How Much to Trust the Eyes? Reliability-Guided Inference for Diffusion-Based Audio-Visual Speech Enhancement

Anonymous authors
Paper under double-blind review

Abstract

Diffusion-based audio-visual speech enhancement (AVSE) iteratively recovers speech from noisy audio conditioned on temporally aligned visual cues. Its real-world deployment may confront distribution shifts, such as visual degradation and audio-visual mismatch, which misguide the reverse diffusion process. Our analysis discloses that visual quality alone does not necessarily determine whether visual conditioning remains useful or misguides enhancement since its effect is coupled with both the pretrained model and the acoustic input. Motivated by these observations, we propose Reliability-Guided Inference (RGI), which takes the departure of both internal visual representations and model responses from those of matched audio-video pairs as evidence of distribution shift inside the pretrained model and converts this evidence into a reliability score that regulates visual conditioning during reverse diffusion. Specifically, at each reverse step, RGI weights the difference between audio-visual and audio-only predictions according to the reliability score, without updating the pretrained models. To retain visual cues that can remain useful at low input signal-to-noise ratio, RGI imposes a lower bound on the conditioning coefficient, the conditioning floor. A diffusion-step schedule further modulates the retained visual contribution along the sampling trajectory. Experiments across four corpora and three frozen diffusion AVSE backbones show that RGI improves enhancement under a range of distribution shifts on the corpora with clean references, with gains that depend on the backbone and corpus. Recordings collected from the Ameca humanoid robot further show that RGI can be deployed for robot interaction, with demonstrations available at `https://anonymous.4open.science/w/ameca-demo/`.

1 Introduction

Audio-visual speech enhancement (AVSE) uses visual speech cues as auxiliary conditioning to recover target speech from noisy audio (Michelsanti et al., 2021). Diffusion-based AVSE (Richter et al., 2023a; Ayilo et al., 2025) applies this conditioning to guide iterative denoising (Ho et al., 2020; Song et al., 2021). Pretrained models typically learn to use this conditioning from matched audio-video pairs (Ephrat et al., 2018; Ayilo et al., 2025), where the visual stream is available, temporally aligned with the audio, and associated with the target speaker. In real-world deployment, however, these conditions may no longer hold, and audio-visual speech recognition on a social robot has been found feasible mainly in controlled and non-interactive settings (Goetzee et al., 2024). For example, robot or speaker motion can change the viewpoint, and visual speech input varies with viewpoint, occlusion, and motion blur in real scenes (Su et al., 2026), while audio-visual asynchrony (Fu et al., 2024) and incorrect speaker association can disrupt audio-visual correspondence. Such conditions introduce distribution shifts relative to the matched audio-video pairs used in training, altering either the appearance of the visual input or its correspondence with the target speech. Fixed audio-visual inference, which applies the visual condition at full strength, continues to use these degraded or mismatched visual cues as conditioning without explicitly accounting for their reduced reliability, potentially steering the reverse diffusion process away from the target speech.

Existing approaches address unreliable visual information through stream confidence weighting (Potamianos & Neti, 2000; Abdelaziz et al., 2015), frame-level reliability scoring under visual

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.

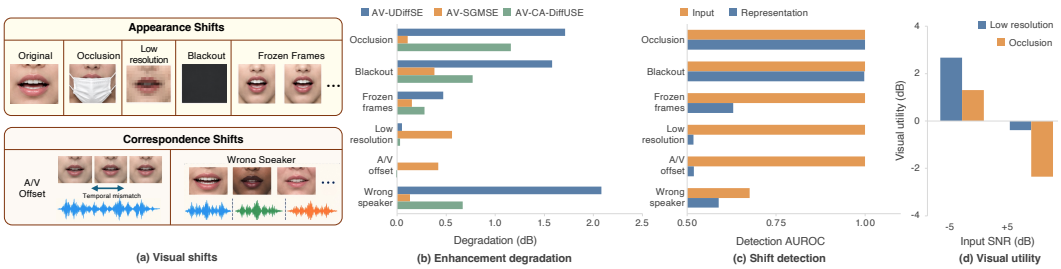


Figure 1: The effect of distribution shifts on frozen AVSE is not determined by the video alone. (a) Appearance and correspondence shifts. (b) Enhancement degradation relative to original aligned video, hereafter original video, across three backbones. (c) Shift detectability, the area under the receiver operating characteristic curve (AUROC), using input frames or internal visual representations. (d) Visual utility relative to audio-only inference at two input signal-to-noise ratios (SNRs).

corruption (Hong et al., 2023), corruption-robust audio-visual representation learning (Kim et al., 2025), visual-quality gating (Su et al., 2026), and dynamic modality selection at inference time (Du et al., 2026). These approaches provide mechanisms for handling degraded or unavailable modalities, but the benefit of visual conditioning also depends on the pretrained enhancer and the acoustic input. Visually degraded cues may retain useful speech information, while visually intact cues may misguide enhancement when they do not correspond to the target speech, much as incongruent lip movements alter perceived speech in human listeners (McGurk & MacDonald, 1976).

Our analysis in Fig. 1 examines the effects of the appearance and correspondence shifts illustrated in Fig. 1(a). Across three frozen backbones, Fig. 1(b) shows that the same perturbation produces markedly different enhancement degradation relative to original video. Comparing these results with shift detectability in Fig. 1(c) reveals a further distinction. On AV-UDiffSE, low-resolution video is readily distinguished from original video using input frames yet causes little degradation, whereas wrong-speaker video is less distinguishable despite causing substantially greater degradation. As shown in Fig. 1(d), low-resolution and occluded video also provide positive visual utility relative to audio-only inference in the group with lower input SNR but negative utility in the higher-SNR group. A central challenge is therefore to determine whether visual conditioning should be retained or suppressed, and to regulate how strongly it influences speech recovery.

Our analysis further suggests that evidence for this assessment is available within the frozen model, as internal visual representations on AV-UDiffSE are most distinguishable from original video under the appearance shifts that cause the largest degradation (Fig. 1(c)). We address this challenge with Reliability-Guided Inference (RGI), which adapts visual conditioning during reverse diffusion without updating the pretrained models. At each reverse step, audio-visual and audio-only predictions evaluated at the same generative state define a visual increment relative to the audio-only prediction. RGI combines deviations in internal visual representations and relative increment magnitudes from matched reference statistics to estimate reliability and weight this increment. Deviations in the internal visual representation provide evidence of appearance shift, and deviations in the relative increment magnitude are intended as indirect evidence of correspondence shift, which can leave the video itself intact. A conditioning floor limits excessive suppression of potentially useful visual cues, while a diffusion-step schedule adjusts their contribution along the sampling trajectory.

Our contributions are summarized as follows:

- We analyze how the effect of visual conditioning under distribution shift depends on the frozen AVSE model and the acoustic input, showing on AV-UDiffSE that shift detectability from input frames does not reflect its impact on enhancement.
- Building on these findings, we propose RGI, which combines a representation score that provides evidence of appearance shift with a model-response score intended as indirect evidence of correspondence shift, and regulates visual conditioning accordingly during reverse diffusion with a conditioning floor and a diffusion-step schedule, without updating the pretrained models.
- We evaluate RGI across four corpora and three frozen diffusion AVSE backbones, showing backbone-dependent improvements under distribution shift on the corpora with clean ref-

erences and testing it on held-out shifts. Recordings collected from the Ameca humanoid robot further show that RGI can be deployed for robot interaction.

2 Visual Conditioning under Distribution Shift

Pretrained AVSE models learn visual conditioning from matched pairs of the noisy speech observation $\mathbf{y}$ and the aligned video of the target speaker. Under distribution shift, the visual condition $\mathbf{v}$ received at deployment departs from this video, so that $(\mathbf{y}, \mathbf{v})$ no longer follows the distribution of matched pairs. An appearance shift alters the video itself, as with occlusion or downsampling, whereas a correspondence shift breaks the relation between $\mathbf{v}$ and $\mathbf{y}$, as with video of another speaker or a temporal offset, while the marginal distribution of $\mathbf{v}$ can remain unchanged. Both shifts can alter the benefit of visual conditioning, with effects that depend on three factors, namely the type of shift, the frozen model, and the acoustic input. To examine these factors, we characterize visual conditioning relative to the audio-only prediction, then investigate evidence in the frozen model for assessing its reliability under each type of shift and the acoustic dependence of its utility.

2.1 Latent Validity of the Visual Increment

To characterize visual conditioning relative to the audio-only prediction, we compare audio-visual and audio-only predictions at the same generative state and diffusion time. Let $\mathbf{x}_t$ denote the generative state at reverse step $t \in \{N, \dots, 1\}$, with corresponding diffusion time τ_t. Evaluating the pretrained score models at $(\mathbf{x}_t, \tau_t)$ yields the audio-visual and audio-only score predictions, denoted by $\mathbf{u}_{\mathrm{av},t}$ and $\mathbf{u}_{\mathrm{a},t}$, respectively. The audio-only prediction is obtained either from a separate pretrained model or by evaluating the audio-visual model with a prescribed null visual input as in classifier-free guidance (Ho & Salimans, 2022), depending on the backbone. Their difference, $\Delta\mathbf{u}_t = \mathbf{u}_{\mathrm{av},t} - \mathbf{u}_{\mathrm{a},t}$, defines the visual increment relative to the chosen audio-only prediction.

The visual increment enters the reverse diffusion process through the score prediction used by the sampler. At reverse step t, an Euler predictor (Song et al., 2021; Richter et al., 2023b) updates the current state $\mathbf{x}_t$ according to

$$\widetilde{\mathbf{x}}_{t-1} = \mathbf{x}_t - \mathbf{f}_t \Delta\tau_t + b_t^2 \Delta\tau_t \, \mathbf{u}_{\mathrm{av},t} + b_t \sqrt{\Delta\tau_t} \, \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{f}_t$ and b_t are the forward diffusion drift and diffusion coefficient evaluated at the current state and diffusion time, $\Delta\tau_t > 0$ is the reverse integration step size with $\tau_{t-1} = \tau_t - \Delta\tau_t$, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise. The predicted state $\widetilde{\mathbf{x}}_{t-1}$ is obtained before any separate acoustic observation update, such as the likelihood step of posterior sampling (Chung et al., 2023; Nortier et al., 2024). Substituting $\mathbf{u}_{\mathrm{av},t} = \mathbf{u}_{\mathrm{a},t} + \Delta\mathbf{u}_t$ into Eq. (1) shows that, with the state, diffusion time, and noise realization held fixed, the visual increment contributes $b_t^2 \, \Delta\tau_t \, \Delta\mathbf{u}_t$ to the predictor update relative to the audio-only prediction. Fixed audio-visual inference therefore includes the full increment with a coefficient of one. Under distribution shift, however, retaining the full increment may no longer be beneficial, motivating an assessment of its validity for the current input.

