RECOGNITION-AWARE SELECTION OF ENHANCEMENT-DERIVED TRANSCRIPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Different speech enhancers supply a frozen recognizer with complementary corrections and conflicting errors. CARE-ASR selects among the resulting transcripts by combining expert risk estimates with transcript agreement and acoustic support. It learns word-error differences and calibrates replacement against a complete default policy. We evaluate four English corpora, then apply the original meeting-speech policy without recalibration to 4,096 project-held-out AMI recordings from 16 meeting–speaker groups. Word error rate falls from 30.040% for reverse ROVER to 27.860%, a 2.180-percentage-point reduction (95% group-bootstrap interval [1.455, 2.919]; adjusted $P = 0.000275$). Fifteen groups improve, with more recovered words, fewer erasures and fewer insertions overall. Same-cohort controls examine the representation: expert inputs help under both pairwise and MSE training, whereas a matched nonlinear selector using only eight expert risks makes 210 more errors. When expert predictions for the AMI fitting sources are generated out of fold, the gain over reverse ROVER persists, and masking the explicit expert inputs adds 240 errors. The findings support learning from expert judgements together with candidate evidence to exploit enhancement diversity without updating the recognizer.

1 INTRODUCTION

An enhancer can make speech easier to hear yet harder for a recognizer to transcribe. Suppressing noise may remove a useful cue; retaining it may obscure another. When several enhanced views are available, the practical question is which resulting transcript to trust. Signal quality alone does not answer that question.

Metric-oriented, flow-based and predictive enhancement models supply diverse hypotheses (Fu et al., 2021; Lee et al., 2025; Wang & Tian, 2025; Rong et al., 2026; Fu et al., 2026). A fixed pipeline commits to one enhanced view; fixed-order word voting exploits agreement across views. Neither learns directly from their recognition errors. In our initial AMI evaluation, raw recognition gives 41.539% word error rate (WER), the best individual enhancer gives 41.315%, and reverse-order ROVER voting gives 38.348%. Learned selection reduces WER to 36.148%. What information lets a selector recover corrections that the strongest individual pipeline and voting leave unused?

CARE-ASR uses trained risk predictors (our experts) both to nominate transcripts and to inform downstream selection. Each candidate is represented through expert risks, transcript relations and acoustic support from the frozen recognizer. A small network combines these signals to predict word-error differences relative to a voting transcript. Whole-transcript and edit-level scores describe acoustic support at two resolutions; calibration determines when to choose another complete transcript (Figure 1).

The default transcript can itself come from a policy that varies across recordings. CARE-ASR separates learning which candidate to prefer from deciding whether to replace this complete default.

We compare the complete system with five released enhancers, two ROVER orders and eight learned risk policies under a common recognizer and scoring procedure. We then apply the original AMI policy to 4,096 recordings from 16 groups absent from the audited project history. With its weights and calibration unchanged, it reduces errors by 7.26% relative to reverse ROVER and improves 15 groups.

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

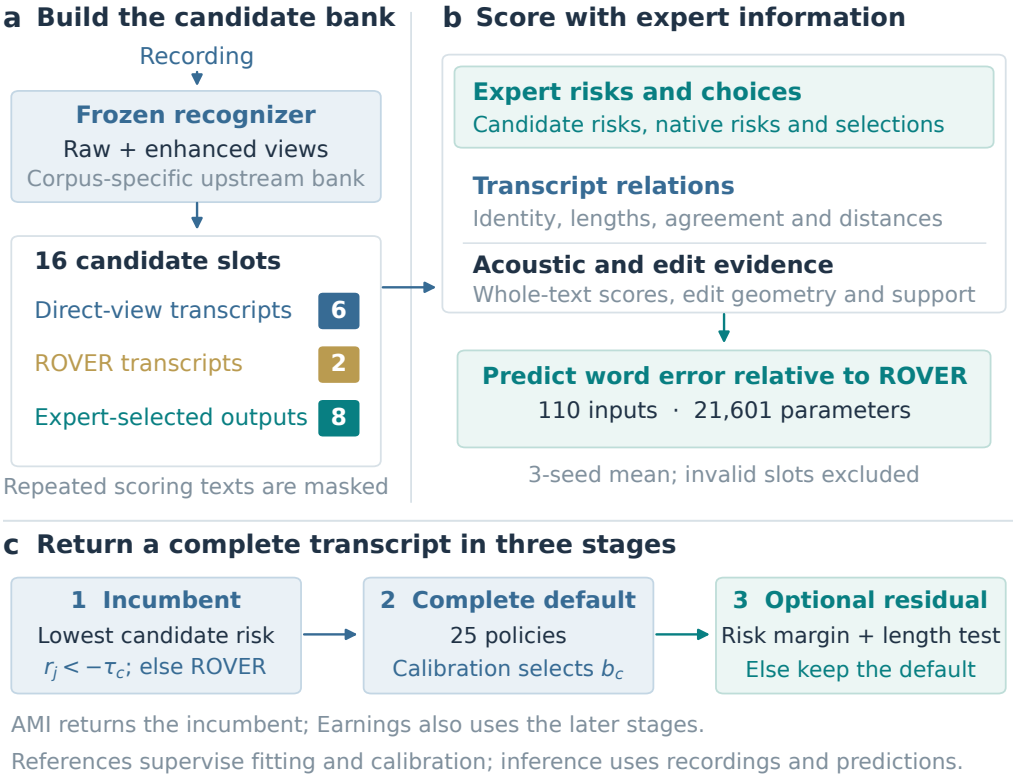


Figure 1: **Recognition-aware selection with CARE-ASR. a,** The interface has six direct-view, two ROVER and eight expert-output slots. Experts select from the corpus-specific upstream bank, which can include views outside the direct slots. **b,** Expert judgements, transcript relations and acoustic evidence inform candidate risks. **c,** Execution proceeds from incumbent selection under margin τ_c, through a calibrated complete default, to an optional residual with margin m and a length test.

Our contribution is a recognition-aware selection procedure built around an expert-informed candidate representation. Matched refits examine what this representation contributes beyond either selected expert texts or expert risks alone. AMI-specific cross-fitting tests its dependence on in-sample expert scores, and shared-risk ROVER compares selection with quality-ranked fusion. Candidate scoring drives the AMI gain; complete-default selection contributes most of the additional improvement on Earnings.

2 Recognition-aware hypothesis selection

2.1 Prediction space and objective

For recording x_i, a frozen recognizer f produces hypotheses $h_{ik} = f(T_k(x_i))$ from frozen enhancement transforms T_k, including the identity transform. Selection chooses among these texts and their voting combinations without updating the recognizer. The final interface contains six fixed direct-view transcripts, two ROVER transcripts and eight expert-selected outputs. Experts draw from the corpus-specific upstream bank and may return a native transcript outside the six direct-view slots. A policy π selects one candidate using the recording, hypothesis texts and model scores; y_i is its training reference.

Corpus WER is the total word-edit distance divided by the reference-word count:

$$\mathrm{WER}(\pi) = \frac{\sum_i \mathrm{ED}(y_i, h_{i,\pi(i)})}{\sum_i |y_i|} = \frac{S_\pi + D_\pi + I_\pi}{N}, \tag{1}$$

Table 1: **Candidate-risk representation.** Every candidate uses the same 110-field layout. Whole masks both edit blocks; Edit retains them. The eight-risk control retains only candidate-specific expert risks within the first block. Counts include masked fields.

Feature block	Fields	Information supplied
Transcript and expert relations	80	Agreement, lengths, slot and expert risks
Whole-audio scores	6	Conditional support for the complete text
Edit geometry	8	Extent and type of transcript changes
Edit support	16	Scores at changed and neighbouring tokens
Total	110	Shared layout across configurations

where S, D and I denote substitutions, deletions and insertions. We also align each selected hypothesis and the raw transcript to the reference using the same deterministic edit rule. Recovery R counts raw-incorrect reference positions made correct; erasure E counts raw-correct positions made incorrect. With $L = S + D + I$,

$$L_{\text{raw}} - L_{\pi} = R - E + I_{\text{raw}} - I_{\pi}. \tag{2}$$

Keeping insertions explicit distinguishes gains at reference-word positions from changes in transcript length. Recovery and erasure describe prediction changes, rather than changes in the underlying acoustic evidence.

2.2 Representing candidate disagreement

The ordered bank has 16 slots: two ROVER hypotheses, raw plus five enhanced hypotheses, and eight risk-expert outputs. Slot identity records a candidate's origin; repeated scoring texts are invalidated before selection. Each expert selects an existing transcript from the corpus-specific upstream bank. The experts cross four feature families—output geometry, global summaries, paired aligned features and circular-refit features—with regression or decision-cost training (Appendix C).

Each candidate has 110 input fields. The first 80 describe its slot identity, transcript lengths, edit distances, view agreement and expert predictions. Six fields summarize its conditional token scores under the original recording. The remaining 24 contain eight edit-geometry fields and 16 edit-support fields. The feature definition is shared across corpora.

Edit support is constructed by aligning each candidate's tokens to slot zero, the raw-first ROVER transcript. Scores are summarized over changed tokens, their neighbours and unchanged token pairs. Conditional log probabilities under the original recording and a silence control provide an audio–silence contrast at the changed regions. These regions are defined by model hypotheses, not by reference text. Appendix C lists the fields and specifies how empty regions are handled.

