# SpeechIEM: An Admissibility Criterion and Gate for Denoiser-Based Distances on Speech Latents

Anonymous authors
Paper under double-blind review

## Abstract

We present SpeechIEM, a label-free full-reference distance on self-supervised speech latents, and—the point of the paper—the criterion that decides whether such a construction is a distance at all. The construction is the Information-Estimation Metric of Ohayon et al. (2026): a denoiser trained on the data becomes a distance by comparing residuals across noise levels. Its failure mode is silent. The distance becomes a fixed rescaling of the Euclidean distance on the same representation, ordering every pair identically to it, while every training curve improves and every monotonicity plot looks healthy. We give the exact condition for that collapse—the denoiser's action on the difference integrating to a multiple of the identity—and show a scalar action to be sufficient but not necessary, which alone rules out gating on the residual's direction. Three routes reach the regime in ordinary practice, one of them an untrained preconditioned denoiser, so a gate is a mandatory step rather than a diagnostic nicety. Which diagnostic to gate on we settle by measurement over 27 denoisers from three grids, scoring each candidate against what it stands in for: the measure-weighted residual cosine rejects two metrics carrying a full advantage, the dispersion of the distance-to-Euclidean ratio rejects none and separates the classes with a factor-of-fifteen empty margin. Admissible, SpeechIEM beats the Euclidean baseline the construction itself contains by $+0.058$ $\tau_b$ across six distortion families against PESQ, with no quality labels—but the sign of that advantage depends on the reference it is scored against ($-0.012$ against ESTOI), and on twelve of the thirteen hidden states the trained denoiser adds almost nothing. Hence a reporting standard for this family of metrics: beat the zero-denoiser baseline on the same representation, pair the comparison, and name the reference quality and the corruption distribution it was measured on.

## 1 Introduction

Learned perceptual distances such as LPIPS (Zhang et al., 2018) compare signals in the feature space of a pretrained network. The Information-Estimation Metric (Ohayon et al., 2026, IEM) takes a different route: it needs only a denoiser $D(y, \sigma)$ trained on the data distribution and compares two signals $x_1, x_2$ through the squared difference of their denoiser residuals integrated over noise levels,

$$\mathrm{IEM}^2(x_1, x_2) \;=\; \int_{\gamma_{\min}}^{\gamma_{\max}} \big\| \,\varepsilon_1(\gamma) - \varepsilon_2(\gamma)\,\big\|^2 \, d\gamma, \qquad \gamma = \sigma^{-2}, \tag{1}$$

where $\varepsilon_i(\gamma) = y_i - D(y_i, \sigma)$ is the residual at the common corruption $y_i = x_i + \sigma\, w$, with one Brownian path $w$ shared by both points. Under a nonlinear denoiser it inherits the local geometry of the data manifold; on images it ranks distortions better than LPIPS-style baselines, and optimising it keeps images plausible (Ohayon et al., 2026). It is an attractive recipe: no similarity labels, one trained component, an interpretation that comes with the construction.

This paper makes that recipe usable on speech—SpeechIEM—which is where the two choices that decide whether it works at all become visible. There is no widely used pixel space for



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.

audio, so the signal space is the latent space of a self-supervised encoder such as WavLM (Chen et al., 2022), the denoiser is trained on latents, and one must pick which of the 13 hidden states; and because such a denoiser is cheap to train, it is tempting to add terms shaping the metric directly—keeping $\|\Delta\varepsilon\|$ near $\|\Delta z\|$ so it is "well scaled". Both decide whether the result is a perceptual distance at all, so the first half of this paper is about getting them right and being able to check that you did.

Throughout, the comparison that matters is not against a waveform or spectrogram distance—those are easy to beat, and beating them says nothing about the denoiser—but against the latent Euclidean distance on the same representation, exactly the IEM of a zero denoiser (Ohayon et al., 2026): the control that makes "the denoiser helped" falsifiable. Our contributions are:

1. An exact admissibility criterion (§4): for a denoiser locally linear on the pair, the metric carries no ordering information beyond latent $L_2$ exactly when its action on the difference integrates to a multiple of the identity; drop that hypothesis and the condition is still necessary, which is the form we gate on. Acting as a scalar is sufficient but not necessary, and that gap alone rules out gating on the residual's direction; three routes into the regime occur in ordinary practice, one of them an untrained preconditioned denoiser—so a monotone curve against SNR certifies nothing.
2. A gate chosen by measurement (§4.3): we score both candidate diagnostics against what they stand in for—whether the ranking differs from $L_2$'s—on 27 denoisers from three grids. The dispersion of the metric-to-$L_2$ ratio never rejects a metric with a real advantage; the residual cosine does, with a threshold inside the overlap of the two classes that moving cannot fix.
3. A protocol that discriminates (§5): ranking degradations of one utterance across six distortion families, paired against closed-form baselines on the same latents. Single-axis probes cannot separate a perceptual distance from a Euclidean one, and we measure that they do not.
4. Measured gains and their regime (§5, §6): against three reference qualities, across all thirteen hidden states, against head capacity and step budget, and in use as a training objective—where the advantage survives and where it does not.
5. SpeechIEM itself (§3) as a six-step recipe, released with the diagnostics, the probe and the run manifests.

## 2 Background: the Information-Estimation Metric

Let $x \sim p$ and $y = x+\sigma w$ with $w \sim \mathcal{N}(0, I)$. Tweedie's identity $D(y, \sigma) = y+\sigma^2\nabla_y \log p_\sigma(y)$ makes the MMSE denoiser's residual $\varepsilon = y-D(y, \sigma)$ a scaled score (Miyasawa, 1961; Vincent, 2011), and the information-estimation relations (Guo et al., 2005) are what make estimation error accumulated over noise levels an informational quantity—hence an integral rather than one denoiser output. Ohayon et al. (2026) define the IEM as equation 1 over $\gamma = \sigma^{-2}$ with a common corruption path, and show that (i) for Gaussian $p$ it equals a Mahalanobis distance, (ii) for the zero denoiser it reduces to $(\gamma_{\max}-\gamma_{\min})\|x_1-x_2\|^2$, i.e. scaled Euclidean distance, and (iii) in general it weights differences by the denoiser's sensitivity.

Two of those are load-bearing here, and both are the original work's: the zero denoiser already gives a monotone, $L_2$-like metric, so baseline and failure mode are both inside the construction; and $\gamma = \sigma^{-2}$ puts almost all the integration weight at the smallest noise levels, which decides how a diagnostic over $\sigma$ must be weighted (§4.3).

## 3 SpeechIEM: the recipe

SpeechIEM is equation 1 instantiated on speech. The definition is Ohayon et al. (2026)'s, not ours, but every choice below had to be made to obtain a distance rather than a rescaled Euclidean one, and the last step certifies that we did. Figure 1 is the whole object in one pass.

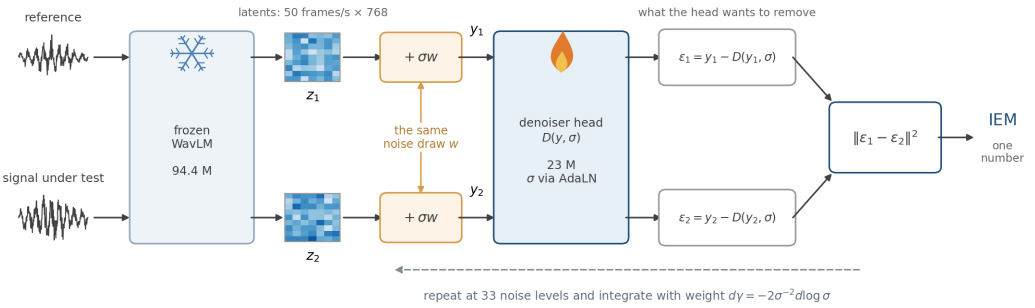


Figure 1: SpeechIEM end to end. Both signals pass through the same frozen encoder, take the same noise draw, and are denoised by the same head; the distance reads the difference of the residuals—what the head wants to remove—not of the outputs. Only the head is trained.