We represent the validity of the visual increment for the current audio-video pair under the frozen model by a latent validity $\beta \in [0, 1]$, the unobserved weight, shared across reverse steps, for which the interpolated score prediction $\mathbf{u}_{\beta,t} = \mathbf{u}_{\mathrm{a},t} + \beta \, \Delta\mathbf{u}_t$ is best suited to recovering the target speech. The endpoints $\beta = 1$ and $\beta = 0$ correspond to retaining the full increment and recovering the audio-only prediction, respectively. Let $\mathcal{O}_t$ denote the information available through reverse step t, including the noisy speech, video, current generative state, and model responses. Since β is unobserved, we consider the posterior average of $\mathbf{u}_{\beta,t}$ conditional on $\mathcal{O}_t$. With $\mathbf{u}_{\mathrm{a},t}$ and $\Delta\mathbf{u}_t$ fixed given $\mathcal{O}_t$, this average is

$$\bar{\mathbf{u}}_t = \mathbb{E}[\mathbf{u}_{\beta,t} \mid \mathcal{O}_t] = \mathbf{u}_{\mathrm{a},t} + r_t \, \Delta\mathbf{u}_t, \qquad r_t = \mathbb{E}[\beta \mid \mathcal{O}_t]. \tag{2}$$

Under this formulation, the posterior mean r_t defines the reliability of the visual increment and determines its weight in the averaged score prediction $\bar{\mathbf{u}}_t$.

Audio-only inference also provides the basis for distinguishing visual utility from enhancement degradation. A visual perturbation can reduce enhancement performance relative to original video while still providing a benefit over audio-only inference. We therefore define the enhancement

degradation as the decrease in enhancement performance of fixed audio-visual inference when original video is replaced by $\mathbf{v}$, and the visual utility as the difference between the performance of fixed audio-visual inference with $\mathbf{v}$ and that of audio-only inference, both measured as in Sec. 4.1.

2.2 Evidence for Visual Reliability

Estimating the reliability r_t in Eq. (2) requires observable evidence about the latent validity of the visual increment $\Delta\mathbf{u}_t$. The model dependence in Fig. 1(b) motivates evaluating this evidence relative to the same frozen model on matched pairs. We consider two sources of evidence available during inference, namely *internal visual representations* and *model responses* to the audio-video pair.

To capture the effect of visual perturbations within the frozen model, we first examine the visual conditioning module, which maps the video $\mathbf{v}$ to the internal representation consumed by the audio-visual model. The representation reflects the visual information passed to the predictor and can respond differently to perturbations than the input frames. Evaluation on AV-UDiffSE supports this distinction. We measure the detectability of each shift by the AUROC for separating shifted from original video, applying the same Mahalanobis detector, fitted on the visual reference set of Sec. 3.1, to temporal statistics of either the input frames or the internal representation. As shown in Fig. 1(b,c), input detectability is close to one for all four appearance shifts, whereas representation detectability is highest for the two appearance shifts that cause the largest enhancement degradation. The observation supports comparing internal visual representations with their reference distribution as evidence of appearance shifts relevant to the frozen model, as in feature-space detectors of distribution shift (Lee et al., 2018; Müller & Hein, 2025). However, these deviations reflect representation changes, while visual utility also depends on the acoustic input.

The dependence on the acoustic input becomes particularly relevant under correspondence shift, where the marginal distribution of the video can be independent of the latent validity. Assuming $p(\mathbf{v} \mid \beta) = p(\mathbf{v})$, Bayes' rule gives

$$p(\beta \mid \mathbf{v}) = p(\beta), \tag{3}$$

where $p(\beta)$ denotes the prior distribution of β. Under this assumption, the posterior remains equal to the prior after observing the video or any representation computed solely from it. App. B derives this result and shows that the averaged score prediction in Eq. (2) minimizes the posterior expected squared error to the interpolated score prediction $\mathbf{u}_{\beta,t}$. Assessing validity in this setting therefore requires evidence involving the relation between visual and acoustic observations.

To incorporate such evidence, we further examine the response of the frozen model to the audio-video pair, similar to diffusion-based detectors that score inputs by statistics of their diffusion paths (Heng et al., 2024). The magnitude of $\Delta\mathbf{u}_t$ relative to the audio-only prediction $\mathbf{u}_{\mathrm{a},t}$ characterizes how strongly the audio-visual prediction departs from it. Comparing this relative magnitude with reference statistics at the same reverse step provides evidence of departures from model responses on matched pairs. The statistic is intended as indirect evidence of correspondence shift through the model response and can also respond to appearance shifts. Together, internal visual representations and model responses provide two sources of evidence for estimating reliability.

2.3 Acoustic Dependence of Visual Utility

Beyond appearance and correspondence shifts, acoustic conditions, characterized here by the input SNR, also affect visual utility. To examine this dependence, we compare mean visual utility relative to audio-only inference across two input SNR groups for the same perturbations. As shown in Fig. 1(d), low-resolution and occluded video provide positive mean utility in the lower-SNR group but negative utility in the higher-SNR group. These results indicate that visual utility varies with the input SNR even under the same frozen model, in line with the larger visual benefit at low SNR reported for human listeners (Sumby & Pollack, 1954) and for audio-visual enhancement (Richter et al., 2023a). In particular, degraded video can still improve enhancement when acoustic evidence is weak, so suppressing it on the basis of visual degradation alone risks lowering speech recovery.

The acoustic dependence is reflected in the formulation of r_t in Eq. (2), where the available evidence $\mathcal{O}_t$ includes the noisy speech observation $\mathbf{y}$, which enters the reliability estimate through the model response. It also motivates evaluating the visual increment relative to the audio-only prediction. The

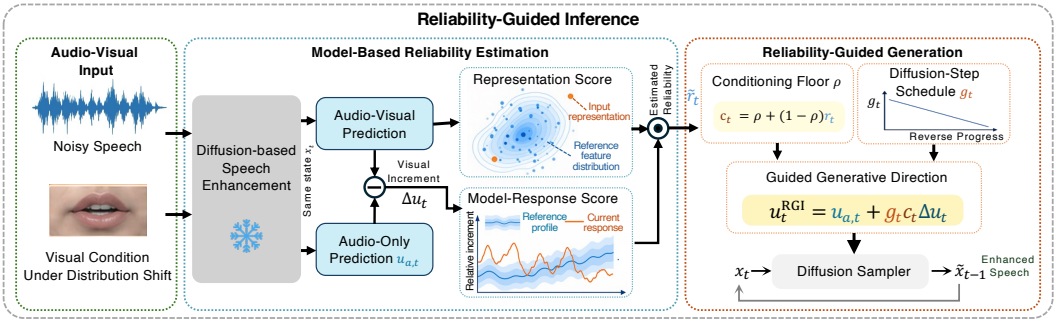


Figure 2: Overview of RGI. Audio-visual and audio-only predictions evaluated at the same generative state define the visual increment $\Delta \mathbf{u}_t$. The representation score measures deviations in internal visual features, while the model-response score accumulates deviations in the log magnitude ratio h_t, both using fixed statistics from matched reference pairs. The two scores yield a reliability estimate $\hat{r}_t$, which is transformed into a conditioning coefficient c_t with floor ρ. The diffusion-step schedule g_t further adjusts the visual contribution in the guided score prediction $\mathbf{u}_t^{\mathrm{RGI}}$.

resulting magnitude ratio provides a model-response statistic that expresses the size of the increment relative to the scale of the audio-only score.

The positive visual utility observed under weak acoustic conditions also highlights a risk of excessive suppression. In Eq. (2), the averaged score prediction approaches $\mathbf{u}_{\mathrm{a},t}$ as r_t approaches zero, reducing the contribution of visual conditioning. A reliability estimate that is too low can therefore suppress visual cues that remain useful for speech recovery. The risk of excessive suppression motivates a lower bound on the conditioning coefficient of the visual increment, called the conditioning floor, to retain a residual visual contribution. The floor serves as an empirical safeguard against excessive suppression, balancing the retention of useful cues against harmful conditioning.

3 Reliability-Guided Inference

RGI approximates the reliability r_t of the visual increment in Eq. (2) from the internal visual representation and the relative magnitude of the increment. As illustrated in Fig. 2, these observations are compared with reference statistics from matched pairs to obtain $\hat{r}_t$. Departures from these statistics serve as evidence of distribution shift inside the frozen model, so $\hat{r}_t$ decreases as either observation departs further from its reference. The estimate then controls the visual contribution through a conditioning floor and a diffusion-step schedule.

3.1 Estimation from Reference Evidence

RGI constructs two scores by comparing internal visual representations and relative visual increment magnitudes with reference statistics from the frozen model. The reference inputs consist of matched audio-video pairs, with the visual reference set drawn from the same corpus as the inputs and the trajectory reference set shared across corpora. These comparisons yield a **representation score** $z_{\mathrm{I}}(\mathbf{v})$, computed once for each video, and a **model-response score** $\bar{z}_{\mathrm{D},t}$, updated during reverse diffusion. The visual reference set $\mathcal{R}_{\mathrm{I}} = \{\mathbf{v}_i\}_{i=1}^{M_{\mathrm{I}}}$ contains M_{I} videos used to characterize the distribution of internal visual features. The trajectory reference set $\mathcal{R}_{\mathrm{D}} = \{(\mathbf{y}_i, \mathbf{v}_i)\}_{i=1}^{M_{\mathrm{D}}}$ contains M_{D} matched audio-video pairs used to characterize the relative magnitude of the visual increment at each reverse step. All reference statistics are computed and scalar parameters are set before inference, and both remain fixed during sampling.

The representation score quantifies deviations of internal visual features from their reference distribution. We construct $\boldsymbol{\phi}(\mathbf{v})$ by concatenating the temporal mean and standard deviation of the internal visual representation, then compute its Mahalanobis distance $d_{\mathrm{I}}(\mathbf{v})$ using the reference feature statistics (Lee et al., 2018). We map this distance to an adjusted empirical upper-tail proportion $q_{\mathrm{I}}(\mathbf{v})$, which measures the fraction of reference distances at least as large as $d_{\mathrm{I}}(\mathbf{v})$, with the finite-sample adjustment used for conformal p-values (Bates et al., 2023). Smaller values indicate greater deviation from the reference distribution. The representation score is defined as $z_{\mathrm{I}}(\mathbf{v}) = \operatorname{logit} q_{\mathrm{I}}(\mathbf{v})$,

where $\mathrm{logit}(w) = \log[w/(1-w)]$, so greater deviations yield lower scores. App. C details the feature statistics and the cross-fitted reference distances.

The model-response score measures deviations in the relative magnitude of the visual increment during reverse diffusion. To characterize this relative magnitude, we evaluate the score predictions $\mathbf{u}_{\mathrm{av},t}$ and $\mathbf{u}_{\mathrm{a},t}$ at the same state $\mathbf{x}_t$ and diffusion time τ_t, and express their difference relative to the audio-only prediction through the log magnitude ratio $h_t = \log[\max\{\mathrm{RMS}(\Delta\mathbf{u}_t), \delta\}/\max\{\mathrm{RMS}(\mathbf{u}_{\mathrm{a},t}), \delta\}]$. Here, RMS denotes the root mean square over all entries, and the fixed lower bound $\delta > 0$ keeps the ratio positive and finite.

To account for variation in this statistic across reverse steps, reference trajectories generated from $\mathcal{R}_{\mathrm{D}}$ provide the mean $\mu_{\mathrm{D},t}$ and standard deviation $s_{\mathrm{D},t}$ of the log magnitude ratio at each step. A fixed lower bound $s_{\min} > 0$ on $s_{\mathrm{D},t}$ prevents division by a vanishing reference scale. These statistics standardize the deviation of h_t from the matched reference behavior. We then bound each standardized deviation to $[-L, L]$ and average its absolute value over the steps evaluated so far. For N reverse steps indexed from N to 1, the model-response score is

$$\bar{z}_{\mathrm{D},t} = -\frac{1}{N-t+1}\sum_{j=t}^{N}\left|\mathrm{clip}\left(\frac{h_j - \mu_{\mathrm{D},j}}{\max\{s_{\mathrm{D},j}, s_{\min}\}}, -L, L\right)\right|, \tag{4}$$

where $L > 0$ and $\mathrm{clip}(w, -L, L) = \min\{L, \max\{-L, w\}\}$ replaces values outside the interval with the nearest boundary. The clipping limits the contribution of extreme deviations, while averaging summarizes departures from the reference behavior along the sampling trajectory. The negative sign assigns lower scores to larger departures.

