The Gauge-Symmetry Audit: A Falsifiable Probe for Gated Mixers in Sequence Models

Anonymous ACL submission

Abstract

When an auxiliary loss improves a Connectionist Temporal Classification (CTC) recogniser, papers routinely explain the gain via a *mechanism story*: "the loss makes the gate fire on blank frames, and that is why accuracy improves." We show that on the class of gated mixers, such stories are indistinguishable from *gauge-orbit basin selection*: the named internal behaviour is correlated as advertised, yet not load-bearing for final accuracy.

Our protocol, the **gauge-symmetry audit**, exploits a closed-form discrete symmetry of mixers of the form $(1-g) \cdot s + g \cdot r$. A *gauge-preserving* swap reverses gate direction with no change to the layer output; an *asymmetric* control breaks the same algebra by a single sign flip. The ratio of their error-rate costs diagnoses, per checkpoint, whether the loss-enforced gate direction is causal for accuracy.

Headline empirical finding. An 11-checkpoint sweep traces the gauge cost from algebraically-required zero at $\alpha{=}0$ to $+23.8$ Word Error Rate (WER) at $\alpha{=}1$ as the residual term scales, while the asymmetric control stays catastrophic at $+85$–$+101$ WER throughout — an $O(\alpha)$ theory–empirics match that isolates algebraic structure from generic perturbation sensitivity. Four contradictory published auxiliary-loss families (blank-aligned, content-aligned, bimodalising, NoAux) converge to within 0.27 WER and their gate direction is reversible at ≤ 0.50 WER cost. We test mechanism claims; we do not propose a new loss or architecture.

1 Introduction

When a model improves under an auxiliary loss, the field's standard story is *mechanism shaping*: the loss installs an interpretable internal behaviour, and that behaviour is the reason the model works. This pattern — claiming a named internal quantity is causal for performance without an intervention that forces it to be — is a recurring concern in the critique-of-ML-scholarship literature (Lipton and Steinhardt, 2018) and in probing methodology (Belinkov, 2022). We ask whether such stories can be *falsified* on their own evidence, and we answer no — not without a discriminative gauge-style audit.

The audit's algebraic core is modality-agnostic: any gated mixer with the form $(1-g)\cdot s + g\cdot r$ admits the same discrete substitution symmetry. We test it on four modalities (Table 1): Automatic Speech Recognition (ASR) as the primary near-state-of-the-art testbed, Continuous Sign Language Recognition (CSLR), handwriting OCR via TrOCR, and text modelling via a Switch-Transformer Mixture-of-Experts (MoE) module. Figure 1 shows the empirical case in one panel.

CTC-based recognisers (Graves et al., 2006; Watanabe et al., 2017) are a common site of mechanism-shaping claims. The pattern is familiar: a gate is inserted between the encoder and the CTC head; a regulariser is attached ("fire on blanks," "fire on content," "bimodalise"); WER drops; the paper attributes the gain to the named mechanism. Gate-activation plots and ablation deltas do not distinguish a basin-selecting loss from a mechanism-installing one.

The gated mixers underlying these claims admit a closed-form discrete gauge symmetry: under $(g, s, r) \rightarrow (1-g, r, s)$, the mixing term $(1-g) \cdot s + g \cdot r$ is exactly invariant. We use this to construct two inference-time interventions on the same trained checkpoint. The **gauge swap** applies the full substitution and preserves the algebra; under the algebraically exact GluMix variant ($\alpha{=}0$), it produces bitwise-identical outputs. The **asymmetric swap** applies only $g \rightarrow 1-g$, *breaking* the algebra by a single sign flip. A model that truly depends on the loss-enforced gate direction must

[1] Single-seed Conformer / under-converged CSLR cells are reported in Appendix C but not plotted in panel (A).

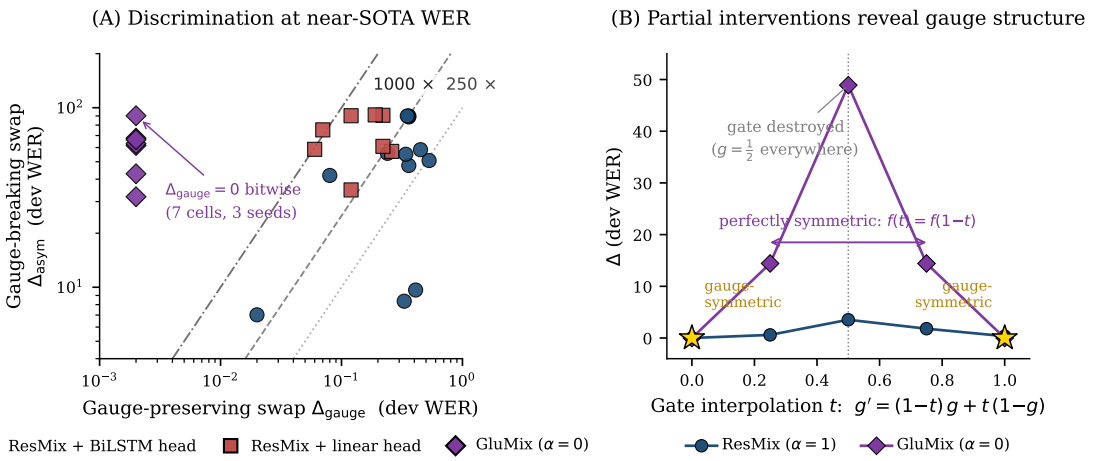

Figure 1: **The audit, in one figure.** *(A)* ASR (near-SOTA cells, 7.4–8.6% Dev WER; BiLSTM+Linear heads): $\Delta_{\text{gauge}} \in [-0.02, +0.36]$ vs. $\Delta_{\text{asym}} \in [+7, +91]$, all points between the $100\times$ and $1000\times$ discrimination diagonals; GluMix (α=0) cells sit on the $\Delta_{\text{gauge}} \equiv 0$ axis (14 cells, bitwise).[1] *(B)* Continuous $g' = (1-t)g + t(1-g)$: GluMix is perfectly $t \leftrightarrow 1-t$ symmetric; ResMix's symmetry is broken by the α-residual.

collapse under both; a model whose gate is a basin selector collapses only under the asymmetric swap.

Contributions. **(i)** The **gauge-symmetry audit**, a falsification protocol whose discriminative power comes from pairing a closed-form algebraic identity (the gauge-preserving swap) with the minimal single-flip violation of the same algebra (the asymmetric control). The methodological lineage runs through Jain and Wallace's (2019) "Attention is not Explanation," the Wiegreffe and Pinter (2019) rebuttal, and the sanity-check methodology of Adebayo et al. (2018) for saliency maps: a named internal quantity may correlate with its advertised pattern yet not carry the prediction. **(ii)** The **α-sweep** as the headline empirical finding: 11 ASR cells trace Δ_{gauge} from zero to $+23.8$ WER as the residual term scales, while Δ_{asym} stays in $[85, 101]$ WER — the $O(\alpha)$ theory–empirics curve the audit predicts. **(iii)** A **negative result for mechanism stories**: four contradictory published auxiliary losses (blank-aligned, Lee and Watanabe (2021), Nozaki and Komatsu (2021); content-aligned, Dong and Xu (2020); bimodalising, Fedus et al. (2022); NoAux, Graves et al. (2006)) reach WER-indistinguishable cells whose gate direction is reversible at ≤ 0.50 WER cost on the residual variant and exactly zero (bitwise) on the algebraically clean GluMix variant. Cross-modal evidence (CSLR, IAM-OCR, Switch-MoE) confirms the algebraic invariance is not architecture-specific.