1. Representation. A frozen wavlm-base-plus encoder (Chen et al., 2022) exposes 13 hidden states ($\ell = 0$ the feature-projection output, $\ell = 1..12$ the transformer layers). We sweep $\ell$; the advantage over Euclidean concentrates almost entirely on $\ell = 0$ (Appendix E), which is what we report. A frame-level validity mask travels with every latent batch, so padded frames never enter a distance.
2. Denoiser. A 4-block transformer head (8 heads, dropout 0.1) on the frozen encoder, with EDM preconditioning (Karras et al., 2022): $D(y,\sigma) = c_{\text{skip}}(\sigma)\, y + c_{\text{out}}(\sigma)\, F(c_{\text{in}}(\sigma)\, y, \frac{1}{4}\log\sigma)$. Inject $\sigma$ through AdaLN (Peebles & Xie, 2023; Perez et al., 2018)—per-block scale, shift and gate—not one additive embedding: a $\sigma$-independent map is linear on the pair by construction, so this buys admissibility more reliably than accuracy (§4).
3. Noise schedule, calibrated per layer. WavLM hidden states have RMS 0.20–0.34, so an absolute schedule such as $\sigma \in [1, 1000]$ puts every sample at $-14$ to $-74$ dB latent SNR and the denoiser never sees a mildly noisy latent. Calibrate $\sigma_{\text{data}}$ on the layer first, take the EDM range relative to it, and store it in the checkpoint so training and evaluation share one schedule.
4. Objective: plain denoising, nothing else. Use the EDM denoising loss on clean latents. Do not restore noisy-speech latents "as if clean" (that puts degraded speech into the learned manifold, and the distance is to the clean one), and add no term aligning $\Delta\varepsilon$ with $\Delta z$ to make the metric "well scaled": such terms have the degenerate metric as their fixed point (§4).
5. Estimators. Evaluate equation 1 on an endpoint-inclusive grid uniform in $\log\sigma$ with the exact Jacobian $d\gamma = -2\sigma^{-2} d\log\sigma$. As a training objective, use the unbiased Monte-Carlo estimator uniform in $\log\sigma$ with importance weight $2\log(\sigma_{\max}/\sigma_{\min})\,\sigma^{-2}$.
6. Gate before you report. Run the dispersion diagnostic of §4.3 and refuse the checkpoint if it fails—what makes the previous five a distance rather than a reparameterised $L_2$, at one forward pass over a few dozen pairs.

Training hyperparameters, data and compute are in Appendix A.

## 4 Admissibility: when the IEM carries information beyond latent $L_2$

### 4.1 An exact criterion, and three routes that violate it

Under the common corruption the noise cancels in the residual difference, $\Delta\varepsilon(\sigma) = \Delta z - (D(y_1,\sigma) - D(y_2,\sigma))$. If the denoiser is locally linear on the pair, $D(y_1,\sigma) - D(y_2,\sigma) =$

$A(\sigma)\Delta z$, then $\Delta\varepsilon = (I - A(\sigma))\Delta z$ and

$$\mathrm{IEM}^2 = \Delta z^\top \Big[ \int (I - A(\sigma))^\top (I - A(\sigma))\, d\gamma \Big] \Delta z. \tag{2}$$

This is a Mahalanobis form, and $M = \int (I - A(\sigma))^\top (I - A(\sigma))\, d\gamma$ is the only thing about the denoiser that survives into the metric. It collapses onto latent $L_2$ exactly when $M$ is a multiple of the identity (Figure 2).

Proposition 1 (Exact collapse criterion). Let the denoiser be locally linear on the pairs, $D(y_1, \sigma) - D(y_2, \sigma) = A(\sigma)\,\Delta z$. Then $\mathrm{IEM}(x_1, x_2) = c\,\|\Delta z\|$ with $c$ independent of the pair—so IEM and latent $L_2$ induce the same ordering on every pair—if and only if $M = c^2 I$ on the span of the differences $\Delta z$. A scalar action $A(\sigma) = a(\sigma)I$ is sufficient, with $c^2 = \int (1 - a(\sigma))^2 d\gamma$; in particular EDM preconditioning with a zero network output gives $D(y, \sigma) = c_{\mathrm{skip}}(\sigma)\, y$, hence $a = c_{\mathrm{skip}}$.

Scalar action is not necessary, and that gap is where the two candidate diagnostics of §4.3 part company. Take $A(\sigma) = b(\sigma)J$ with $J^\top = -J$, $J^\top J = I$ — a pure rotation, as non-scalar as an action gets; such a $J$ exists in every even dimension, and these latents have 768. Then $(I - bJ)^\top (I - bJ) = (1 + b^2)I$, so the metric is a fixed rescaling of latent $L_2$ and orders every pair identically to it—while the residual cosine reads $(1 + b^2)^{-1/2} < 1$ and calls the denoiser nonlinear. A diagnostic on the residual's direction can therefore admit a metric that is exactly $L_2$; the dispersion reads zero on it, being about the quantity the criterion is about. §4.3 reaches the same verdict by measurement.

Dropping the linearity hypothesis costs exactly one direction of the criterion, and it is worth stating which, because the gate of §4.3 is a test on the ratio rather than on $M$.

Proposition 2 (What a ratio test can and cannot see). For any denoiser, if $\mathrm{IEM}(x_1, x_2) = c\,\|\Delta z\|$ with $c$ independent of the pair, then IEM and latent $L_2$ order every pair identically. Conversely, identical ordering on a set of pairs implies only that IEM is a strictly increasing function of $\|\Delta z\|$ there, not that the function is linear. A constant ratio is therefore sufficient for the collapse and non-constancy only necessary for the ordering to differ: a metric that is a nonlinear monotone function of $\|\Delta z\|$ has a dispersed ratio and still carries no ordering information.

We gate on the necessary condition, and this is why §4.3 scores it against the ordering itself over 27 denoisers instead of arguing from the definition: what the measurement adds is that the gap the proposition leaves open stays empty in practice — no checkpoint we measured passed the ratio test while ranking identically to latent $L_2$, and the two the probe cannot separate from it at 64 utterances carry small positive advantages rather than none (Appendix C).

The sufficient direction of Proposition 1 is a corollary of item (ii) of §2. What is new is the criterion—which names what to measure—and that the regime is no corner case. Three routes reach it, all three in ordinary practice:

(a) Doing nothing. By the last clause of the proposition an untrained preconditioned denoiser is already scalar-acting, with $a = c_{\mathrm{skip}}$. A plot showing that "the IEM falls as SNR rises" is therefore not evidence of anything: an untrained head produces exactly that curve.

(b) Asking for a well-scaled metric. A natural-looking auxiliary loss keeps the metric calibrated to the latent difference: $\mathcal{L}_\delta = (\|\Delta\varepsilon\| - \|\Delta z\|)^2$ and $\mathcal{L}_\angle = 1 - \cos(\Delta\varepsilon, \Delta z)$. Together they are minimised exactly when $\Delta\varepsilon = \Delta z$, i.e. when $D(y_1) - D(y_2) = 0$ on the pair, and then $\mathrm{IEM}^2 \to (\gamma_{\max} - \gamma_{\min})\|\Delta z\|^2$: the zero-denoiser baseline. Such terms train the metric to be latent $L_2$ while every training curve improves.

(c) Conditioning the denoiser weakly on $\sigma$. A map with little $\sigma$-dependence is linear on the pair by construction, so how $\sigma$ enters the architecture is an admissibility choice.

The proposition and its corollaries are verified numerically in the released test suite: constant ratio to five digits and identical argsort for a scalar-acting denoiser, a channel-reweighting one correctly not flagged, and the rotational case above admitted by the cosine while the dispersion rejects it.

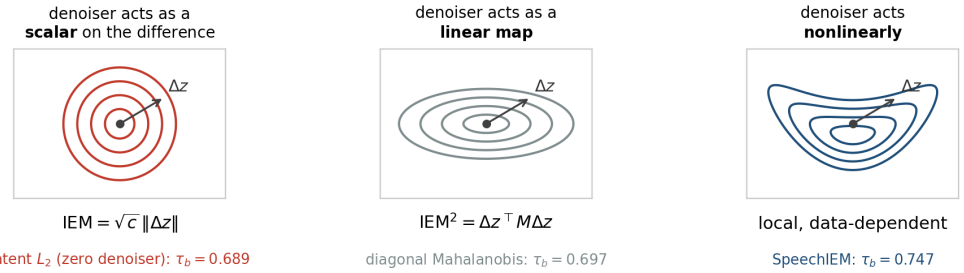


Figure 2: Iso-distance contours of the metric a denoiser induces; points on one contour are indistinguishable to it. Scalar action gives circles—Euclidean up to a constant, ordering every pair identically; linear action a fixed quadratic form; only nonlinear action bends with the data. Numbers measured: cross-family $\tau_b$ against PESQ, 200 utterances.

Nothing in this section is about speech. Both propositions are statements about the map $\Delta z \mapsto \Delta\varepsilon$ and mention neither the encoder nor the signal domain, so the criterion and the gate transfer unchanged to an IEM built on a preconditioned denoiser over images, video or any other learned latent space; what is speech-specific is only which of the three routes one lands on, which §4 measures here.

### 4.2 A declared grid, and what it settles

Whether routes (b) and (c) bite is empirical, so we answer it on hidden state 0 with a grid whose arms differ only in objective (denoise-only versus denoise + align), $\sigma$-injection head (additive versus AdaLN) and step budget (500 as a short probe against the recipe's 3000), every other override byte-identical. Table 1 puts the two diagnostics of §4.3 beside what they stand in for—the cross-family advantage over latent $L_2$ of §5, on one probe for all eight cells.

The outcome is cleanly bimodal. Two cells—the alignment-augmented objective at the short budget, either head—have essentially no advantage: $+0.0000$ with an interval of zero width for the AdaLN one, whose metric ranks identically to latent $L_2$ ($\tau_b = 0.6972$ for both), and $+0.0001$ ($[-0.001, +0.001]$) for the additive one. The collapse is exact, as the proposition requires, and both reproduce at a second seed (Appendix D). The other six land between $+0.0436$ and $+0.0587$, a spread of 0.015 against per-cell interval widths of $\approx$0.013; those intervals are against latent $L_2$ rather than paired between cells, so no two admissible cells are resolved here. What the grid establishes is that once a metric is admissible at all, the recipe barely matters: we expected AdaLN to buy accuracy, and what it buys is admissibility, more reliably.