Both the representation score and the model-response score decrease with greater deviation from their respective reference statistics. We therefore combine them as $z_t = z_{\mathrm{I}}(\mathbf{v}) + \bar{z}_{\mathrm{D},t}$ and map the combined score to a reliability estimate through

$$\hat{r}_t = \sigma(\kappa(z_t - \theta_{\mathrm{r}})), \qquad \kappa > 0, \tag{5}$$

where σ is the logistic function, κ controls the slope, and θ_{r} specifies the combined score at which $\hat{r}_t = 1/2$. To anchor this midpoint to the reference statistics, we set $\theta_{\mathrm{r}} = \mathrm{logit}(\alpha) + m_{\mathrm{D}}$, where $\alpha \in (0, 1)$ is a fixed lower-tail level for the tail proportion $q_{\mathrm{I}}(\mathbf{v})$, and m_{D} is the median model-response score at the final reverse step $t = 1$ over reference trajectories from $\mathcal{R}_{\mathrm{D}}$. Thus, a tail proportion $q_{\mathrm{I}}(\mathbf{v}) = \alpha$ combined with a model-response score $\bar{z}_{\mathrm{D},t} = m_{\mathrm{D}}$ yields $\hat{r}_t = 1/2$. With either score held fixed, decreasing the other reduces the reliability estimate. The resulting $\hat{r}_t$ takes the place of the reliability r_t defined in Eq. (2). It equals r_t when β takes only the endpoint values and its posterior log-odds are $\kappa(z_t - \theta_{\mathrm{r}})$, and otherwise serves as a bounded surrogate.

3.2 Guidance Along the Sampling Trajectory

To retain a residual visual contribution when the reliability estimate is low, we map $\hat{r}_t$ to the conditioning coefficient $c_t = \rho + (1-\rho)\hat{r}_t$, where $\rho \in [0, 1)$ is the conditioning floor. The mapping bounds c_t between ρ and 1 before applying the diffusion-step schedule, limiting the suppression of visual cues that may remain useful.

The conditioning coefficient c_t reflects the reliability estimate, while a prescribed diffusion-step schedule g_t further adjusts the visual contribution across sampling stages, similar to step-dependent conditioning in diffusion sampling (Sadat et al., 2024; Kynkäänniemi et al., 2024). For $N > 1$, we define the sampling progress as $\xi_t = (N-t)/(N-1)$ and set $g_t = g_{\mathrm{hi}} + (g_{\mathrm{lo}} - g_{\mathrm{hi}})\xi_t$, where $g_{\mathrm{hi}} \geq g_{\mathrm{lo}} > 0$ specify the initial and final scales. The schedule linearly reduces the scale from g_{hi} to g_{lo}, becoming constant when the two values are equal. It modulates the conditioning strength without changing the reliability estimate.

Combining the schedule with c_t yields the guided score prediction $\mathbf{u}_t^{\mathrm{RGI}} = \mathbf{u}_{\mathrm{a},t} + g_t c_t\, \Delta\mathbf{u}_t$. Substituting this prediction into Eq. (1) gives

$$\widetilde{\mathbf{x}}_{t-1} = \mathbf{x}_t + \left[b_t^2\left(\mathbf{u}_{\mathrm{a},t} + g_t c_t\, \Delta\mathbf{u}_t\right) - \mathbf{f}_t\right]\Delta\tau_t + b_t\sqrt{\Delta\tau_t}\,\boldsymbol{\epsilon}_t. \tag{6}$$

The effective weight of the visual increment, $g_t c_t$, is the conditioning strength, with lower bound $g_t\rho$. It is the analogue of the guidance scale in classifier-free guidance (Ho & Salimans, 2022). Setting $g_t = c_t = 1$ recovers fixed audio-visual inference.

RGI replaces the score input while retaining the original diffusion coefficients, integration rule, and noise convention. For samplers with predictor and corrector substeps (Song et al., 2021), the audio-visual and audio-only predictions are recomputed at each substep state, and the reliability estimate is updated using the corresponding model response. The resulting $\mathbf{u}_t^{\mathrm{RGI}}$ is then supplied to that substep. Any separate acoustic observation update retains its original form and is applied at its prescribed position in the sampling procedure.

4 EXPERIMENTS

We evaluate RGI across three frozen diffusion AVSE backbones and four corpora to assess its effectiveness under distribution shift, with a no-reference check on naturally recorded speech.

4.1 EXPERIMENTAL SETUP

Datasets and conditions. We evaluate RGI under distribution shift and retain original video to assess performance without shift. **(i)** TCD-TIMIT (Harte & Gillen, 2015) is mixed with DEMAND noise (Thiemann et al., 2013) for controlled evaluation on 900 utterances following Ayilo et al. (2025). **(ii)** GRID (Cooke et al., 2006) and Lombard GRID (Alghamdi et al., 2018) provide 300 utterances each to assess performance across corpora and speaking conditions. **(iii)** LRS3 (Afouras et al., 2018) supports evaluation on naturally recorded speech without clean target references. Following the deployment challenges in Li et al. (2026), the controlled evaluation includes five development perturbations, low resolution, occlusion, frozen frames, audio-video offset, and blackout.

Backbones and baselines. We apply RGI to three frozen diffusion AVSE backbones with different visual conditioning designs. **(i)** AV-UDiffSE combines a visually conditioned speech diffusion prior with a non-negative matrix factorization noise model (Ayilo et al., 2025). **(ii)** AV-SGMSE incorporates lipreading features into a conditional score network (Richter et al., 2023a). **(iii)** AV-CA-DiffUSE introduces contrastive audio-visual alignment during diffusion training (Mboungou et al., 2026). For each backbone, fixed audio-visual inference is the reference for alternative inference rules. These include modality selection based on input quality (Sadeghi & Alameda-Pineda, 2021; Du et al., 2026), condition-annealed diffusion sampling (CADS) (Sadat et al., 2024), guidance interval (GI) (Kynkäänniemi et al., 2024), state-dependent feedback guidance (FBG) (Koulischer et al., 2025), and test-time dynamic fusion (TTD) adapted from image fusion (Cao et al., 2024). We also evaluate continuous quality gating based on an input quality score (Hong et al., 2023; Su et al., 2026).

Evaluation metrics. We measure enhancement performance using scale-invariant signal-to-distortion ratio (SI-SDR) improvement relative to the noisy input (Le Roux et al., 2019). Results are summarized by *Mean*, the average over the five development perturbations, *Worst*, the lowest mean across these perturbations, and *Original*, the mean performance with original aligned video. Higher values indicate better performance for all three measures.

4.2 ENHANCEMENT UNDER DISTRIBUTION SHIFT ACROSS BACKBONES AND CORPORA

To validate adaptive visual conditioning under distribution shift, we compare inference methods in Table 1. RGI improves Mean, Worst, and Original over fixed audio-visual inference in most combinations. As shown in Fig. 3, these improvements hold for original video and under every shift on AV-UDiffSE with TCD-TIMIT/DEMAND, including wrong-speaker video and sham padding held out from parameter selection. Gains under shift are consistent with regulating visual contributions according to model-specific evidence, allowing useful conditioning to be retained while limiting its influence under large reference deviations. Together with the improvements on original video, these gains suggest that RGI can improve robustness without sacrificing the benefit of aligned video.

Modality selection frequently falls back to audio-only inference, producing the repeated or near-identical values marked in gray. The fallback limits exposure to harmful visual conditioning but also discards useful cues. By retaining a graded visual contribution, RGI generally achieves a better balance between these effects. Scheduled guidance and dynamic fusion remain competitive, with CADS and TTD leading on particular backbone and corpus combinations. The most effective control mechanism thus depends on the pretrained model and evaluation conditions.

Table 1: Enhancement performance across frozen backbones and corpora, measured by SI-SDR improvement in dB and reported as mean ± standard deviation. Mean averages the five development perturbations, Worst is the lowest perturbation mean, and Original uses original aligned video. **Bold** and underlined values indicate the best and second-best results within each backbone and corpus. Gray cells indicate performance equal or close to audio-only inference when modality selection suppresses visual conditioning. All results are averaged over three independent runs.

	TCD-TIMIT/DEMAND			GRID			Lombard GRID		
Method	Mean↑	Worst↑	Original↑	Mean↑	Worst↑	Original↑	Mean↑	Worst↑	Original↑
AV-UDiffSE	9.45 ±0.017	8.50 ±0.018	10.21 ±0.019	4.69 ±0.030	4.03 ±0.061	4.88 ±0.056	6.13 ±0.016	4.67 ±0.031	6.75 ±0.047
+ Modality selection	9.02 ±0.017	9.02 ±0.017	9.22 ±0.044	4.03 ±0.063	4.03 ±0.063	4.04 ±0.095	5.33 ±0.014	5.33 ±0.014	5.33 ±0.052
+ CADS [ICLR'24]	9.38 ±0.019	8.43 ±0.022	10.20 ±0.014	4.68 ±0.042	4.11 ±0.080	4.85 ±0.055	6.12 ±0.045	4.73 ±0.067	6.72 ±0.062
+ GI [NeurIPS'24]	9.33 ±0.008	8.39 ±0.026	10.12 ±0.024	4.56 ±0.066	3.86 ±0.036	4.78 ±0.022	6.11 ±0.071	4.75 ±0.053	6.73 ±0.020
+ FBG [NeurIPS'25]	9.37 ±0.010	8.45 ±0.022	10.15 ±0.032	4.66 ±0.055	4.02 ±0.056	4.84 ±0.042	6.08 ±0.086	4.58 ±0.106	6.70 ±0.076
+ TTD [NeurIPS'24]	9.53 ±0.022	8.65 ±0.040	10.23 ±0.024	4.74 ±0.057	**4.15** ±0.036	4.91 ±0.043	6.18 ±0.033	4.81 ±0.056	6.76 ±0.030
+ RGI (Ours)	**9.91** ±0.020	**9.23** ±0.021	**10.48** ±0.020	**4.76** ±0.070	4.14 ±0.058	**5.03** ±0.077	**6.41** ±0.024	**5.55** ±0.035	**6.87** ±0.053
AV-SGMSE	13.30 ±0.007	13.07 ±0.023	13.63 ±0.009	9.71 ±0.026	9.58 ±0.041	9.84 ±0.046	8.85 ±0.020	8.29 ±0.015	**8.91** ±0.015
+ Modality selection	12.57 ±0.015	12.57 ±0.015	12.73 ±0.005	9.57 ±0.023	9.57 ±0.023	9.57 ±0.037	8.38 ±0.026	8.38 ±0.026	8.40 ±0.044
+ CADS [ICLR'24]	13.38 ±0.005	13.09 ±0.017	13.64 ±0.009	**10.24** ±0.017	**10.16** ±0.027	**10.25** ±0.033	8.79 ±0.050	8.34 ±0.056	8.90 ±0.022
+ GI [NeurIPS'24]	13.31 ±0.004	13.06 ±0.014	13.62 ±0.016	9.70 ±0.029	9.56 ±0.015	9.83 ±0.019	**8.86** ±0.021	8.30 ±0.051	8.89 ±0.029
+ FBG [NeurIPS'25]	13.27 ±0.007	13.02 ±0.008	13.57 ±0.017	9.64 ±0.019	9.54 ±0.032	9.82 ±0.037	8.81 ±0.016	8.21 ±0.054	8.83 ±0.019
+ TTD [NeurIPS'24]	13.29 ±0.009	13.05 ±0.016	13.61 ±0.011	9.73 ±0.017	9.60 ±0.046	9.86 ±0.018	8.84 ±0.024	8.32 ±0.033	8.88 ±0.017
+ RGI (Ours)	**13.40** ±0.007	**13.12** ±0.013	**13.66** ±0.024	9.88 ±0.016	9.79 ±0.043	9.95 ±0.018	8.82 ±0.018	**8.41** ±0.018	**8.91** ±0.038
AV-CA-DiffUSE	8.28 ±0.021	7.57 ±0.017	8.72 ±0.018	5.60 ±0.062	4.93 ±0.047	5.76 ±0.085	5.62 ±0.014	3.93 ±0.052	6.30 ±0.073
+ Modality selection	9.02 ±0.017	9.02 ±0.017	8.98 ±0.042	4.03 ±0.063	4.03 ±0.061	4.03 ±0.093	5.33 ±0.014	5.33 ±0.014	5.35 ±0.040
+ CADS [ICLR'24]	8.09 ±0.020	7.41 ±0.028	8.52 ±0.050	5.48 ±0.054	4.82 ±0.064	5.65 ±0.031	5.49 ±0.065	3.87 ±0.074	6.16 ±0.068
+ GI [NeurIPS'24]	8.27 ±0.027	7.59 ±0.038	8.73 ±0.031	5.56 ±0.033	4.86 ±0.037	5.72 ±0.033	5.50 ±0.042	3.85 ±0.076	6.19 ±0.043
+ FBG [NeurIPS'25]	8.16 ±0.021	7.47 ±0.033	8.59 ±0.023	5.54 ±0.063	4.87 ±0.049	5.71 ±0.092	5.57 ±0.030	3.84 ±0.022	6.23 ±0.068
+ TTD [NeurIPS'24]	8.57 ±0.017	7.99 ±0.035	8.92 ±0.027	**5.69** ±0.063	**5.09** ±0.061	**5.84** ±0.046	5.82 ±0.035	4.32 ±0.025	6.40 ±0.056
+ RGI (Ours)	**9.48** ±0.035	**9.24** ±0.013	**9.69** ±0.012	5.63 ±0.080	4.95 ±0.077	5.83 ±0.041	**6.29** ±0.048	**5.55** ±0.063	**6.75** ±0.046

