 | fails WER and quality |
| DNSMOS-argmax selection (five cand.) | 0.456 | 0.399 | 1.454 | 0.578 | +3.34 | fails WER, PESQ |
| noisy input admitted to the pool | 0.270 | 0.085 | 1.364 | 0.547 | +1.38 | fails all quality |
| *Alternative weighting and guidance* | | | | | | |
| $\epsilon$-set avg., $\epsilon$=0.05 | 0.267 | 0.167 | 1.479 | 0.582 | +3.39 | PASS; ESTOI marginal |
| $\epsilon$-set avg., $\epsilon$=0.10 | 0.269 | 0.175 | 1.482 | 0.584 | +3.42 | PASS, strict |
| $\epsilon$-set avg., $\epsilon$=0.20 | 0.281 | 0.192 | 1.485 | 0.586 | +3.46 | PASS, strict |
| per-clip $\tau$ from DIVERSITY (0.05/0.30) | 0.274 | 0.196 | 1.484 | 0.585 | +3.44 | PASS, strict |
| lip-guided cand. at $\lambda_{\text{guide}}$=0.025 (half) | 0.291 | 0.205 | 1.490 | 0.587 | +3.49 | PASS, strict |
| lip-guided cand. at $\lambda_{\text{guide}}$=0.1 (double) | 0.264 | 0.177 | 1.460 | 0.579 | +3.32 | fails PESQ |
| means at $K$=4 instead of 8 | 0.284 | 0.201 | 1.483 | 0.585 | +3.43 | PASS, strict |

**Guidance strength is the lip-guided candidate's own dial.** The lip-guided candidate is generated at $\lambda_{\text{guide}}$=0.05, validated on trainval (Appendix C.1). Regenerating only that candidate at half and double strength, with everything else unchanged, moves it along the same content–quality axis as the rule itself (at $\lambda_{\text{guide}}$=0.025 WER 0.398, PESQ 1.464; at $\lambda_{\text{guide}}$=0.1 WER 0.376, PESQ 1.282), and the average inherits the trade. Half strength still satisfies all four criteria with intervals excluding zero (0.291; PESQ +0.020 [+.013, +.027]), and double strength recovers more content (0.264) but ends below always-enhance on PESQ ($-0.010$ [$-.018$, $-.002$]). The frozen $\lambda_{\text{guide}}$ sits inside, not at the edge of, the region where the rule passes.

### D.3 Compute budget and candidate count

**Matched trajectory-budget controls without guidance.** Table VII compares unguided means, evidence-weighted averages, and hard selection at the same 12/36 trajectory budgets as

AVEC/AVEC$^+$, without retuning. At 36 trajectories, the unguided mean improves PESQ by 0.031 over always-enhance but retains WER 0.450; evidence weighting and hard selection reach 0.390 and 0.360, still above the input. Guided pools reach 0.313 for AVEC (no significant input-relative gain) and 0.279 for AVEC$^+$. Thus, the tested unguided controls do not reproduce the content gain at matched trajectory counts. Runtime also includes transcript scoring: each trajectory costs about 2.8 s on one A6000, plus one Whisper pass per candidate.

Table VII: Matched trajectory-budget controls on the primary LRS3–NTCD $-5$ dB test set ($n$=657). Unguided policies use trajectory budgets matched to AVEC (12 trajectories) and AVEC$^+$ (36 trajectories), with $\tau$=0.10 and $\lambda_{\text{aud}}$=0.5 kept fixed. The noisy-input WER is 0.335. Markers indicate whether the paired 95% interval against the noisy input for WER, or against always-enhance for quality, excludes zero favorably ($^*$), unfavorably ($^\dagger$), or includes zero ($^\circ$).

| Policy | Traj. | WER ↓ | Harm ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR ↑ |
|---|---|---|---|---|---|---|
| *12-trajectory budget* | | | | | | |
| always enhance ($K$=4 batch mean) | 4 | 0.504$^\dagger$ | 0.446 | 1.470 | 0.580 | +3.24 |
| unguided mean, $K$=8 | 8 | 0.459$^\dagger$ | 0.412 | 1.484$^*$ | 0.586$^*$ | +3.43$^*$ |
| unguided mean, $K$=12 | 12 | 0.455$^\dagger$ | 0.393 | 1.490$^*$ | 0.589$^*$ | +3.50$^*$ |
| evidence-weighted, 12 unguided | 12 | 0.409$^\dagger$ | 0.355 | 1.484$^*$ | 0.585$^*$ | +3.35$^*$ |
| hard selection, 12 unguided | 12 | 0.402$^\dagger$ | 0.341 | 1.467$^\circ$ | 0.572$^\dagger$ | +2.94$^\dagger$ |
| **AVEC** (three deployed outputs) | 12 | **0.313**$^\circ$ | **0.224** | 1.469$^\circ$ | 0.579$^\circ$ | +3.33$^*$ |
| *36-trajectory budget* | | | | | | |
| unguided mean, $K$=36 | 36 | 0.450$^\dagger$ | 0.402 | **1.501**$^*$ | **0.591**$^*$ | **+3.58**$^*$ |
| evidence-weighted, 36 unguided | 36 | 0.390$^\dagger$ | 0.330 | 1.498$^*$ | 0.590$^*$ | +3.51$^*$ |
| hard selection, 36 unguided | 36 | 0.360$^\dagger$ | 0.294 | 1.477$^\circ$ | 0.574$^\dagger$ | +3.05$^\dagger$ |
| **AVEC$^+$** (six candidates) | 36 | **0.279**$^*$ | **0.181** | 1.485$^*$ | 0.586$^*$ | +3.46$^*$ |

**Sampler budget.** On one A6000, AVEC and AVEC$^+$ cost approximately 37 and 105 s per clip, with VSR and Whisper adding seconds. Reducing all three additional means from $K$=8 to 4 lowers the budget from 36 to 24 trajectories and yields WER 0.284, harm 20.1%, and significant improvements on all four criteria (Table VIII). Two disjoint cache halves both pass and differ by 0.003 WER; this is only a limited seed-sensitivity check.

### D.4 Does the evidence carry the content? Substitutions, degradation, and a clean-input control

With all candidates and aggregation parameters fixed, mismatching visual/acoustic/both evidence sources gives WER 0.331/0.294/0.368, respectively; only the acoustic-only mismatch retains a significant input-relative gain (Table IX). Reference-quality gains persist. Fixed trainval-mean weights (0.11, 0.21, 0.17 for deployed/lip-/audio-guided outputs and 0.22, 0.12, 0.16 for their respective eight-sample means) yield WER 0.337, worse than per-clip weighting by 0.058 [.041, .075] under Whisper and 0.041 [.025, .058] under Parakeet; this is not an optimized static-mixture baseline. Ground-truth visual evidence gives WER 0.271 (0.260 with hard selection), versus 0.219 for the candidate-selection oracle. Better visual evidence alone therefore does not close the candidate-utilization gap.

**Degraded visual evidence erodes the content gain gracefully.** Table X degrades only the VSR input, leaving all candidates unchanged. Blur and halved frame rate (12.5 fps) raise VSR WER to 0.379/0.424, while AVEC$^+$ remains significantly below input WER at 0.291/0.286. Severe occlusion raises VSR WER to 0.906 and removes this significant margin (0.328, comparable to mismatched evidence in Table IX). PESQ/ESTOI vary little. These tested degradations show tolerance to moderate visual-evidence errors, not a general accuracy threshold for useful VSR.

**A clean-input control: when there is nothing to fix.** On all 657 clean-input clips, WER remains 0.016–0.019, with roughly 2% input-relative harm across policies (Table XI). Enhancement nevertheless lowers PESQ/ESTOI from 4.64/1.00 to 4.03/0.96, and guidance reduces quality further. AVEC$^+$ partly recovers this loss but remains 0.13 PESQ below always-enhance, despite 0.25 dB higher SI-SDR. DNSMOS instead ranks enhanced speech above the clean input (3.75 versus 3.53).