### 4.3 Why the weighting matters, and a gate that works

Every denoiser linearises at large $\sigma$, where the input is dominated by the shared noise, so an unweighted average of $\cos(\Delta\varepsilon, \Delta z)$ over $\sigma$ is uninformative. The $\gamma$ measure resolves this: over 4.6 decades the Jacobian $2\sigma^{-2}$ puts $\approx 0.87$ of the weight on the smallest of the nine levels the cosine is read on and $\approx 10^{-9}$ on the largest (Figure 5). Such weights make a diagnostic quoted without its grid meaningless, so we state ours: nine levels for the cosine, 65 for the dispersion and mass profile, 33 for the probe of §5.

The cosine measures direction, but a ranking follows magnitude, and the two can disagree: a per-example scalar $\Delta\varepsilon = a(x)\,\Delta z$ gives $\cos = 1$ while the ratio $\mathrm{IEM}/\|\Delta z\| = |a(x)|$ varies across pairs, and that variation is information a ranking can use. So we gate instead on its dispersion, $\mathrm{CV}[\,\mathrm{IEM}(x_1, x_2)/\|\Delta z\|\,]$ over pairs spanning all distortion families and severities, which is zero exactly when the metric is a fixed rescaling of latent $L_2$ — by Proposition 2 a necessary condition for the ordering to differ rather than an equivalent one, hence measured against the ordering below rather than asserted. Alongside it we report a mass profile: the smallest contiguous share of the $\log\sigma$ range holding 90% of $\mathrm{IEM}^2$.

Table 1: The degeneracy grid: hidden state 0, arms differing only in objective, $\sigma$-injection head and step budget, ordered by the quantity the diagnostics should predict. CV and cos come from 108 pairs on nine and 65 levels respectively; $\Delta\tau_b$ is the paired cross-family advantage over scaled latent $L_2$ against PESQ on 64 utterances. The last two columns give each diagnostic's verdict at its threshold, ! marking a disagreement with that advantage.

| objective | head | steps | CV % | cos | $\Delta\tau_b$ vs $L_2$ | CV gate | cos gate |
|---|---|---|---|---|---|---|---|
| denoise + align | AdaLN | 500 | 0.00 | 0.999 | +0.0000 | degen. | degen. |
| denoise + align | additive | 500 | 0.34 | 0.970 | +0.0001 | degen. | degen. |
| denoise only | additive | 3000 | 6.77 | 0.627 | +0.0436 | adm. | adm. |
| denoise + align | AdaLN | 3000 | 14.14 | 0.995 | +0.0438 | adm. | degen. ! |
| denoise only | additive | 500 | 5.85 | 0.662 | +0.0486 | adm. | adm. |
| denoise + align | additive | 3000 | 7.70 | 0.962 | +0.0505 | adm. | degen. ! |
| denoise only | AdaLN | 3000 | 7.55 | 0.074 | +0.0520 | adm. | adm. |
| denoise only | AdaLN | 500 | 7.09 | 0.423 | +0.0587 | adm. | adm. |

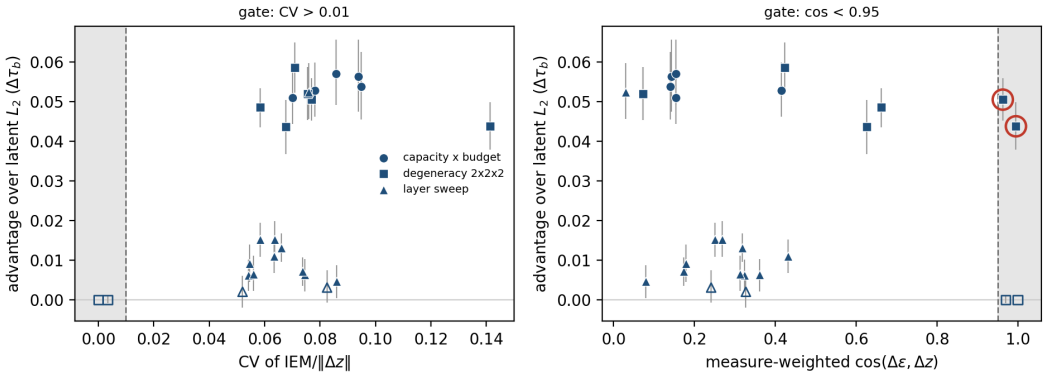


Figure 3: Each candidate diagnostic against what it stands in for, over 27 checkpoints from three grids. Shaded: the threshold's reject strip; hollow: cells the probe cannot separate from latent $L_2$; red rings: false rejections. Left, the dispersion's classes are separated by an empty margin (no cell between 0.34% and 5.2%), so any threshold across a factor of fifteen agrees. Right, the cosine's threshold falls inside the overlap—a collapsed cell at 0.970 beside an admissible one at 0.962—which moving it cannot fix (Appendix C).

Table 1 settles which diagnostic to trust, and the answer is not the obvious one. Thresholding the dispersion at 1% classifies all eight cells correctly; the weighted cosine at 0.95 gets six of eight, calling the alignment-augmented objective at 3000 steps degenerate under both heads (cos = 0.995 and 0.962) where both carry a full advantage (+0.0438, +0.0505). With budget that objective keeps aligning the residual difference's direction while letting its magnitude depend on the example—and an ordering follows magnitude. The cosine is not useless as a description (over the 27 cells below it ranks the advantage at −0.34 against the dispersion's +0.59), but it is the threshold that fails, and a threshold is what a pipeline acts on. Joining every checkpoint here with both a gate reading and a probe—27 cells from three grids—keeps that verdict: the dispersion rule is right on 25 of 27 and never rejects a real advantage, the cosine on 23 and rejects two (Appendix C).

The $\sigma$-conditioning axis behaves as route (c) predicts: at equal budget AdaLN yields a markedly more nonlinear denoiser than a single additive embedding (cos 0.074 against 0.627 under the plain objective at 3000 steps) and spreads the metric's mass over more of the schedule (2.01 of 4.6 decades against 1.73).

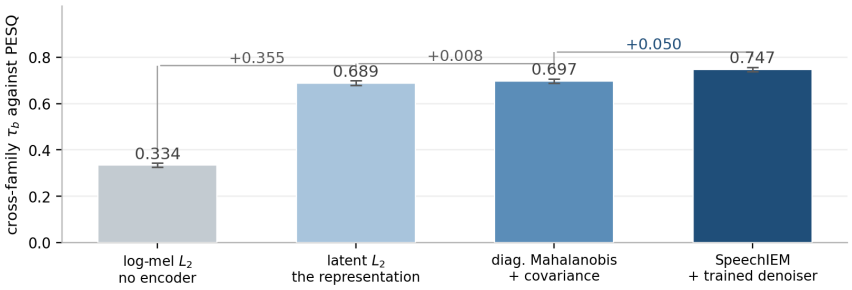


Figure 4: Where the agreement comes from, one change at a time on the same pairs: the representation is worth +0.355, and on top of it the best linear reweighting +0.008 against the trained denoiser's +0.050. Bootstrap intervals over utterances; full table in Table 8.

## 5 A probe that can separate SpeechIEM from latent $L_2$

Why single-axis probes cannot work. Varying only the severity of one degradation, every sensible distance is monotone, so a correlation with quality cannot distinguish a perceptual distance from a Euclidean one. Table 9 (appendix) makes this concrete on the same probe data: within each of the six families the paired difference $\tau_b(\text{IEM}) - \tau_b(L_2)$ lies between $-0.001$ and 0.000 with intervals containing zero, while every latent distance correlates with PESQ at 0.83 to 0.98 and log-mel $L_2$ at 0.80 to 0.90; pooling the families gives +0.050 ($[+0.047, +0.054]$). The information is in the comparison between axes.

Design. Each held-out utterance gets six distortion families at five severities each—additive noise using the utterance's own recorded noise, reverberation, clipping, band limitation, $\mu$-law quantisation, packet loss—over documented severity ranges (Appendix A); every operator preserves length and alignment and is deterministic given a seed. Per degraded clip we record SpeechIEM, three closed-form baselines—scaled latent $L_2$ (the exact zero-denoiser IEM), a diagonal Mahalanobis distance whitened by clean training-latent variances, and log-mel $L_2$ (no encoder)—and PESQ (Rix et al., 2001), ESTOI (Jensen & Taal, 2016) and SI-SDR (Le Roux et al., 2019) as reference qualities.

Statistic. Kendall's $\tau_b$ (Kendall, 1945) against the reference quality, split into within-family pairs (same family, different severity—any distance scores here) and cross-family pairs restricted to one utterance (a clipped versus a reverberant version of the same speech), which is the question a perceptual distance exists to answer. Both distances see the same pairs, so we bootstrap the difference, which is what makes the comparison decidable (Appendix A).