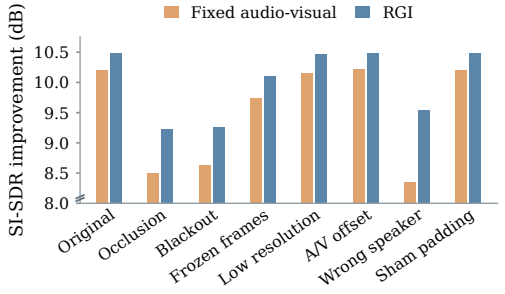


Figure 3: Condition-level SI-SDR improvement on AV-UDiffSE with TCD-TIMIT/DEMAND. Wrong speaker and sham padding are held out.

Table 2: Ablation of RGI on AV-CA-DiffUSE with TCD-TIMIT/DEMAND, measured by SI-SDR improvement in dB.

Variant	Mean↑	Worst↑	Original↑
Fixed audio-visual inference	8.28	7.57	8.72
Audio-only inference	9.02	9.02	9.02
w/o model-response score	9.35	8.62	**9.69**
w/o representation score	9.34	9.10	9.68
w/o conditioning floor	9.33	9.05	9.68
w/o diffusion-step schedule	9.01	8.74	8.93
Schedule only	9.05	8.55	9.62
Constant coefficient	9.32	8.93	9.64
RGI (full)	**9.48**	**9.24**	**9.69**

4.3 Ablation Study

We examine the contributions of reliability evidence and conditioning control using AV-CA-DiffUSE on TCD-TIMIT/DEMAND. As shown in Table 2, fixed audio-visual inference falls below audio-only inference on all three measures in this setting, so each component is evaluated by whether it limits harmful conditioning while retaining useful visual cues. Removing either score lowers Mean and Worst while leaving Original nearly unchanged, and neither source of evidence alone matches the full method under shift. The larger decline in Worst without the model-response score suggests that the model-response score provides evidence complementary to the representation score. Combining the two scores yields the best Mean and Worst without sacrificing performance on original video.

The floor and schedule determine how the reliability estimate shapes generation. Without the floor, Worst falls to nearly the audio-only level while Original is essentially unchanged, indicating that the floor prevents excessive suppression of visual cues that remain useful under the most severe shift. Removing the schedule lowers all three measures, most on original video, which suggests that the schedule is important for converting reliability weighting into enhancement gains.

The input-independent variants further isolate the role of reliability estimation. Schedule-only guidance remains close to the full method on original video but stays below audio-only inference on Worst, so a prescribed schedule alone is insufficient under severe shift. A constant conditioning

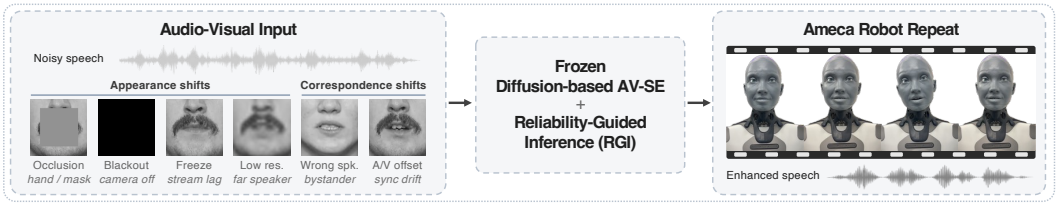


Figure 4: Ameca demonstration under distribution shifts. Noisy speech and visual inputs with appearance or correspondence shifts are processed by a frozen diffusion AVSE model with RGI. The enhanced speech is then used for the spoken-repeat demonstration with Ameca.

coefficient improves Mean through uniform attenuation, yet its Worst also stays below audio-only inference. In contrast, the full method exceeds audio-only inference on Worst while achieving the best Original, indicating that adapting the coefficient to each audio-video pair limits harmful conditioning without discarding useful visual cues.

4.4 Demonstration on the Ameca Robot

To demonstrate RGI in a robot interaction setting, we present noisy speech and visual speech inputs with controlled shifts to the Ameca humanoid robot. As illustrated in Fig. 4, the shifts include occlusion, blackout, frozen frames, and low resolution, together with wrong-speaker video and audio-video offsets. A frozen diffusion AVSE model equipped with RGI processes the audio-video pair, and the enhanced speech is used for the spoken-repeat demonstration. Recordings of these spoken repeats collected from the robot show that RGI can be deployed in humanoid robot interaction, where movement, occlusion, and competing speakers can compromise visual cues.

5 Related Work

Generative AVSE recovers speech through resynthesis (Hsu et al., 2023; Mira et al., 2023) or visually conditioned diffusion (Richter et al., 2023a; Ayilo et al., 2025; Mboungou et al., 2026). Robust visual conditioning has been explored through switching generative models (Sadeghi & Alameda-Pineda, 2021), confidence weighting and reliability scoring (Potamianos & Neti, 2000; Abdelaziz et al., 2015; Hong et al., 2023), corruption-robust representations (Kim et al., 2025), and modality gating, selection, or fusion (Su et al., 2026; Du et al., 2026; Zhang et al., 2023). Unlike these input-level or training-based approaches, RGI regulates visual contributions within frozen enhancers using internal visual representations and model responses.

Diffusion guidance controls prediction differences (Ho & Salimans, 2022) through restricted intervals (Kynkäänniemi et al., 2024), state-dependent feedback (Koulischer et al., 2025), or weaker reference models (Karras et al., 2024), while CADS anneals the condition itself (Sadat et al., 2024). Test-time adaptation addresses target-domain shifts through parameter updates (Wang et al., 2021), normalization-statistic estimation (Schneider et al., 2020), forward-only adaptation (Niu et al., 2024), or diffusion-based input adaptation (Gao et al., 2023). RGI instead adjusts the visual increment during sampling according to reference-based reliability evidence. Model parameters remain frozen, although rebuilding target-domain reference statistics constitutes reference adaptation.

6 Conclusion

We investigated visual conditioning in diffusion-based AVSE under distribution shift and showed that its utility depends on both the frozen model and the acoustic input. Based on these findings, we proposed RGI to regulate the visual increment during reverse diffusion using internal visual representations and model responses, without updating the pretrained models. We evaluated RGI across four corpora and three backbones and found improvements under distribution shift on the corpora with clean references, with gains that depend on the backbone and corpus. We further show with the Ameca demonstration that RGI can be deployed for humanoid robot interaction with noisy speech and shifted video. Future work will extend this approach through more accurate reliability estimation for individual inputs and reduced dependence on deployment-specific reference statistics.

AI USE STATEMENT

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

REFERENCES

Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa. Learning Dynamic Stream Weights for Coupled-HMM-Based Audio-Visual Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy J. Brown. A Corpus of Audio-Visual Lombard Speech with Frontal and Profile Views. *The Journal of the Acoustical Society of America*, 2018.

Jean-Eudes Ayilo, Mostafa Sadeghi, Romain Serizel, and Xavier Alameda-Pineda. Diffusion-Based Unsupervised Audio-Visual Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for Outliers with Conformal p-Values. *The Annals of Statistics*, 2023.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Test-Time Dynamic Image Fusion. In *Advances in Neural Information Processing Systems*, 2024.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 2018.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *International Conference on Learning Representations*, 2023.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. *The Journal of the Acoustical Society of America*, 2006.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Inference-Time Dynamic Modality Selection for Incomplete Multimodal Classification. In *International Conference on Learning Representations*, 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*, 2018.

Dongjie Fu, Xize Cheng, Xiaoda Yang, Hanting Wang, Zhou Zhao, and Tao Jin. Boosting Speech Recognition Robustness to Modality-Distortion with Contrast-Augmented Prompts. In *ACM International Conference on Multimedia*, 2024.

Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the Source: Diffusion-Driven Adaptation to Test-Time Corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Sander Goetzee, Konstantin Mihhailov, Roel Van De Laar, Kim Baraka, and Koen V. Hindriks. Audio-Visual Speech Recognition for Human-Robot Interaction: A Feasibility Study. In *IEEE International Conference on Robot and Human Interactive Communication*, 2024.

Naomi Harte and Eoin Gillen. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia*, 2015.

Alvin Heng, Alexandre H. Thiery, and Harold Soh. Out-of-Distribution Detection with a Single Unconditional Diffusion Model. In *Advances in Neural Information Processing Systems*, 2024.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Regeneration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a Diffusion Model with a Bad Version of Itself. In *Advances in Neural Information Processing Systems*, 2024.

Sungnyun Kim, Sungwoo Cho, Sangmin Bae, Kangwook Jang, and Se-Young Yun. Multi-Task Corrupted Prediction for Learning Robust Audio-Visual Speech Representation. In *International Conference on Learning Representations*, 2025.