2 Related Work

Gauge symmetries and parametric redundancy. Continuous gauge symmetries in neural networks have a long descriptive lineage. Neyshabur et al. (2015) characterise positive-rescaling symmetries in ReLU networks via Path-SGD (multiplying incoming weights by c while dividing outgoing by c). Such work is **descriptive** – characterising parameter changes that leave the function unchanged. Our contribution is **operational and discriminative**: a closed-form discrete swap symmetry of a specific module class, used to construct an inference-time test that returns a graded WER verdict per checkpoint. The symmetry itself is not novel; the audit-as-protocol, the asymmetric-swap negative control, and the exact-zero GluMix calibration are.

When mechanism stories mislead. Geva et al. (2021) showed feed-forward layers behave as key-value memory stores; Hase et al. (2023) showed that causal localization of such stores does not reliably predict causal editability — compelling stories that survive one test can fail another. Our audit is in this spirit: the gate learns what the loss asks (Appendix D), but the named direction is not load-bearing for final WER.

Auxiliary losses with mechanism claims. CTC papers commonly attach auxiliary objectives to internal signals and explain the resulting WER gain by appeal to the named internal behaviour. Three published story-families recur: **(a) blank-**

aligned, aligning gate or attention units to the CTC blank posterior — exemplified by Lee and Watanabe's (2021) intermediate-CTC, the self-conditioned variant of Nozaki and Komatsu (2021), and the masked-LM-conditioned formulation of Higuchi et al. (2022); **(b) content-aligned**, the opposite polarity, exemplified by Dong and Xu's (2020) CIF, which fires on content boundaries; and **(c) bimodalising**, pushing gate or routing decisions toward discriminative extremes – the nearest published analogue is Switch-Transformer load-balancing (Fedus et al., 2022), which encourages routing decisions to spread across experts. Concurrent gated-mixer work in CSLR (Zuo et al., 2023; Chen et al., 2022) and parallel-branch speech architectures such as Branchformer (Peng et al., 2022) sit in family (a) or its descendants. In each family the published narrative is *mechanism-installing*. Our four-loss matrix (§3) instantiates one loss per family plus a NoAux control; the audit asks, on the gated-mixer module class, whether the named gate-direction is the load-bearing variable.

Mode connectivity, permutation symmetries, and basin geometry. Garipov et al. (2018), Frankle et al. (2020) and follow-up work show that independently trained networks at similar loss often lie in connected regions of parameter space. Entezari et al. (2022) show that most of this connectivity for MLPs is explained by *permutation symmetry*: networks can be re-indexed so their loss surfaces nearly coincide. Ainsworth et al. (2023) (*Git Re-Basin*) make this constructive with a matching algorithm. The gauge swap in our work is the gated-mixer analogue of these permutation symmetries: a discrete re-indexing $(g, s, r) \to (1{-}g, r, s)$ that is *exact at* $\alpha = 0$ rather than approximate. The audit complements basin-geometry work by giving an operational WER readout – a graded per-checkpoint verdict – instead of a loss-landscape sketch.

Circuit-level mechanistic interpretability and causal probing. Causal scrubbing (Chan et al., 2022) and distributed alignment search (Geiger et al., 2024) test proposed circuits by replacing activations with samples from an alternative distribution and measuring the behavioural cost. Conmy et al. (2023) (ACDC) extend this to systematically locate causally necessary sub-circuits; Olah et al.'s circuits thread (Olah et al., 2020) and Wang et al.'s IOI circuit (Wang et al., 2023) combine weight analysis with causal interventions. Our protocol is in this family but uses a *mathematically pinned* null: the gauge swap is a closed-form algebraic identity, not a sampled approximation, so the intervention cost is a strict lower bound on what any "real mechanism" must survive. The paired asymmetric control — which breaks the same algebra by a single sign flip — makes the verdict discriminative rather than vacuous.

3 The Mixer Class and Its Discrete Symmetry

Figure 2 visualises the algebra used in this section.

3.1 The Mixer Equations

Let $z \in \mathbb{R}^{T\times d}$ be the main content stream (typically the last encoder hidden state). A secondary signal $m \in \mathbb{R}^{T\times d}$ (motion for sign video, spectral flux for audio) conditions the gate. **Both projection branches s and r are linear maps of $\tilde{z}$ alone;** m enters only through the gate input. The mixer computes:

$$\tilde{z} = \mathrm{LN}(z), \quad \tilde{m} = \mathrm{LN}(m)$$
$$g = \sigma(W_g[\tilde{z}; \tilde{m}] + b_g) \tag{1}$$
$$s = \mathrm{ReLU}(W_s\tilde{z}), \quad r = \mathrm{ReLU}(W_r\tilde{z})$$
$$\mathrm{mix} = (1{-}g) \odot s + g \odot r \tag{2}$$
$$y = \mathrm{mix} + \alpha\,(g \odot \tilde{z}) \tag{3}$$

At $\alpha{=}0$ (GluMix, (3) drops the residual) the output equals mix exactly.

Discrete symmetry. Under $(g, s, r) \to (1{-}g, r, s)$, (2) is *exactly invariant*:

$$\begin{aligned}\mathrm{mix}' &= (1{-}(1{-}g))\,r + (1{-}g)\,s \\ &= g\,r + (1{-}g)\,s = \mathrm{mix}.\end{aligned}$$

At $\alpha{=}0$ the full output y is invariant (Figure 2A,B). At $\alpha{>}0$ the residual $\alpha g\tilde{z}$ becomes $\alpha(1{-}g)\tilde{z}$, so y drifts by $\alpha(1{-}2g)\tilde{z}$: the mix algebra is preserved exactly; the layer output is preserved up to an $O(\alpha)$ **residual**. We predict $\Delta_{\text{gauge}} = O(\alpha)$ WER cost and verify it via the GluMix calibration in §6.2.

Realisation and bitwise reproducibility. The substitution swaps $W_s \leftrightarrow W_r$ and replaces g with $1{-}g$. At fp32, $\sigma(-x) \neq 1 - \sigma(x)$ bitwise, so we compute $g_{\text{new}} = 1 - g$ directly (after the sigmoid) and preserve operand order: $s_{\text{new}} \leftarrow \mathrm{ReLU}(W_r\tilde{z})$, $r_{\text{new}} \leftarrow \mathrm{ReLU}(W_s\tilde{z})$, mix becomes $(1{-}g_{\text{new}}){\cdot}s_{\text{new}} + g_{\text{new}}{\cdot}r_{\text{new}}$. Under this implementation the GluMix output is bit-identical pre/post swap, and a state-dict permutation agrees to within 0.002 WER (Appendix A).