Results. On 200 held-out utterances (74 975 cross-family pairs) the admissible layer-0 metric agrees within-family at $\geq 0.96$ for every distance and family—single-axis probes are uninformative—while across families the picture is consistent and modest (Table 8): against PESQ SpeechIEM beats scaled latent $L_2$ by +0.058 $\tau_b$ ($[+0.054, +0.063]$) and diagonal Mahalanobis by +0.050; against SI-SDR by +0.026 and +0.023; against ESTOI it is worse than both ($-0.012$, $-0.011$), moving toward perceived quality and away from intelligibility—the trade a perceptual distance should make. All three latent distances beat log-mel $L_2$ by +0.32 to +0.41: the representation carries most of the agreement, and of what can be added on top of it the denoiser is worth six times the best linear reweighting (Figure 4). Appendix F runs the same paired comparison against crowdsourced MOS instead of a model reference, and the perceived-quality side of that trade survives there on recorded conditions.

Hidden state 0 was chosen on a 64-utterance probe, and the probe walks the test index by position, so those 64 are the first of these 200 — the selected quantity would otherwise be reported partly on the data that selected it. Split accordingly, the advantage against PESQ is +0.0576 ($[+0.052, +0.063]$) on the 136 utterances that took no part in the choice against +0.0600 on the 64 that did, and all three references replicate on the held-out part with overlapping intervals (Appendix E).

Not a matter of scale. The obvious reply—a four-block head at 3000 steps is simply too small—is testable, and the answer is flat: sweeping head depth against step budget with all else fixed, six cells land between +0.0510 and +0.0571 against interval widths of 0.014–0.018, and quadrupling depth buys +0.002 at the short budget and nothing at the long one (Appendix G). Within a factor of four in capacity and three in budget this is a property of the construction on these latents, not of our compute.

## 6 The metric as a training objective

A better ranker is not automatically a better gradient. In a design where only the loss varies (Appendix I, with the referenced denoiser gated first, since against a degenerate one the comparison is $L_2$ vs $L_2$), SpeechIEM replacing waveform MSE loses in the waveform domain ($-0.44$ dB SI-SDR, $p = 0.046$) and is level on log-spectral distance, while added to a waveform anchor it reduces log-spectral distance by 0.127 at no waveform cost ($p = 0.077$)—the one-sided pattern a phase-insensitive latent objective predicts. Six paired cells admit no $p$ below 0.031, so both are directions rather than settled gains.

## 7 Related work

Full-reference audio distances. The signal-domain standards are hand-designed—PESQ (Rix et al., 2001) and ViSQOL (Chinen et al., 2020) model the auditory periphery rather than learn from data—and we use PESQ throughout as a reference quality. Among learned ones, DPAM (Manocha et al., 2020) and CDPAM (Manocha et al., 2021) need just-noticeable-difference labels; SpeechBERTScore (Saeki et al., 2024) compares self-supervised features token-wise and needs none, and NOMAD (Ragano et al., 2024) needs no matching reference. The IEM is in the label-free half but reaches it differently: nothing is fitted to similarity judgements and the only trained component is a denoiser of the data. We do not rank ourselves against either half here. Our claims are about when the construction is a distance at all, and a comparison across published distances changes the representation together with the metric—so every comparison in this paper is against the zero-denoiser baseline on the same latents, which is the only one that isolates the denoiser from the encoder.

Non-intrusive predictors. MOSNet (Lo et al., 2019), DNSMOS (Reddy et al., 2021), NISQA (Mittag et al., 2021) and UTMOS (Saeki et al., 2022) regress MOS directly and need no reference, a framing the VoiceMOS challenges (Huang et al., 2022) made standard. They are a different tool: they cannot say how far two signals are from each other, which is what a training loss or a retrieval criterion requires.

Denoisers as priors, and how metrics get validated. That the MMSE denoiser encodes the score (Miyasawa, 1961; Vincent, 2011) underpins diffusion models (Karras et al., 2022), plug-and-play priors, and score-based speech enhancement (Richter et al., 2023), where a denoiser on speech is trained to generate; Ohayon et al. (2026) make the same object a metric on images instead. We carry it over to learned speech representations with what it needs there: the criterion for when it is a distance at all, a gate that enforces it, and a measurement of how much advantage survives a change of corruption distribution. The image community validates perceptual distances against forced-choice data (Zhang et al., 2018; Ding et al., 2022); speech has no comparable JND corpus with references, so we lean on paired statistics against three reference qualities and on the zero-denoiser baseline on the same representation, and we are explicit (§8) that this leaves the absolute perceptual claim untested.

## 8 Limitations

Our denoisers are small (a 4-block head on a frozen encoder) at 3000 steps, which bounds our two kinds of claim differently. The criterion and its gate are scale-free and domain-free: statements about the map $\Delta z \mapsto \Delta\varepsilon$ that name neither the encoder nor the signal domain,

verified in closed form, so they apply to any IEM built on a preconditioned denoiser and a larger denoiser only makes the gate more useful. The measured gains are not: more capacity, a longer schedule, or a prior trained on speech directly might raise the cross-family advantage, and the sign already moves with the reference it is scored against (§5). What we claim at this scale is the conditionality itself, which no capacity removes. The axis we did not vary is the one our own measurements point at: every distance here reads wavlm-base-plus, the representation accounts for $+0.32$ to $+0.41$ $\tau_b$ against log-mel $L_2$ while the denoiser adds $\approx 0.05$ on the one hidden state where it adds anything (Appendix E), so a SpeechIEM on wavlm-large latents is the experiment most likely to move the absolute numbers and it is not here.

The larger limit is the reference. Our primary one, PESQ, is itself a metric, so "SpeechIEM agrees with PESQ better than latent $L_2$ does" is in part a statement about PESQ; the criterion, the gate and the discriminating protocol do not depend on it, since all three are about whether the metric differs from latent $L_2$ at all. Against human ratings we can say where the advantage holds and where it does not: on the recorded live and anchor conditions of the NISQA corpus (Mittag et al., 2021) it survives against crowdsourced MOS, and on simulated conditions it is carried by whether the condition contains a codec (Appendix F). That is a narrower claim than a perceptual one — a 480-file positive result on one corpus does not establish that SpeechIEM tracks perception in general, and a purpose-built listening test with references is still the evaluation this construction needs. Two further limits: the URGENT train/held-out boundary is speaker-disjoint for only $\approx 62\%$ of utterances (Common Voice is indexed per clip), and our probe partition is a grouped hold-out of the same simulation, not an independent run (Appendix A).

## 9 Conclusion

A denoiser-based perceptual distance is only as good as its denoiser's nonlinearity, and on self-supervised speech latents that nonlinearity is easy to lose silently. SpeechIEM is therefore a recipe whose last step is a gate: an exact criterion for the collapse onto a rescaled latent $L_2$, and the measurement picking the diagnostic to gate on—the dispersion of IEM$/\|\Delta z\|$, which over 27 denoisers never rejected a real advantage. Admissible, it beats the baseline the construction contains across six distortion families against PESQ, with no quality labels—but only on one of thirteen hidden states, and with a sign that moves with the reference it is scored against. Both the advantage and its sign are therefore regime-dependent, so the recipe ships with a reporting standard: beat the zero-denoiser baseline on the same representation, pair the comparison, name the reference and the distribution.

## References

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6):1505–1518, 2022.

Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. ViSQOL v3: An open source production ready objective speech and audio metric. In International Conference on Quality of Multimedia Experience (QoMEX), 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5):2567–2581, 2022.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. IEEE Transactions on Information Theory, 51 (4):1261–1282, 2005.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The VoiceMOS challenge 2022. In Interspeech, 2022.

ITU-T. Subjective evaluation of speech quality with a crowdsourcing approach. Recommendation P.808, 2018.

Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11):2009–2022, 2016.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Advances in Neural Information Processing Systems, 2022.

Maurice G. Kendall. The treatment of ties in ranking problems. Biometrika, 33(3):239–251, 1945.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In ICASSP, 2019.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. MOSNet: Deep learning-based objective assessment for voice conversion. In Interspeech, 2019.

Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J. Bryan, Gautham J. Mysore, and Zeyu Jin. A differentiable perceptual audio metric learned from just noticeable differences. In Interspeech, 2020.

Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In ICASSP, 2021.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In Interspeech, 2021.

Koichi Miyasawa. An empirical Bayes estimator of the mean of a normal population. Bulletin of the International Statistical Institute, 38:181–188, 1961.

Guy Ohayon, Pierre-Etienne H. Fiquet, Florentin Guth, Johannes Ballé, and Eero P. Simoncelli. Learning a distance measure from the information-estimation geometry of data. In International Conference on Learning Representations (ICLR), 2026. arXiv:2510.02514.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In AAAI Conference on Artificial Intelligence, 2018.

Alessandro Ragano, Jan Skoglund, and Andrew Hines. NOMAD: Unsupervised learning of perceptual embeddings for speech enhancement and non-matching reference audio quality assessment. In ICASSP, 2024.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In ICASSP, 2021.

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

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In ICASSP, 2001.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In Interspeech, 2022.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics. In Interspeech, 2024.

Pascal Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7):1661–1674, 2011.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.

Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, Zhaoheng Ni, Anurag Kumar, Jan Pirklbauer, Marvin Sach, Shinji Watanabe, Tim Fingscheidt, and Yanmin Qian. URGENT challenge: Universality, robustness, and generalizability for speech enhancement. In Interspeech, 2024.