Felix Koulischer, Florian Handke, Johannes Deleu, Thomas Demeester, and Luca Ambrogioni. Feedback Guidance of Diffusion Models. In *Advances in Neural Information Processing Systems*, 2025.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. TorchAudio-Squim: Reference-Less Speech Quality and Intelligibility Measures in TorchAudio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models. In *Advances in Neural Information Processing Systems*, 2024.

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

Olivier Ledoit and Michael Wolf. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 2004.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, 2018.

Kai Li, Wenze Ren, Junjie Li, Cheng Yu, Peijun Yang, Chien-yu Huang, Haibin Wu, Szu-Wei Fu, Wen-Chin Huang, Hsin-Min Wang, Xiaolin Hu, Ming Li, DeLiang Wang, and Yu Tsao. The ISCSLP 2026 Real-World Audio-Visual Speech Enhancement Challenge. *arXiv preprint arXiv:2608.23759*, 2026.

Colombe Mboungou, Mostafa Sadeghi, Jean-Eudes Ayilo, and Romain Serizel. Audio-Visual Contrastive Alignment for Diffusion-Based Visual-Conditioned Speech Enhancement. In *Annual Conference of the International Speech Communication Association*, 2026.

Harry McGurk and John MacDonald. Hearing Lips and Seeing Voices. *Nature*, 1976.

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*, 2021.

Rodrigo Mira, Buye Xu, Jacob Donley, Anurag Kumar, Stavros Petridis, Vamsi Krishna Ithapu, and Maja Pantic. LA-VocE: Low-SNR Audio-Visual Speech Enhancement Using Neural Vocoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD Detection via Feature Normalization. In *International Conference on Machine Learning*, 2025.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-Time Model Adaptation with Only Forward Passes. In *International Conference on Machine Learning*, 2024.

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

Gerasimos Potamianos and Chalapathy Neti. Stream Confidence Estimation for Audio-Visual Speech Recognition. In *International Conference on Spoken Language Processing*, 2000.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

Julius Richter, Simone Frintrop, and Timo Gerkmann. Audio-Visual Speech Enhancement with Score-Based Generative Models. In *ITG Conference on Speech Communication*, 2023a.

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*, 2023b.

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

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the Diversity of Diffusion Models Through Condition-Annealed Sampling. In *International Conference on Learning Representations*, 2024.

Mostafa Sadeghi and Xavier Alameda-Pineda. Switching Variational Auto-Encoders for Noise-Agnostic Audio-Visual Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving Robustness Against Common Corruptions by Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*, 2020.

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.

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

Fei Su, Cancan Li, Ming Li, and Juan Liu. M2S-AVSR: Modality-Aware Multi-View Self-Supervised Representation for Robust Audio-Visual Speech Recognition. *arXiv preprint arXiv:2606.05763*, 2026.

W. H. Sumby and Irwin Pollack. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 1954.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The Diverse Environments Multi-Channel Acoustic Noise Database (DEMAND): A Database of Multichannel Environmental Noise Recordings. In *Proceedings of Meetings on Acoustics*, 2013.

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

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable Dynamic Fusion for Low-Quality Multimodal Data. In *International Conference on Machine Learning*, 2023.

APPENDIX CONTENTS

A Notation **15**

B Derivations **16**

B.1 Visual Increment in the Sampler Update 16

B.2 Posterior-Mean Optimality 16

B.3 Video-Only Non-Identifiability 16

B.4 Properties of the Reliability Estimate and Conditioning Strength 17

C Reliability Estimation Details **17**

C.1 Representation Score 17

C.2 Model-Response Score 18

C.3 Reference Sets 18

C.4 Inference Procedure 18

D Backbones **19**

E Comparison Methods **20**

F Datasets, Visual Shifts, and Metrics **20**

G Implementation Details **22**

H Additional Results **22**

H.1 Ablation Variants 22

H.2 Condition-Level Results on AV-UDiffSE 22

H.3 Held-Out Visual Shifts 23

H.4 Quality Gating 23

H.5 Visual Reference Size 23

H.6 No-Reference Evaluation on LRS3 24

A NOTATION

Table 3 lists the symbols used in the main text and in the appendix. Vectors and matrices are set in bold, and upright subscripts and superscripts are labels. Italic subscripts index reverse steps and reference samples, and the superscript (i) indexes reference trajectories when the subscript denotes the reverse step. Symbols local to a single derivation are defined where they appear.

Table 3: Notation used in the main text and the appendix.

Symbol	Description
Signals, states, and sampler	
$\mathbf{y}$, $\mathbf{v}$	Noisy speech observation and visual condition
$\mathbf{x}_t$, $t \in \{N, \dots, 1\}$	Generative state at reverse step t of N reverse steps
τ_t, $\Delta\tau_t$	Diffusion time and reverse integration step size, with $\tau_{t-1} = \tau_t - \Delta\tau_t$
$\mathbf{f}_t$, b_t, $\boldsymbol{\epsilon}_t$	Forward diffusion drift, diffusion coefficient, and standard Gaussian noise
$\widetilde{\mathbf{x}}_{t-1}$	Predicted state before any acoustic observation update
Predictions and latent validity	
$\mathbf{u}_{\mathrm{av},t}$, $\mathbf{u}_{\mathrm{a},t}$	Audio-visual and audio-only score predictions at $(\mathbf{x}_t, \tau_t)$
$\Delta\mathbf{u}_t$	Visual increment $\mathbf{u}_{\mathrm{av},t} - \mathbf{u}_{\mathrm{a},t}$
β, $p(\beta)$	Latent validity of the visual increment and its prior
$\mathcal{O}_t$	Information available through reverse step t
$\mathbf{u}_{\beta,t}$, $\bar{\mathbf{u}}_t$	Interpolated score prediction and its posterior average
r_t	Reliability $\mathbb{E}[\beta \mid \mathcal{O}_t]$ of the visual increment
Reliability evidence	
$\boldsymbol{\phi}(\mathbf{v})$	Temporal mean and standard deviation of the internal visual representation
$d_{\mathrm{I}}(\mathbf{v})$, $q_{\mathrm{I}}(\mathbf{v})$	Mahalanobis distance and adjusted empirical upper-tail proportion
$z_{\mathrm{I}}(\mathbf{v})$	Representation score $\mathrm{logit}\, q_{\mathrm{I}}(\mathbf{v})$
h_t	Log magnitude ratio of $\Delta\mathbf{u}_t$ to $\mathbf{u}_{\mathrm{a},t}$
$\bar{z}_{\mathrm{D},t}$	Model-response score in Eq. (4)
z_t, $\hat{r}_t$	Combined score and reliability estimate in Eq. (5)
σ, logit	Logistic function and its inverse
clip, RMS, $\mathbf{1}\{\cdot\}$	Clipping to an interval, root mean square over all entries, and indicator function
Reference statistics	
$\mathcal{R}_{\mathrm{I}}$, M_{I}	Visual reference set and its size
$\mathcal{R}_{\mathrm{D}}$, M_{D}	Trajectory reference set and its size
$\boldsymbol{\mu}_{\mathrm{I}}$, $\mathbf{P}$, $\boldsymbol{\Sigma}_{\mathrm{I}}$	Reference mean, principal projection, and shrinkage covariance of the visual features
$d^{\mathrm{cf}}_{\mathrm{I},i}$	Cross-fitted distance of the reference video $\mathbf{v}_i$
$\mu_{\mathrm{D},t}$, $s_{\mathrm{D},t}$	Reference mean and standard deviation of h_t at reverse step t
m_{D}	Median reference model-response score at the final reverse step
Conditioning control and parameters	
c_t, ρ	Conditioning coefficient and conditioning floor
g_t, ξ_t	Diffusion-step schedule and sampling progress
g_{hi}, g_{lo}	Initial and final scales of the diffusion-step schedule
$g_t c_t$, $\mathbf{u}_t^{\mathrm{RGI}}$	Conditioning strength and guided score prediction
κ, θ_{r}, α	Slope, midpoint, and lower-tail level of the reliability mapping
L, $s_{\min}$, δ	Clipping bound, lower bound on the reference scale, and lower bound on the RMS magnitudes in h_t
Comparison methods and evaluation	
$\gamma(\mathbf{v})$, η	Input quality score and selection threshold
$\mathbf{x}^\star$, $\hat{\mathbf{x}}$	Clean reference and enhanced output in the time domain
$\bar{e}_k$	Mean SI-SDR improvement under the k-th development perturbation

B DERIVATIONS

The derivations below expand the formulation in Sec. 2 and the guidance rule in Sec. 3.2. App. B.1 derives the contribution of the visual increment to the sampler update, Apps. B.2 and B.3 derive the results referenced in Sec. 2.2, and App. B.4 collects properties of the reliability estimate and the conditioning strength.

B.1 VISUAL INCREMENT IN THE SAMPLER UPDATE

Consider the Euler predictor in Eq. (1) with the state $\mathbf{x}_t$, the diffusion time τ_t, and the noise realization $\boldsymbol{\epsilon}_t$ held fixed. Denote by $\widetilde{\mathbf{x}}^{\mathrm{av}}_{t-1}$ and $\widetilde{\mathbf{x}}^{\mathrm{a}}_{t-1}$ the predicted states obtained with $\mathbf{u}_{\mathrm{av},t}$ and with $\mathbf{u}_{\mathrm{a},t}$ as the score input, and by $\widetilde{\mathbf{x}}^{\mathrm{RGI}}_{t-1}$ the predicted state in Eq. (6). The three updates share $\mathbf{x}_t$, $\mathbf{f}_t$, b_t, $\Delta\tau_t$, and $\boldsymbol{\epsilon}_t$, so their differences depend only on the score input,

$$\widetilde{\mathbf{x}}^{\mathrm{av}}_{t-1} - \widetilde{\mathbf{x}}^{\mathrm{a}}_{t-1} = b_t^2 \Delta\tau_t \, \Delta\mathbf{u}_t, \qquad \widetilde{\mathbf{x}}^{\mathrm{RGI}}_{t-1} - \widetilde{\mathbf{x}}^{\mathrm{a}}_{t-1} = g_t c_t \, b_t^2 \Delta\tau_t \, \Delta\mathbf{u}_t. \tag{7}$$

The conditioning strength $g_t c_t$ therefore scales the displacement that the visual increment contributes to the predictor update. Combining the two identities gives

$$\widetilde{\mathbf{x}}^{\mathrm{RGI}}_{t-1} = \big(1 - g_t c_t\big)\, \widetilde{\mathbf{x}}^{\mathrm{a}}_{t-1} + g_t c_t \, \widetilde{\mathbf{x}}^{\mathrm{av}}_{t-1}. \tag{8}$$

For $g_t c_t \in [0, 1]$, the RGI update lies on the segment between the audio-only and fixed audio-visual updates, and $g_t c_t = 0$ and $g_t c_t = 1$ recover the two endpoints. Classifier-free guidance (Ho & Salimans, 2022), written as $\mathbf{u}_{\mathrm{a},t} + \omega\, \Delta\mathbf{u}_t$ with a guidance scale ω for which $\omega = 1$ recovers the conditional prediction (Kynkäänniemi et al., 2024), has the same form, so $g_t c_t$ takes the place of ω and varies with the input and the reverse step. The same argument applies to corrector substeps whose step size does not depend on the score input, as for the annealed Langevin correctors of the three backbones (Richter et al., 2023b).