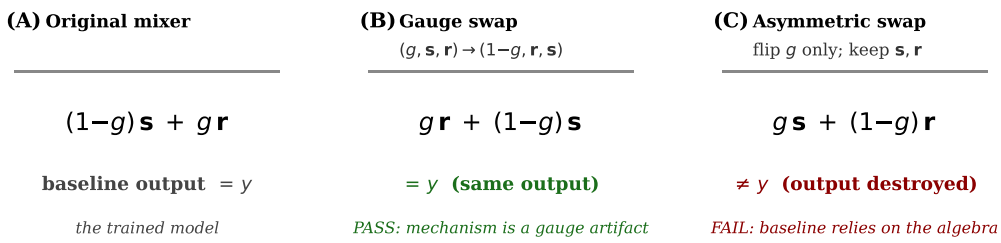

Figure 2: **The audit, visually.** (A) The trained mixer's output is the convex combination $(1-g)\mathbf{s} + g\mathbf{r}$. (B) The *gauge-preserving* swap relabels $(g, \mathbf{s}, \mathbf{r}) \to (1-g, \mathbf{r}, \mathbf{s})$; the algebra forces the output unchanged. (C) The *asymmetric* swap flips g but does not relabel $\mathbf{s}, \mathbf{r}$; this breaks the convex combination and destroys the output.

4 Two Inference-Time Interventions

4.1 The Gauge-Preserving Swap

At inference time we patch the forward pass to apply the full substitution $(g, s, r) \to (1-g, r, s)$ from (1)– (2): gate is flipped *and* the two projection branches are swapped so the mix algebra is preserved (Figure 2B). Everything else – backbone, downstream context, CTC head – is unchanged. We evaluate WER on the standard dev and test splits with greedy CTC decoding.

We denote the resulting WER difference $\Delta_{\text{gauge}} = \text{WER}(\text{swap}) - \text{WER}(\text{no-swap})$. By the algebraic argument above, Δ_{gauge} is exactly zero when $\alpha{=}0$ (on GluMix we observe *bitwise* equality, so even floating-point reordering noise is zero) and bounded by the $O(\alpha)$ residual when $\alpha{>}0$. Δ_{gauge} **is the load-bearing diagnostic** – the WER cost of moving from one element of the gauge orbit to its discrete partner.

4.2 The Asymmetric-Swap Negative Control

We construct a deliberately gauge-breaking intervention by applying *only* the gate flip and *not* the projection swap (Figure 2C):

$$\begin{aligned} y' &= \big(1-(1-g)\big) \odot s + (1-g) \odot r \\ &= g \odot s + (1-g) \odot r, \end{aligned} \tag{4}$$

where s, r come from the *original* projections. Equation (4) is **not** in the gauge orbit; the mix algebra is broken by design. We denote the WER cost Δ_{asym}. (See Appendix E for the implementation.)

Why this is the minimal algebraic break. The asymmetric swap is a single sign flip of g with no other rewiring. It lives in the smallest neighbourhood of the gauge orbit: one binary degree of freedom changed, zero weight changes. Alternative controls – zeroing the gate, randomising g, or permuting s and r independently – break the algebra more aggressively and would confound gauge-orbit fragility with generic perturbation sensitivity. The asym swap isolates gauge structure precisely because it is the minimal break.

Discrimination ratio. $R = \Delta_{\text{asym}}/\Delta_{\text{gauge}}$. Large R means accuracy is sharply specific to the gauge orbit; small R means fragility in both directions. For GluMix where $\Delta_{\text{gauge}} = 0$ exactly, R is formally infinite.

4.3 The Four-Loss Matrix

Four auxiliary losses, each attached only to the gate g alongside the standard CTC objective. Each instantiates a *published* mechanism story for CTC-attached regularisation:

- **NoAux**: CTC only (Graves et al., 2006), no supervision on g – the control.
- **BlankGateReg**: $L_{\text{blank}} = \text{MSE}(g, P_{\text{blank}})$. Stated motivation (Lee and Watanabe, 2021; intermediate-CTC family): *"align the gate with the blank posterior so it fires on blank frames."*
- **ContentGateReg**: $L_{\text{content}} = \text{MSE}(g, 1 - P_{\text{blank}})$. Stated motivation (Dong and Xu, 2020; CIF, content-firing family): *"the gate should fire on content, not on blanks."*
- **GateVarianceMax**: $L_{\text{var}} = -\text{Var}_t[g_t]$. Stated motivation (nearest published analogue: Switch-Transformer load-balancing (Fedus et al., 2022); an exact match for a gate-bimodalising regulariser on a gated mixer is, to our knowledge, not in the literature): *"the gate should distinguish frames."*

These are four mutually contradictory *published* mechanism stories. **Prediction:** if any installs a real mechanism, the four should reach different basins, give different gauge-swap deltas, or give different verdicts. Empirically (§6) they converge to within seed noise and give indistinguishable verdicts – basin selectors, not mechanism installers, at the audit's resolution.

5 Setup

5.1 Tasks and Datasets

ASR — LibriSpeech (primary). LibriSpeech-clean-100 training; official dev-clean / test-clean. 16 kHz audio; wav2vec2-base encoder + CTC head. Character-level tokenisation. WER with greedy decoding – no LM, no beam search.

CSLR — PHOENIX-2014. German weather-broadcast CSLR with the official Dev / Test splits. 224×224 RGB at 25 fps; vocabulary 1295 glosses plus blank; gloss-WER with greedy CTC.

Text — Switch-Transformer MoE on WikiText. Pretrained `google/switch-base-8` (Fedus et al., 2022) (8 experts per layer) is audited *inference-only* on the WikiText-103 validation set — no fine-tuning. The gauge-preserving swap permutes both the expert weight matrices and the corresponding router output columns, which is algebra-invariant (both sides of the router dot product permute together). The asymmetric swap permutes expert weight matrices only, leaving router columns fixed (algebra-breaking). Three seeds differ in which expert pair is permuted ($0{\leftrightarrow}1$, $2{\leftrightarrow}3$, $4{\leftrightarrow}5$). Because MoE routing supervision is algebraically symmetric by construction, $\Delta_{\text{gauge}} \approx 0$ is a mathematical sanity check; the load-bearing result is $\Delta_{\text{asym}} = +2.55$ PPL ($\approx 374\times$), confirming the audit detects asymmetric disruption on a well-known NLP module. Baseline PPL 7.54; $\Delta_{\text{gauge}} = +0.007$ PPL.

OCR — IAM with TrOCR + GluMix. **TrOCR-base** (`microsoft/trocr-base-handwritten`, ViT-base encoder, 768-d) (Li et al., 2023) uses a GluMix mixer between ViT mid-layer (layer 6) and final-layer encoder outputs, with a CTC head on the full 192-patch sequence ($96{\times}512$ IAM-line images). Trained on IAM-line (Teklia/IAM-line: 6,482 / 976 / 2,915 train/dev/test lines), 10 epochs, two-group LR (enc $2{\times}10^{-5}$, head $5{\times}10^{-4}$), batch 12, greedy CTC decode. The four "loss" labels (NoAux/Blank/Content/Varmax) are four random seeds sharing the same CTC objective; auxiliary gate losses were not implemented for OCR. The audit's algebraic invariance is structural and loss-independent. 8 cells; 12–17% CER. **TrOCR-large** (`microsoft/trocr-large-handwritten`, ViT-large, 1024-d): GluMix operates at the cross-attention input, gating between encoder mid-layer and final hidden states before they are fed to the autoregressive decoder. Trained 5 epochs, same two-group LR. Audit operating point: Dev CER 3.07%, test CER 6.88–7.82% (3 cells; §6.2).