Table VIII: Sampler-budget analysis on LRS3–NTCD at $-5$ dB ($n$=657). The upper block varies the number $K$ of trajectories averaged into individual default, lip-guided, and audio-guided candidates; the lower block reports complete AVEC/AVEC$^+$ policies and their total trajectory budgets. The deployed enhancer output is a $K$=4 unguided batch mean; its headline values in Table 1 differ slightly from the corresponding $K$=4 row because the runs use different seeds and batching. The two 24-trajectory AVEC$^+$ rows use disjoint halves of the sample cache and therefore different sampler seeds, providing a limited seed-sensitivity check.

| Candidate / policy | Traj. | WER ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR ↑ |
|---|---|---|---|---|---|
| *Individual candidate construction* | | | | | |
| single sample, $K$=1 | 1 | 0.614 | 1.413 | 0.557 | +2.53 |
| unguided mean, $K$=2 | 2 | 0.529 | 1.446 | 0.573 | +3.03 |
| unguided mean, $K$=4 (= deployed output up to seed) | 4 | 0.486 | 1.465 | 0.581 | +3.28 |
| unguided mean, $K$=8 | 8 | 0.459 | 1.484 | 0.586 | +3.43 |
| lip-guided mean, $K$=4 | 4 | 0.372 | 1.425 | 0.567 | +3.13 |
| lip-guided mean, $K$=8 | 8 | 0.350 | 1.433 | 0.572 | +3.30 |
| audio-guided mean, $K$=4 | 4 | 0.402 | 1.415 | 0.564 | +3.13 |
| audio-guided mean, $K$=8 | 8 | 0.382 | 1.432 | 0.568 | +3.27 |
| *Aggregate policies* | | | | | |
| AVEC (three deployed outputs) | 12 | 0.313 | 1.469 | 0.579 | +3.33 |
| AVEC$^+$, means at $K$=4 (samples 0–3) | 24 | 0.284 | 1.483 | 0.585 | +3.43 |
| AVEC$^+$, means at $K$=4 (samples 4–7) | 24 | 0.281 | 1.481 | 0.584 | +3.43 |
| **AVEC$^+$, means at $K$=8 (frozen)** | 36 | 0.279 | 1.485 | 0.586 | +3.46 |

The wrapper shows little content change in this control but cannot return the unmodified input. Deciding whether enhancement is needed requires a separate gate.

### D.5 The final policies under an independent recognizer

**Independent re-scoring with Parakeet.** Parakeet-TDT 0.6B v2 (NVIDIA NeMo Team, 2025), an independent FastConformer transducer, reads the input comparably to Whisper (WER 0.339 versus 0.335). Under Parakeet, AVEC$^+$ improves input-relative WER by 0.032 $[-.056, -.008]$ with harm 22.8%, versus always-enhance's 41.7%; AVEC remains at input level with harm 25.7%, and uniform averaging is not significantly better than input (Table XII). Effect sizes are evaluator-dependent, without identifying a uniquely evaluator-specific component. Substituting only Parakeet's input transcript in the score, with candidates unchanged and the audio-guided target still Whisper-derived,

Table IX: Evidence substitutions for AVEC$^+$ on the primary LRS3–NTCD $-5$ dB test set ($n$=657), with all six candidates fixed and the aggregation parameters frozen at $\tau$=0.10 and $\lambda_{\text{aud}}$=0.5. "Mismatched" denotes the transcript from the next different-speaker clip in a fixed ordering, whereas "ground truth" uses the reference transcript as visual evidence. ΔWER is paired against the noisy input; ΔPESQ and ΔESTOI are relative to always-enhance. The final two rows are aggregation controls using fixed trainval-mean weights and an in-pool oracle that selects the lowest-reference-WER candidate.

| Evidence (visual / audio) | WER ↓ | Harm ↓ | ΔWER vs. input | ΔPESQ | ΔESTOI |
|---|---|---|---|---|---|
| *Source-evidence substitutions* | | | | | |
| **VSR / Whisper (AVEC$^+$)** | 0.279 | 0.181 | $-0.056$ $[-.079, -.033]$ | +0.015 | +0.006 |
| mismatched / Whisper | 0.331 | 0.228 | $-0.005$ $[-.026, +.016]$ | +0.017 | +0.007 |
| VSR / mismatched | 0.294 | 0.222 | $-0.042$ $[-.067, -.016]$ | +0.016 | +0.008 |
| mismatched / mismatched | 0.368 | 0.295 | $+0.033$ $[+.007, +.059]$ | +0.017 | +0.007 |
| same, hard selection | 0.422 | 0.373 | $+0.087$ $[+.060, +.115]$ | +0.004 | −0.002 |
| ground truth / Whisper | 0.271 | 0.177 | $-0.064$ $[-.087, -.041]$ | +0.014 | +0.006 |
| same, hard selection | 0.260 | 0.163 | $-0.075$ $[-.098, -.053]$ | +0.008 | +0.001 |
| *Aggregation controls* | | | | | |
| fixed global weights (trainval mean) | 0.337 | 0.280 | $+0.002$ $[-.023, +.027]$ | +0.015 | +0.009 |
| oracle: lowest-WER candidate | 0.219 | 0.135 | – | – | – |

Table X: Robustness to degraded visual evidence on the primary LRS3–NTCD $-5$ dB test set ($n$=657). The six enhancement candidates are held fixed while only the mouth video supplied to the VSR evidence extractor is degraded; $\tau$=0.10 and $\lambda_{\text{aud}}$=0.5 remain frozen. "VSR WER" is the VSR transcript's WER against the reference. Weighted-average and hard-selection content results are shown separately, while PESQ and ESTOI are reported for the weighted output. Markers denote paired 95% comparisons against the noisy input for WER and against always-enhance for quality, as in Table IX.

| | | Weighted average | | Hard selection | | Quality (weighted) | |
|---|---|---|---|---|---|---|---|
| Mouth video | VSR WER | WER ↓ | Harm ↓ | WER ↓ | Harm ↓ | PESQ ↑ | ESTOI ↑ |
| Intact | 0.255 | 0.279* | 0.181 | 0.267* | 0.167 | 1.485* | 0.586* |
| Gaussian blur, $\sigma$=1.5 px | 0.313 | 0.283* | 0.181 | 0.275* | 0.167 | 1.485* | 0.586* |
| Gaussian blur, $\sigma$=3 px | 0.379 | 0.291* | 0.195 | 0.284* | 0.184 | 1.486* | 0.586* |
| frame rate halved (12.5 fps) | 0.424 | 0.286* | 0.187 | 0.280* | 0.169 | 1.485* | 0.586* |
| 40% of the frame occluded | 0.906 | 0.328° | 0.227 | 0.316* | 0.183 | 1.486* | 0.587* |

Table XI: Clean-input control on the LRS3 test clips ($n$=657), using the same candidate-generation and aggregation settings as the primary experiment ($\tau$=0.10, $\lambda_{\text{aud}}$=0.5). Here the clean speech itself is the input and therefore defines the ceiling for reference-based quality. Harm is relative to that clean input. Markers indicate whether paired 95% intervals against the input for WER, or against always-enhance for quality, exclude zero favorably (*), unfavorably (†), or include zero (°).

| Policy | WER ↓ | Harm ↓ | PESQ ↑ | ESTOI ↑ | SI-SDR ↑ | DNSMOS ↑ |
|---|---|---|---|---|---|---|
| clean input | 0.019 | – | 4.644 | 1.000 | – | 3.53 |
| always enhance | 0.017° | 0.021 | 4.026 | 0.962 | +19.29 | 3.75 |
| always lip-guided | 0.019° | 0.027 | 3.552† | 0.942† | +19.07† | – |
| always audio-guided | 0.016° | 0.017 | 3.551† | 0.943† | +19.12† | – |
| AVEC | 0.016° | 0.017 | 3.833† | 0.955† | +19.47* | 3.73 |
| **AVEC$^+$** | 0.017° | 0.018 | 3.894† | 0.958† | +19.54* | 3.74 |

gives AVEC$^+$ WER 0.272 under Whisper and 0.307 under Parakeet. The content gain therefore persists when acoustic evidence and evaluation use different recognizers.