B.2 POSTERIOR-MEAN OPTIMALITY

The main text uses posterior averaging to define the averaged score prediction in Eq. (2), and the same prediction minimizes the expected squared error under the latent validity model. Consider the interpolated score prediction $\mathbf{u}_{\beta,t} = \mathbf{u}_{\mathrm{a},t} + \beta\, \Delta\mathbf{u}_t$ with $\beta \in [0, 1]$ and $r_t = \mathbb{E}[\beta \mid \mathcal{O}_t]$. Conditional on $\mathcal{O}_t$, the audio-only prediction $\mathbf{u}_{\mathrm{a},t}$ and the increment $\Delta\mathbf{u}_t$ are fixed, so β is the only random quantity in $\mathbf{u}_{\beta,t}$, and linearity of the conditional expectation gives

$$\mathbb{E}[\mathbf{u}_{\beta,t} \mid \mathcal{O}_t] = \mathbf{u}_{\mathrm{a},t} + r_t \, \Delta\mathbf{u}_t. \tag{9}$$

For a candidate score prediction $\mathbf{u}$ that depends on $\mathcal{O}_t$ only, the squared error $\|\mathbf{u} - \mathbf{u}_{\beta,t}\|_2^2$ has the posterior expectation

$$\mathbb{E}\big[\|\mathbf{u} - \mathbf{u}_{\beta,t}\|_2^2 \mid \mathcal{O}_t\big] = \big\|\mathbf{u} - \mathbb{E}[\mathbf{u}_{\beta,t} \mid \mathcal{O}_t]\big\|_2^2 + \mathbb{E}\big[\|\mathbf{u}_{\beta,t} - \mathbb{E}[\mathbf{u}_{\beta,t} \mid \mathcal{O}_t]\|_2^2 \mid \mathcal{O}_t\big]. \tag{10}$$

The second term does not depend on $\mathbf{u}$, so the expected squared error is uniquely minimized by the conditional mean in Eq. (9), which is the averaged score prediction $\bar{\mathbf{u}}_t$ of Eq. (2). The optimal score prediction therefore depends on the posterior of β only through its mean r_t. Since $\mathbf{u}_{\beta,t} - \mathbb{E}[\mathbf{u}_{\beta,t} \mid \mathcal{O}_t] = (\beta - r_t)\, \Delta\mathbf{u}_t$, the minimum value equals $\mathrm{Var}[\beta \mid \mathcal{O}_t]\, \|\Delta\mathbf{u}_t\|_2^2$. Estimating the averaged score prediction thus reduces to estimating r_t from the available evidence.

B.3 VIDEO-ONLY NON-IDENTIFIABILITY

Estimating r_t requires evidence about β, and the derivation below shows when the video alone cannot provide it. Let $p(\beta)$ denote the prior law of the latent validity marginalized over the noisy speech, $p(\beta) = \int p(\beta \mid \mathbf{y})\, p(\mathbf{y})\, \mathrm{d}\mathbf{y}$, and let $p(\mathbf{v} \mid \beta)$ denote the corresponding video marginal at validity level β. A correspondence shift that preserves the video marginal gives $p(\mathbf{v} \mid \beta) = p(\mathbf{v})$ for every β. The posterior density of β given $\mathbf{v}$ is proportional to $p(\mathbf{v} \mid \beta)\, p(\beta)$, and a likelihood that is constant in β cancels in the normalization, so $p(\beta \mid \mathbf{v}) = p(\beta)$ almost everywhere, which is Eq. (3). The same argument extends to every measurable statistic $\psi(\mathbf{v})$ of the video alone, whether deterministic or randomized with randomization independent of β given $\mathbf{v}$. Such a statistic is a

transformation of $\mathbf{v}$ whose law does not depend on β, so conditioning on it also leaves the prior law unchanged, with $p(\beta \mid \psi(\mathbf{v})) = p(\beta)$ and $\mathbb{E}[\beta \mid \psi(\mathbf{v})] = \mathbb{E}[\beta]$.

Because the representation score is a statistic of the video alone, it is subject to the same limitation, which motivates the model-response score of Sec. 3.1 as a second source of evidence that depends on the audio-video pair through the frozen model.

B.4 Properties of the Reliability Estimate and Conditioning Strength

The mapping in Eq. (5) and the control rule in Sec. 3.2 have the following properties, each of which follows directly from the definitions.

Range. Since σ maps the real line to $(0, 1)$, the reliability estimate satisfies $\hat{r}_t \in (0, 1)$ and the conditioning coefficient satisfies $c_t = \rho + (1-\rho)\hat{r}_t \in (\rho, 1)$. Since $\xi_t \in [0, 1]$, the schedule satisfies $g_t \in [g_{\mathrm{lo}}, g_{\mathrm{hi}}]$, and the conditioning strength is bounded as

$$g_t\rho < g_t c_t < g_t \leq g_{\mathrm{hi}}. \tag{11}$$

With $g_{\mathrm{hi}} \leq 1$, as in the configuration of App. G, the strength remains in $(0, 1)$, and Eq. (8) is a convex combination at every reverse step.

Monotonicity. The derivative of the reliability estimate with respect to the combined score is

$$\frac{\partial \hat{r}_t}{\partial z_t} = \kappa\, \hat{r}_t\big(1 - \hat{r}_t\big) > 0, \tag{12}$$

and $z_t = z_{\mathrm{I}}(\mathbf{v}) + \bar{z}_{\mathrm{D},t}$ increases with each score. With either score held fixed, the reliability estimate, the conditioning coefficient, and the conditioning strength therefore decrease as the other score indicates a larger deviation from its reference statistics. The slope κ sets this sensitivity, which is largest at the midpoint, where $\partial \hat{r}_t / \partial z_t = \kappa/4$.

Posterior interpretation. Suppose that $\beta \in \{0, 1\}$ and that the posterior log-odds of $\beta = 1$ given $\mathcal{O}_t$ equal $\kappa(z_t - \theta_{\mathrm{r}})$. Then $r_t = \mathbb{E}[\beta \mid \mathcal{O}_t] = p(\beta = 1 \mid \mathcal{O}_t) = \sigma(\kappa(z_t - \theta_{\mathrm{r}})) = \hat{r}_t$. Under this model, κ scales the evidence carried by the combined score and θ_{r} absorbs the prior log-odds, so the midpoint condition places the decision boundary at the reference-anchored combined score. When β takes intermediate values, $\hat{r}_t$ retains the range and monotonicity stated above without a likelihood model for β.

Running update. The model-response score in Eq. (4) is a running average and can be updated at each reverse step without storing earlier model responses. Starting from $\bar{z}_{\mathrm{D},N+1} = 0$,

$$\bar{z}_{\mathrm{D},t} = \frac{N-t}{N-t+1}\,\bar{z}_{\mathrm{D},t+1} - \frac{1}{N-t+1}\left|\mathrm{clip}\left(\frac{h_t - \mu_{\mathrm{D},t}}{\max\{s_{\mathrm{D},t}, s_{\min}\}}, -L, L\right)\right|. \tag{13}$$

Each averaged term lies in $[0, L]$, so $\bar{z}_{\mathrm{D},t} \in [-L, 0]$, and the change from $\bar{z}_{\mathrm{D},t+1}$ to $\bar{z}_{\mathrm{D},t}$ is at most $L/(N-t+1)$ in magnitude. A single extreme deviation therefore has a bounded effect that decreases as sampling proceeds. For samplers with predictor and corrector substeps, the same recursion runs over all evaluated substeps, with $N - t + 1$ replaced by the number of substeps evaluated so far.

C Reliability Estimation Details

The construction of the reference statistics in Sec. 3.1 is detailed below, and Alg. 1 summarizes the inference procedure. All reference statistics are computed once before inference from original aligned inputs and remain fixed during sampling.

C.1 Representation Score

The visual conditioning module of each backbone maps the video $\mathbf{v}$ to a sequence of frame-level features that the score network consumes, as described in App. D. The feature vector $\boldsymbol{\phi}(\mathbf{v})$ concatenates the temporal mean and the temporal standard deviation of this sequence, so its dimension is twice the feature dimension and does not depend on the clip length.

From $\{\boldsymbol{\phi}(\mathbf{v}_i)\}_{i=1}^{M_{\mathrm{I}}}$, we compute the reference mean $\boldsymbol{\mu}_{\mathrm{I}}$, a projection $\mathbf{P}$ onto the leading principal directions of the centered features, and a covariance $\boldsymbol{\Sigma}_{\mathrm{I}}$ of the features after projection, shrunk

toward a scaled identity (Ledoit & Wolf, 2004). The projection restricts the distance to directions supported by the reference data, and the shrinkage keeps $\mathbf{\Sigma}_{\mathrm{I}}$ well conditioned. The Mahalanobis distance with respect to this single reference distribution (Lee et al., 2018) is

$$d_{\mathrm{I}}(\mathbf{v}) = \sqrt{\big(\mathbf{P}(\boldsymbol{\phi}(\mathbf{v}) - \boldsymbol{\mu}_{\mathrm{I}})\big)^{\top} \mathbf{\Sigma}_{\mathrm{I}}^{-1}\, \mathbf{P}(\boldsymbol{\phi}(\mathbf{v}) - \boldsymbol{\mu}_{\mathrm{I}})}, \tag{14}$$

which measures the deviation of $\boldsymbol{\phi}(\mathbf{v})$ from the reference mean in units of the reference variability.

Reference distances are obtained through cross-fitting (Chernozhukov et al., 2018). We partition $\mathcal{R}_{\mathrm{I}}$ into disjoint folds, refit $\boldsymbol{\mu}_{\mathrm{I}}$, $\mathbf{P}$, and $\mathbf{\Sigma}_{\mathrm{I}}$ without the fold containing $\mathbf{v}_i$, and evaluate Eq. (14) with the refitted statistics to obtain $d_{\mathrm{I},i}^{\mathrm{cf}}$. Each reference video is therefore excluded from the statistics used to evaluate its distance. The adjusted empirical upper-tail proportion is

$$q_{\mathrm{I}}(\mathbf{v}) = \frac{1 + \sum_{i=1}^{M_{\mathrm{I}}} \mathbf{1}\{d_{\mathrm{I},i}^{\mathrm{cf}} \geq d_{\mathrm{I}}(\mathbf{v})\}}{M_{\mathrm{I}} + 1}, \tag{15}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The added one in the numerator and the denominator is the finite-sample adjustment used for conformal p-values (Bates et al., 2023), so $q_{\mathrm{I}}(\mathbf{v}) \in [1/(M_{\mathrm{I}} + 1), 1]$. The representation score is $z_{\mathrm{I}}(\mathbf{v}) = \operatorname{logit} q_{\mathrm{I}}(\mathbf{v})$, as defined in Sec. 3.1. Because $q_{\mathrm{I}}(\mathbf{v}) = 1$ when $d_{\mathrm{I}}(\mathbf{v})$ does not exceed any cross-fitted reference distance, the argument of the logit is clipped to $[10^{-9}, 1 - 10^{-9}]$ for numerical stability.