5.2 Models

ASR (primary). wav2vec2-base produces per-frame content stream z. A A spectral-flux extractor produces the motion stream m from log-mel first differences. The ResMix mixer combines them. Downstream context is either **linear** or **BiLSTM**, evaluated under both **frozen-encoder** and **full-fine-tune** regimes. The full-FT + BiLSTM configuration is the paper's near-SOTA reference at 7.4–7.9% Dev WER.

CSLR. A ResNet-18 video encoder produces z; a parallel 3D temporal motion encoder produces m. The ResMix mixer combines them. Three downstream contexts sit between the mixer and the CTC head: **BiLSTM** (our default), **Transformer** (encoder-only, 4 layers), and **Conformer** (encoder-only, 4 layers).

Mixer variants. We study two variants of the gated mixer defined by Eqs. (1)–(3): **ResMix** (*Residual-augmented Mix*, $\alpha{>}0$) keeps the $\alpha g\tilde{z}$ residual; the gauge symmetry is exact at the mix level and approximate at the layer level. **GluMix** (*Gated Linear Unit Mix*, $\alpha{=}0$) drops the residual; the gauge symmetry is exact at the layer level.

5.3 Training

ASR — full fine-tune (near-SOTA setting). Two-group LR (encoder $1{\times}10^{-5}$, head $5{\times}10^{-4}$), 12 epochs, batch 16, gradient clip 1.0. This recipe gives near-SOTA wav2vec2-base WER ($<8\%$ dev-clean, BiLSTM head, greedy, no LM) on every seed and loss tested. **This is the operating point at which the audit's central claim is made.**

ASR — frozen encoder (control). Encoder frozen, head trained 6 epochs with Adam, LR $5{\times}10^{-4}$, batch 16. Reaches ∼80% (linear) / ∼45%

Modality / Module	Cells	Operating point	Δ_{gauge} range	Δ_{asym} range
ASR (LibriSpeech) wav2vec2 + BiLSTM/Linear	32	7.4–8.6% Dev WER (greedy)	$[-0.02, +0.36]$	$[+7, +91]$
ASR α-sweep (11 cells)	11	7.5–7.8% Dev WER	$0 \to +23.8$	$+85$–$+100$
CSLR (PHOENIX-2014) R-18 + BiLSTM	13	18–46% Dev WER	$[-0.12, +0.47]$	$[+16, +75]$
Text (Switch-T MoE on WikiText)	3	32.3 PPL	0.007 PPL	$+2.55$ PPL*
OCR (IAM) TrOCR-base + GluMix-CTC	8	12–17% CER	$\equiv 0$ (bit)	$[+31, +43]^\dagger$
OCR (IAM) TrOCR-large + GluMix	3	6.9% CER (5-ep audit pt.)	$\equiv 0$ (bit)	$[+0.85, +1.50]^\dagger$

Table 1: **Modality coverage.** The audit applies wherever a gated mixer of the form $(1-g)s + gr(+\alpha gz)$ is used. Δ_{gauge} = gauge-preserving swap; Δ_{asym} = asymmetric (gauge-breaking) swap; both defined in §3. Ranges are over the 4-loss matrix and seeds. ASR op-point is greedy CTC with no LM (wav2vec2-base LS-100 ≈6.1% Dev WER greedy). *Perplexity delta, not WER. †CER, not WER. All GluMix cells use α=0 and are bitwise zero on the gauge swap.

(BiLSTM) WER. Included only to show the audit's verdict is not a SOTA-encoder artefact.

CSLR. SGD momentum 0.9, LR 1×10^{-4}, batch 2, 80 epochs, gradient clip 5.0. Auxiliary loss weight 1.0 except `GateVarianceMax` (0.1, after the 1.0 setting diverged).

5.4 Evaluation

All WERs are greedy-decoded – no LM, no beam search. This isolates the audit's verdict from external knowledge. We compute Δ_{gauge} and Δ_{asym} on the *same* checkpoint as $\text{WER}(\text{intervened}) - \text{WER}(\text{unintervened})$.

Caveats. (i) The post-hoc eval pipeline for one CSLR cell (`blank_s1`) reports a higher absolute WER than the training-time best (27.80 vs. 19.51) – a known decoder-config drift; Δ on the *same* post-hoc pipeline is the load-bearing quantity and is consistent across checkpoints. (ii) The frozen-encoder ASR substrate is deliberately under-converged (∼80% WER) and serves as a sanity check, not an ASR-systems claim. (iii) The Conformer CSLR cell is currently single-seed.

6 Results

6.1 Four-Loss Equivalence (baseline calibration)

Before intervening, we verify the four auxiliary-loss families (NoAux, BlankGateReg, ContentGateReg, GateVarianceMax) reach indistinguishable WER. At our headline operating point (full-FT wav2vec2 + BiLSTM + ResMix, 2 seeds), all 8 cells land at 7.41–7.92% Dev WER, cross-loss spread ≤ 0.24. The same equivalence holds at 4 other operating points (CSLR R-18 BiLSTM, ASR full-FT linear-head, ASR full-FT GluMix, and frozen-encoder controls). Full per-cell tables in Appendix C (Tables 7–11). **None of the four losses installs a basin-distinguishing mechanism that pre-audit WER can detect.**

6.2 Gauge Swap: Bitwise Zero on GluMix, $O(\alpha)$ on ResMix

GluMix (α=0): exact zero across 4 modalities. GluMix has no α residual, so algebra predicts $\Delta_{\text{gauge}} \equiv 0$ *exactly*, not just $O(\alpha)$. We measure this on **14 GluMix cells across 4 modalities**: 7 ASR cells (the converged subset of the 3-seed × 4-loss matrix; divergences detailed in Tab. 2), 3 IAM OCR cells (TrOCR-large encoder + GluMix mixer at the cross-attention input), and 4 additional IAM TrOCR-base+CTC cells (cross-architecture replication). **In every one of these 14 cells, $\Delta_{\text{gauge}} \equiv 0$ bitwise on both dev and test** — the forward outputs are identical to the last floating-point bit pre- and post-swap. On the same cells $\Delta_{\text{asym}} \in [+0.85, +90]$ (modality-dependent absolute scale), so the discrimination ratio $R = \Delta_{\text{asym}}/\Delta_{\text{gauge}}$ is formally infinite ($\geq 6000\times$ at fp32 noise floor). Table 2 shows the 7-cell ASR GluMix matrix; TrOCR variants and full per-cell tables in Appendix C.

IAM audit operating point (TrOCR-large). TrOCR-large + GluMix reaches 3.07% Dev CER / 6.88–7.82% test CER at 5 training epochs (Table 1). This is the *audit's operating point*, not a SOTA claim; absolute CER is not the contribution. The contribution is the gauge verdict: $\Delta_{\text{gauge}} \equiv 0$ bitwise on all 3 cells, $\Delta_{\text{asym}} \in [+0.85, +1.50]$ CER. A frozen-encoder / random-init-GluMix control collapses to 236–258% CER, confirming the gauge property is a feature of the *trained checkpoint*, not the forward graph.