An encoder with widths $110 \to 64 \to 32$ and GELU activations maps each candidate to an embedding. The set-context variant concatenates this embedding, the slot-zero embedding, their difference, and masked mean and maximum embeddings over candidates. A $160 \to 64 \to 32 \to 1$ head then predicts relative risk. The independent variant zeros the four context blocks while retaining the same head. Each scorer has 21,601 trainable parameters. Duplicate scoring texts are masked, and scores are centred on slot zero.

Whole-audio scores pool disputed and undisputed tokens; edit fields describe the changes and their acoustic support separately. Both variants retain transcript agreement and expert predictions. References supply fitting and calibration targets, whereas inference features use only recordings and candidate outputs.

2.3 Learning relative word error

Let $e_{ij} = \text{ED}(y_i, h_{ij})$, $d_i = \max(1, |h_{i,\text{raw}}|)$ and $t_{ij} = (e_{ij} - e_{i0})/d_i$. The target expresses how many errors a candidate adds or removes relative to the voting transcript, normalized by raw-hypothesis length. Lower scores indicate preferable candidates.

For valid slots V_i, let $\mathcal{P}_i = \{(j,k) \in V_i^2 : t_{ik} > t_{ij}\}$. Pairwise training weights each ordering by its word-error difference:

$$M_i = \frac{\sum_{j \in V_i \setminus \{0\}} (r_{ij} - t_{ij})^2}{\max(1, |V_i \setminus \{0\}|)}, \tag{3}$$

$$\ell_i^{\text{pair}} = \frac{\sum_{(j,k) \in \mathcal{P}_i} (t_{ik} - t_{ij}) \operatorname{softplus}((r_{ij} - r_{ik})/.03)}{\max(1, |\mathcal{P}_i|)} + .05 M_i. \tag{4}$$

The pairwise term weights ordering mistakes by their word-error cost, with fixed score scale .03; the $.05M_i$ term penalizes inaccurate risk magnitudes. A soft-target alternative uses cross-entropy from $\operatorname{softmax}(-t_i/.01)$ to $\operatorname{softmax}(-r_i/.03)$, plus $.05M_i$. Both exclude invalid slots. Matched controls also fit M_i alone with the same inputs and architecture.

The raw-length denominator is shared across candidates and available at inference; negative risk favours replacing the voting transcript.

Training is corpus balanced. Scaling is estimated on training data, and three seed scores are averaged for selection. Whole-audio configurations mask the 24 edit fields; edit-support configurations retain them, with the network shape unchanged. The initial study crosses input variant, loss, candidate context and two learning rates, giving 16 configurations. Short-form settings subsequently allow bounded corpus adaptation. Earnings uses complete-call calibration. Training and adaptation settings appear in Appendix D.

2.4 Calibrating the deployed policy

Calibration turns candidate scores into a complete output rule in three stages. First, the ensemble forms an *incumbent*: among valid nonzero slots, it chooses the lowest risk only if that risk is below $-\tau_c$; otherwise it returns slot zero (raw-first ROVER). The corpus-calibrated incumbent margin τ_c is distinct from the residual margin m below. A no-change setting always returns slot zero.

Second, calibration chooses a complete default b_c from 25 policies: the 16 fixed-slot policies, eight guarded expert policies and the incumbent. Selection minimizes calibration word errors; ties favour the incumbent, then fewer changes from raw, then a deterministic identifier.

Third, an optional residual compares candidates with the selected default. Let b be the index of the first valid candidate whose donor string equals $b_c(x)$, so $h_b = b_c(x)$; b is an index, not a policy. When this match exists, length admissibility is

$$\mathcal{A}_{q,k}(x) = \{j \in V : |b_c(x)| < k \text{ or } |h_j| \geq q|b_c(x)|\} \cup \{b\}. \tag{5}$$

For $j^* = \arg\min_{j \in \mathcal{A}_{q,k}} r_j$, the returned transcript is

$$\pi(x) = \begin{cases} h_{j^*}, & r_{j^*} < r_b - m, \\ b_c(x), & \text{otherwise.} \end{cases} \tag{6}$$

The rule replaces the default only when an admissible candidate has a predicted advantage exceeding m; the length test limits contraction. Without a valid default score, it returns the original donor string. Calibration considers five (q,k) settings crossed with seven margins and a no-change option. Appendix D gives the exact settings and text-identity convention.

Candidate generation and conditional scoring precede these policy stages. Appendix D.5 specifies execution with the stored scalers, masks, margins and donor identities.

3 Experimental design

3.1 Data and recognition pipeline

We use English subsets of FLEURS, TED-LIUM 3, AMI and Earnings-22 (Conneau et al., 2022; Hernandez et al., 2018; Carletta, 2006; Del Rio et al., 2022). The initial evaluation includes 80 FLEURS recordings, 95 TED segments, 3,772 AMI segments and two complete Earnings calls, with 1,694, 2,320, 22,314 and 22,796 reference words. Parameters are fitted on training sources.

Separate calibration sources select checkpoints, configurations, defaults and thresholds. The configuration study reports calibration results; the main recognition comparisons use the evaluation role. Appendix D gives allocations and prior use.

A further evaluation uses 4,096 AMI SDM recordings from 16 meeting–speaker connected groups, 256 recordings per group and 31,565 reference words. Source identifiers, group membership and decoded-audio hashes are checked against the audited project history. The original AMI policy, matched controls and analysis are fixed before inference on this cohort. This project-held-out sample retains the publisher's split labels; it is not the complete official test split. Appendix F gives admission rules, the execution timeline and the three prespecified comparisons.

All pipelines use frozen Qwen3-ASR-0.6B (Shi et al., 2026). Compared methods within a corpus receive the same sources and use the same references, normalization and scoring. For Earnings, selected chunk strings are concatenated in their original order before whole-call normalization and alignment. Reference words are scored at the complete-call level.

3.2 Baselines and analysis

External comparisons include raw recognition, FlowSE, MetricGAN+, ZipEnhancer, PASE, RE-USE and both ROVER orders. Internal comparisons include eight learned risk policies crossing the feature families and objectives above, plus two earlier configurations of our system: the earlier reranker and pooled audio-supported ranker. These internal variants distinguish candidate modelling from later corpus calibration and policy composition.

WER is the primary metric; S/D/I, recovery, erasure and per-source changes provide complementary error accounting. We report exact counts and rounded WER over the recorded sources. Inference cost includes candidate generation and acoustic scoring, so single-enhancer rows use fewer resources. The 21,601-parameter count refers to the trainable scorer.

Statistical inference on the new cohort uses the 16 connected groups as paired units. The three prespecified comparisons use exact two-sided sign-flip tests with Holm correction, accompanied by 50,000 paired group-bootstrap percentile intervals for micro-WER differences. Model, threshold and mixture choices are fixed before this evaluation.

Subsequent information controls, cross-fitting and fusion comparisons reuse this cohort, with training and policy choices confined to the original fitting and calibration roles. Their comparison families are reported separately from the original evaluation.

4 Results

4.1 Selection improves on individual enhancement

CARE-ASR achieves WERs of 3.601%, 2.155%, 36.148% and 11.410% on FLEURS, TED, AMI and Earnings, respectively (Table 2), outperforming FlowSE and MetricGAN+ in all four settings. It also outperforms PASE and RE-USE on three corpora and ties both on TED.

The largest gain is on AMI, where the best individual enhancer, RE-USE, changes raw recognition's 9,269 errors to 9,219. CARE-ASR instead gives 8,066 errors: 1,203 fewer than raw and 1,153 fewer than RE-USE, corresponding to relative WER reductions of 12.98% and 12.51%. Earnings errors fall from 2,728 to 2,601, a 4.66% relative reduction. FLEURS and TED start at much lower WERs and show smaller changes.

4.2 More recovery with less erasure than voting

Reverse ROVER is the stronger voting order on AMI, with 8,557 errors. CARE-ASR reduces this by 491, or 5.74% relative. The change combines three improvements (Figure 2): recovery rises from 1,196 to 1,408, erasure falls from 671 to 538, and insertions fall from 1,028 to 882. Thus

$$491 = \underbrace{212}_{\text{additional recovery}} + \underbrace{133}_{\text{fewer erasures}} + \underbrace{146}_{\text{fewer insertions}} . \tag{7}$$

Table 2: **Recognition with enhancement, voting and learned selection.** Corpus WER (%, lower is better) on the initial evaluation subsets. Bold indicates the column minimum, including ties. Single-enhancer pipelines and multi-candidate policies have different inference budgets.

System	FLEURS	TED-LIUM 3	AMI	Earnings-22
Raw	3.837	2.241	41.539	11.967
FlowSE	4.250	3.966	67.662	18.266
MetricGAN+	4.191	4.353	54.293	22.530
ZipEnhancer	3.837	3.491	42.874	55.676
PASE	6.494	2.155	46.159	13.656
RE-USE	3.719	2.155	41.315	11.871
ROVER (raw first)	**3.601**	2.155	38.604	11.594
ROVER (reverse)	**3.601**	**2.112**	38.348	11.620
CARE-ASR	**3.601**	2.155	**36.148**	**11.410**

ZipEnhancer uses cached segmented-inference outputs with the tail-writing issue described in Appendix C.2; Appendix F.8 tests the rebuilt pipeline without this path.