## A Data, splits and compute

Paired statistic. SpeechIEM and a baseline are evaluated on the same pairs, so their sampling errors are strongly correlated and we bootstrap the difference $\tau_b(\text{IEM}) - \tau_b(\text{baseline})$, resampling whole utterances identically for both (2000 resamples). Separate intervals would be several times wider than the paired one and could not decide the comparison.

Denoiser training. AdamW, lr $2 \cdot 10^{-4}$ with cosine decay, batch 8, 4 s random crops, 3000 steps over 24k utterances of the URGENT 2026 Track-1 training simulation (Zhang et al., 2024) unless a sweep says otherwise; one denoiser per hidden state under identical settings. The shared Brownian path is the exact reverse-time conditional, generated one noise level at a time.

Distortion families. Severity $t \in [0, 1]$ maps onto a documented range per family: additive noise 30–0 dB SNR (the utterance's own recorded noise, rescaled), reverberation $\text{RT}_{60}$ 0.15–1.2 s, clipping quantile 0.999–0.5, band limitation cutoff 7–1 kHz, $\mu$-law quantisation 8–2 bits, and packet loss 1–30%. The released probe also carries an Opus family at $32 \to 6$ kbps with its algorithmic delay compensated so the alignment contract holds, which no result here reports. Every operator preserves length and dtype and is deterministic given a seed.

Speech. Denoisers are trained on the URGENT 2026 Track-1 training simulation (Zhang et al., 2024): 436 240 noisy/clean pairs, from which we index a fixed pool of 24 000 utterances so every arm of every grid draws the same speech in the same order. Probes read a held-out partition of the same simulation. Two properties of that split have to be stated rather than assumed. It is grouped by speaker where the speaker is derivable from utt2spk (VCTK, EARS, LibriTTS, LibriVox/DNS, MLS, Elderly), which covers ≈62% of the training partition; the remaining ≈38% is Common Voice, indexed per clip, so for those rows the boundary is per-utterance rather than speaker-disjoint. And grouping on the clean file would protect nothing: the simulation has one clean recording per mixture (436 240 of 436 240 distinct), so only the speaker groups anything. Our held-out probe set is therefore a grouped hold-out of the training simulation, not an independent simulation run.

Compute. One RTX A4000 (16 GB). A denoiser is ∼35 min at 3000 steps; the cross-family probe is ∼1 h per checkpoint at 64 utterances × 6 families × 5 severities with 33 quadrature points, with reference metrics cached across checkpoints since they do not depend on one. The enhancer comparison of §6 is 6 runs of $8 \times 256$ steps.

## B Diagnostics per hidden state

Table 2 lists both diagnostics for every checkpoint we measured. Under the denoise-only AdaLN recipe all thirteen hidden states are admissible, with the ratio's coefficient of variation between 5.2 and 8.6% (so the layer comparison of Appendix E is between admissible

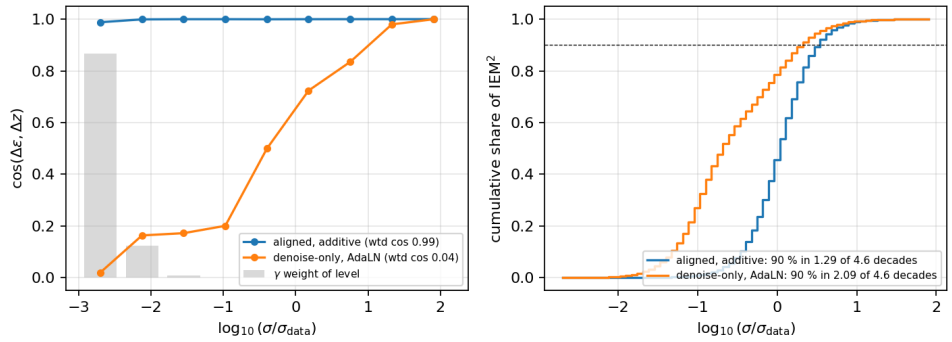


Figure 5: Layer 0, two training recipes. Left: $\cos(\Delta\varepsilon, \Delta z)$ per noise level, with the $\gamma$-weights of the integration grid behind (grey). The aligned/additive denoiser is linear on the pair at every level, including the smallest $\sigma$ where $\approx$0.87 of the weight sits; the denoise-only AdaLN one is strongly nonlinear there and linearises only where the weight is negligible. Right: cumulative share of $\text{IEM}^2$ over $\log \sigma$; the degenerate metric holds 90% of its mass in 1.29 of 4.6 decades, the admissible one in 2.09.

metrics throughout). The least linear residuals are at the two extremes (weighted cos of 0.030 at hidden state 0 and 0.080 at 12) and the middle layers are markedly more linear (0.31–0.43), which is worth keeping in mind when reading a layer ranking. The upper block is an earlier alignment-augmented sweep at a 500-step budget: its weighted cosines were 0.968–0.991 on every layer, and the dispersion could not be measured on those checkpoints, so their verdict rests on the cosine alone—which §4.3 shows is not a sound basis for one. That gap is what the declared grid of §4 was built to close, and every claim in the main text rests on the grid rather than on this block.

## C The gate outside its own grid

Table 1 validates the two candidate gates on the eight cells of the declared grid, which is one hidden state, one seed and two recipes. Every checkpoint in this repository that carries both a gate reading and a cross-family probe can be joined the same way, on the checkpoint path rather than on a parsed name, and that gives 27 cells from three grids built for different purposes: the $2 \times 2 \times 2$ above, the 13-layer sweep (13 representations rather than 13 recipes) and the capacity sweep. The criterion a gate is judged against is the one it stands in for—whether the paired interval of $\tau_b(\text{IEM}) - \tau_b(L_2)$ on the same pairs excludes zero.

The pooled verdict (Table 3) agrees with the grid and locates the difference between the two rules precisely. Four of the 27 cells are not separable from latent $L_2$: the two collapsed cells of the grid (+0.0000 and +0.0001) and hidden states 5 and 11, whose advantages are +0.0020 ($[-0.002, +0.006]$) and +0.0031 ($[-0.001, +0.007]$)—too small for 64 utterances to resolve rather than absent. Both rules admit those two, so they are wrong in the same way and the difference between them is entirely in false rejections: none for the dispersion, two for the cosine, both of them the alignment-augmented objective at 3000 steps carrying +0.0438 and +0.0505.

Figure 3 shows why the two rules behave differently, and it is not that one threshold is better tuned. The dispersion separates the classes with an empty margin around its threshold—no cell lands between 0.34% and 5.2%, so any threshold in a factor of fifteen gives the same answer—while the cosine has to cut through a region where a collapsed metric ($\cos = 0.970$) and an admissible one ($\cos = 0.962$) sit next to each other. A gate whose threshold falls inside the overlap of the two classes cannot be fixed by moving it. The column that carries the claim also does not depend on the reference: against SI-SDR and ESTOI the cosine rejects the same two cells and the dispersion none, while the number of cells the probe

Table 2: Admissibility diagnostics per checkpoint: measure-weighted $\cos(\Delta\varepsilon, \Delta z)$, the coefficient of variation of IEM/$\|\Delta z\|$ in per cent, the ratio of head output to skip contribution (which separates a degenerate metric from an untrained head), and the contiguous share of the 4.6-decade $\log\sigma$ range holding 90% of IEM$^2$. A dash in the CV column marks a report that predates that diagnostic.

| set | layer | head | wtd. cos | CV % | head/skip | mass | adm. |
|---|---|---|---|---|---|---|---|
| aligned/add., 500 st. | 0 | additive | 0.990 | – | 3.60 | 1.29 | – |
| | 1 | additive | 0.979 | – | 4.07 | 1.29 | – |
| | 2 | additive | 0.969 | – | 4.02 | 1.29 | – |
| | 3 | additive | 0.968 | – | 3.25 | 1.29 | – |
| | 4 | additive | 0.976 | – | 3.35 | 1.29 | – |
| | 5 | additive | 0.973 | – | 3.42 | 1.29 | – |
| | 6 | additive | 0.979 | – | 2.81 | 1.29 | – |
| | 7 | additive | 0.982 | – | 2.96 | 1.29 | – |
| | 8 | additive | 0.986 | – | 3.12 | 1.29 | – |
| | 9 | additive | 0.986 | – | 3.28 | 1.29 | – |
| | 10 | additive | 0.984 | – | 3.48 | 1.29 | – |
| | 11 | additive | 0.980 | – | 5.63 | 1.37 | – |
| | 12 | additive | 0.991 | – | 4.07 | 1.29 | – |
| denoise/AdaLN | 0 | adaln | 0.030 | 7.60 | 4.60 | 2.01 | yes |
| | 1 | adaln | 0.179 | 5.47 | 4.66 | 1.94 | yes |
| | 2 | adaln | 0.251 | 5.83 | 4.44 | 1.87 | yes |
| | 3 | adaln | 0.269 | 6.36 | 3.86 | 1.80 | yes |
| | 4 | adaln | 0.313 | 5.60 | 3.98 | 1.73 | yes |
| | 5 | adaln | 0.327 | 5.21 | 3.56 | 1.73 | yes |
| | 6 | adaln | 0.361 | 7.45 | 3.27 | 1.73 | yes |
| | 7 | adaln | 0.325 | 5.42 | 3.39 | 1.73 | yes |
| | 8 | adaln | 0.431 | 6.35 | 3.44 | 1.80 | yes |
| | 9 | adaln | 0.318 | 6.61 | 3.68 | 1.80 | yes |
| | 10 | adaln | 0.174 | 7.37 | 4.09 | 1.94 | yes |
| | 11 | adaln | 0.242 | 8.25 | 4.80 | 1.73 | yes |
| | 12 | adaln | 0.080 | 8.61 | 4.73 | 1.94 | yes |
| budget | 0 | adaln | 0.402 | 7.43 | 4.39 | 1.87 | yes |
| | 0 | adaln | 0.156 | 9.33 | 4.65 | 2.01 | yes |
| x22 | 0 | additive | 0.633 | 6.80 | 4.08 | 1.65 | yes |
| | 0 | adaln | 0.991 | 6.87 | 3.86 | 1.44 | yes |
| | 0 | additive | 0.967 | 8.99 | 3.86 | 1.29 | yes |

can separate from latent $L_2$ at all falls from 23 to 22 and then to 12—which says what a reference resolves, not what a gate does.