C.2 Model-Response Score

The trajectory reference set $\mathcal{R}_{\mathrm{D}}$ provides per-step statistics of the log magnitude ratio. For each matched pair $(\mathbf{y}_i, \mathbf{v}_i) \in \mathcal{R}_{\mathrm{D}}$, we record the log magnitude ratio $h_t^{(i)}$ along its reference trajectory at every reverse step. The reference mean is

$$\mu_{\mathrm{D},t} = \frac{1}{M_{\mathrm{D}}} \sum_{i=1}^{M_{\mathrm{D}}} h_t^{(i)}, \tag{16}$$

and $s_{\mathrm{D},t}$ is the corresponding standard deviation of $\{h_t^{(i)}\}_{i=1}^{M_{\mathrm{D}}}$. The lower bound $s_{\min}$ in Eq. (4) applies when the reference responses at a step are nearly identical. For samplers with predictor and corrector substeps, the statistics are indexed by both step and substep, and the running average in Eq. (4) includes all evaluated substeps.

The same trajectories provide the reference model-response scores $\bar{z}_{\mathrm{D},1}^{(i)}$ at the final reverse step, and m_{D} is their median. The midpoint $\theta_{\mathrm{r}} = \operatorname{logit}(\alpha) + m_{\mathrm{D}}$ is therefore determined by the reference statistics of each backbone and requires no shifted inputs.

C.3 Reference Sets

The visual reference set contains $M_{\mathrm{I}} = 500$ clips of original video, and the trajectory reference set contains $M_{\mathrm{D}} = 30$ matched audio-video pairs. Both sets are drawn from speakers outside the test set. Neither set requires clean target speech, since the reference statistics depend only on the video and on model responses to noisy inputs.

For GRID, Lombard GRID, and LRS3, the visual reference statistics are rebuilt from original videos of the corresponding corpus, while the trajectory reference set and all scalar parameters are shared across corpora. Rebuilding the reference statistics leaves all model parameters unchanged. App. H.5 examines the sensitivity of RGI to the size of the visual reference set.

C.4 Inference Procedure

Alg. 1 summarizes RGI for one input. The representation score and the midpoint are computed once, and each reverse step evaluates the two predictions, updates the model-response score by Eq. (13), and supplies the guided score prediction to the sampler of the backbone.

Algorithm 1 RGI with a frozen diffusion AVSE backbone

Require: noisy speech $\mathbf{y}$, video $\mathbf{v}$, frozen audio-visual and audio-only score predictors, reference statistics of Apps. C.1 and C.2, and parameters α, κ, L, $s_{\min}$, δ, ρ, g_{hi}, and g_{lo}

1: Compute $d_{\mathrm{I}}(\mathbf{v})$, $q_{\mathrm{I}}(\mathbf{v})$, and $z_{\mathrm{I}}(\mathbf{v})$ by Eqs. (14) and (15)
2: Set $\theta_{\mathrm{r}} \leftarrow \mathrm{logit}(\alpha) + m_{\mathrm{D}}$ and initialize $\mathbf{x}_N$ as prescribed by the backbone
3: **for** $t = N, \ldots, 1$ **do**
4: Evaluate $\mathbf{u}_{\mathrm{av},t}$ and $\mathbf{u}_{\mathrm{a},t}$ at $(\mathbf{x}_t, \tau_t)$ and set $\Delta\mathbf{u}_t \leftarrow \mathbf{u}_{\mathrm{av},t} - \mathbf{u}_{\mathrm{a},t}$
5: $h_t \leftarrow \log\big[\max\{\mathrm{RMS}(\Delta\mathbf{u}_t), \delta\} / \max\{\mathrm{RMS}(\mathbf{u}_{\mathrm{a},t}), \delta\}\big]$
6: Update $\bar{z}_{\mathrm{D},t}$ by Eq. (13) and set $z_t \leftarrow z_{\mathrm{I}}(\mathbf{v}) + \bar{z}_{\mathrm{D},t}$
7: $\hat{r}_t \leftarrow \sigma(\kappa(z_t - \theta_{\mathrm{r}}))$, $c_t \leftarrow \rho + (1-\rho)\hat{r}_t$, and $g_t \leftarrow g_{\mathrm{hi}} + (g_{\mathrm{lo}} - g_{\mathrm{hi}})\xi_t$
8: $\mathbf{u}_t^{\mathrm{RGI}} \leftarrow \mathbf{u}_{\mathrm{a},t} + g_t c_t \, \Delta\mathbf{u}_t$
9: Obtain $\widetilde{\mathbf{x}}_{t-1}$ by Eq. (6), repeating lines 4 to 8 at each corrector substep
10: Apply the acoustic observation update of the backbone, if any, to obtain $\mathbf{x}_{t-1}$
11: **end for**
12: **return** the enhanced speech obtained from $\mathbf{x}_0$

D BACKBONES

The three backbones differ in how visual features condition the score network, how the audio-only prediction is obtained, and whether the sampler includes an acoustic observation update. Table 4 summarizes these properties. All backbones remain frozen, and RGI changes only the score input supplied to their samplers.

AV-UDiffSE. AV-UDiffSE (Ayilo et al., 2025) is an unsupervised enhancer. Its clean-speech score network is trained on clean audio-visual speech only, and the noise is modeled at inference by non-negative matrix factorization (NMF). Visual features from AV-HuBERT fine-tuned for lipreading (Shi et al., 2022) enter the score network through cross-attention. Enhancement combines the clean-speech prior with the noise model through posterior sampling (Nortier et al., 2024), which interleaves predictor and corrector substeps with a likelihood update of the generative state and of the noise model parameters. The noisy speech enters both score predictions through the generative state, which the likelihood update ties to $\mathbf{y}$. We use the single-pass variant introduced by Ayilo et al. (2025), which updates the noise model along one reverse trajectory. The audio-only prediction $\mathbf{u}_{\mathrm{a},t}$ comes from the separately trained audio-only counterpart of the same prior.

AV-SGMSE. AV-SGMSE denotes the audio-visual score-based enhancer of Richter et al. (2023a), which extends SGMSE+ (Richter et al., 2023b) with visual conditioning. The score network is conditioned on the noisy speech and on features from a frozen AV-HuBERT lipreading model (Shi et al., 2022), which are combined across layers with learned weights and aligned in time with the audio. The backbone is trained with supervision on pairs of noisy and clean speech, so the reverse process starts from the noisy speech perturbed by Gaussian noise and requires no separate acoustic observation update. Sampling uses predictor and corrector substeps (Song et al., 2021). The audio-only prediction $\mathbf{u}_{\mathrm{a},t}$ is obtained by evaluating the same network with a prescribed null visual input.

AV-CA-DiffUSE. AV-CA-DiffUSE (Mboungou et al., 2026) adopts the architecture of AV-UDiffSE and trains its prior with an additional contrastive objective, which aligns audio embeddings of the denoised speech with pooled visual features. The contrastive branch is used only during training, so inference follows the same posterior sampling procedure as AV-UDiffSE. The two backbones also share the audio-only prior of AV-UDiffSE, which is why their audio-only results coincide, as reflected by the Mean and Worst of modality selection in Table 1.

Internal visual representation. For each backbone, the internal visual representation used by the representation score is the output of the visual conditioning module that the score network consumes, as defined in Sec. 2.2. It depends on neither the generative state nor the diffusion step, so $z_{\mathrm{I}}(\mathbf{v})$ is computed once for each video.

Table 4: Frozen diffusion AVSE backbones, all sampled with predictor and corrector substeps.

Backbone	Training	Visual conditioning	Audio-only prediction	Observation update
AV-UDiffSE	Unsupervised, clean speech	AV-HuBERT features via cross-attention	Separate audio-only prior	NMF likelihood
AV-SGMSE	Supervised, noisy and clean pairs	Layer-weighted AV-HuBERT features	Null visual input	None
AV-CA-DiffUSE	Unsupervised, clean speech with contrastive alignment	AV-HuBERT features via cross-attention	Audio-only prior of AV-UDiffSE	NMF likelihood

E Comparison Methods

Each comparison method is an inference rule for the same frozen backbone and sampler, and the methods differ in how visual conditioning enters the reverse process. Table 5 summarizes the evidence each method uses, the quantity it modifies, and its control rule. Methods acting on the visual increment replace the unit conditioning strength of fixed audio-visual inference in Eq. (1), whereas CADS and TTD modify the conditioning input or the features before the score network.

Fixed audio-visual and audio-only inference. Fixed audio-visual inference supplies $\mathbf{u}_{\mathrm{av},t}$ at every reverse step, which corresponds to a conditioning strength of one on $\Delta\mathbf{u}_t$. Audio-only inference supplies $\mathbf{u}_{\mathrm{a},t}$, which corresponds to a strength of zero. The two rules define the endpoints of Eq. (8).

Modality selection and quality gating. Both methods rely on an input quality score $\gamma(\mathbf{v}) \in [0, 1]$, which measures how typical the input is of original aligned inputs, in the spirit of stream confidence and visual-quality measures (Hong et al., 2023; Su et al., 2026). Modality selection chooses between fixed audio-visual inference and audio-only inference for each input, applying the strength $\mathbf{1}\{\gamma(\mathbf{v}) \geq \eta\}$ with threshold η, which follows the switching and selection principles of Sadeghi & Alameda-Pineda (2021) and Du et al. (2026). Quality gating applies the continuous strength $\gamma(\mathbf{v})$ instead. Both strengths are fixed for each input along the sampling trajectory.

Condition-annealed diffusion sampling. CADS (Sadat et al., 2024) replaces the condition with a noisy version whose noise is annealed along the sampling trajectory. A piecewise-linear schedule in diffusion time interpolates from a condition dominated by Gaussian noise at the start of sampling to the clean condition toward the end, and the noisy condition is rescaled to the statistics of the clean one. CADS thus acts on the conditioning input instead of the visual increment.

Guidance interval. GI (Kynkäänniemi et al., 2024) applies classifier-free guidance only within an interval of noise levels. In the notation of App. B.1, the strength on $\Delta\mathbf{u}_t$ equals a guidance scale inside the interval and one outside it, so GI coincides with fixed audio-visual inference at reverse steps outside the interval.

Feedback guidance. FBG (Koulischer et al., 2025) models the conditional prediction as a mixture of conditional and unconditional components and adapts the guidance scale along the trajectory from a running comparison of how well the conditional and unconditional predictions explain the observed transitions. Its strength on $\Delta\mathbf{u}_t$ therefore depends on the state and the reverse step.

Test-time dynamic fusion. TTD (Cao et al., 2024) computes fusion weights for each source at test time through a softmax over negative exponentiated uni-source reconstruction losses, so that sources with lower loss receive larger weights. It requires no training and adapts the fusion to each input.

RGI. RGI acts on $\Delta\mathbf{u}_t$ with the strength $g_t c_t$, which combines the representation and model-response scores with the conditioning floor and the diffusion-step schedule, as described in Sec. 3.

F Datasets, Visual Shifts, and Metrics

Corpora. TCD-TIMIT/DEMAND follows the evaluation protocol of Ayilo et al. (2025). Clean utterances from the test speakers of TCD-TIMIT (Harte & Gillen, 2015) are mixed with DEMAND noise (Thiemann et al., 2013) at input SNRs of -5 and $+5$ dB, the two SNR groups compared in

Table 5: Comparison methods as inference rules for a frozen backbone.