Table XII: Final policies re-scored with Parakeet-TDT 0.6B v2 on the primary test set ($n$=657). Harm is the fraction of clips whose Parakeet WER is higher than that of the corresponding noisy input. The two ΔWER columns report paired differences from the noisy input and always-enhance, respectively, with 95% clip-bootstrap intervals. The lower block additionally replaces Whisper with Parakeet only as the acoustic evidence source while leaving the remaining policy unchanged.

| Policy | WER ↓ | Harm ↓ | ΔWER vs. input | ΔWER vs. always-enhance |
|---|---|---|---|---|
| noisy input | 0.339 | – | – | −0.142 [−.171, −.113] |
| always enhance | 0.481 | 0.417 | +0.142 [+.113, +.171] | – |
| always lip-guided | 0.374 | 0.311 | +0.035 [+.008, +.062] | −0.107 [−.131, −.082] |
| always audio-guided | 0.402 | 0.349 | +0.063 [+.040, +.087] | −0.078 [−.102, −.055] |
| uniform average, AVEC$^+$ candidates | 0.353 | 0.283 | +0.014 [−.011, +.039] | −0.127 [−.149, −.106] |
| fixed global weights (trainval mean) | 0.348 | 0.275 | +0.010 [−.016, +.035] | −0.132 [−.155, −.110] |
| AVEC, weighted average (frozen $\tau$) | 0.334 | 0.257 | −0.004 [−.029, +.020] | −0.146 [−.169, −.124] |
| **AVEC$^+$, weighted average (frozen $\tau$)** | **0.307** | **0.228** | **−0.032** [−.056, −.008] | **−0.173** [−.198, −.149] |
| *Using Parakeet as the acoustic evidence source* | | | | |
| AVEC, acoustic evidence from Parakeet | 0.331 | 0.256 | −0.008 [−.033, +.017] | −0.149 [−.172, −.127] |
| AVEC$^+$, acoustic evidence from Parakeet | 0.307 | 0.234 | −0.031 [−.055, −.008] | −0.173 [−.197, −.149] |

# E INDEPENDENT VALIDATION BEYOND THE PRIMARY EVALUATOR

## E.1 LISTENER CORROBORATION PROTOCOL

The listening test is meant only as corroboration of content and quality. Automatic metrics remain the primary evaluation since they cover all clips, policies, backbones and transfer conditions, whereas the listening test uses 30 clips. We collected responses from 53 Amazon Mechanical Turk workers. Each worker transcribed 20 clips drawn at random from the 30, each in one of the seven conditions of Table XIII (one condition per clip). This yields five or six transcripts for each of the 210 clip-condition cells, 1,060 in total. Workers first completed a headphone screening (Woods et al., 2017) of three trials (the original uses six and requires five) and had to get at least two right. In each trial the worker typed the words (at most five replays, empty responses not accepted) and rated overall quality on a five-point scale. Three clean attention-check clips were interleaved with the test clips. A submission was included in the analysis if at least two of the three checks had word accuracy of $0.80$ or higher and the responses indicated a good-faith attempt. Listener WER was computed against the gold transcript, clipped to $[0, 1]$ per transcript, then averaged per cell and over clips. Details on consent, compensation, screening and rejection are in the Ethics statement.

Waveforms, policies, thresholds and the clip subset were taken from the automatic protocol and frozen before annotation. Human labels did not influence any parameter, including the weights, $\tau$ and $\lambda_{\mathrm{aud}}$. To keep a clean or enhanced version from revealing the words of a noisy version heard later, no worker heard more than one rendition of a source utterance. Item order and file names were randomized, and all waveforms RMS-normalized. Gold transcripts were checked independently against the clean audio and never shown to workers. Normalization rules (spelling, contractions, fillers, partial words) were fixed before scoring. The two confirmatory questions, on mean listener WER per clip-condition, are whether always-enhance increases input-relative loss and whether frozen AVEC reduces it. Hard selection was not rated, so its quality comparison remains automatic.

## E.2 RESULTS

Among the processed conditions, listener WER follows the content ordering Whisper gives on the same clips, from AVEC$^+$ through AVEC and the always-on lip- and audio-guided candidates to always-enhance (Tables 4 and XIII). Even clean speech has listener WER $0.147$, since freely transcribing short, out-of-context utterances is hard on top of any acoustic degradation. Every enhanced condition is rated above the noisy input and below the clean source, and AVEC$^+$ is rated highest among the enhancement policies while having the lowest listener WER, so the content gain of evidence-guided aggregation does not cost the quality gain of enhancement.

Table XIII: Listener WER (free transcription) and quality (1 to 5) for the seven conditions (30 clips, $n$=53 listeners), mean $\pm$ standard deviation across clips.

| Condition | Listener WER ↓ | Quality ↑ (1–5) |
|---|---|---|
| Clean source | $0.147 \pm 0.145$ | $4.12 \pm 0.39$ |
| Noisy input | $0.608 \pm 0.312$ | $2.28 \pm 1.24$ |
| Always enhance | $0.586 \pm 0.307$ | $2.83 \pm 1.00$ |
| Always lip-guided | $0.465 \pm 0.366$ | $2.92 \pm 1.25$ |
| Always audio-guided | $0.468 \pm 0.371$ | $2.97 \pm 1.20$ |
| AVEC (3 cand.) | $0.455 \pm 0.343$ | $2.84 \pm 1.03$ |
| **AVEC$^+$ (6 cand.)** | $\mathbf{0.433 \pm 0.344}$ | $\mathbf{3.03 \pm 1.11}$ |

## E.3 QUALITATIVE EXAMPLES

Table XIV shows recovery from a correct input transcript, two improvements beyond an incorrect input transcript using lip-guided candidates, and a failure when all candidates are wrong. Waveform averaging need not preserve any candidate transcript: among 316 clips with errors in every candidate, it beats the best candidate on 30 and is error-free on 3; across the full set, it is worse than every

Table XIV: Qualitative examples on LRS3–NTCD at $-5$ dB. Each block lists the reference, the Whisper transcript of the noisy input, the six AVEC$^+$ candidates with their frozen-temperature evidence weights $\omega$, and the transcript of the returned weighted waveform. The first example starts from a correct noisy-input transcript and shows recovery from an always-enhance error; the next two start from incorrect noisy-input transcripts yet return outputs better than both the input and the deployed enhancer. The final block is a failure case in which all candidates have lost the reference content, so aggregation cannot recover it.