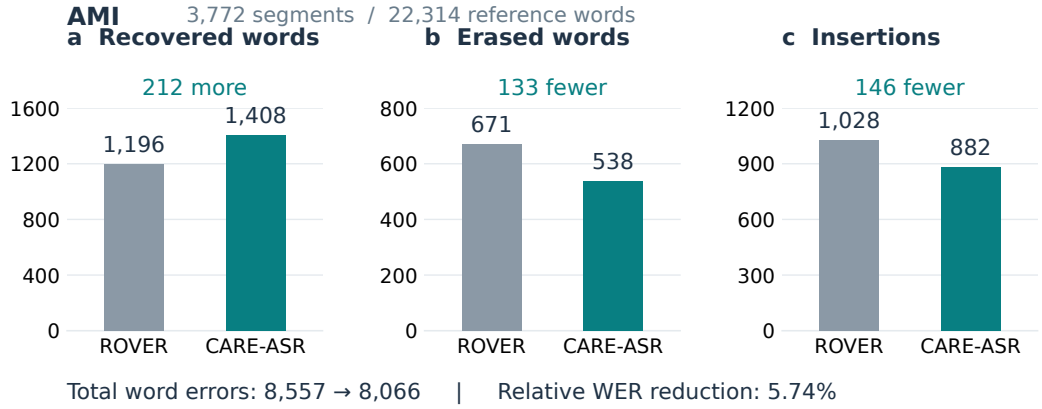


Figure 2: **Beyond voting on AMI.** CARE-ASR and reverse ROVER on the same 3,772 segments and 22,314 reference words. Recovery and erasure are defined relative to raw recognition; insertions are total insertion errors. The three differences sum to 491 fewer word errors. These are complete-system comparisons.

The gain combines more corrected reference-word positions with fewer erasures and insertions. Relative to raw recognition, CARE-ASR improves 839 AMI segments, leaves 2,634 unchanged and worsens 299. The improved segments remove 1,709 errors and the worsened segments add 506, yielding the same net reduction of 1,203.

On Earnings, CARE-ASR also improves both voting orders, with 42 fewer errors than raw-first ROVER and 48 fewer than reverse ROVER. FLEURS ties both voting orders. Reverse ROVER makes one fewer error on TED. The AMI gain is distributed across all four sessions: 71, 123, 114 and 183 fewer errors than reverse ROVER. Both Earnings calls also improve, by 26 and 22 errors (Appendix B.2).

4.3 Learned selectors and policy composition

Table 3 compares learned selectors on the same sources. CARE-ASR removes 391 AMI errors relative to the output-geometry decision-loss control and 47 relative to the aligned decision-loss control. The pooled audio-supported precursor gives 8,069 errors, close to the final system's 8,066. For Earnings, output-geometry, aligned and pooled selectors give 2,689, 2,822 and 2,779 errors, respectively, compared with 2,601 for CARE-ASR. The lowest TED count is 46, from the output-geometry MSE control.

Table 3: **Learned-selector variants.** WER (%) on identical sources. The risk policies are internal controls; the earlier reranker and pooled ranker are internal precursors to corpus-calibrated CARE-ASR. Bold indicates the column minimum, including ties. Appendix A lists exact counts.

Decision policy	FLEURS	TED-LIUM 3	AMI	Earnings-22
Global: decision loss	3.837	2.069	36.757	11.875
Output: decision loss	**3.601**	2.112	37.900	11.796
Aligned: decision loss	3.719	2.198	36.358	12.379
Circular-refit: decision	3.719	2.155	36.672	11.717
Global: MSE	3.660	2.586	36.815	12.090
Output: MSE	**3.601**	**1.983**	37.909	11.704
Aligned: MSE	3.660	2.241	37.282	12.353
Circular-refit: MSE	3.719	2.112	37.048	12.335
Earlier reranker	**3.601**	2.241	36.421	11.906
Pooled audio-supported ranker	3.719	2.155	36.161	12.191
CARE-ASR	**3.601**	2.155	**36.148**	**11.410**

Table 4: **Where policy composition changes recognition.** Word errors with fitted candidate scores held fixed. Each row adds one policy stage; calibration determines whether that stage changes the output.

Cumulative configuration	FLEURS	TED-LIUM 3	AMI	Earnings-22
Incumbent policy	61	50	8066	2676
+ Complete-default selection	61	50	8066	2609
+ Bounded residual (CARE-ASR)	61	50	8066	2601

On AMI, CARE-ASR improves on the aligned decision-loss control in every session, by 15, 4, 24 and 4 errors. Its final output comes from the calibrated candidate model: edit support, pairwise loss, independent candidate encoding, adaptation at 10^{-5} and margin .01. Table 4 isolates the subsequent policy stages by holding these scores fixed.

FLEURS, TED and AMI retain the incumbent outputs through both later stages. Earnings follows a different pattern: changing the complete default reduces errors from 2,676 to 2,609, and the residual reaches 2,601. The residual recovers eight additional reference words and erases six fewer, with six extra insertions. Default selection therefore accounts for most of the improvement beyond the Earnings incumbent; the adapted candidate model already supplies AMI's final result.

4.4 Frozen evaluation on new source groups

Applied without recalibration to the project-held-out AMI cohort, the original CARE-ASR policy produces 8,794 errors (27.860% WER), compared with 10,048 (31.833%) for raw recognition and 9,482 (30.040%) for reverse ROVER (Figure 3). The 688 errors removed relative to voting comprise 368 additional recoveries, 239 fewer erasures and 81 fewer insertions. The reduction has a group-bootstrap 95% interval of [1.455, 2.919] WER percentage points and Holm-adjusted $P = 0.000275$. Fifteen source groups improve and one worsens.

4.5 What supports selection within the common bank?

Matched feature controls share architecture, candidates, seeds and calibration rules. Edit, Whole, MSE and the archived tree yield similar aggregate error counts (Appendices F and F.5). The following same-cohort analyses examine the shared representation; they were specified after the original results, with fitting and selection confined to the original roles.

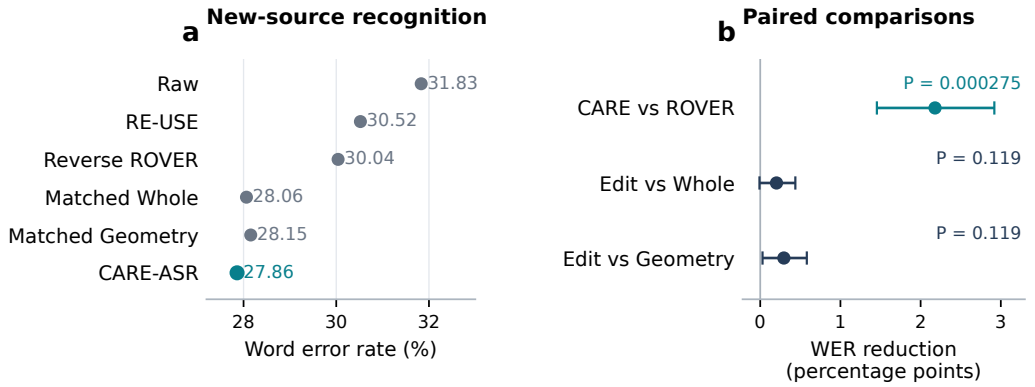


Figure 3: **Recognition on project-held-out AMI sources. a,** WER on 4,096 recordings; CARE-ASR and matched Edit return identical transcripts. **b,** Positive differences favour CARE or Edit. Bars are unadjusted 95% paired-bootstrap intervals over 16 connected groups; P values are Holm-adjusted exact sign-flip tests for the three prespecified comparisons. Appendix F reports all methods and group counts.

Table 5: **Learning from the same candidate bank.** All rows use 4,096 recordings and 31,565 words. Masked rows remove 33 expert fields; eight-risk rows retain only candidate-specific risks. Masking effects are 285 errors (pairwise) and 309 (MSE); the last column compares every row with CARE. See Appendices F.5–F.10.

Method	Errors	WER (%)	Excess errors vs CARE
CARE-ASR	8,794	27.860	0
Matched MSE	8,855	28.053	+61
Archived tree	8,885	28.148	+91
Full-input linear (110 fields)	8,935	28.307	+141
Eight-risk nonlinear	9,004	28.525	+210
Calibration-selected expert	9,040	28.639	+246
Equal expert average	9,044	28.652	+250
Eight-risk linear combination	9,046	28.658	+252
Pairwise, expert inputs masked	9,079	28.763	+285
MSE, expert inputs masked	9,164	29.032	+370
Calibrated candidate MBR	9,574	30.331	+780
Uniform candidate MBR	12,155	38.508	+3,361

Expert slots add no distinct transcripts on this cohort: all reproduce native candidate texts. Masking their 33 explicit input fields and retraining nevertheless increases errors from 8,794 to 9,079 under pairwise training, and from 8,855 to 9,164 under MSE. These effects are 0.903 WER points [0.263, 1.592] ($P = 0.0152$) and 0.979 points [0.519, 1.435] (Holm $P = 0.004211$). Both benefits are dominated by fewer erasures (323 and 385), despite more insertions (92 and 88).

Are the risk scores themselves enough? Retaining only the eight candidate-specific risks in the same nonlinear architecture gives 9,004 errors, 210 more than full inputs: 0.665 WER points [0.339, 1.013], exact group $P = 0.001221$ in a separate single-contrast follow-up. Thirteen groups improve. Here CARE gains 165 recoveries and removes 91 insertions, offset by 46 additional erasures. The comparison tests the remaining fields jointly, not a particular acoustic block (Appendix F.10).

Full-input linear selection leaves 141 additional errors (0.447 points [0.160, 0.767], Holm $P = 0.017822$); the calibration-selected expert leaves 246. Appendix F.9 reports the joint tests, separate from the risk-only contrast.