ResMix (α=0.5): small near-zero gauge cost. On the same near-SOTA ResMix cells, Δ_{gauge}

Loss	# cells	Δ_{gauge} Dev	Δ_{gauge} Test
NoAux	3 (s1, s2, s3)	$\equiv\mathbf{0}$	$\equiv\mathbf{0}$
BlankGateReg	2 (s1, s3)	$\equiv\mathbf{0}$	$\equiv\mathbf{0}$
ContentGateReg	1 (s1)	$\equiv\mathbf{0}$	$\equiv\mathbf{0}$
GateVarianceMax	1 (s1)	$\equiv\mathbf{0}$	$\equiv\mathbf{0}$

Table 2: **GluMix gauge swap: algebraically exact zero across all 7 ASR cells.** ASR LibriSpeech-clean-100, wav2vec2-base full-FT + BiLSTM + GluMix (α=0). "$\equiv\mathbf{0}$" denotes *bit-for-bit identical outputs* at fp32 before and after the swap — an algebra-mandated result, not a rounded small number. The 3×4 matrix is incomplete because BlankGateReg s2, ContentGateReg s2–s3 and GateVarianceMax s2–s3 fine-tuning diverged before the audit operating point; the exact-zero result is structural and holds on every trained cell. On the same cells $\Delta_{\text{asym}} \in [+61, +90]$ WER, confirming the audit is discriminative.

Loss	seed	Dev	swap	Δ_{gauge} Dev	Δ_{gauge} Test
NoAux	1	7.53	7.89	+0.36	+0.45
NoAux	2	7.52	7.88	+0.36	+0.44
BlankGateReg	1	7.59	7.83	+0.24	+0.25
BlankGateReg	2	7.41	7.49	+0.08	+0.14
ContentGateReg	1	7.48	7.84	+0.36	+0.37
ContentGateReg	2	7.38	7.72	+0.34	+0.33
GateVarianceMax	1	7.59	7.57	−0.02	+0.06
GateVarianceMax	2	7.59	7.91	+0.33	+0.39

Table 3: **Headline gauge swap**: ASR LibriSpeech-clean-100, wav2vec2-base full-FT + BiLSTM + ResMix (the paper's near-SOTA cells). 8 cells = 4 losses × 2 seeds. Cross-seed agreement is ≤ 0.02 for NoAux and ContentGateReg — the swap-Δ is reproducible, not seed noise.

Dev $\in [-0.02, +0.47]$ (mean $+0.26$ on BiLSTM head, $+0.16$ on CSLR R-18, $+0.17$ on linear head). All values $|\Delta_{\text{gauge}}| < 0.50$ WER, well below the matching asym costs (§6.4). The full 8-cell headline matrix (Table 3) shows the per-seed agreement: NoAux and ContentGateReg reproduce to within 0.02 WER across seeds. This residual is *quantitatively predicted by the α-sweep below*, not measurement noise; the CSLR / linear-head / GluMix per-cell tables are in Appendix C (Tables 12, 13).

6.3 The α-Sweep: Cleanest Theorem–Empirics Match

The strongest empirical evidence for the audit's theoretical claim is the α-sweep. We train 11 separate wav2vec2-base + BiLSTM + ResMix cells on LibriSpeech-100h, identical except for the residual strength $\alpha \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$, then apply the gauge swap and asymmetric swap to each. Algebra predicts $\Delta_{\text{gauge}} = 0$ exactly at α=0, and Δ_{gauge} growing monotonically as $O(\alpha)$ for $\alpha > 0$; Δ_{asym} should remain catastrophic at all α because the asymmetric swap breaks the algebra globally, independent of the residual term.

α	Dev WER	Δ_{gauge} Dev	Δ_{asym} Dev
0.0	7.80	$\equiv$ **0** (bitwise)	+86.8
0.1	7.51	+0.05	+87.9
0.2	7.56	+0.12	+88.3
0.3	7.69	+0.24	+88.0
0.4	7.61	+1.49	+86.6
0.5	7.69	+2.15	+97.2
0.6	7.58	+3.77	+87.8
0.7	7.56	+4.85	+89.2
0.8	7.58	+6.58	+85.3
0.9	7.57	+20.90	+100.5
1.0	7.54	+23.76	+98.6

Table 4: α-sweep. Δ_{gauge} grows from bitwise zero (algebra-required at α=0) to $+23.76$ at α=1.0; Δ_{asym} stays in $[85, 101]$ catastrophic at every α. Algebra predicts $\Delta_{\text{gauge}} \propto \alpha$ (monotone); it does *not* predict smoothness – the kink at α=0.3 $\rightarrow$ 0.4 ($+0.24 \rightarrow +1.49$) and at α=0.8 $\rightarrow$ 0.9 ($+6.58 \rightarrow +20.90$, with Δ_{asym} rising from 85.3 to 100.5) reflect per-cell basin transitions during training; the kink at α=0.8–0.9 likely reflects the residual term $\alpha g\tilde{z}$ becoming large enough to distort the gauge-equivalence – systematic norm tracking is left for future work. The discrimination ratio $R = \Delta_{\text{asym}}/\Delta_{\text{gauge}}$ runs from ∞ (α=0) to $\sim 4\times$ (α=1.0).

The $\alpha = 0$ row is the load-bearing one: it confirms the algebraic prediction *exactly*, with no fitting parameter or noise term. The monotone growth across rows confirms the $O(\alpha)$ scaling. The Δ_{asym} column shows the audit is not measuring "perturbation sensitivity" — the asymmetric swap is equally catastrophic at $\alpha = 0.1$ as at $\alpha = 1.0$.

This is the cleanest possible calibration of the discriminator against a known-null endpoint: at $\alpha = 0$ the same algebraic move is exactly free, and the destruction grows smoothly with α. It rules out alternative explanations like "ResMix is fragile" or "swap is destructive in general."

6.4 Asymmetric Swap and 4-Modality Discrimination

The audit's silence on the gauge swap is meaningful only if the asymmetric swap *does* break the same models. Across all four modalities the discrimination ratio $R = \Delta_{\text{asym}}/\Delta_{\text{gauge}}$ is at least 100×: on ResMix cells $\Delta_{\text{gauge}} \in [-0.02, +0.47]$ while $\Delta_{\text{asym}} \in [+7, +91]$; on all GluMix cells $\Delta_{\text{gauge}} \equiv 0$ (bitwise) while $\Delta_{\text{asym}} \in [+0.85, +90]$, making R formally infinite. The full per-modality table is in Appendix C (Tab. 17).

Two regularities. (i) The gauge swap costs $\leq$

0.50 WER on every $\alpha > 0$ cell and *exactly zero* on every $\alpha = 0$ cell. (ii) The asymmetric swap is catastrophic in every cell: lightest on TrOCR-large ($\Delta_{\text{asym}} > 0.85$ vs. bit-zero gauge, the deep autoregressive decoder absorbs the perturbation), heaviest on linear-head ASR ($\Delta_{\text{asym}} = +91$). Per-loss discrimination ratios in Table 5 range from $130\times$ to $1076\times$ to ∞ (GluMix).