| Output | Weight / WER | Whisper transcript |
|---|---|---|
| *recovered (input already right)* - clip `c9V6OKlY80k__00001` | | |
| Reference | | but i love it for a totally different reason |
| Noisy input | WER 0.00 | but i love it for a totally different reason |
| Deployed | $\omega$=0.00, WER 1.00 | franko de f ljusov |
| Lip-guided | $\omega$=0.01, WER 0.44 | but i love it for the food |
| Audio-guided | $\omega$=0.49, WER 0.00 | but i love it for a totally different reason |
| Guided mean | $\omega$=0.02, WER 0.44 | but i love it for the full release |
| Unguided mean | $\omega$=0.00, WER 0.67 | but i love the food |
| Audio-guided mean | $\omega$=0.49, WER 0.00 | but i love it for a totally different reason |
| **Returned average** | WER 0.00 | but i love it for a totally different reason |
| *improved beyond the input* - clip `6wNif5SlN08__00001` | | |
| Reference | | the boys reminded me of home |
| Noisy input | WER 1.17 | of course it's like the game of cards |
| Deployed | $\omega$=0.00, WER 1.33 | would you mind if we have a minute |
| Lip-guided | $\omega$=0.00, WER 1.00 | and what it might mean |
| Audio-guided | $\omega$=0.00, WER 1.00 | of what you might believe in |
| Guided mean | $\omega$=0.98, WER 0.00 | the boys reminded me of home |
| Unguided mean | $\omega$=0.00, WER 1.17 | what you might do me but i love |
| Audio-guided mean | $\omega$=0.02, WER 0.50 | it always reminded me of home love |
| **Returned average** | WER 0.00 | the boys reminded me of home |
| *improved beyond the input* - clip `jggHw5PdQV0__00002` | | |
| Reference | | and who does that benefit |
| Noisy input | WER 1.00 | and move this back a bit |
| Deployed | $\omega$=0.00, WER 1.40 | i knew this was going to happen |
| Lip-guided | $\omega$=0.19, WER 0.60 | and who does that point ever get |
| Audio-guided | $\omega$=0.01, WER 0.80 | i knew just that |
| Guided mean | $\omega$=0.42, WER 0.20 | and who does that play |
| Unguided mean | $\omega$=0.22, WER 0.40 | and who is that player |
| Audio-guided mean | $\omega$=0.16, WER 0.60 | and who does that play for us |
| **Returned average** | WER 0.20 | and who does that play |
| *failure* - clip `1PNX6MSdVsk__00010` | | |
| Reference | | because it's given by other people |
| Noisy input | WER 0.00 | because it's given by other people |
| Deployed | $\omega$=0.01, WER 1.50 | i am a beautiful beautiful beautiful smiley tired obeseus |
| Lip-guided | $\omega$=0.34, WER 0.83 | this is useless my other pieces |
| Audio-guided | $\omega$=0.06, WER 0.83 | for this yes even by other reasons |
| Guided mean | $\omega$=0.15, WER 1.00 | this is my father jesus |
| Unguided mean | $\omega$=0.22, WER 0.83 | you are my other pieces |
| Audio-guided mean | $\omega$=0.22, WER 0.83 | this is my other season |
| **Returned average** | WER 0.83 | this is my other piece |

candidate on 9 clips (1.4%). Thus waveform convexity implies no corresponding transcript-level guarantee.

Table XV: Full condition-transfer results for the primary enhancer with the trainval-selected $\tau$ and $\lambda_{\text{aud}}$ frozen; candidates are regenerated independently for every set. Panel (a) reports WER / input-relative harm (%), with the noisy-input WER included in each set label. Panel (b) reports PESQ / ESTOI / SI-SDR (dB). Markers on AVEC and AVEC$^+$ denote paired 95% intervals that exclude zero favorably ($^*$), unfavorably ($^\dagger$), or include zero ($^\circ$), comparing WER with the noisy input and quality with always-enhance. TCD–DEMAND intervals use a cluster bootstrap over the 90 utterances. LRS3–NTCD 0 dB and TCD–DEMAND 0 dB are remixes; the TCD–NTCD remix rows use the primary-set mixing recipe at the labeled global SNR. Independent Parakeet and DNSMOS results for the same sets are given in Table XVI.

**(a) Content preservation**

| Set ($n$; noisy WER) | Always enhance WER / Harm | AVEC WER / Harm | AVEC$^+$ WER / Harm |
|---|---|---|---|
| LRS3-DEMAND $-5$ (657; .284) | .374 / 30.0 | .265$^\circ$ / 18.3 | **.240**$^*$ / 15.8 |
| LRS3-DEMAND 0 (657; .090) | .148 / 23.7 | .077$^*$ / 8.8 | **.075**$^*$ / 7.8 |
| LRS3-DEMAND $+5$ (656; .039) | .050 / 9.5 | **.028**$^*$ / 3.2 | .028$^*$ / 2.4 |
| TCD-DEMAND $-5$ (450; .210) | .240 / 26.0 | .198$^\circ$ / 20.2 | **.180**$^*$ / 15.1 |
| TCD-DEMAND 0 (450; .083) | .104 / 21.1 | .077$^\circ$ / 13.1 | **.075**$^\circ$ / 12.2 |
| TCD-DEMAND $+5$ (450; .040) | .049 / 13.6 | **.042**$^\circ$ / 8.7 | .042$^\circ$ / 9.6 |
| TCD-NTCD $-5$ (880; 1.000) | 1.212 / 46.8 | 1.097$^\dagger$ / 33.6 | **1.070**$^\dagger$ / 29.2 |
| TCD-NTCD 0 (882; .868) | .975 / 47.7 | .817$^*$ / 28.9 | **.782**$^*$ / 25.1 |
| TCD-NTCD $+5$ (882; .567) | .606 / 41.3 | .474$^*$ / 20.7 | **.459**$^*$ / 19.6 |
| TCD-NTCD remix $-5$ (882; .628) | .662 / 34.6 | .541$^*$ / 20.4 | .526$^*$ / 19.3 |
| TCD-NTCD remix 0 (882; .397) | .417 / 34.4 | .316$^*$ / 18.4 | .301$^*$ / 17.5 |
| TCD-NTCD remix $+5$ (882; .213) | .229 / 27.3 | .161$^*$ / 13.2 | .157$^*$ / 11.0 |
| LRS3-NTCD 0 (657; .089) | .206 / 34.4 | .092$^\circ$ / 12.3 | **.083**$^\circ$ / 10.5 |
| LRS3-NTCD $+5$ (664; .038) | .073 / 16.9 | .033$^\circ$ / 5.0 | **.031**$^*$ / 3.6 |

**(b) Reference-based quality**

| Set ($n$; noisy WER) | Always enhance PESQ / ESTOI / SI-SDR | AVEC PESQ / ESTOI / SI-SDR | AVEC$^+$ PESQ / ESTOI / SI-SDR |
|---|---|---|---|
| LRS3-DEMAND $-5$ (657; .284) | 1.904/.654/+5.66 | 1.894$^\dagger$/.652$^\dagger$/+5.80$^*$ | 1.913$^*$/.657$^*$/+5.89$^*$ |
| LRS3-DEMAND 0 (657; .090) | 2.201/.764/+9.82 | 2.166$^\dagger$/.763$^\dagger$/+9.90$^*$ | 2.189$^\dagger$/.766$^*$/+9.96$^*$ |
| LRS3-DEMAND $+5$ (656; .039) | 2.564/.845/+13.14 | 2.507$^\dagger$/.841$^\dagger$/+13.22$^*$ | 2.538$^\dagger$/.844$^\dagger$/+13.29$^*$ |
| TCD-DEMAND $-5$ (450; .210) | 1.942/.737/+9.81 | 1.944$^\circ$/.730$^\dagger$/+9.81$^\circ$ | 1.960$^*$/.735$^\dagger$/+9.90$^*$ |
| TCD-DEMAND 0 (450; .083) | 2.322/.823/+14.02 | 2.301$^\dagger$/.815$^\dagger$/+13.98$^\circ$ | 2.318$^\circ$/.819$^\dagger$/+14.05$^\circ$ |
| TCD-DEMAND $+5$ (450; .040) | 2.669/.877/+16.99 | 2.627$^\dagger$/.870$^\dagger$/+16.96$^\circ$ | 2.649$^\dagger$/.873$^\dagger$/+17.04$^*$ |
| TCD-NTCD $-5$ (880; 1.000) | 1.210/.238/−2.21 | 1.212$^\circ$/.236$^\dagger$/−2.41$^\dagger$ | 1.220$^*$/.240$^\circ$/−2.29$^\dagger$ |
| TCD-NTCD 0 (882; .868) | 1.391/.414/+4.05 | 1.397$^*$/.412$^\dagger$/+3.86$^\dagger$ | 1.409$^*$/.417$^*$/+4.00$^\dagger$ |
| TCD-NTCD $+5$ (882; .567) | 1.698/.585/+8.70 | 1.699$^\circ$/.581$^\dagger$/+8.55$^\dagger$ | 1.710$^*$/.586$^*$/+8.65$^\dagger$ |
| TCD-NTCD remix $-5$ (882; .628) | 1.688/.566/+7.28 | 1.680$^\dagger$/.562$^\dagger$/+7.19$^\dagger$ | 1.696$^\circ$/.568$^*$/+7.31$^\circ$ |
| TCD-NTCD remix 0 (882; .397) | 1.982/.685/+11.30 | 1.961$^\dagger$/.680$^\dagger$/+11.20$^\dagger$ | 1.984$^\circ$/.685$^\circ$/+11.30$^\circ$ |
| TCD-NTCD remix $+5$ (882; .213) | 2.329/.779/+14.72 | 2.289$^\dagger$/.774$^\dagger$/+14.66$^\dagger$ | 2.309$^\dagger$/.777$^\dagger$/+14.75$^*$ |
| LRS3-NTCD 0 (657; .089) | 1.737/.713/+7.96 | 1.729$^\dagger$/.712$^\circ$/+7.99$^\circ$ | 1.750$^*$/.717$^*$/+8.07$^*$ |
| LRS3-NTCD $+5$ (664; .038) | 2.106/.801/+11.42 | 2.075$^\dagger$/.795$^\dagger$/+11.48$^*$ | 2.097$^\dagger$/.799$^\dagger$/+11.55$^*$ |