Table 3: Both candidate gates over the 27 joined cells. "Correct" counts cells where the verdict matches whether the probe can separate the metric from latent $L_2$; the two columns after it split the errors by kind. The two rules share their two admissions (hidden states 5 and 11, where the probe lacks power at 64 utterances), so only the false-rejection column distinguishes them—and that is the one a pipeline acts on.

| gate | rule | correct | rejects a real advantage | admits an unresolved cell | Spearman with $\Delta\tau_b$ |
|---|---|---|---|---|---|
| dispersion of IEM/$\|\Delta z\|$ | CV $> 1\%$ | 25/27 | 0 | 2 | $+0.59$ |
| residual cos | cos $< 0.95$ | 23/27 | 2 | 2 | $-0.34$ |

## D The grid at a second training seed

The grid of §4 is one seed, and two of its cells carry claims a single realisation cannot settle: the exact collapse (denoise + align with AdaLN at 500 steps, whose metric ranks identically to latent $L_2$) and the cell where the residual-cosine gate is wrong (cos = 0.995 at 3000 steps

with a full advantage). We therefore re-ran those two cells and one admissible anchor at seed = 2, everything else byte-identical, and probed them on the same 64 utterances and the same degradations, so the two seeds are compared on identical pairs (Table 4).

What we expected was asymmetric, which is what made the check worth running, and that is what came out. The collapse is structural—a scalar-acting denoiser gives a fixed rescaling of latent $L_2$ whatever the seed—so that cell should land in the same place rather than merely near it, and it does: at seed 2 the metric again ranks identically to latent $L_2$ ($\tau_b = 0.6972$ for both, a paired difference of +0.0000 with an interval of zero width), with the same 0.00% dispersion and the same 1.29 of 4.6 decades of mass. The cell that embarrasses the cosine gate reproduces too, which is the point of including it: cos = 0.995 again, above the threshold, while the advantage is +0.0448 against +0.0438 at seed 1—so the gate's failure is a property of what the alignment objective learns with budget, not of one run. The admissible anchor moves as a measured quantity should, within its interval (+0.0480 against +0.0520; the two intervals overlap). Both gates return the same verdict at both seeds in all three cells.

Table 4: The decisive cells of the degeneracy grid at two training seeds, as seed 1 / seed 2. Diagnostics and the paired cross-family advantage over scaled latent $L_2$ against PESQ, on identical probe pairs. The last column is whether both gates return the same verdict at both seeds.

| cell (objective, head, steps) | CV % | cos | $\Delta\tau_b$ vs $L_2$ | same verdict |
|---|---|---|---|---|
| denoise + align, AdaLN, 500 | 0.00 / 0.00 | 0.999 / 0.999 | +0.0000 / +0.0000 | yes |
| denoise + align, AdaLN, 3000 | 14.14 / 13.22 | 0.995 / 0.995 | +0.0438 / +0.0448 | yes |
| denoise only, AdaLN, 3000 | 7.55 / 7.16 | 0.074 / 0.028 | +0.0520 / +0.0480 | yes |

## E Which hidden state the denoiser helps on

We trained one denoiser per WavLM hidden state under identical settings (3000 steps, 24k utterances, batch 8, 4 s crops; ∼35 minutes each on one RTX A4000) and probed all thirteen with the six-family design of §5 on 64 held-out utterances. Every layer passes the dispersion gate (Table 2), so the comparison is between admissible metrics throughout.

The advantage is concentrated almost entirely on hidden state 0—the feature-projection output, before any transformer block. It gains +0.0524 $\tau_b$ over scaled latent $L_2$; the twelve transformer layers gain between +0.0020 and +0.0152, averaging +0.0083, so layer 0's advantage is 6.3× the mean of the rest (Figure 6). Within the transformer stack there is no clear depth trend (Spearman −0.45 over layers 1–12), and the drop happens immediately: layer 1 already gives only +0.0092. This is the sharpest form of the point that the representation does most of the work: on twelve of thirteen of them, the trained denoiser adds almost nothing to the Euclidean distance on the same features.

Two readings we checked and can rule out. The first is that layer 0 simply has more room to improve because its linear baseline is weaker: it is not that, since the correlation between the $L_2$ baseline and the advantage across the thirteen layers is +0.15, and layers 0 and 6 have near-identical baselines (0.6972 and 0.6975) with advantages of +0.0524 and +0.0063. The second is that the advantage is really a covariance effect: against the diagonal Mahalanobis baseline only four layers (0, 2, 3, 9) have an interval clearing zero, and only layer 0's is substantial (+0.0465; the next largest is +0.0095). So beyond the earliest representation the denoiser's nonlinearity adds little that a covariance-aware linear distance does not already capture.

The selection set sits inside the evaluation set, so we report the split. The ranking above is decided on 64 utterances and the headline number of §5 is read off 200 whose first 64 are exactly those, so we recompute the paired advantage on each part of that probe separately. It replicates: against PESQ +0.0576 ($[+0.0521, +0.0631]$) on the 136 held-out utterances against +0.0600 ($[+0.0535, +0.0668]$) on the selection set, against SI-SDR +0.0254 against +0.0282, and against ESTOI −0.0118 against −0.0125 — every pair of intervals overlaps,

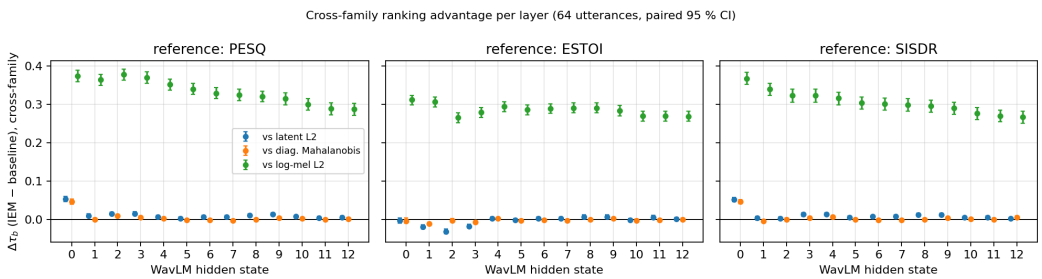


Figure 6: Paired cross-family advantage $\tau_b(\text{IEM}) - \tau_b(\text{baseline})$ per WavLM hidden state with 95% bootstrap intervals over utterances, for the denoise-only AdaLN recipe at a fixed budget, against three references. Zero is the line to beat; hidden state 0 is the outlier.

and the ESTOI reversal is as reproducible as the PESQ advantage. What this does not check is the ranking of the other twelve states out of sample, which would need twelve more probes; the claim it supports is that the layer-0 number is not an artefact of the utterances that chose layer 0.

Finally, "best layer" depends on the question. The highest absolute agreement belongs to layer 2 ($\tau_b = 0.7527$ against layer 0's 0.7496), because its Euclidean baseline is stronger to begin with. Layer 0 is where the denoiser contributes most, which is why it is the layer used throughout the main text.

## F Human ratings: the advantage on recorded conditions

Every reference in §5 is a model, so we also score the admissible layer-0 metric against human ratings. The NISQA Speech Quality Corpus (Mittag et al., 2021) is the corpus that makes this possible without collecting data: 2980 of its files carry both a crowdsourced P.808 rating (ITU-T, 2018) and a clean reference recording, which a full-reference distance needs. Files are resampled to 16 kHz mono and aligned to their reference; the statistic is the same paired bootstrap of $\tau_b(\text{IEM}) - \tau_b(\text{scaled latent } L_2)$ on identical pairs.

The corpus is three populations rather than one — recorded live and anchor conditions, and simulated conditions with and without a codec — so pooling them would report whichever group is largest. We split on NISQA's own condition labels and report all three, with a Holm–Bonferroni correction over the three tests (Table 5).