Method	Evidence	Modified quantity	Control rule
Fixed audio-visual inference	None	$\Delta \mathbf{u}_t$	Strength 1
Audio-only inference	None	$\Delta \mathbf{u}_t$	Strength 0
Modality selection	Input quality score	$\Delta \mathbf{u}_t$	Strength $\mathbf{1}\{\gamma(\mathbf{v}) \geq \eta\}$
Quality gating	Input quality score	$\Delta \mathbf{u}_t$	Strength $\gamma(\mathbf{v})$
CADS	Diffusion time	Conditioning input	Noise annealed toward the clean condition
GI	Noise level	$\Delta \mathbf{u}_t$	Guidance scale inside an interval, one outside
FBG	Conditional and unconditional predictions	$\Delta \mathbf{u}_t$	Guidance scale adapted by feedback
TTD	Uni-source losses	Source features	Softmax weights from the losses
RGI	$z_{\mathrm{I}}(\mathbf{v})$ and $\bar{z}_{\mathrm{D},t}$	$\Delta \mathbf{u}_t$	Strength $g_t c_t$

Fig. 1(d), which gives the 900 noisy utterances of the controlled evaluation with clean references and mouth-region video. GRID (Cooke et al., 2006) and Lombard GRID (Alghamdi et al., 2018) contribute 300 utterances each and differ from TCD-TIMIT in speakers, vocabulary, and recording conditions. Lombard GRID additionally contains speech produced in noise, which changes articulation relative to plain speech. LRS3 (Afouras et al., 2018) contains naturally recorded speech from talks without clean references, so its results rely on no-reference measures.

Development perturbations. The five development perturbations follow the deployment challenges considered in Li et al. (2026) and are applied to the mouth-region video while the audio is left unchanged, as illustrated in Fig. 1(a). *Low resolution* downsamples each frame and upsamples it to its original size, which removes fine articulatory detail. *Occlusion* masks a region of the mouth area in every frame. *Frozen frames* repeats the preceding frame over a contiguous segment of the clip. *Audio-video offset* shifts the video relative to the audio by a fixed number of frames and fills the vacated frames by repeating the boundary frame. *Blackout* replaces a contiguous segment of frames with black frames.

Held-out shifts. The two held-out shifts are excluded from parameter selection. *Wrong speaker* replaces the video with that of a different speaker uttering the same sentence, which preserves plausible visual speech while breaking audio-visual correspondence. Among the 900 utterances, 550 have such a donor among the test speakers, and comparisons under this shift use these 550 utterances, while every other condition uses all 900. *Sham padding* replaces the frames that the audio-video offset would vacate with the boundary frame and keeps all other frames aligned, which separates the effect of the padding operation from that of temporal misalignment.

Reference-based metric. For a clean reference $\mathbf{x}^\star$ and an enhanced output $\hat{\mathbf{x}}$ in the time domain, both normalized to zero mean, SI-SDR (Le Roux et al., 2019) projects the output onto the reference to remove the scale ambiguity,

$$\text{SI-SDR}(\hat{\mathbf{x}}, \mathbf{x}^\star) = 10 \log_{10} \frac{\|\zeta\, \mathbf{x}^\star\|_2^2}{\|\zeta\, \mathbf{x}^\star - \hat{\mathbf{x}}\|_2^2}, \qquad \zeta = \frac{\hat{\mathbf{x}}^\top \mathbf{x}^\star}{\|\mathbf{x}^\star\|_2^2}, \tag{17}$$

where ζ is the optimal scaling of the reference. SI-SDR improvement subtracts the SI-SDR of the noisy input, $\text{SI-SDR}(\hat{\mathbf{x}}, \mathbf{x}^\star) - \text{SI-SDR}(\mathbf{y}, \mathbf{x}^\star)$, with $\mathbf{y}$ taken in the time domain. Let $\bar{e}_k$ denote the SI-SDR improvement averaged over all utterances under the k-th development perturbation, $k \in \{1, \ldots, 5\}$. The summary measures in Table 1 are

$$\text{Mean} = \frac{1}{5} \sum_{k=1}^{5} \bar{e}_k, \qquad \text{Worst} = \min_{k \in \{1,\ldots,5\}} \bar{e}_k, \tag{18}$$

and Original is the SI-SDR improvement averaged over all utterances with original video. Original video and the held-out shifts are excluded from Mean and Worst. Because Worst is taken over perturbation means, it identifies the perturbation under which a method is least effective on average.

No-reference measures. LRS3 provides no clean references. DNSMOS (Reddy et al., 2022) predicts listener ratings of perceptual quality from the output alone, and TorchAudio-SQUIM (Kumar

et al., 2023) predicts STOI (Taal et al., 2011), wideband PESQ (Rix et al., 2001), and SI-SDR without a reference. These measures characterize enhancement on naturally recorded speech and are reported on their own scales.

G Implementation Details

Table 6 lists the parameters of RGI in the main configuration. The midpoint θ_{r} is not set by hand, since it follows from α and the reference statistic m_{D} of each backbone.

Table 6: Parameters of RGI in the main configuration.

Symbol	Role	Value
M_{I}	Clips in the visual reference set	500
M_{D}	Matched pairs in the trajectory reference set	30
α	Lower-tail level anchoring the midpoint	0.05
κ	Slope of the reliability mapping	3
L	Clipping bound on standardized deviations	8
$s_{\min}$	Lower bound on the reference scale	10^{-3}
δ	Lower bound on the RMS magnitudes in h_t	10^{-20}
ρ	Conditioning floor	0.1
g_{hi}, g_{lo}	Initial and final scales of the diffusion-step schedule	1, 0.5

Computation. RGI requires the audio-visual and audio-only predictions at every reverse step, as guidance methods based on the difference of two predictions do (Ho & Salimans, 2022). For AV-SGMSE, both predictions come from the same network, and for AV-UDiffSE and AV-CA-DiffUSE, the audio-only prediction comes from the audio-only prior. The representation score requires one evaluation of the visual conditioning module for each video, and the model-response score adds two root-mean-square reductions and the scalar operations of Eq. (13) at each step. The reference statistics are computed once before inference and are reused for all inputs of a corpus.

H Additional Results

The results below complement Tables 1 and 2 and Fig. 3 and include the no-reference evaluation on LRS3. All values are SI-SDR improvements in dB unless stated otherwise.

H.1 Ablation Variants

Each variant in Table 2 modifies one element of RGI. The two single-score variants omit the model-response score or the representation score from the combined score z_t. Removing the conditioning floor sets $\rho = 0$, and removing the diffusion-step schedule sets $g_t = 1$. Schedule-only guidance sets $c_t = 1$, so that the visual increment is weighted by the diffusion-step schedule alone. The constant-coefficient variant replaces c_t with a single value shared by all inputs and reverse steps and retains the diffusion-step schedule. These last two variants apply the same weighting to every input and therefore isolate the contribution of input-dependent reliability estimation.

H.2 Condition-Level Results on AV-UDiffSE

Table 7 reports SI-SDR improvement under each development perturbation on AV-UDiffSE with TCD-TIMIT/DEMAND. The fixed audio-visual and RGI rows give the values shown in Fig. 3, and the Mean and Worst columns, computed from the five perturbation columns, match Table 1. RGI improves over fixed audio-visual inference under every perturbation and on original video. The largest gains arise under occlusion and blackout, where fixed audio-visual inference falls below audio-only inference and RGI exceeds it. Under low resolution and audio-video offset, fixed audio-visual inference already exceeds audio-only inference on average, and RGI retains this benefit while further increasing enhancement.

Table 7: Condition-level SI-SDR improvement in dB on AV-UDiffSE with TCD-TIMIT/DEMAND.

Method	Original	Low res.	Occlusion	Frozen	Offset	Blackout	Mean	Worst
Fixed audio-visual inference	10.21	10.16	8.50	9.74	10.21	8.63	9.45	8.50
Audio-only inference	9.02	9.02	9.02	9.02	9.02	9.02	9.02	9.02
RGI	10.48	10.48	9.23	10.10	10.49	9.26	9.91	9.23

H.3 Held-Out Visual Shifts

Table 8 reports the two held-out shifts on AV-UDiffSE with TCD-TIMIT/DEMAND, which Fig. 3 displays for fixed audio-visual inference and RGI. The wrong-speaker comparison uses the 550 utterances with an eligible donor described in App. F. Under wrong-speaker video, fixed audio-visual inference falls below audio-only inference, and RGI exceeds both as well as modality selection, quality gating, and CADS. Under sham padding, RGI retains its performance on original video, whereas modality selection and quality gating fall to the audio-only level. The padded boundary frames thus lower the input quality score without degrading enhancement, which parallels the low-resolution case in Fig. 1(b,c), where input frames readily reveal a shift that causes little degradation.

Table 8: SI-SDR improvement in dB under the held-out shifts on AV-UDiffSE with TCD-TIMIT/DEMAND.

Method	Wrong speaker	Sham padding
Fixed audio-visual inference	8.35	10.21
Audio-only inference	9.03	9.02
Modality selection	9.00	9.02
Quality gating	9.25	9.02
CADS	8.50	10.20
RGI	**9.54**	**10.48**

H.4 Quality Gating

Table 9 summarizes continuous quality gating, which Sec. 4.1 includes as an input quality baseline, with the measures of Table 1. Its Mean and Worst coincide with audio-only inference on every backbone and corpus, since the input quality score suppresses visual conditioning under all five development perturbations, which mirrors the fallback of modality selection discussed in Sec. 4.2. On original video, quality gating matches or exceeds modality selection in every combination.

Table 9: Quality gating on each backbone and corpus, measured by SI-SDR improvement in dB.

	TCD-TIMIT/DEMAND			GRID			Lombard GRID		
Backbone	Mean	Worst	Original	Mean	Worst	Original	Mean	Worst	Original
AV-UDiffSE	9.02	9.02	9.73	4.03	4.03	4.04	5.33	5.33	5.38
AV-SGMSE	12.57	12.57	12.97	9.57	9.57	9.57	8.38	8.38	8.41
AV-CA-DiffUSE	9.02	9.02	9.35	4.03	4.03	4.04	5.33	5.33	5.41

H.5 Visual Reference Size

Table 10 varies the number of visual reference clips on AV-UDiffSE with TCD-TIMIT/DEMAND while keeping the trajectory reference set and all scalar parameters fixed. Each value is averaged over original video and the five development perturbations on a fixed 180-utterance subset and over repeated reference draws, which resample the reference clips. The 500-clip setting is the visual reference size M_{I} of the main configuration in Table 6. Enhancement varies little across the tested sizes, and the variation stays within the spread across repeated reference draws at a fixed size, which is at most 0.07 dB. The representation score therefore requires only a modest number of original clips from the deployment conditions.

Table 10: Sensitivity to the size of the visual reference set on AV-UDiffSE with TCD-TIMIT/DEMAND, measured by SI-SDR improvement in dB.

Visual reference clips	Reference draws	SI-SDR improvement
50	3	9.88
100	3	9.87
200	3	9.86
500	1	9.82

H.6 No-Reference Evaluation on LRS3

Table 11 reports the no-reference measures of App. F on LRS3 with AV-UDiffSE, where the visual reference statistics of RGI are rebuilt from LRS3 clips as described in App. C.3. These measures are not comparable with the reference-based SI-SDR improvement in Table 1. RGI attains the highest DNSMOS and SQUIM-SI-SDR and matches fixed audio-visual inference in SQUIM-STOI, while quality gating remains below fixed audio-visual inference on all four measures.

Table 11: No-reference evaluation on LRS3 with AV-UDiffSE. Higher is better for all four measures.

Method	DNSMOS↑	SQUIM-STOI↑	SQUIM-SI-SDR↑	SQUIM-PESQ↑
Quality gating	3.493	0.880	11.95	1.901
Fixed audio-visual inference	3.516	0.893	15.06	2.170
RGI	3.535	0.893	15.12	2.122