# F Generalization: full results

## F.1 Condition transfer

**Condition transfer: full table.** Table XV lists every condition-transfer set with Whisper WER, harm, PESQ, ESTOI, and SI-SDR for always-enhance, AVEC, and AVEC$^+$, with paired-interval markers; Figure 4 in the main text plots the WER column as the difference from the noisy input, and the PESQ column, for the nine LRS3 and TCD-DEMAND sets and the three TCD-NTCD remixes. The three official TCD-NTCD sets appear here only: they are the NTCD-TIMIT release mixtures, whose SNR labels are speech-active, so the measured global SNR is about 5 dB below the label (label $+5 \approx 0$ dB, $0 \approx -5$ dB, $-5 \approx -10$ dB), and their test noise is Babble and Cafe only. Their inputs are therefore close to unreadable (noisy WER 0.57–1.00), which is the regime where the acoustic evidence carries the least information. To place the TCD-TIMIT speech on the same footing as the

other sets, the *TCD-NTCD remixed* rows mix the same 882 test utterances by the primary-set recipe: the noise is recovered from the NTCD-TIMIT release (noisy minus clean) for a randomly drawn test speaker and file, tiled and cropped to the utterance, and scaled to the labeled global SNR, one of the five NTCD noise types per clip (seed 42, the same noise segment across the three SNRs). At the labeled SNR the inputs are readable (noisy WER 0.21–0.63), the enhancer harms 27–35% of clips, and the two wrappers reduce WER below the input at every SNR (AVEC$^+$: $-0.056$, $-0.096$, $-0.102$ at $+5$, 0, $-5$ dB) while keeping PESQ and SI-SDR at always-enhance's level; ESTOI is below always-enhance at $+5$ dB and at or above it at 0 and $-5$ dB. The absolute WERs remain well above those of the LRS3 sets at the same SNR because the TIMIT sentences are short (8 words) and phonetically balanced rather than predictable, so the recognizer's language prior recovers little once the masking noise is strong: at $-5$ dB the noisy-input WER is 0.04 under Car noise on both corpora but 0.88–0.92 under White and Babble on TCD-TIMIT against 0.32–0.62 on LRS3; under the stationary DEMAND noises the two corpora are equally readable (Table XV). The other sets are mixed at the labeled global SNR (the LRS3 sets and the two 0 dB remixes by the primary-set recipe; TCD-DEMAND by the enhancer authors' release).

Table XVI: Independent-recognizer and perceptual-quality checks on the condition-transfer sets. Panel (a) re-scores all waveforms with Parakeet-TDT 0.6B v2: the noisy column reports WER, and enhanced-policy entries report WER / input-relative harm (%). Markers compare WER with the noisy input using paired clip bootstraps, except TCD–DEMAND, which uses an utterance-cluster bootstrap. Panel (b) reports DNSMOS (OVRL) for the same sets; markers on AVEC and AVEC$^+$ compare DNSMOS with always-enhance. Here $^*$ is favorable, $^\dagger$ unfavorable, and $^\circ$ indicates an interval containing zero.

**(a) Parakeet-TDT**

| Set ($n$) | Noisy WER | Always WER / Harm | AVEC WER / Harm | AVEC$^+$ WER / Harm |
|---|---|---|---|---|
| LRS3-DEMAND $-5$ (657) | .293 | .352$^\dagger$ / 27.7 | .287$^\circ$ / 19.9 | .265$^*$ / 17.2 |
| LRS3-DEMAND 0 (657) | .100 | .148$^\dagger$ / 23.1 | .089$^\circ$ / 13.4 | .088$^\circ$ / 12.0 |
| LRS3-DEMAND $+5$ (656) | .040 | .060$^\dagger$ / 11.0 | .044$^\circ$ / 5.9 | .036$^\circ$ / 4.9 |
| TCD-DEMAND $-5$ (450) | .204 | .228$^\dagger$ / 25.8 | .192$^\circ$ / 20.0 | .168$^*$ / 15.3 |
| TCD-DEMAND 0 (450) | .097 | .111$^\circ$ / 15.8 | .096$^\circ$ / 12.2 | .094$^\circ$ / 12.0 |
| TCD-DEMAND $+5$ (450) | .062 | .064$^\circ$ / 10.2 | .063$^\circ$ / 10.4 | .061$^\circ$ / 9.1 |
| TCD-NTCD $-5$ (880) | 1.003 | 1.000$^\circ$ / 16.9 | .977$^*$ / 13.8 | .973$^*$ / 11.5 |
| TCD-NTCD 0 (882) | .826 | .860$^\dagger$ / 36.7 | .783$^*$ / 27.4 | .768$^*$ / 25.2 |
| TCD-NTCD $+5$ (882) | .528 | .555$^\dagger$ / 40.2 | .461$^*$ / 26.4 | .448$^*$ / 25.3 |
| TCD-NTCD remix $-5$ (882) | .596 | .588$^\circ$ / 27.8 | .515$^*$ / 18.1 | .509$^*$ / 18.7 |
| TCD-NTCD remix 0 (882) | .374 | .382$^\circ$ / 28.9 | .321$^*$ / 20.6 | .305$^*$ / 18.1 |
| TCD-NTCD remix $+5$ (882) | .187 | .215$^\dagger$ / 26.9 | .170$^*$ / 17.8 | .162$^*$ / 16.2 |
| LRS3-NTCD 0 (657) | .108 | .218$^\dagger$ / 36.2 | .108$^\circ$ / 15.8 | .099$^\circ$ / 12.8 |
| LRS3-NTCD $+5$ (664) | .040 | .080$^\dagger$ / 17.2 | .049$^\circ$ / 8.6 | .044$^\circ$ / 7.5 |