The advantage is real on recorded conditions. On the 480 live and anchor files SpeechIEM beats latent $L_2$ against human MOS by $+0.009$ $\tau_b$ ($[+0.001, +0.017]$, $p = 0.033$ after correction). These are recordings of real transmission and environment conditions rather than simulations, and they are the group where a distance trained on real speech latents should help. It is a small effect on one corpus, but it is measured against listeners, on data that had no part in any choice in this paper, and with the multiple comparison paid for.

Where it reverses, and which factor causes that. On NISQA's simulated conditions the comparison goes the other way, and not uniformly: the files whose condition contains no codec give $+0.001$ ($[-0.005, +0.007]$), indistinguishable from latent $L_2$, and those with one give $-0.029$ ($[-0.034, -0.024]$). Changing one factor at a time locates the cause (Table 6), and it is none of the three a reader would reach for first. It is not the reference: on NISQA's own conditions PESQ and human MOS agree to within 0.001 ($-0.022$ against $-0.021$), so a model reference and listeners say the same thing here. It is not codec artefacts as such: adding an Opus family at 32 → 6 kbps to our own probe, with its algorithmic delay compensated so the alignment contract holds, leaves the advantage at $+0.052$ ($[+0.046, +0.060]$). And it is not the speech corpus alone: scoring our seven families on NISQA's reference recordings costs about a third of the advantage but keeps its sign ($+0.034$). What reverses

it is NISQA's condition set as a whole — real transmission chains, concatenated impairments and unknown front ends, rather than one operator at a controlled severity.

Two things follow, and they are the reason the paper states a reporting standard rather than a number. The advantage of §5 is a property of a metric and a corruption distribution, so it has to be quoted with one; and a probe built from single operators, however many families it has, does not cover the distribution a deployed distance meets. That is a limit of our protocol as much as of the metric.

Table 5: SpeechIEM against scaled latent $L_2$ on human MOS, by condition type, for the admissible layer-0 metric. $\Delta\tau_b$ is the paired bootstrap on identical pairs; $p$ is Holm-corrected over the three groups; $^*$ marks an interval excluding zero.

| condition type | $n$ | $\tau_b$(IEM) | $\tau_b(L_2)$ | $\Delta\tau_b$ vs $L_2$ | $p$ (Holm) |
|---|---|---|---|---|---|
| live & anchor | 480 | 0.496 | 0.487 | $+0.009$ $[+0.001, +0.017]^*$ | 0.033 |
| simulated, no codec | 382 | 0.615 | 0.614 | $+0.001$ $[-0.005, +0.007]$ | 0.642 |
| simulated, codec | 2118 | 0.574 | 0.603 | $-0.029$ $[-0.034, -0.024]^*$ | $<0.001$ |

Table 6: Decomposing the gap between our cross-family probe and NISQA, one change at a time; every row is the paired bootstrap of $\tau_b$(IEM) $-$ $\tau_b$(scaled latent $L_2$) for the same layer-0 denoiser. The speech corpus costs about a third of the advantage and a real codec family costs little; NISQA's condition set is what reverses the sign, and the reference is not the cause.

| speech | corruptions | reference | $n$ | $\Delta\tau_b$ (IEM $-L_2$) |
|---|---|---|---|---|
| URGENT | 6 families | PESQ | 6000 | $+0.058$ $[+0.054, +0.063]$ |
| URGENT | 7 families (+Opus) | PESQ | 2240 | $+0.052$ $[+0.046, +0.060]$ |
| NISQA refs | 7 families (+Opus) | PESQ | 2240 | $+0.034$ $[+0.028, +0.040]$ |
| NISQA | NISQA conditions | PESQ | 2499 | $-0.022$ $[-0.026, -0.018]$ |
| NISQA | NISQA conditions | human MOS | 2500 | $-0.021$ $[-0.025, -0.017]$ |

## G Capacity and budget

Every negative result in the main text invites the same reply: the denoiser is a four-block head trained for 3000 steps, so perhaps the construction works and ours was merely too small. That is testable rather than arguable. We swept head depth against step budget with everything else fixed—same hidden state, same objective, same 24 000-utterance pool, so a short run is a prefix of a long one—and probed each cell with the design of §5.

The surface is flat (Table 7). All six cells lie between $+0.0510$ and $+0.0571$, a spread of 0.006 against per-cell interval widths of 0.014–0.018, so no pair of cells is resolved by this design (each interval is against latent $L_2$, not against another cell). The two axes behave differently in what little they do move: quadrupling depth buys $+0.0019$ at 3000 steps and $-0.0032$ at 9000, while tripling the budget buys $+0.0061$ at depth 2, $+0.0043$ at depth 4 and $+0.0010$ at depth 8—if anything the budget helps less as the head grows, which is the opposite of the "too small" reading.

So the advantage is a property of this construction on these latents, not of our compute, and the reply above does not hold at the scales we can reach. Two honest caveats: the range is a factor of four in depth and three in budget, which does not exclude a different regime an order of magnitude further out; and depth is the capacity axis because the head's width is tied to the encoder's 768 channels, so widening it would also change the input projection.

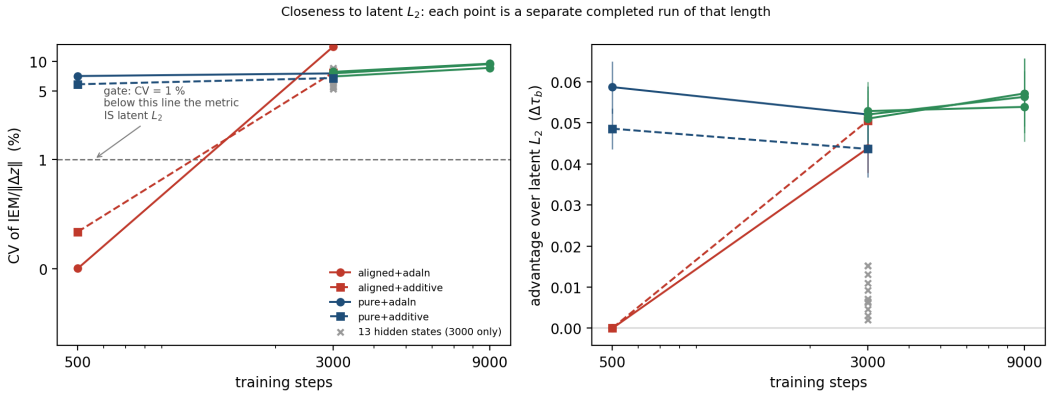


Figure 7: Closeness to latent $L_2$ (left) and advantage over it (right) against the step budget. Each point is a separate completed run of that length rather than a snapshot of one trajectory, since every arm anneals its learning rate over its own budget; lines join runs differing only in budget. The budget axis moves only the alignment-augmented objective: it starts below the gate with zero advantage and leaves that regime by 3000 steps, while the plain objective is already admissible at 500 steps and does not improve.

Table 7: Head depth against step budget on hidden state 0, everything else identical. $\Delta\tau_b$ is the paired cross-family advantage over scaled latent $L_2$ against PESQ on 64 held-out utterances; * marks an interval excluding zero. Quadrupling capacity buys +0.002 at the short budget and nothing at the long one.

| depth | steps | $\tau_b$(IEM) | $\tau_b(L_2)$ | $\Delta\tau_b$ vs $L_2$ |
|---|---|---|---|---|
| 2 | 3000 | 0.748 | 0.697 | $+0.0510\ [+.044, +.058]^*$ |
| 2 | 9000 | 0.754 | 0.697 | $+0.0571\ [+.049, +.066]^*$ |
| 4 | 3000 | 0.749 | 0.697 | $+0.0520\ [+.045, +.059]^*$ |
| 4 | 9000 | 0.754 | 0.697 | $+0.0563\ [+.048, +.066]^*$ |
| 8 | 3000 | 0.750 | 0.697 | $+0.0529\ [+.046, +.060]^*$ |
| 8 | 9000 | 0.751 | 0.697 | $+0.0539\ [+.045, +.063]^*$ |

## H Further tables

## I The enhancement comparison in full

Does an IEM loss train a better enhancer than MSE? A metric that ranks quality better is not automatically a better gradient, and our answer at a fixed budget is no.

A matched design. Only the loss varies: an identity-initialised dilated convolutional enhancer (330k parameters, output loudness matched to the input so neither arm can win by changing gain), URGENT speech, lr $3{\cdot}10^{-4}$, batch 8, and—because any selection rule is itself a metric that favours one loss—no early stopping and no best-checkpoint selection, so the final checkpoint after a fixed number of steps is what is evaluated. Arms are paired within (seed, training subset) cells, so both losses see the same initialisation and the same data; we report the bootstrap interval of the paired difference over cells with an exact sign-flip permutation test over its $2^6$ sign assignments. Earlier versions of this comparison differed between arms in learning rate, batch size, initialisation and checkpoint-selection metric at once, and ran on synthetic tones with a 2.6k-parameter model; the difference between arms was then smaller than the difference between seeds.