Candidate-path sensitivity. Without ZipEnhancer, rebuilding the bank and refitting all experts and selectors gives 8,902 CARE errors versus 9,614 for reverse ROVER: a +2.256-point WER reduction [1.746, 2.754], $P = 0.000061$. CARE also improves on the refitted linear control; matched MSE remains close (Appendix F.8).

4.6 Cross-fitting and quality-ranked fusion

We partition the 7,095 AMI fitting sources into three meeting-key folds and generate their expert predictions out of fold, then retrain the downstream scorer. Non-AMI fitting features and all calibration records remain unchanged. With cross-fitted AMI expert inputs, the retrained scorer produces 8,793 errors, compared with 8,794 for the original-bank refit and 9,033 when the explicit expert fields are masked. The 240-error information benefit is 0.760 WER points [0.228, 1.282], with Holm $P = 0.016235$; the 689-error advantage over reverse ROVER is 2.183 points [1.557, 2.760], $P = 0.000122$ in the same two-test family (Appendix G.1).

ROVER uses the original CARE risk scores to rank its six native views. Full six-view fusion gives 9,343 errors, 549 more than CARE: 1.739 WER points [1.030, 2.476], Holm $P = 0.000671$. When calibration selects the number of ranked views, it chooses one—minimum-risk native transcript selection, not fusion—which gives 8,817 errors. CARE's 23-error advantage has interval $[-0.057, 0.221]$ points and $P = 0.356079$ in this comparison's two-test family. The calibrated single-view alternative remains close (Appendix G.2).

5 Related work

Enhancement for downstream recognition. Metric-optimized, flow-matching and Zipformer enhancement, PASE and RE-USE provide diverse views (Fu et al., 2021; Lee et al., 2025; Wang & Tian, 2025; Rong et al., 2026; Fu et al., 2026). Pretrained recognizers such as Whisper and Qwen3-ASR (Radford et al., 2023; Shi et al., 2026) permit selection without updating recognition weights.

Transcript combination and risk. ROVER combines hypotheses by word voting (Fiscus, 1997); quality estimation can rank its inputs before fusion (Jalalvand et al., 2015). Confusion networks connect consensus to recognition loss (Mangu et al., 2000); minimum Bayes-risk decoding targets expected task loss (Goel & Byrne, 2000; Jinnai, 2026). HypR studies hypothesis revision (Wang et al., 2024), and N-best T5 uses constrained decoding (Ma et al., 2023). CARE-ASR scores complete enhancement-derived candidates using expert, transcript and acoustic information, then calibrates replacement against a default policy.

Candidate learning and defaults. The components draw on pairwise ranking (Burges et al., 2005), set encoders (Zaheer et al., 2017; Lee et al., 2019) and stacked generalization (Wolpert, 1992). CARE-ASR combines lower-level risks (Ting & Witten, 1999) with candidate evidence. Its default calibration relates to selective prediction (Geifman & El-Yaniv, 2017; 2019) and baseline policy improvement (Laroche et al., 2019) through empirical risk and length margins.

6 Discussion and conclusion

The original AMI policy transfers without recalibration, combining more recovery, less erasure and fewer insertions than voting. Candidate scoring supplies this gain; Earnings also benefits from complete-default selection. Expert information remains useful after AMI cross-fitting and under two learning objectives. Eight risks alone lose accuracy, but full-input MSE and calibrated native selection remain close to CARE. These comparisons support expert-informed candidate scoring across the tested objectives and decision rules.

The study uses one frozen recognizer on English speech. Candidate generation and acoustic scoring remain necessary; the small scorer does not eliminate these costs. The method suits settings where hypotheses are available or recognition quality justifies their generation.

AI USE STATEMENT

AI tools assisted with literature searches, data verification, and manuscript polishing. The authors were responsible for implementation, data analysis, referencing, and the writing of the final manuscript.

REPRODUCIBILITY STATEMENT

The supplementary package provides selected checkpoints and scalers, forward and loss implementations, source-level scores, policy inputs and replay scripts. It includes the original AMI predictions, matched-model weights, expert-information controls and calibration grids. The cross-fitting and shared-risk fusion modules supply their selected downstream models, predictions, group counts and executable checks; the complete cross-fitting training archive is preserved separately. Appendices C, D, F and G specify construction, training, evaluation and replay scope. Included data regenerate the tables and figures. Recordings, upstream enhancement weights and waveform inference resources retain their original distribution channels.

ETHICS STATEMENT

The study uses public speech corpora. Speech data and pretrained models remain subject to their original licences and access terms. The supplementary release provides aggregate and source-level recognition counts; original recordings remain with their respective data providers.

REFERENCES

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, 2005. URL https://www.microsoft.com/en-us/research/publication/learning-to-rank-using-gradient-descent/.

Jean Carletta. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5, 2006. URL https://groups.inf.ed.ac.uk/ami/corpus/overview.shtml.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. arXiv:2205.12446, 2022. URL https://arxiv.org/abs/2205.12446.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. Earnings-22: A practical benchmark for accents in the wild. arXiv:2203.15591, 2022. URL https://arxiv.org/abs/2203.15591.

Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. MetricGAN+: An improved version of MetricGAN for speech enhancement. In *Interspeech*, 2021. URL https://arxiv.org/abs/2104.03538.

Szu-Wei Fu, Rong Chao, Xuesong Yang, Sung-Feng Huang, Ryandhimas E. Zezario, Rauf Nasretdinov, Ante Jukić, Yu Tsao, and Yu-Chiang Frank Wang. Rethinking training targets, architectures and data quality for universal speech enhancement. arXiv:2603.02641, 2026. URL https://arxiv.org/abs/2603.02641.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1705.08500.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Vaibhava Goel and William J. Byrne. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135, 2000. doi: 10.1006/csla.2000.0138. URL https://www.sciencedirect.com/science/article/pii/S0885230800901384.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. arXiv:1805.04699, 2018. URL https://arxiv.org/abs/1805.04699.

Shahab Jalalvand, Matteo Negri, Daniele Falavigna, and Marco Turchi. Driving ROVER with segment-based ASR quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1095–1105, 2015. doi: 10.3115/v1/P15-1106. URL https://aclanthology.org/P15-1106/.

Yuu Jinnai. Re-evaluating minimum bayes risk decoding for automatic speech recognition. arXiv:2510.19471, version 2, 2026. URL https://arxiv.org/abs/2510.19471v2.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Seonggyu Lee, Sein Cheong, Sangwook Han, and Jong Won Shin. FlowSE: Flow matching-based speech enhancement. In *ICASSP*, 2025. URL https://arxiv.org/abs/2508.06840.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2021. URL https://arxiv.org/abs/2108.10257.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Rao Ma, Mark J. F. Gales, Kate M. Knill, and Mengjie Qian. N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. In *Interspeech 2023*, pp. 3267–3271, 2023. doi: 10.21437/Interspeech.2023-1616. URL https://www.isca-archive.org/interspeech_2023/ma23e_interspeech.html.

Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, 2000. doi: 10.1006/csla.2000.0152. URL https://arxiv.org/abs/cs/0010012.

Vaishakh Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one blind image restoration. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.13090.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Xiaobin Rong, Qinwen Hu, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. PASE: Leveraging the phonological prior of WavLM for low-hallucination generative speech enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32826–32834, 2026. doi: 10.1609/aaai.v40i39.40562. URL https://ojs.aaai.org/index.php/AAAI/article/view/40562.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 1214–1223, 2021. URL https://arxiv.org/abs/2011.11156.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report. arXiv:2601.21337, 2026. URL https://arxiv.org/abs/2601.21337.

Kai Ming Ting and Ian H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999. doi: 10.1613/jair.594. URL https://doi.org/10.1613/jair.594.

Haoxu Wang and Biao Tian. ZipEnhancer: Dual-path down-up sampling-based Zipformer for monaural speech enhancement. In *ICASSP*, 2025. URL https://arxiv.org/abs/2501.05183.

Yi-Wei Wang, Ke-Han Lu, and Kuan-Yu Chen. HypR: A comprehensive study for ASR hypothesis revising with a reference corpus. In *Interspeech*, 2024. URL https://www.isca-archive.org/interspeech_2024/wang24j_interspeech.html.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. arXiv:1804.03209, 2018. URL https://arxiv.org/abs/1804.03209.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1. URL https://www.sciencedirect.com/science/article/pii/S0893608005800231.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.06114.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.09881.

A COMPLETE RECOGNITION COMPARISON

Table 6 lists exact word-error counts for all comparator families. Each corpus uses a single reference denominator. The earlier reranker and pooled audio-supported ranker are internal precursor configurations of CARE-ASR; the final system uses the corpus-calibrated configurations in Appendix D.

Table 6: **All recognition comparisons.** Exact word errors.

System	FLEURS	TED-LIUM 3	AMI	Earnings-22
Raw	65	52	9269	2728
FlowSE	72	92	15098	4164
MetricGAN+	71	101	12115	5136
ZipEnhancer	65	81	9567	12692
PASE	110	50	10300	3113
RE-USE	63	50	9219	2706
ROVER (raw first)	61	50	8614	2643
ROVER (reverse)	61	49	8557	2649
Global: decision loss	65	48	8202	2707
Output: decision loss	61	49	8457	2689
Aligned: decision loss	63	51	8113	2822
Circular-refit: decision	63	50	8183	2671
Global: MSE	62	60	8215	2756
Output: MSE	61	46	8459	2668
Aligned: MSE	62	52	8319	2816
Circular-refit: MSE	63	49	8267	2812
Earlier reranker	61	52	8127	2714
Pooled audio-supported ranker	63	50	8069	2779
CARE-ASR	61	50	8066	2601
Reference words	1694	2320	22314	22796

B ERROR CHANGES AND POLICY COMPOSITION

Table 7 reports the terms used in Equation 2. Recovery and erasure refer to reference-word positions under deterministic edit alignment. Insertions contribute separately to WER.