**(b) DNSMOS (OVRL)**

| Set ($n$) | Noisy | Always enhance | AVEC | AVEC$^+$ |
|---|---|---|---|---|
| LRS3-DEMAND $-5$ (657) | 2.70 | 3.56 | 3.53$^\dagger$ | 3.53$^\dagger$ |
| LRS3-DEMAND 0 (657) | 2.91 | 3.62 | 3.59$^\dagger$ | 3.60$^\dagger$ |
| LRS3-DEMAND $+5$ (656) | 3.13 | 3.64 | 3.62$^\dagger$ | 3.62$^\dagger$ |
| TCD-DEMAND $-5$ (450) | 3.12 | 3.49 | 3.48$^\circ$ | 3.48$^\circ$ |
| TCD-DEMAND 0 (450) | 3.26 | 3.58 | 3.58$^\circ$ | 3.59$^\circ$ |
| TCD-DEMAND $+5$ (450) | 3.38 | 3.65 | 3.65$^\circ$ | 3.65$^\circ$ |
| TCD-NTCD $-5$ (880) | 2.47 | 3.40 | 3.32$^\dagger$ | 3.31$^\dagger$ |
| TCD-NTCD 0 (882) | 2.60 | 3.49 | 3.43$^\dagger$ | 3.44$^\dagger$ |
| TCD-NTCD $+5$ (882) | 2.76 | 3.53 | 3.50$^\dagger$ | 3.51$^\dagger$ |
| TCD-NTCD remix $-5$ (882) | 2.72 | 3.60 | 3.55$^\dagger$ | 3.56$^\dagger$ |
| TCD-NTCD remix 0 (882) | 2.91 | 3.66 | 3.63$^\dagger$ | 3.64$^\dagger$ |
| TCD-NTCD remix $+5$ (882) | 3.10 | 3.67 | 3.66$^\dagger$ | 3.67$^\circ$ |
| LRS3-NTCD 0 (657) | 2.66 | 3.56 | 3.50$^\dagger$ | 3.50$^\dagger$ |
| LRS3-NTCD $+5$ (664) | 2.87 | 3.60 | 3.55$^\dagger$ | 3.56$^\dagger$ |

**The transfer sets under Parakeet and DNSMOS.** Under Parakeet-TDT, AVEC$^+$ has lower WER than always-enhance on all fourteen transfer sets (Table XVI), with input-relative gains concentrated in the harder conditions and smaller margins than Whisper's. On mild TCD-DEMAND sets, always-enhance does not significantly worsen input WER. DNSMOS does not mirror the content gains: wrapper scores are generally similar to or below always-enhance, with decreases up to $0.09$; interval markers are reported in the table.

**The enhancer's own training corpus (matched domain).** The released AV-UDiffSE+ checkpoint is an unsupervised model whose clean-speech prior was trained on TCD-TIMIT, and its authors evaluate it on their own TCD-TIMIT+DEMAND mixtures; the primary LRS3–NTCD condition differs in both speech corpus and noise family. To assess this mismatch, we rebuild the enhancer's own test condition: the $90$ TCD-TIMIT test utterances of the $9$ test speakers, each mixed with all five DEMAND noise types of the authors' release at $-5$ dB ($n$=450 clips). Nothing is retuned: the six candidates are regenerated on this corpus with the same guidance settings and $K$=8, and $\tau$ and $\lambda_{\mathrm{aud}}$ are the LRS3-frozen values. Because each utterance appears five times, every interval is a cluster bootstrap over utterances ($90$ clusters, $10{,}000$ resamples).

**Harm persists in the matched condition.** On its own speech corpus and its authors' noise family the enhancer still returns $26.0\%$ of clips worse than their input ($[.22, .31]$) while raising DNSMOS by $+0.36$ and every reference metric by a wide margin (PESQ $1.36 \to 1.94$, SI-SDR $-5.0 \to +9.8$ dB). The cleaner-but-wrong regime thus exists in-domain. The rate is lower than the $44.6\%$ of LRS3-NTCD and varies across the tested noise types (harm $11\%$ under bus noise, $16\%$ office, $29\%$ square, $34\%$ metro, $40\%$ traffic).

**The average transfers.** Read once with the LRS3-frozen parameters (Table XV), AVEC$^+$ returns WER $0.180$ against the input's $0.210$ and always-enhance's $0.240$, harm $15.1\%$ against $26.0\%$, PESQ and SI-SDR above always-enhance ($+0.017$, $+0.09$ dB), and ESTOI $0.002$ below it; AVEC reaches $0.198$ and $20.2\%$, with PESQ and SI-SDR near always-enhance and lower ESTOI. Hard selection over the six candidates reaches WER $0.176$ and harm $10.7\%$ but loses the PESQ and SI-SDR margins (intervals include zero) and $0.007$ ESTOI, as on the primary set.

### F.2 Non-diffusion backbones

**A deterministic enhancer: what the rule can and cannot add.** PASE (Rong et al., 2026b) resynthesizes audio from WavLM-derived phonological features through a vocoder and is designed to reduce linguistic hallucination. Its inference is deterministic, admitting neither guidance nor posterior samples. On LRS3–NTCD $-5$ dB ($n$=657), PASE lowers WER to $0.294$ ($\Delta = -0.042$ $[-.065, -.018]$ against the input), yet harms $16.4\%$ $[13.7, 19.3]$ of clips. Its PESQ, ESTOI, and DNSMOS are $1.608$, $0.685$, and $3.70$. We exclude SI-SDR because vocoder resynthesis is not sample-aligned: the measured $-11.9$ dB is not explained by a constant delay (median lag $1.5$ samples).

The candidates are PASE's output and the mean of eight input-perturbed outputs. The mean is worse on WER, PESQ, and ESTOI ($0.302$, $1.526$, $0.650$). With frozen parameters, AVEC reaches WER $0.278$ ($\Delta = -0.016$ $[-.025, -.008]$ against PASE) and harm $14.0\%$, sacrificing $0.013$ $[.009, .018]$ PESQ and $0.009$ $[.007, .011]$ ESTOI; hard selection reaches WER $0.269$ and harm $13.4\%$ (Table 3). Untuned perturbations of $-15$ and $-25$ dB instead of $-20$ yield similar trade-offs: weighted WER $0.283$ ($\Delta = -0.011$ $[-.020, -.003]$) and $0.286$ ($\Delta = -0.008$ $[-.015, +.000]$), with PESQ losses $0.020$ and $0.012$. Thus input-relative harm also occurs in this deterministic enhancer, and evidence weighting reduces it without a quality gain for the tested pool. We do not claim superiority over PASE: it has higher PESQ and ESTOI than the wrapped primary system on this benchmark.

**An audiovisual wrapper on an audio-only backbone: which ingredient does what.** Wrapping audio-only PASE adds video and ASR/VSR computation, so Table 3 is not an equal-input comparison. To isolate these contributions, Table XVII fixes the audio-only perturbation candidates and compares uniform, audio-only ($w_{\mathrm{lip}}$=0), audio-visual, and lip-only ($\lambda_{\mathrm{aud}}$=0) weighting. Uniform averaging does not improve WER over PASE (both intervals include zero) and reduces quality. Audio-only weighting improves WER over uniform by $0.010$/$0.018$ for the two-/nine-candidate pools (intervals exclude zero). Adding video reduces WER further at $\tau$=0.10 by $0.008$ $[.003, .014]$/$0.014$ $[.006, .023]$, and by $0.011$/$0.028$ under hard selection. Joint weighting has the lowest WER point estimate at every tested operating point, consistent with complementary evidence sources (Table VI). PESQ losses

remain similar across weightings within each pool (0.013–0.016 for two candidates; 0.07–0.08 for nine), linking the observed quality cost primarily to candidate construction.