The IEM loss must be admissible before the comparison means anything. An "IEM vs $L_2$" comparison against a degenerate denoiser (§4) is $L_2$ vs $L_2$ by construction and returns a null result that looks like a finding. The grid therefore runs the diagnostics of §4.3 on the

Table 8: Cross-family Kendall $\tau_b$ against a reference quality (higher is better) for the admissible layer-0 metric, and the paired bootstrap of IEM minus each baseline on the same pairs. 200 held-out utterances, 6 families × 5 severities, 74 975 cross-family pairs; * marks an interval excluding zero.

| layer | reference | $\tau_b$(IEM) | $\tau_b(L_2)$ | $\Delta$ vs $L_2$ | $\Delta$ vs Mahalanobis |
|---|---|---|---|---|---|
| 0 | PESQ | 0.747 | 0.689 | $+0.058\ [+0.054, +0.063]^*$ | $+0.050\ [+0.046, +0.054]^*$ |
| 0 | ESTOI | 0.733 | 0.745 | $-0.012\ [-0.016, -0.008]^*$ | $-0.011\ [-0.015, -0.007]^*$ |
| 0 | SI-SDR | 0.641 | 0.614 | $+0.026\ [+0.022, +0.030]^*$ | $+0.023\ [+0.019, +0.026]^*$ |

Table 9: Agreement with PESQ within a single degradation family, layer-0 IEM against its zero-denoiser baseline on 200 held-out utterances. Signed Pearson, sign-corrected so higher is better; the last column is the paired bootstrap of the Kendall difference on the same pairs. Within a family the two are indistinguishable; pooling the families separates them.

| degradation axis | $n$ | IEM | latent $L_2$ | log-mel | $\Delta\tau_b$ vs $L_2$ |
|---|---|---|---|---|---|
| additive noise, 30–0 dB | 1000 | 0.904 | 0.897 | 0.797 | $-0.001\ [-0.003, +0.000]$ |
| reverberation, 0.15–1.2 s | 1000 | 0.976 | 0.964 | 0.903 | $+0.000\ [+0.000, +0.000]$ |
| clipping, quantile 0.999–0.5 | 1000 | 0.948 | 0.932 | 0.826 | $-0.001\ [-0.003, +0.000]$ |
| band limit, 7–1 kHz | 1000 | 0.831 | 0.829 | 0.881 | $+0.000\ [+0.000, +0.000]$ |
| $\mu$-law, 8–2 bit | 1000 | 0.921 | 0.914 | 0.826 | $+0.000\ [+0.000, +0.000]$ |
| packet loss, 1–30% | 1000 | 0.937 | 0.927 | 0.824 | $-0.001\ [-0.003, +0.000]$ |
| all six pooled | 6000 | 0.892 | 0.851 | 0.384 | $+0.050\ [+0.047, +0.054]^*$ |

referenced checkpoint before any training starts and refuses to proceed if it fails; we use the admissible layer-0 denoiser, and the loss is the unbiased Monte-Carlo estimator of §3 with four samples per step.

Result. Table 10 reports both arms against the unprocessed input, which is also the enhancer's score at initialisation. Reading the improvements rather than the absolute numbers changes what the comparison says. In the waveform domain MSE dominates: it gains +0.58 dB SI-SDR and +0.44 dB SNR over the input where SpeechIEM loss gains +0.14 and +0.02 dB, and the paired difference is against SpeechIEM on both (−0.44 dB SI-SDR, 95% CI $[-0.54, -0.34]$, permutation $p = 0.046$). But on log-spectral distance the two are level and the sign flips in SpeechIEM's favour (22.07 vs 22.22 from an input of 25.19; $p = 0.17$), and PESQ separates them by 0.006 on a 0.05 improvement. Neither arm improves ESTOI. This is the behaviour a phase-insensitive latent objective should have: it moves the spectral envelope about as well as MSE and the waveform alignment hardly at all, so a comparison scored on SI-SDR is a comparison MSE was always going to win.

How far it generalises. Two limits. The headroom is small—the best arm recovers 0.6 dB of SI-SDR—because the enhancer is a 330k-parameter identity-initialised network on 2048 utterances; and validation SI-SDR peaks by epoch 4 and then flattens, so the ranking is set well before the budget ends. We therefore state the negative result narrowly: at this scale an IEM loss is not a drop-in replacement for waveform MSE, and the metric being a better ranker across distortion families does not transfer to being a better gradient for waveform reconstruction. Whether the spectral gain survives being added to a waveform anchor—an $\mathrm{MSE} + \lambda\,\mathrm{IEM}^2$ objective—is the natural next question and is not settled here.

The anchored variant, at double the budget. Replacing MSE is not the only way to use the metric, and the comparison above does not test the natural alternative of adding it to a waveform anchor. We therefore ran the same design at $16 \times 256$ steps with a third arm, $\mathrm{MSE} + \lambda\,\mathrm{IEM}^2$ (Table 11). The weight is calibrated on gradient magnitude rather than on loss value: matching the two values leaves SpeechIEM term's gradient 74× MSE's, which trains the enhancer on the latent objective alone and is not the comparison we wanted; $\lambda = 6.7 \cdot 10^{-5}$ makes the two gradients comparable at initialisation and is pinned by a test. The anchored arm keeps MSE's waveform behaviour and improves log-spectral distance,

Table 10: The IEM loss against waveform MSE at $8 \times 256$ steps: three seeds $\times$ two disjoint training subsets, everything else identical, final checkpoint, test partition. "input" is the unprocessed noisy signal on the same 256 utterances, which is also the identity-initialised enhancer's score before training; the arms should be read as improvements on it. The interval is a bootstrap over the six paired cells and $p$ is an exact sign-flip permutation test. Differences are sign-corrected so that positive always favours SpeechIEM arm; $\downarrow$ marks the one metric where lower is better.

| metric | input | IEM arm | MSE arm | IEM − MSE [95% CI] | perm. $p$ |
|---|---|---|---|---|---|
| SI-SDR | 3.338 | 3.477 | 3.913 | $-0.436$ $[-0.543, -0.339]$ | 0.046 |
| SNR | 4.004 | 4.020 | 4.444 | $-0.424$ $[-0.720, -0.262]$ | 0.046 |
| PESQ | 1.288 | 1.333 | 1.339 | $-0.006$ $[-0.009, -0.003]$ | 0.046 |
| ESTOI | 0.636 | 0.635 | 0.635 | $+0.000$ $[-0.001, +0.001]$ | 1.000 |
| LSD $\downarrow$ | 25.187 | 22.066 | 22.216 | $+0.150$ $[-0.019, +0.296]$ | 0.169 |

which is the one-sided pattern a phase-insensitive latent objective predicts; the effect is small and the sign-flip test over $2^6$ assignments cannot put it under $p = 0.077$, so it is a direction to follow up rather than a settled gain.

Table 11: The anchored objective $\mathrm{MSE} + \lambda\,\mathrm{IEM}^2$ against waveform MSE at $16 \times 256$ steps, same three seeds $\times$ two disjoint subsets and same protocol as Table 10. Differences are sign-corrected so positive favours SpeechIEM-augmented arm; $\downarrow$ marks the metric where lower is better.

| metric | input | IEM arm | MSE arm | IEM − MSE [95% CI] | perm. $p$ |
|---|---|---|---|---|---|
| SI-SDR | 3.338 | 3.943 | 3.921 | $+0.021$ $[-0.006, +0.046]$ | 0.231 |
| SNR | 4.004 | 4.471 | 4.451 | $+0.020$ $[+0.000, +0.040]$ | 0.169 |
| PESQ | 1.288 | 1.340 | 1.338 | $+0.001$ $[-0.002, +0.004]$ | 0.508 |
| ESTOI | 0.636 | 0.634 | 0.635 | $-0.001$ $[-0.001, -0.000]$ | 0.169 |
| LSD $\downarrow$ | 25.187 | 22.061 | 22.187 | $+0.127$ $[+0.065, +0.177]$ | 0.077 |

Table 12: The pure IEM loss against waveform MSE at $16 \times 256$ steps: the waveform deficit of Table 10 at double the budget.

| metric | input | IEM arm | MSE arm | IEM − MSE [95% CI] | perm. $p$ |
|---|---|---|---|---|---|
| SI-SDR | 3.338 | 3.477 | 3.921 | $-0.445$ $[-0.537, -0.354]$ | 0.046 |
| SNR | 4.004 | 4.183 | 4.451 | $-0.269$ $[-0.324, -0.216]$ | 0.046 |
| PESQ | 1.288 | 1.334 | 1.338 | $-0.005$ $[-0.008, -0.001]$ | 0.108 |
| ESTOI | 0.636 | 0.635 | 0.635 | $+0.000$ $[-0.001, +0.001]$ | 1.000 |
| LSD $\downarrow$ | 25.187 | 22.043 | 22.187 | $+0.145$ $[+0.001, +0.311]$ | 0.169 |

## J Reproducing the tables

Every table in this paper is generated from the experiment outputs by one script, so a number in the text cannot drift from the run that produced it; each experiment writes a manifest with the exact command, the git commit and the final metrics per arm, and the analyses read per-epoch JSON records rather than parsing logs. The admissibility diagnostics run as a pipeline stage between training and analysis and abort by default, because a diagnostic nobody calls is not a gate. Code, the diagnostics, the distortion probe and the run manifests: [repository url withheld for double-blind review].