Table 7: **CARE-ASR error accounting.** Substitutions (S), deletions (D), insertions (I), recovery and erasure.

Corpus	S	D	I	Recovery	Erasure
FLEURS	41	17	3	3	0
TED-LIUM 3	33	12	5	4	3
AMI	3421	3763	882	1408	538
Earnings-22	1149	1075	377	287	102

The nested comparison in Table 4 holds fitted candidate scores fixed. Removing the residual returns the complete default; removing complete-default calibration returns the incumbent. The three policies have identical outputs on FLEURS, TED and AMI. On Earnings, their totals are 2,676, 2,609 and 2,601 errors.

Relative to the Earnings default, the residual adds 16 substitutions, removes 30 deletions and adds six insertions. It recovers eight additional reference positions and erases six fewer, giving $8+6-6=8$ fewer word errors. Of 448 chunks, 51 change under this rule. Call-level totals are given in Table 8.

B.1 TEXT-IDENTITY SENSITIVITY

Deduplication and default lookup use different notions of text identity. Candidates are deduplicated by scoring text, but the original default lookup requires the literal donor string. In 231 of 448

Table 8: **Earnings-22 composition by complete call.**

Call ID	Words	Incumbent	Default	Residual
4453225	11304	1479	1462	1464
4469088	11492	1197	1147	1137

Earnings chunks, a valid scoring-text equivalent exists without a valid literal match, so the original policy retains the donor output.

This distinction matters because risk belongs to a candidate record, not to its word sequence alone: slot and expert fields also enter the score. Using the first scoring-text equivalent can therefore supply a different risk for the same words. We evaluate that alternative lookup while keeping the original policy as the main result.

The cached comparison first reproduces the original actions and scores, then changes only the lookup to the first valid scoring-text equivalent. Scores, validity, lengths, default outputs and the .02 residual margin remain fixed. Selected original strings are concatenated before complete-call normalization.

Using scoring-text equivalence changes 42 additional chunk outputs and gives 2,605 errors, compared with 2,601 for the original lookup and 2,609 for the complete default. The two calls incur $+1$ and $+3$ errors. Relative to the original residual, the alternative adds 25 recoveries, 14 erasures and 15 insertions. Both lookup results are supplied in the supplement; all main comparisons use the original executed policy.

B.2 Meeting- and call-level comparisons

Table 9 aggregates the main predictions by AMI session and complete Earnings call, without refitting or further selection. CARE-ASR has fewer errors than raw recognition, reverse ROVER and the aligned decision-loss control in all six rows. Relative to aligned selection, the four AMI sessions together remove 47 errors; the Earnings calls remove 77 and 144, totalling 221. Each recorded session or call contributes to the gain.

Table 9: **Recognition changes by recorded source group.** Exact word-error counts on the same sources. ROVER is the reverse order; Aligned is the decision-loss control in Table 3, including its recorded guard. Words gives the reference denominator. AMI sessions share the same four participants; these are source-wise descriptions of the reported evaluation, not additional independent cohorts.

Session/call	Words	Raw	ROVER	Aligned	CARE-ASR
AMI					
TS3004a	3,006	1,214	1,145	1,089	1,074
TS3004b	6,269	2,397	2,182	2,063	2,059
TS3004c	6,836	2,814	2,515	2,425	2,401
TS3004d	6,203	2,844	2,715	2,536	2,532
Total	22,314	9,269	8,557	8,113	8,066
Earnings-22					
4453225	11,304	1,510	1,490	1,541	1,464
4469088	11,492	1,218	1,159	1,281	1,137
Total	22,796	2,728	2,649	2,822	2,601

C Candidate features and upstream experts

C.1 Transcript and expert representation

Let a denote raw-first ROVER, h the candidate, $v_1, \ldots, v_K$ the native hypotheses, and $d = \max(1, |h_{\text{raw}}|)$. The first 80 fields are ordered as follows:

1. **Slot identity (16):** a one-hot code for the ordered candidate slot.

2. **Expert risks (32):** eight candidate risks, eight risks at the nearest native hypothesis, eight minimum native-view risks, and eight indicators that an expert selected this candidate. A candidate outside the native bank receives an interpolated risk using weights proportional to $1/(\mathrm{ED}(h, v_k)/d + .05)$. Exact matches and nearest-hypothesis ties use the first matching slot.
3. **Native-bank geometry (14):** normalized edit distances $\mathrm{ED}(h, v_k)/d$ and normalized native lengths. Each vector is sorted over the available native views, then right-padded with zeros to seven entries.
4. **Scalar summaries (18):** $\log(1 + d)$; $|h|/d$; $\log(1 + |h|)$; an empty-hypothesis indicator; native-bank membership; equality with raw and the two ROVER outputs; native-view vote fraction; minimum, mean, maximum and standard deviation of native edit distance; unique-token and numeric-token fractions; normalized distances to the two ROVER outputs and the aligned decision expert's output.

The expert-risk fields are fitted predictions. All distances and identity fields are computed from model hypotheses.

Whole-audio scores (6). Teacher-forced scoring under the original recording supplies the mean, sum and minimum body-token log probability, end-of-sequence (EOS) log probability, body-token count and empty-text indicator. The tokenizer's EOS token is appended to each candidate; prompt tokens are excluded from the scored labels. Empty body-token sequences have zero mean, sum and minimum, while EOS is still scored.

Edit geometry (8). Align the candidate and a by token edit distance, excluding EOS. Record both token counts; substitution, deletion and insertion counts; changed-token fractions for each sequence; and relative edit distance. Alignment ties prefer diagonal, then deletion, then insertion.

Edit support (16). The silence control is an all-zero waveform with the same length and dtype as the original 16-kHz mono recording. Both conditions use the same audio-normalization and processor path and English prompt. The ordered fields are candidate changed-token mean, sum and minimum log probability; anchor changed-token mean, sum and minimum; candidate and anchor neighbourhood means; common-token score-difference mean and minimum; changed-token mean and sum differences; candidate and anchor changed-token original-audio-minus-silence means; ordinary EOS difference; and silence-contrasted EOS difference. Neighbourhoods extend changed-token sets by one token. Empty sets give zero summaries. Whole-audio configurations mask fields 86–109, using zero-based indexing, before normalization.

C.2 Upstream encoder and feature construction

Upstream experts receive 2,160 fields ordered as $[z, g, G, L]$: 27 transcript-relation fields, eight native-coordinate fields, 2,080 global encoder fields and 45 paired fields. The first three blocks form the 2,115 base fields. The `upstream_features/` source mapping verifies equivalence to the feature code in the frozen inference lock.

FLEURS and TED use seven native views, including BSRNN-Flow; LibriSpeech, AMI and Earnings use raw plus the five shared enhancers. Experts can select BSRNN-Flow despite its absence from the six direct-view slots. Native-risk, agreement and distance fields use each corpus's full bank.

For FLEURS and TED, raw-first ROVER uses raw, BSRNN-Flow, FlowSE, MetricGAN+, PASE, RE-USE and ZipEnhancer, in that order; reverse ROVER reverses the entire list. The BSRNN-Flow path uses `flow_bsrnn.ckpt` from model repository `lichenda/icassp_2026_urgent_baseline`, revision `d4add2435a74b3f2dd54a9bbd417a058c68983b1`, with EMA weights and 15 Euler steps. The recorded checkpoint SHA-256 and source commit are in `candidate_bank_provenance.json`. Exclusion of ZipEnhancer leaves six native ROVER inputs for these two corpora, distinct from the selector's five remaining direct-view slots.

Native encoder interface. Inputs are saved mono 16-kHz waveforms. Frozen Qwen3-ASR-0.6B uses empty context, automatic language inference, bfloat16, SDPA attention and disabled TF32. We use the audio tower's `last_hidden_state`, returned through `get_audio_features`, with

1,024 channels; no intermediate layer is selected. For unpadded mel length M from the processor's attention-mask sum, the state count is $13\lfloor M/100 \rfloor + \lceil (M \bmod 100)/8 \rceil$; padded states are excluded. Project-held-out AMI uses four-source blocks: raw, MetricGAN+, PASE and ZipEnhancer are grouped by source; FlowSE and RE-USE by method. The original FLEURS and TED paths additionally group BSRNN-Flow by method. Earnings groups the four baseline views and the two extra views separately within each chunk.