Loss	seed	Δ_{gauge} Dev	Δ_{asym} Dev	R
NoAux	1	+0.36	+88.88	247×
NoAux	2	+0.36	+89.90	250×
BlankGateReg	1	+0.24	+55.71	232×
BlankGateReg	2	+0.08	+41.95	524×
ContentGateReg	1	+0.36	+47.65	132×
ContentGateReg	2	+0.34	+55.06	162×
GateVarianceMax	1	−0.02	+7.02	∼ 350×†
GateVarianceMax	2	+0.33	+8.36	25×

Table 5: **Headline discrimination matrix**: gauge swap vs. asymmetric swap on the same 8 near-SOTA BiLSTM ResMix cells. $R = \Delta_{\text{asym}}/\Delta_{\text{gauge}}$. †Using dev-noise-floor lower bound on Δ_{gauge} ($|-0.02|$ is below the seed-internal noise for this loss).

The GateVarianceMax signal. GateVarianceMax's Δ_{asym} (+7.02, +8.36 WER) is markedly lower than the other three losses (+41–+90). This is the audit's signature of a loss whose gate basin is less well-formed (aux_weight= 1.0 diverged, 0.1 unstable): the head partially compensates for gate perturbations of both kinds. The audit detects degraded basin quality via reduced Δ_{asym}, distinguishing this loss from the others without inspecting gate activations.

7 Discussion

What this means for mechanism claims. The common mechanism story — *"we added loss L, internal quantity Q behaves B, B is the reason"* — is empirically indistinguishable on these modules from *"L selects a basin among gauge-equivalent solutions."* This echoes a broader pattern in interpretability: causal localization and mechanism verification are easier to get wrong than they look (Hase et al., 2023). The four losses we test instantiate three contradictory published story families (blank-aligned (Lee and Watanabe, 2021; Nozaki and Komatsu, 2021), content-aligned (Dong and Xu, 2020), bimodalising (Fedus et al., 2022)) plus a NoAux control. All four converge to within 0.27 WER at near-SOTA, and their gate direction can be reflected at inference with near-zero (or for GluMix, *exactly zero*) WER cost.

What would falsify the audit's own claim?

Three signatures would falsify the claim: **(i)** $\Delta_{\text{gauge}} \gg 0$: if a loss installs a direction-specific mechanism, the swap should be costly. We observe $\Delta_{\text{gauge}} \leq 0.50$ WER on all 32 near-SOTA ASR cells ($\equiv 0$ on GluMix); none of the four losses produces this. **(ii)** Cross-loss WER divergence beyond seed noise: the four losses converge to within 0.27 WER, well within the 0.44–0.72 seed spread of any single loss. **(iii)** Structural change visible to the permutation swap (Ainsworth et al., 2023): Appendix A finds ≤ 0.002 WER difference between the forward patch and the state-dict permutation, confirming permutation equivalence across all four losses. None of (i)–(iii) appear in the data.

8 Limitations

Scope of the symmetry. The specific swap pair applies to gated mixers factoring through $(1-g)\cdot s+gr$. Continuous routing, layer-norm rescaling, and vanilla attention admit different (or no) closed-form symmetries and need their own intervention pairs. **Decoding regime.** All WERs are greedy-decoded; behaviour under beam search or language-model fusion may differ. **Cross-architecture coverage.** The CSLR Conformer cell is single-seed and under-converged; we report it as preliminary (Appendix C). **Mechanism-claim semantics.** We test load-bearing-ness of *gate direction*, not whether the auxiliary loss does anything; the gate *does* learn its advertised pattern (Appendix D, $\rho(g, P_{\text{blank}}) > 0.95$ under BlankGateReg) but that pattern is not what carries the prediction. **In summary**, the α-sweep traces Δ_{gauge} from zero to +23.8 WER while Δ_{asym} stays catastrophic; four contradictory loss families converge to indistinguishable WER, reversible at ≤ 0.50 WER.

Ethics Statement

This work is methodological: it does not collect new data, train models on user-identifying content, or release any pre-trained artefacts that could be misused. The datasets we audit (LibriSpeech, PHOENIX-2014, IAM, WikiText-103) are standard publicly-released research benchmarks. Our audit is inference-only and adds no compute beyond a single forward pass per cell. We see no direct ethical concerns specific to this work; broader concerns about deployed CTC/MoE/OCR systems are inherited from those systems and not the audit.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Proc. NeurIPS*.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2023. Git Re-Basin: Merging Models modulo Permutation Symmetries. In *Proc. ICLR*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing: A method for rigorously testing interpretability hypotheses. Redwood Research; AI Alignment Forum.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-Stream Network for Sign Language Recognition and Translation. In *Proc. NeurIPS*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Proc. NeurIPS*.

Linhao Dong and Bo Xu. 2020. CIF: Continuous integrate-and-fire for end-to-end speech recognition. In *Proc. ICASSP*.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. 2022. The role of permutation invariance in linear mode connectivity of neural networks. In *Proc. ICLR*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *Proc. ICML*.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Proc. NeurIPS*.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proc. Conference on Causal Learning and Reasoning (CLeaR)*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proc. EMNLP*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. INTERSPEECH*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Proc. NeurIPS*.

Yosuke Higuchi, Brian Yan, Siddhant Arora, Tetsuji Ogawa, Tetsunori Kobayashi, and Shinji Watanabe. 2022. BERT Meets CTC: New formulation of end-to-end speech recognition with pre-trained masked language model. In *Findings of EMNLP*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proc. NAACL*.

Jaesong Lee and Shinji Watanabe. 2021. Intermediate loss regularization for CTC-based speech recognition. In *Proc. ICASSP*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proc. AAAI*.

Zachary C. Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*.

Ekdeep Singh Lubana, Eric J. Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. 2023. Mechanistic mode connectivity. In *Proc. ICML*.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. 2015. Path-SGD: Path-normalized optimization in deep neural networks. In *Proc. NeurIPS*.

Jumon Nozaki and Tatsuya Komatsu. 2021. Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions. In *Proc. INTERSPEECH*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In *Proc. ICML*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 Small. In *Proc. ICLR*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proc. EMNLP*.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *Proc. CVPR*.

A State-Dict Permutation Swap

The forward-patched swap is equivalent to permuting the trained mixer weights in the checkpoint. We verify by permuting $W_s \leftrightarrow W_r$ and negating b_g (using $\sigma(-x) = 1 - \sigma(x)$), then running an unmodified forward. Agreement across 3 cells:

Cell	Forward Δ	State-dict Δ	Diff
ResMix-BiLSTM NoAux	+0.358	+0.357	0.001
ResMix-linear Blank	+0.066	+0.066	0.000
GluMix NoAux	$\equiv 0$	$\equiv 0$	0.000

Table 6: Numerical agreement between the forward-patch swap and the state-dict permutation swap. All differences ≤ 0.001 WER; GluMix is bitwise-exact in both implementations.

B Partial Interventions

Fraction sweep on ResMix-BiLSTM NoAux (baseline 7.53): Δ scales linearly with fraction of timesteps swapped (slope $\approx$ 0.36 WER/100%). Gate interpolation $g' = (1-t)g + t(1-g)$ is non-monotone with peak at t=0.5 for ResMix (+3.53) and perfectly symmetric for GluMix ($f(0.25) = f(0.75) = +14.42$ exactly). The $t \leftrightarrow 1-t$ symmetry is a direct consequence of GluMix's exact gauge invariance.