Table XVII: Witness ablation on PASE for LRS3–NTCD at $-5$ dB ($n$=657), with each candidate pool held fixed and $\tau$=0.10, $\lambda_{\mathrm{aud}}$=0.5 unless the weighting rule removes a term. The 2-candidate pool contains the deployed output and the mean of eight input-perturbed outputs; the 9-candidate pool additionally exposes the seven individual perturbed outputs. Because PASE is audio-only, visual information enters only through the VSR evidence used by the weighting rule. We compare uniform, acoustic-only, joint audio-visual, and visual-only weighting, each with weighted averaging and, where shown, hard selection. Markers denote paired 95% comparisons with always-enhance (WER 0.294, harm 16.4%, PESQ 1.608, ESTOI 0.685); the last column compares WER with acoustic-only weighting at the same operating point. SI-SDR is omitted because PASE uses vocoder resynthesis.

| Candidates | Weighting | WER ↓ | Harm ↓ | PESQ ↑ | ESTOI ↑ | WER vs. audio-only |
|---|---|---|---|---|---|---|
| 2 candidates | uniform | 0.295$^{\circ}$ | 0.167 | 1.594$^{\dagger}$ | 0.676$^{\dagger}$ | – |
| | audio-only, $\tau$=.10 | 0.286$^{\circ}$ | 0.151 | 1.594$^{\dagger}$ | 0.675$^{\dagger}$ | – |
| | audio-only, hard sel. | 0.280$^{*}$ | 0.135 | 1.592$^{\dagger}$ | 0.674$^{\dagger}$ | – |
| | audio-visual, $\tau$=.10 | 0.278$^{*}$ | 0.140 | 1.594$^{\dagger}$ | 0.676$^{\dagger}$ | $^{*}$ |
| | audio-visual, hard sel. | **0.269**$^{*}$ | **0.134** | 1.592$^{\dagger}$ | 0.674$^{\dagger}$ | $^{*}$ |
| | lip-only, $\tau$=.10 | 0.285$^{*}$ | 0.157 | 1.594$^{\dagger}$ | 0.676$^{\dagger}$ | $^{\circ}$ |
| | lip-only, hard sel. | 0.278$^{*}$ | 0.149 | 1.592$^{\dagger}$ | 0.675$^{\dagger}$ | $^{\circ}$ |
| 9 candidates | uniform | 0.302$^{\circ}$ | 0.190 | 1.526$^{\dagger}$ | 0.650$^{\dagger}$ | – |
| | audio-only, $\tau$=.10 | 0.284$^{\circ}$ | 0.161 | 1.530$^{\dagger}$ | 0.652$^{\dagger}$ | – |
| | audio-only, hard sel. | 0.282$^{\circ}$ | 0.139 | 1.532$^{\dagger}$ | 0.654$^{\dagger}$ | – |
| | audio-visual, $\tau$=.10 | 0.270$^{*}$ | 0.145 | 1.533$^{\dagger}$ | 0.656$^{\dagger}$ | $^{*}$ |
| | audio-visual, hard sel. | **0.254**$^{*}$ | **0.108** | 1.534$^{\dagger}$ | 0.658$^{\dagger}$ | $^{*}$ |
| | lip-only, $\tau$=.10 | 0.279$^{*}$ | 0.163 | 1.531$^{\dagger}$ | 0.655$^{\dagger}$ | $^{\circ}$ |
| | lip-only, hard sel. | 0.267$^{*}$ | 0.143 | 1.537$^{\dagger}$ | 0.662$^{\dagger}$ | $^{*}$ |

**Enlarging the pool where there are no posterior samples.** Table XVIII expands both non-sampling backbones to nine candidates: the deployed output, seven perturbed outputs, and their mean, with frozen $\tau$ and $\lambda_{\mathrm{aud}}$. Hard selection reduces FlowAVSE WER/harm from 0.256/15.4% to 0.218/9.6% and PASE from 0.294/16.4% to 0.254/10.8% (paired intervals exclude zero). These content gains cost about 0.07 PESQ and 0.01–0.03 ESTOI against always-enhance at either operating point: perturbations offer useful clip-specific alternatives but are worse on average. The uniform nine-candidate average equals the perturbed mean. These results establish a trade-off for this pool; alternative candidate constructions were not tested.

Table XVIII: Perturbation-pool ablation for FlowAVSE and PASE, which do not expose posterior samples, on the primary test set ($n$=657; frozen $\tau$=0.10, $\lambda_{\mathrm{aud}}$=0.5). The 2-candidate pool contains the deployed output and the mean of eight outputs obtained from the original input and seven perturbed copies; the 9-candidate pool additionally exposes those seven perturbed outputs individually. ΔWER is paired against always-enhance, and the quality deltas are also relative to always-enhance. Rows with "frozen $\tau$ / hard sel." report the two operating points in that order. SI-SDR is omitted for PASE because vocoder resynthesis limits waveform-based interpretation.

| | | Content | | | Quality change vs. always-enhance | | |
|---|---|---|---|---|---|---|---|
| System | Policy | WER ↓ | Harm ↓ | ΔWER | ΔPESQ | ΔESTOI | ΔSI-SDR |
| FlowAVSE | always enhance | 0.256 | 0.154 | – | – | – | – |
| | perturbed mean (= uniform over 9) | 0.276 | 0.193 | +0.019 [+.005, +.034] | −0.076 | −0.012 | −0.74 |
| | AVEC, 2 cand. (frozen $\tau$ / hard sel.) | 0.242 / 0.238 | 0.140 / 0.132 | −0.014 [−.024, −.004] | −0.012 | −0.003 | −0.34 |
| | AVEC, 9 cand., frozen $\tau$ | 0.228 | 0.116 | −0.028 [−.042, −.014] | −0.072 | −0.011 | −0.71 |
| | AVEC, 9 cand., hard selection | **0.218** | **0.096** | −0.039 [−.054, −.024] | −0.074 | −0.011 | −0.71 |
| PASE | always enhance | 0.294 | 0.164 | – | – | – | – |
| | perturbed mean (= uniform over 9) | 0.302 | 0.190 | +0.008 [−.005, +.022] | −0.081 | −0.034 | – |
| | AVEC, 2 cand. (frozen $\tau$ / hard sel.) | 0.278 / 0.269 | 0.140 / 0.134 | −0.016 [−.025, −.007] | −0.013 | −0.009 | – |
| | AVEC, 9 cand., frozen $\tau$ | 0.270 | 0.145 | −0.024 [−.036, −.012] | −0.075 | −0.029 | – |
| | AVEC, 9 cand., hard selection | **0.254** | **0.108** | −0.040 [−.053, −.028] | −0.073 | −0.026 | – |

# G DIAGNOSTICS BEHIND THE ANALYSIS OF SECTION 3

This section gives further analyses behind Section 3: the sample-disagreement diagnostic, the counterfactual-video probe, the speech-corpus by noise-family view, and the sensitivity of the harm event to the evaluator. Evidence controls for the wrapper itself are in Appendix D.4.

**Sample disagreement as a diagnostic.** For each input, DIVERSITY is the mean pairwise viseme distance among transcripts of $K = 8$ stochastic restorations. This distance-based diagnostic differs from the distinct-transcript count shown in Figure 2(c). It correlates with the WER of a separately generated output (Spearman $\rho = 0.79$ on LRS3–NTCD and $0.86$ on LRS3–DEMAND), but does not identify which candidate is correct. An audio-only transcript-change score flags both improvements and degradations (harm AUC $0.779$). For detecting the input-relative harm event, LIP, DIVERSITY, and output DNSMOS have AUCs of $0.75$, $0.76$, and $0.51$, respectively. LIP denotes the visual consistency score defined in Equation (3). These diagnostic comparisons are separate from the evaluation of the combination rule, which does not use DIVERSITY.

## G.1 COUNTERFACTUAL-VIDEO PROBE: CONSTRUCTION AND FULL TABLES

**Construction.** For each clip we hold the $K{=}8$ posterior samples drawn under the true mouth video (Section 3) and redraw $K{=}8$ samples under a broken one: *shift* rolls the mouth frames circularly by 12 frames ($0.48$ s, which destroys synchronization but preserves every marginal statistic), *mismatch* substitutes the mouth video of the next different-speaker clip in a fixed ordering. The grounding score is an energy-style set divergence between the two transcript sets, $G = 2\,D(\mathcal{T},\mathcal{C}) - D(\mathcal{T},\mathcal{T}) - D(\mathcal{C},\mathcal{C})$, where $D$ averages pairwise viseme-level distances; $G \approx 0$ means breaking the video left the output content distribution unchanged. The probe shows that per-clip video reliance predicts per-clip harm.