Transcript relations (27). Let w_j be the whitespace-tokenized, normalized transcript of native view j, with raw at $j = 0$, $n_j = |w_j|$, $d = \max(1, n_0)$ and K available views. Write $D_{j\ell}$ for unit-cost word-edit distance, u_j for the number of distinct tokens, and v_j for the number of views whose complete token sequence equals w_j. Let $q_a(w)$ be the fraction of repeated contiguous a-grams: the number of a-grams minus the number of distinct a-grams, divided by the maximum of one and the number of a-grams. The ordered fields for candidate j are:

1. Fields 0–4: $\log(1 + n_0)$, $\log(1 + n_j)$, $\log(1 + K)$, $\mathbf{1}[n_0 = 0]$, $\mathbf{1}[n_j = 0]$.
2. Fields 5–8: n_j/d, $(n_j - n_0)/d$, $u_j/\max(1, n_j)$, u_0/d.
3. Fields 9–12: $q_2(w_j)$, $q_2(w_0)$, $q_3(w_j)$, $q_3(w_0)$.
4. Fields 13–19: D_{j0}/d, $D_{j0}/\max(1, n_j, n_0)$, then the mean, median, minimum over $\ell \neq j$, maximum and population standard deviation of $D_{j\ell}/d$. Only the minimum excludes the candidate itself.
5. Fields 20–24: v_j/K, v_0/K, the fraction of distinct complete transcripts, and the mean and population standard deviation of $\{n_\ell/d : \ell \neq j\}$.
6. Fields 25–26: $(\mathrm{mean}_\ell\, D_{0\ell} - \mathrm{mean}_\ell\, D_{j\ell})/d$ and $\mathrm{mean}_\ell[D_{j\ell}/\max(1, n_\ell)]$.

Native views are ordered raw first, followed by method name. Empty token sets and a-gram sets use the denominators specified above. These features use candidate strings only, not reference transcripts.

Native coordinates (8). For mel blocks starting at $a = 0, 100, 200, \ldots < M$, state j has nominal sample coordinate $160(a + 8j)$, for $0 \leq j < \lceil \min(100, M - a)/8 \rceil$. These are contextual encoder coordinates, not word timestamps or isolated receptive fields. Raw and candidate states are paired by exact coordinate intersection, retaining coordinates below both waveform sample counts. There is no interpolation. Let p be the number of pairs, T_0, T_j the state counts, and s_0, s_j the waveform sample counts. The eight fields are $\log(1 + s_0)$, $\log(1 + s_j)$, $(s_j - s_0)/s_0$, p/T_0, p/T_j, the raw and candidate unwritten-tail fractions, and $\mathbf{1}[\lfloor p/2 \rfloor > 0]$.

The unwritten tail denotes positions in the zero-initialized ZipEnhancer output buffer left unwritten by segmented inference, rather than a measurement of trailing acoustic silence. Its fraction divides the predicted span by the corresponding saved waveform's sample count; it does not remove states. The recorded ZipEnhancer tail allocation for $s_0 > 96000$ is $\max(0, s_0 - (P - 4000))$, where $P = 32000 + 24000\lceil (s_0 - 32000)/24000 \rceil$; other views have zero allocation. A CPU audit of 13 fitting records found all four predicted tails entirely zero while their raw counterparts contained nonzero samples. Cached waveforms remained unchanged. Appendix F.8 reports a matched pipeline sensitivity analysis that excludes ZipEnhancer and refits the upstream experts and downstream selectors.

Global summaries (2,080). Set $\epsilon = 10^{-8}$. For an encoder sequence $H = (h_t)$, let $n_t = \|h_t\|_2$, $u_t = h_t/\max(n_t, \epsilon)$, and define $S(v)$ as mean, population standard deviation and quantiles at $.10, .25, .50, .75, .90$, in that order. Quantiles use linear interpolation. The 1,039-field marginal vector is

$$M(H) = [\mathrm{mean}_t\, h_t,\ S(\log \max(n_t, \epsilon)),\ S(u_t^\top u_{t-1}),\ \mathrm{mean}_t\, \mathbf{1}[n_t < \epsilon]],$$

with a circular predecessor at the first state. For the complete raw and candidate sequences, G concatenates their two marginal vectors, relative mean-vector displacement $\|\bar{h}_0 - \bar{h}_j\|_2/\max(\|\bar{h}_0\|_2 + \|\bar{h}_j\|_2, \epsilon)$ and mean-vector cosine $\bar{h}_0^\top \bar{h}_j/\max(\|\bar{h}_0\|_2 \|\bar{h}_j\|_2, \epsilon)$.

Aligned and circular summaries (45). On the paired states (r_t, c_t), form three series: cosine similarity clipped to $[-1, 1]$, relative displacement $\|r_t - c_t\|_2 / \max(\|r_t\|_2 + \|c_t\|_2, \epsilon)$, and log norm ratio $\log[\max(\|c_t\|_2, \epsilon) / \max(\|r_t\|_2, \epsilon)]$. Each contributes $S(v)$ followed by eight window means. Window $a \in \{0, \dots, 7\}$ starts at $\lfloor ap/8 \rfloor$ and ends at $\min(p, \max(\lfloor ap/8 \rfloor + 1, \lfloor (a+1)p/8 \rfloor))$. Concatenating the three 15-field summaries gives L. With no paired positions, L is zero. Circular-refit rolls the candidate's paired sequence by $\lfloor p/2 \rfloor$ before these calculations and retrains the expert; z, g, G are unchanged. For $p = 1$, the circular and aligned summaries coincide.

C.3 Expert models and guarded outputs

Eight risk experts cross four feature families with two losses. Each uses a $2160 \to 128 \to 64 \to 1$ ReLU network with 2,115 base fields and 45 local fields. Output geometry retains the first 35 base fields; global retains the base fields; paired aligned and circular-refit variants add their respective local fields. Unused fields are zeroed.

Scaling uses fitting sources and valid non-raw slots only. For each source, average the feature first and second moments over those slots, then weight sources proportionally to raw-hypothesis length within each corpus, with equal total corpus mass. Standard deviations below 10^{-8} are replaced by one and the corresponding standardized fields are set to zero. The upstream scaler does not clip features. The network's raw-slot risk is fixed to zero; non-raw targets are $(e_j - e_{\text{raw}}) / \max(1, n_{\text{raw}})$. Each objective averages the same three seeds. Padded native-view slots are invalidated before selection. The final 110-field selector uses its separate normalization described in Appendix D.

Regression experts minimize non-anchor normalized-risk squared error. Decision-loss experts minimize expected regret under $\text{softmax}(-r/.05)$ plus .1 times that regression loss. Each family averages three seeds, giving 24 fitted experts. Training uses AdamW (Loshchilov & Hutter, 2019), learning rate .001, weight decay .0001, batch size 128, at most 32 epochs, patience five and gradient clipping at norm five. Training weights balance corpora; calibration data select checkpoints.

A guarded expert policy applies its calibrated risk margin and then checks contraction against raw. The admissibility library includes the identity option and 20 combinations of minimum raw length $\{1, 4, 8, 16, 32\}$ and minimum candidate/raw length ratio $\{.25, .5, .75, 1\}$. A guard can revert to raw on each source, so its complete output policy differs from the fixed expert-output slot.

D Training, calibration and scoring

D.1 Data roles and grouping

Parameters are fitted on `fit`; checkpoints and policy settings are selected on `early_stop`; final evaluation uses `selection`. These original role names also identify the supplementary records. Fitting/calibration counts are 239/75 for FLEURS, 230/182 for TED and 7,095/2,231 for AMI. Pooled training additionally uses 1,557 LibriSpeech fitting sources and 599 calibration sources. The paper evaluates FLEURS, TED, AMI and Earnings. The original training bank uses in-sample expert predictions; Appendix G.1 tests fold-generated predictions for AMI fitting sources with the remaining roles fixed.

AMI calibration uses ES2011a–d; evaluation uses TS3004a–d. Earnings calibrates complete-call policies on calls 4479944 and 4482383, then evaluates calls 4453225 and 4469088. The archived six-call Earnings fitting allocation is not used to fit this pooled scorer, and no chunk-level reference targets are used for Earnings calibration. Original row indices and the role summary are included in the supplement.

Final policy settings use calibration outcomes only. The evaluation sources had also been assessed in earlier experiments. Evaluation groups comprise 30 FLEURS speaker identifiers, two TED groups, four AMI sessions with the same four participants, and two Earnings calls. The available records do not establish whether these sources were included in the recognizer's pretraining data.

D.2 Optimization

Initial fitting compares independent and set-context encodings, Whole and Edit inputs, pairwise and soft-target losses, and learning rates 3×10^{-4} and 10^{-3}. Seeds are $27101, 27102, 27103$. After variant masking, valid-slot means and standard deviations are estimated on fitting rows. Scales below 10^{-6} are replaced by one; standardized features are clipped to $[-12, 12]$ and invalid slots are zeroed.

For each training or calibration role, the weight of row i is $w_i = d_i n/(CN_c)$, where n is the pooled row count, $C = 4$ the training-corpus count and N_c the corpus reference-word total in that role. The target denominator remains raw-hypothesis length. Training uses AdamW, weight decay .0001, batch size 256, at most 40 epochs, patience seven and gradient clipping at norm five. The final scoring layer starts at zero.

Checkpoint selection first minimizes macro-corpus WER at margins zero and .01, then weighted loss, both on calibration data. Epoch zero remains eligible. The calibration margin is selected from $\{0, .002, .005, .01, .02, .05, .1, \text{no change}\}$. Scores are averaged over all three seeds.

D.3 Adaptation and selected configurations

Short-form corpus adaptation uses learning rates 10^{-5} and 10^{-4}, AdamW weight decay 10^{-4}, batch size 256, at most 24 epochs, patience five and gradient norm clipping at five. Original and adapted ensembles are compared at the same eight margins; ties favour original weights, then fewer switches. Earnings retains its whole-call-calibrated ensemble, without chunk-level reference targets.