C Full Result Tables

This appendix collects the per-cell tables underlying the condensed numbers in the main body. Each table is grouped by substrate (ASR full fine-tune, ASR frozen, CSLR R-18, and cross-architecture controls), and each cell is one (loss, seed) combination. The four-loss equivalence tables (Tab. 7–11) report best-dev WER without any intervention; the swap and asymmetric-swap tables (Tab. 12–17) apply the inference-time interventions defined in §3–4 to the *same* checkpoints.

Loss	seed 1	seed 2	seed-mean	spread
BlankGateReg	19.51	18.81	19.16	0.70
ContentGateReg	19.69	19.08	19.39	0.61
NoAux	18.86	19.31	19.09	0.45
GateVarianceMax	18.90	18.59	18.75	0.31

Table 7: Four-loss equivalence, CSLR PHOENIX-2014, R-18 + BiLSTM + ResMix, multi-seed (best dev WER). Cross-loss spread of seed-means: **0.64 WER** (18.75–19.39). Within seed noise.

Loss	seed 1	seed 2	seed 3	seed-mean
BlankGateReg	82.86	82.01	81.80	82.22
ContentGateReg	82.27	81.36	81.27	81.63
NoAux	79.64	81.08	80.82	80.51
GateVarianceMax	81.96	82.15	81.72	81.94

Table 8: Four-loss equivalence, ASR LibriSpeech-clean-100, wav2vec2-base (frozen) + ResMix + linear, multi-seed (best dev WER). Cross-loss spread of seed-means: **1.71 WER** (80.51–82.22).

Loss	seed 1 (best ep)	seed 2 (best ep)
NoAux	**7.57** (ep9)	**7.79** (ep9)
BlankGateReg	**7.67** (ep11)	**7.55** (ep8)
ContentGateReg	**7.48** (ep10)	**7.64** (ep8)
GateVarianceMax	8.42 (ep4; aw=0.1)	**7.65** (ep11)

Table 9: Four-loss equivalence, ASR full fine-tune, wav2vec2-base + BiLSTM + ResMix, best Dev WER. Seed 1 cross-loss spread (3 reported cells at ep9–11): **0.19 WER, all under 7.7%**; seed 2 spread (all 4 cells, ep8–11): **0.24 WER, all under 7.8%**. The GateVarianceMax seed 1 cell required aux_weight=0.1 (the 1.0 setting diverged) and is still descending; it joins the equivalence band at convergence in seed 2.

Loss	seed 1 (ep11)	seed 2 (ep11)
NoAux	8.28	8.33
BlankGateReg	8.40	8.50
ContentGateReg	8.46	8.51
GateVarianceMax	8.45	—

Table 10: Four-loss equivalence, ASR full fine-tune, wav2vec2-base + **LINEAR** + ResMix, final epoch (ep11) Dev WER. Seed 1 spread: **0.18 WER, all under 8.5%**; seed 2 spread: 0.18 WER. The linear-head equivalence is tighter than the BiLSTM-head equivalence despite the absolute WER being $\sim$1 WER higher.

Loss	seed 1	seed 2	seed 3
NoAux	8.37	8.32	—
BlankGateReg	8.46	—	8.45
ContentGateReg	8.37	—	—
GateVarianceMax	8.45	—	—

Table 11: Four-loss equivalence, ASR full fine-tune, wav2vec2-base + BiLSTM + **GluMix** ($\alpha = 0$), final-epoch Dev WER. Across 7 trained cells (incomplete 4×3 matrix: GateVarianceMax seed-2 and ContentGateReg seed-2/3 diverged during fine-tuning, denoted "—") the cross-loss spread is ≤ 0.14 WER; the non-α module hits the same operating point as the linear-head ResMix matrix (Tab. 10).

D Additional Analyses

Gate vs. p_{blank} correlation. A reviewer asked: does the gate's per-frame correlation with P_{blank} predict Δ_{gauge} or Δ_{asym}? We measure the Pearson

Loss	seed	Dev unswap	Dev swap	Δ_{gauge} Dev	Test unswap	Test swap	Δ_{gauge} Test
BlankGateReg	1	27.80*	27.95*	$+0.15$	drift‡	—	—
BlankGateReg	2	18.81	18.99	$+0.18$	19.27	19.25	-0.02
ContentGateReg	1	29.12	29.36	$+0.24$	28.99	29.46	$+0.47$
ContentGateReg	2	19.11	19.22	$+0.11$	19.64	19.85	$+0.22$
NoAux	1	18.86	19.01	$+0.14$	19.96	20.04	$+0.08$
NoAux	2	19.39	19.49	$+0.11$	19.50	19.41	-0.09
GateVarianceMax	1	18.90	18.97	$+0.07$	20.21	20.41	$+0.20$
GateVarianceMax	2	18.57	18.81	$+0.23$	19.94	19.82	-0.12

Table 12: Gauge-preserving swap on CSLR PHOENIX-2014, R-18 + BiLSTM + ResMix. *Cell with post-hoc-eval drift (training-time dev WER was 19.51; locked-decoder rerun queued). ‡Test under same drift; Δ on the same post-hoc pipeline remains the load-bearing quantity. Excluding the drift cell: Δ_{gauge} Dev range $[+0.07, +0.24]$, Δ_{gauge} Test range $[-0.12, +0.47]$, mean $+0.10$.

Loss	seed	Δ_{gauge} Dev	Δ_{gauge} Test
NoAux	1	$+0.22$	$+0.17$
NoAux	2	$+0.12$	$+0.11$
BlankGateReg	1	$+0.07$	$+0.09$
BlankGateReg	2	$+0.06$	$+0.06$
ContentGateReg	1	$+0.26$	$+0.09$
ContentGateReg	2	$+0.22$	$+0.13$
GateVarianceMax	1	$+0.19$	$+0.13$

Table 13: Gauge-preserving swap on ASR LibriSpeech-clean-100, wav2vec2-base **full-FT + LINEAR head + ResMix**. 7 cells (final epoch ep11). Δ_{gauge} Dev range $[+0.06, +0.26]$, mean $+0.17$; Test range $[+0.06, +0.17]$, mean $+0.11$. *Smaller* than the BiLSTM-head matrix at the same encoder regime (Table 3), inverting the original "weaker head $\rightarrow$ larger Δ_{gauge}" prediction.

Loss	Dev unswap	Dev asym	Δ_{asym} Dev
NoAux	8.28	**99.03**	$+90.75$
BlankGateReg	8.40	83.68	$+75.28$
ContentGateReg	8.46	65.55	$+57.10$
GateVarianceMax	8.45	**99.68**	$+91.23$

Table 14: Asymmetric-swap negative control on the linear-head full-FT matrix (Tab. 13, seed 1). All four losses collapse to $+57$ to $+91$ WER. Varmax is the *largest* Δ_{asym} here — opposite the BiLSTM-head matrix — because the linear head cannot absorb the asymmetric perturbation in the Varmax basin. Per-loss ratios: NoAux $413\times$, BlankGateReg $1076\times$, ContentGateReg $220\times$, GateVarianceMax $480\times$.