**Thresholds.** "Confident" = below the median DIVERSITY over all $657$ clips; "video-sensitive" = above the median $G$ *within* the confident half. Both medians were fixed after seeing the data (post hoc), so this probe is reported as an in-sample audit rather than a confirmatory test.

Table XIX: The probe's $2{\times}2$ on the full set ($n{=}657$; cells = harmed/total (rate)). Fisher exact tests compare the two video-sensitivity columns within the confident row. $\rho(G, \text{DIVERSITY}) = 0.30$ (mismatch) / $0.22$ (shift); $\rho(G, \text{LIP}) = 0.26$; within the confident half, AUC for harm: $G$ $0.84$, LIP $0.77$ (mismatch counterfactual).

| Counterfactual | Evidence regime | Video-insensitive | Video-sensitive |
|---|---|---|---|
| mismatch | confident (low DIVERSITY) | 5/165 (0.030) | 81/164 (0.494) |
| | uncertain (high DIVERSITY) | 207/328 (0.631) | |
| shift | confident (low DIVERSITY) | 14/182 (0.077) | 72/147 (0.490) |
| | uncertain (high DIVERSITY) | 207/328 (0.631) | |

**Catch rates and the lower-harm cell.** Of the $86$ confident-but-harmed clips, the video-sensitivity flag catches $81$ ($94.2\%$) under mismatch and $72$ ($83.7\%$) under shift; LIP above its own within-confident median catches $72$ ($83.7\%$). The confident, video-insensitive cell covers $25\%$ of clips at harm $3.0\%$ $[0.6, 6.1]$ (mismatch; $7.7\%$ under shift), against $44.6\%$ for always-enhancing the corpus. The lip analogue of that cell is harmed $8.5\%$: the probe buys a threefold cleaner audit cell than the free signal, at the price of $2K$ extra generations per clip.

## G.2 SPEECH-CORPUS BY NOISE-FAMILY ANALYSIS

At $-5$ dB, the exploratory corpus–noise $2{\times}2$ (Figure I) shows similar degradation for TCD+NTCD and LRS3+NTCD ($\Delta$WER $+0.173$/$+0.169$); the fully matched TCD+DEMAND condition also has $26\%$ harm (Appendix F.1). On the matched subsets, NTCD degradation is 2.6–3.6$\times$ DEMAND's; the full LRS3 sets retain this ordering ($+0.169$ vs. $+0.090$, $1.9\times$). These two noise families support a condition-specific difference, not a general law about speech-like noise. DNSMOS increases in all four cells. The disagreement–error correlation replicates on DEMAND ($\rho = 0.86$), where the proposed average meets all four criteria (Table XV).

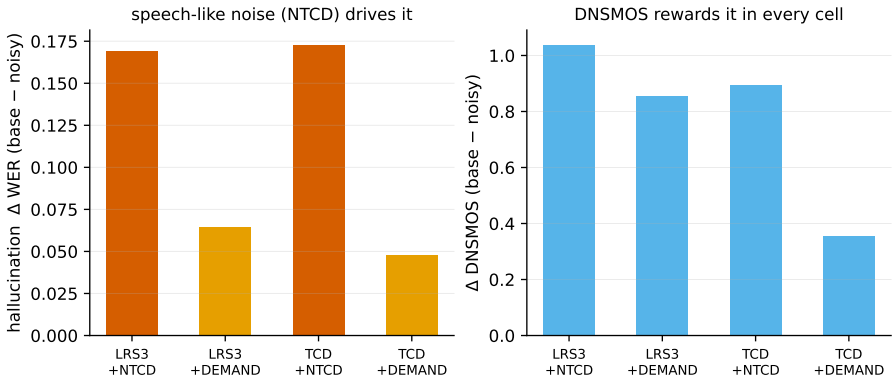



Figure I: 2×2 (speech corpus × noise type, −5 dB; matched exploratory subsets, LRS3+NTCD $n$=657, LRS3+DEMAND $n$=240, TCD cells $n$=90 each, one mixing pipeline throughout). ASR-measured degradation is larger with NTCD than DEMAND on these subsets under both speech corpora; DNSMOS increases with enhancement in all four cells.

### G.3 Evaluator sensitivity

**The coupling concern.** Both DIVERSITY and LIP use Whisper output transcripts, as does the WER-based harm event, potentially inflating their association through shared evaluator behavior. The phoneme- and semantic-level analyses below are therefore transcript-loss sensitivity checks rather than independent waveform evaluations. Independent Parakeet re-scoring is reported for the final policies in Appendix D.5.

**Phoneme-level scoring.** G2P-derived PER increases from $0.274$ to $0.398$ after enhancement ($\Delta$ interval $[+.101, +.149]$; Spearman correlation with WER $0.968$). Derived from Whisper text, this checks transcript-loss sensitivity, not the waveform's independent phonetic accuracy.

**Semantic similarity.** Mean-pooled `all-mpnet-base-v2` embeddings show lower reference–transcript similarity after enhancement ($0.699 \to 0.572$; $\Delta$ interval $[-.151, -.102]$), especially on the 293 clips with increased WER ($0.706 \to 0.345$). Similarity correlates with WER at Spearman $-0.916$, suggesting semantic as well as lexical changes. This remains a Whisper-transcript analysis, not an independent listening test.

## H Limitations and scope

**Evaluator scope.** Harm is Whisper-defined, and Whisper also supplies the acoustic evidence and the audio-guided target, so the rule steers toward what Whisper reads and Whisper scores the result. Parakeet-TDT independently re-scores the final policies (Appendix D.5) and reads the input comparably to Whisper (WER $0.339$). It retains the input-relative result (AVEC$^+$ $-0.032$ $[-.056, -.008]$ below the input; harm $41.7\% \to 22.8\%$), although with a smaller margin than Whisper's $-0.056$. The listener study (Appendix E.2) is an independent supporting check; it is not used to fit the method and is limited to its sampled language, noise condition, clips, and listeners.

**Two operating points.** Under Whisper, the hard-selection point of AVEC$^+$ ($\tau \to 0$) has the lowest harm ($16.7\%$) of any primary-system policy we report that returns only enhancer outputs (Table VI), while the weighted-average point trades $1.4$ points of harm and $0.012$ WER $[.001, .024]$ for $+0.008$ PESQ, $+0.006$ ESTOI, and $+0.11$ dB. The preferred operating point differs under Parakeet; averaging nevertheless retains the independent-recognizer gain over always-enhance and adds reference quality. On FlowAVSE and PASE, hard selection gives lower Whisper WER than weighted averaging for the evaluated pools.

**Quality, cost, and access.** The primary AVEC$^+$ gains over always-enhance are modest ($+0.015$ PESQ, $+0.006$ ESTOI, and $+0.23$ dB SI-SDR; DNSMOS $0.06$ lower), and some reference metrics decrease at milder SNRs and on the TCD conditions (Table XV). AVEC uses about 37 s per clip on

one A6000 and AVEC$^{+}$ about $105$ s (Appendix D.3). Guided candidates require access to an iterative sampler's intermediate estimates and a frozen phoneme recognizer.

**Scope.** The method requires mouth video, and the evaluation covers English speech, synthetic mixtures, and frontal video. All evaluation sets informed earlier diagnostic analyses; fixed combination parameters limit subsequent tuning but do not make those sets independent of method development (Appendix B). The wrapper never returns the input, so it cannot decide not to enhance clean speech (Appendix D.4). It only limits the enhancer's damage once enhancement is run.