FLEURS selects independent edit support, soft-target loss, initial learning rate .001, original weights and margin .05. TED selects independent edit support with pairwise loss and initial learning rate .0003; its no-change margin returns raw-first ROVER. AMI selects independent edit support, pairwise loss, initial learning rate .001, adaptation rate 10^{-5} and margin .01. Earnings uses set context, whole-audio features, soft-target loss, initial learning rate .001, original weights and incumbent margin .05.

The complete-policy library has 25 defaults. Residual calibration tests five length settings: $(q, k) = (0, 1)$, $(.75, 1)$, $(.75, 8)$, $(1, 1)$ and $(1, 8)$. Each is crossed with seven margins, $m \in \{0, .002, .005, .01, .02, .05, .1\}$; no-change is also available. FLEURS, TED and AMI retain the incumbent without residual changes. Earnings selects the output-geometry MSE guard, then $q = 1, k = 1, m = .02$.

D.4 Text identity and word scoring

Candidate order is the two ROVER variants; raw, FlowSE, MetricGAN+, PASE, RE-USE and ZipEnhancer; then eight expert outputs. The first two slots retain their recorded ROVER text treatment, and later duplicate scoring texts are invalidated. Default lookup uses the original donor string. Earnings length tests use the normalized word vocabulary employed in complete-call scoring. The effect of alternate lookup identity is reported in Appendix B.1.

The word aligner uses exact dynamic programming with diagonal, deletion, then insertion tie preference. Earnings hypotheses are joined in the archived chunk order and normalized once per complete call. Summaries use the sum of S/D/I divided by the sum of reference words. The source records retain per-method S/D/I, recovery, erasure and character-error counts.

D.5 Three-stage execution

For a recording, first compute the ordered candidate strings, valid mask and seed-averaged risks r_j, centred on slot zero. The stored settings contain an incumbent margin τ_c, a default-policy identifier, and either no residual or a residual tuple (q, k, m). Execution is:

1. If no valid nonzero candidate remains, set the incumbent to h_0. Otherwise, let $j_0 = \arg\min_{j>0,\, j\in V} r_j$, with the first slot breaking ties: the incumbent is h_{j_0} if $r_{j_0} < -\tau_c$, and h_0 otherwise. The no-change setting always selects h_0. This stage has no residual length constraint.

2. Apply the selected complete policy $b_c(x)$: a fixed-slot output, a guarded expert output, or the incumbent from step 1. Its identifier was selected on the calibration role and is fixed for evaluation.

3. If no residual was selected, return $b_c(x)$. Otherwise locate its original donor string in the valid candidate bank. If absent, return that donor string. If present at index b, restrict candidates by the stored length rule and replace h_b only when the lowest admissible risk is strictly below $r_b - m$.

AMI uses $\tau_c = .01$, incumbent as complete default, and no residual, so step 1 produces its reported result. Earnings uses incumbent margin $\tau_c = .05$, the guarded output-geometry MSE policy as default, and $(q, k, m) = (1, 1, .02)$ for the residual. Although risks are learned relative to slot zero, target differences against a valid default index satisfy $t_j - t_b = (e_j - e_b)/d$. The residual can therefore compare against that default on the same target scale; its replacement margin remains empirically calibrated.

E TRAINED FEATURE AND OBJECTIVE COMPARISONS

Tables 10 and 11 compare all 16 training configurations on calibration data before corpus adaptation: 75 FLEURS recordings, 182 TED segments and 2,231 AMI segments. Each averages three seeds with the same 110-input architecture and uses its own incumbent margin τ_c, selected by the common rule. Whole masks the edit fields, Edit retains them, and Independent masks the explicit set-context blocks. The main performance tables instead report the evaluation role.

Table 10: **Trained input and objective comparison at learning rate 10^{-3}.** Calibration WER (%), with three seeds per configuration and a shared 110-field architecture. Each row uses its own incumbent margin τ_c selected by the same rule.

Context	Input	Loss	τ_c	FLEURS	TED	AMI
Independent	Whole	Pairwise	0.02	3.641	1.700	26.227
Independent	Whole	Soft target	0.02	3.641	1.700	26.293
Independent	+ Edit	Pairwise	0.01	3.771	1.700	26.058
Independent	+ Edit	Soft target	0.02	3.706	1.700	26.317
Set context	Whole	Pairwise	0.02	3.641	1.700	26.311
Set context	Whole	Soft target	0.02	3.641	1.700	26.444
Set context	+ Edit	Pairwise	0.01	3.771	1.682	26.100
Set context	+ Edit	Soft target	0.05	3.641	1.700	26.438

On AMI calibration data, adding the edit block under pairwise training reduces errors by 28, 35, 48 and 31 across the four combinations of context and learning rate. The effects of soft targets and set context vary with the configuration. Appendix F complements this calibration study with separately fitted Whole, Geometry and Edit models evaluated on new sources.

Table 11: **Training configurations at learning rate 3×10^{-4}.** Same definitions as Table 10.

Context	Input	Loss	τ_c	FLEURS	TED	AMI
Independent	Whole	Pairwise	0.02	3.641	1.647	26.450
Independent	Whole	Soft target	0.02	3.706	1.700	26.287
Independent	+ Edit	Pairwise	0.002	3.771	1.700	26.160
Independent	+ Edit	Soft target	0.02	3.706	1.700	26.377
Set context	Whole	Pairwise	0.02	3.641	1.700	26.365
Set context	Whole	Soft target	0.02	3.641	1.700	26.383
Set context	+ Edit	Pairwise	0.0	3.771	1.700	26.178
Set context	+ Edit	Soft target	0.02	3.641	1.700	26.287

F FROZEN EVALUATION ON PROJECT-HELD-OUT AMI SOURCES

F.1 COHORT AND PROSPECTIVE SPECIFICATION

The additional evaluation comprises 4,096 single-distant-microphone recordings: 256 per meeting–speaker connected component, with 16 components, 57 session identifiers and 65 speaker identifiers. The admission process audits source IDs, meeting and speaker membership, and decoded PCM content against the available local and restored remote project history. All 4,096 admitted PCM hashes are distinct, with zero matches to that historical union. Seventy-two invalid or duplicate candidate records are excluded by the admission rules before recognition. The audio totals 10,911.51 seconds at 16 kHz.

These sources are held out from the project's fitting, calibration and method selection. Their publisher labels are retained: 3,584 recordings belong to the released training split and 512 to the test split. Project-held-out therefore describes their role in this experiment, not an evaluation on the full official test split. Membership in the pretrained recognizer's training data is not established. The source manifest includes IDs, connected components, shard locations and audio-content hashes.

The cohort size, primary comparisons, thresholds and analysis were specified internally and hash-bound before new-cohort inference; they were not publicly preregistered. The recorded UTC+8 timeline places the execution lock at 09:50:42 on 16 September 2026, before inference. The complete prediction seal follows at 11:06:04, and reference-based scoring first accesses text at 11:06:05.

All admitted sources enter the analysis. An empty normalized reference contributes insertion errors and zero reference words. For this prespecified evaluation, no cohort outcome is used to choose thresholds, seeds, models or subgroups.

F.2 FROZEN SYSTEM AND EQUAL-OPPORTUNITY CONTROLS

The original AMI CARE-ASR policy uses its three existing adapted checkpoints, the original feature scalers, margin .01, and the incumbent default with no residual changes. Its six input views, upstream experts, candidate order, recognizer, normalization and inference grouping are unchanged. No new-source fitting is performed.

Three matched controls use Whole (86 active fields), Whole plus edit geometry (94), and full Edit support (110). All use the same 110-input architecture with 21,601 parameters, independent candidate encoding, pairwise objective and three seeds, 27101, 27102 and 27103. Unused fields are zero-masked before estimating fitting-only scaling. Every variant receives the same fitting data, initial random states, epoch budgets, stopping procedure, AMI adaptation rate and calibration grid. There are 18 fitting stages: three variants, three seeds, and pooled fitting followed by AMI adaptation. The fitting and calibration pools contain 9,121 and 3,087 sources, respectively; calibration includes the original 2,231 AMI segments.

All three retain the first 80 transcript and expert fields, including risk predictions from aligned and circular-refit experts. These contrasts measure the addition of explicit edit-geometry and edit-support blocks to that shared representation. Whole already has access to expert information, including predictions based on local paired features.

The common rule minimizes calibration word errors, breaking ties by fewer switches and then the fixed grid order. It selects margins .02, .01 and .01 for Whole, Geometry and Edit. Scores are averaged across all three seeds within each variant, with no seed selection on the new cohort. Full matched Edit and original CARE-ASR return identical transcripts on all 4,096 recordings. Their weights and training histories are recorded separately, but these identical predictions constitute one recognition result.

F.3 PRIMARY ANALYSIS AND COMPLETE RESULTS

Micro-WER uses all 31,565 reference words. For each comparison, let d_g be comparator errors minus method errors in connected component g, and n_g its reference-word count. The reported reduction is $100 \sum_g d_g / \sum_g n_g$. The exact two-sided test enumerates all $2^{16} = 65,536$ sign assignments and compares the absolute signed sum to $|\sum_g d_g|$. It assumes sign exchangeability of

paired component differences under the null. Holm correction covers the fixed three-comparison family. The 50,000 paired component-bootstrap replicates use seed 260916, resampling groups with replacement and recomputing the ratio of total differences to total words. Percentile intervals are not multiplicity-adjusted.

Table 12: **Prespecified comparisons on new AMI sources.** Reductions and intervals are in WER percentage points. Groups $+/-$ count improvements and deteriorations; Edit versus Whole also has one tie.