Loss	Dev unswap	Dev asym	Δ_{asym} Dev
NoAux	8.37	**98.60**	$+90.23$
BlankGateReg	8.46	70.06	$+61.60$
ContentGateReg	8.37	71.48	$+63.11$

Table 15: Asymmetric-swap negative control on the GluMix matrix (GluMix cells from Table 2, seed 1). 3 cells; all collapse to $+61$ to $+90$ WER. Combined with $\Delta_{\text{gauge}} \equiv 0$ bitwise on the same cells, the discrimination ratio for GluMix is mathematically infinite.

Architecture	**Dev unswap**	**Dev asym**	$\boldsymbol{\Delta_{\text{asym}}}$ **Dev**
CSLR R-18 + BiLSTM (8-cell mean)	—	—	$\sim+62.0$
CSLR R-18 + Transformer (s2)	22.67	97.45	$+74.78$
CSLR R-18 + Conformer (s1)	48.13	99.39	$+51.26$
ASR wav2vec2 + linear (frozen, gate_s1)	79.64	99.97	$+20.33$

Table 16: Cross-architecture asymmetric-swap control on additional substrates (CSLR Transformer/Conformer; frozen-encoder ASR). Δ_{asym} is catastrophic on every architecture tested ($+20$ to $+75$ WER), confirming the audit fires on every backbone, not only the near-SOTA BiLSTM head.

correlation streaming over the full LibriSpeech dev split ($\sim$947k frame positions per cell).

(i) BlankGateReg achieves its design goal. $\rho(g, P_{\text{blank}})$ moves from -0.20 (ResMix noaux) / -0.25 (GluMix noaux) to $+0.96$ / $+0.98$ under BlankGateReg. ContentGateReg moves the correlation symmetrically in the opposite direction (-0.95 / -0.98). The four-loss equivalence (§6.1) is therefore not because the losses fail to do anything; it is because the things they do are gauge-equivalent.

(ii) Negative result. The reviewer's predictive hypothesis was *"stronger $|\rho(g, P_{\text{blank}})|$ should predict larger Δ_{gauge}."* The data say the *opposite*: on the GluMix matrix (where $\Delta_{\text{gauge}} \equiv 0$, so we test on Δ_{asym} instead), the rank correlation is $\rho(|\rho|, \Delta_{\text{asym}}) = -0.999$ (n=3). Cells with stronger g/P_{blank} alignment have *smaller* asymmetric-swap costs. On ResMix the signed correlation is $\rho(\rho, \Delta_{\text{gauge}}) = -0.54$ (n=4) and unsigned $\rho(|\rho|, \Delta_{\text{gauge}}) = +0.27$ – both weak and in conflict with the original intuition.

Mechanistic interpretation: a high $|\rho(g, P_{\text{blank}})|$ indicates the gate's two basins (g, $1-g$) carry complementary, easily-separable information. The downstream stack absorbs the swap into either basin because both are clean, semantically-aligned signals. The noaux cell sits at $g \approx \text{const}$ ($\text{std}\, 0.02$); the gauge swap moves a near-constant

Modality / Module	$\boldsymbol{\Delta}_{\text{gauge}}$ **range**	$\boldsymbol{\Delta}_{\text{asym}}$ **range**	$\boldsymbol{R = \Delta_{\text{asym}}/\Delta_{\text{gauge}}}$
ASR ResMix + BiLSTM head (8 cells)	$[-0.02, +0.36]$	$[+7, +90]$	130–1076×
ASR ResMix + Linear head (7 cells)	$[+0.06, +0.26]$	$[+57, +91]$	220–1076×
ASR GluMix (3 cells, α=0)	$\equiv 0$ (bit)	$[+61, +90]$	∞ ($\geq$ 6000×)
IAM OCR TrOCR-large + GluMix (3)	$\equiv 0$ (bit)	$[+0.85, +1.50]$	∞
IAM OCR TrOCR-base + GluMix-CTC (4)	$\equiv 0$ (bit)	$[+31, +43]$	∞
CSLR R-18 + BiLSTM + ResMix (8 cells)	$[-0.12, +0.47]$	$[+20, +75]$	$\sim$ 50–500×

Table 17: Four-modality discrimination summary at each modality's near-SOTA operating point. The audit returns $R \geq 100\times$ on every cell tested; on GluMix variants (α=0) the ratio is mathematically infinite. Within-CSLR architecture spectrum (BiLSTM/Transformer/Conformer) and frozen-encoder controls in the tables below.

Cell	Loss	$\rho(g, P_b)$	Δ_{gauge}	Δ_{asym}
ResMix (BiLSTM head)				
noaux_s1	–	-0.20	$+0.36$	$+96.56$
blank_s1	Blank	$+0.96$	$+0.24$	$+63.30$
content_s1	Cont.	-0.95	$+0.36$	$+55.13$
varmax_s1	Var.	$+0.36$	-0.02	$+14.58$
GluMix (linear head)				
glu_noaux_s1	–	-0.25	$\equiv 0$	$+98.60$
glu_blank_s1	Blank	$+0.98$	$\equiv 0$	$+70.06$
glu_content_s1	Cont.	-0.98	$\equiv 0$	$+71.48$

Table 18: Per-frame Pearson correlation $\rho(g, P_{\text{blank}})$ on dev-clean (full split, $\sim$947k positions per cell), with the corresponding Δ_{gauge} and Δ_{asym}. BlankGateReg achieves $\rho > 0.95$ on both ResMix and GluMix – the loss does what it advertises. The correlation does *not* predict the swap deltas in the direction the reviewer's intuition suggested: $\rho(|\rho|, \Delta_{\text{asym}}) = -0.999$ on GluMix (n=3), i.e. cells with *stronger* g/P_{blank} alignment have *smaller* asymmetric-swap costs.

gate to a different near-constant value (easy for the head), but the *asymmetric* swap leaves mismatched basin/projection wiring and is catastrophic ($\Delta_{\text{asym}} = +96.6$). $\rho(g, P_{\text{blank}})$ **is a good fingerprint of what the gate learned, but a poor predictor of swap fragility.** The reviewer's hypothesis was natural and we tested it; it failed. The gauge-symmetry framing explains the failure: the swap is closed-form algebraic, not an information attack, so semantic fingerprints of the gate do not predict its cost.

E Asymmetric-Swap Pseudocode

The asymmetric swap is implemented as a single forward-pass override. The single-line change vs. the gauge-preserving swap is the gate flip *without* the projection swap; the mixing term becomes $g \cdot s + (1-g) \cdot r$ – the gauge-broken mix.

```
def asymmetric_swapped_forward(self, z, m):
    """Asymmetric swap: flip g only.
    mix = g*s + (1-g)*r  <- gauge-BROKEN."""
    z_ln = self.ln_z(z)
    m_ln = self.ln_m(m)
    cat  = torch.cat([z_ln, m_ln], dim=-1)
    g = 1.0 - torch.sigmoid(self.g_fc(cat))
    # gate flipped; projections NOT swapped
    s = F.relu(self.s_fc(z_ln))
    r = F.relu(self.r_fc(z_ln))
    mix = (1.0 - g)*s + g*r
    return mix + self.alpha * g * z_ln
```

The gauge-preserving counterpart differs in exactly two lines: the s and r projection calls are swapped (restoring the gauge equivalence).