Prespecified comparison	Δ WER	95% interval	Holm P	Groups $+/-$
CARE vs. reverse ROVER	2.180	[1.455, 2.919]	0.000275	15/1
Edit vs. Whole	0.203	[-0.010, 0.438]	0.118896	8/7
Edit vs. Geometry	0.295	[0.029, 0.581]	0.118896	11/5

The comparison with reverse ROVER remains significant after correction. Both feature contrasts favour Edit in aggregate but do not meet the corrected threshold. For Geometry, the percentile bootstrap interval excludes zero while the Holm-adjusted exact test does not reject; these are different uncertainty procedures, and the interval is unadjusted. Table 13 reports every prespecified method family, Table 14 gives all source-group counts, and the supplementary JSON includes individual-seed outputs.

Table 13: **Complete new-source recognition comparison.** Exact counts over 4,096 recordings. Recovery and erasure are measured against raw recognition. Single-view, expert and multi-candidate methods have different input and compute budgets; the three matched feature variants have equal budgets.

Method	Errors	WER (%)	Recovery	Erasure	Insertions
Raw recognition	10,048	31.833	0	0	1,716
FlowSE	19,821	62.794	597	10,497	1,589
MetricGAN+	16,112	51.044	522	7,309	993
PASE	11,741	37.196	1,184	2,775	1,818
RE-USE	9,634	30.521	1,883	1,615	1,570
ZipEnhancer	11,018	34.906	1,163	2,366	1,483
ROVER (raw first)	9,574	30.331	685	680	1,247
ROVER (reverse)	9,482	30.040	1,011	876	1,285
Output: decision loss	9,562	30.293	1,000	838	1,392
Output: MSE	9,573	30.328	989	895	1,335
Global: decision loss	9,040	28.639	1,166	550	1,324
Global: MSE	9,105	28.845	1,091	648	1,216
Aligned: decision loss	8,882	28.139	1,507	751	1,306
Aligned: MSE	9,274	29.381	848	538	1,252
Circular-refit: decision loss	9,039	28.636	1,329	541	1,495
Circular-refit: MSE	9,190	29.115	941	569	1,230
Matched Whole	8,858	28.063	1,298	590	1,234
Matched Geometry	8,887	28.155	1,280	550	1,285
Matched Edit	8,794	27.860	1,379	637	1,204
Original CARE-ASR	8,794	27.860	1,379	637	1,204

F.4 Reproduction from saved predictions

The supplementary release provides per-source hypotheses, references, S/D/I and correctness alignments; cached 110-field inputs and risk arrays; original and matched scorer weights; selected margins; group assignments; and portable analysis code. Verification reproduces all 32 output summaries, 16,384 ensemble policy actions and the three primary tests. A separate scalar dynamic-programming implementation reproduces 18,917 distinct source–transcript alignments. These checks operate on saved outputs and cached features. Repeating upstream waveform inference additionally requires source audio and large pretrained models from their original distribution channels. Execution and result-serialization recovery records are retained with the original hashes.

Table 14: **Counts by meeting–speaker component.** Component names are canonical meeting identifiers, not individual sessions. ROVER uses reverse order; CARE is the original frozen policy and also matches the full Edit outputs. Δ ROVER is ROVER minus CARE errors. All components enter the paired analysis.

Component	Words	ROVER	CARE	Whole	Geometry	Δ ROVER
EN2001a	2,500	677	546	567	582	+131
EN2002a	1,722	724	695	708	701	+29
EN2004a	1,997	674	625	632	632	+49
EN2006a	1,422	632	592	586	593	+40
ES2003a	2,363	673	606	605	617	+67
ES2004a	2,348	646	607	608	605	+39
ES2008a	2,355	501	442	442	446	+59
ES2009a	2,113	579	551	549	545	+28
ES2010a	2,015	449	379	378	390	+70
ES2012a	2,554	647	662	651	660	-15
ES2016a	1,714	466	413	435	432	+53
IS1007a	1,784	532	510	519	522	+22
TS3005a	1,836	662	626	633	613	+36
TS3007a	1,575	590	543	541	550	+47
TS3011a	1,547	420	393	402	403	+27
TS3012a	1,720	610	604	602	596	+6

F.5 Additional same-candidate comparisons

Four additional comparisons examine risk regression and candidate consensus within the same 16-slot bank. Specified after the original cohort results, they retain CARE-ASR and all 4,096 evaluated sources while fitting and selecting comparators only on the original fitting and calibration roles. They form a separate Holm family using the same connected-group tests and bootstrap procedure; the original three-test analysis is unchanged.

Matched risk regression. The MSE model shares full Edit's 110 fields, 21,601-parameter independent-context network, relative-error targets, fitting-only scaling, source weights and seeds. It replaces Equation 4 by M_i, retaining the AdamW settings and pooled/AMI budgets in Appendix D: initial learning rate .001 and adaptation rate 10^{-5}. The original checkpoint-selection rule uses MSE for its secondary loss tie-break. All three seed scores are averaged; the original AMI calibration pool selects margin .01. The pooled and adapted checkpoints, six fitting histories and complete calibration grid are supplied.

Candidate-consensus selection. For each recording, let U_i contain the first valid representative of each unique whitespace-token sequence. Finite-candidate minimum-risk selection uses

$$q_{ij} = \sum_{k \in U_i} w_{ik} \frac{\mathrm{ED}(h_{ik}, h_{ij})}{\max(1, |h_{ik}|)}. \tag{8}$$

Uniform MBR sets $w_{ik} = 1/|U_i|$ and chooses the lowest-risk valid slot, favouring the first-slot ROVER anchor on ties. Calibrated MBR instead permits $w_{ik} \propto \exp(s_{ik}/T)$, where s_{ik} is the stored whole-audio mean token log probability. It searches $T \in \{.1, .3, 1, 3, 10\}$ plus uniform weights and the eight-margin grid in Appendix D. The margin compares $q_{ij} - q_{i0}$ with the anchor. Original AMI calibration errors, then switch count and fixed grid order, select uniform weights and no change, returning raw-first ROVER. These finite-bank consensus controls use proxy weights, not calibrated transcript posteriors, and differ from published sampling-based MBR systems.

Archived nonlinear regression. The tree comparator uses the previously selected three-seed gradient-boosted risk ensemble: 15 leaves, 200 boosting iterations and margin .02. It reads the same full candidate features, with its archived normalization and multi-corpus policy choice unchanged. This supplies another regression reference, but its architecture and historical search budget differ from the matched neural models.

Table 15: **Additional comparisons with a common candidate bank.** Exact counts over 31,565 reference words. Recovery and erasure are relative to raw recognition. MSE and CARE share inputs, architecture and fitting budgets; MBR uses candidate texts and, for its calibrated form, sequence scores. The archived tree has a different fitting design.

Method	Errors	WER (%)	Recovery	Erasure	Insertions
CARE-ASR	8,794	27.860	1,379	637	1,204
Matched MSE	8,855	28.053	1,332	576	1,279
Archived tree	8,885	28.148	1,414	631	1,336
Uniform candidate MBR	12,155	38.508	637	3,690	770
Calibrated candidate MBR	9,574	30.331	685	680	1,247

Table 16: **CARE-ASR versus the four additional comparators.** Positive WER reductions favour CARE. Intervals are unadjusted 95% paired group-bootstrap intervals; Holm correction covers these four comparisons only. Groups $+/0/-$ count improvements, ties and deteriorations.

Comparator	Δ WER	95% interval	Holm P	Groups $+/0/-$
Matched MSE	0.193	[-0.030, 0.447]	0.233643	8/2/6
Archived tree	0.288	[-0.025, 0.623]	0.233643	9/2/5
Uniform candidate MBR	10.648	[8.666, 12.587]	0.000122	16/0/0
Calibrated candidate MBR	2.471	[1.802, 3.144]	0.000183	15/0/1

CARE-ASR has 61 fewer errors than matched MSE and 91 fewer than the archived tree (Tables 15–16). Both paired intervals include zero and neither corrected test rejects. Relative to MSE, CARE adds 47 recoveries and 61 erasures while removing 75 insertions, for a net reduction of 61 errors. The larger separation is from candidate consensus: learned regression approaches CARE-ASR under the shared representation, while the incremental contribution of pairwise training remains uncertain.

All seeds and comparator outputs are included. Records are joined by unique source ID, independently of partition order. Verification recomputes the four tests, candidate decisions and calibration choices from saved inputs; repeating the six MSE fitting stages reproduces all weights and training histories on the recorded setup. The original execution logs and source-ID join record are retained in the archive.

F.6 Explicit expert-input ablation

This analysis asks what the explicit expert inputs add when their chosen transcripts remain available. It was specified after the original cohort and matched-regression results, with protocol, code and input hashes recorded before retraining. The analysis uses the existing cohort and has its own contrast, separate from the original prospective comparisons.

Information intervention. Zero-based fields 16–47 contain eight direct candidate risks, eight nearest-native risks, eight minimum native risks and eight expert-selection indicators. Field 79 is normalized distance to the aligned decision expert's transcript. We mask these 33 fields before fitting-only standardization and at inference. The 110-input, 21,601-parameter network retains 77 active fields, including slot identity, transcript relations, whole-audio and edit features. All candidate transcripts, valid-slot masks and their ordering remain identical. This isolates the explicit input block while preserving the expert-selected candidates and their slot identities.

Matched retraining. The unchanged pairwise training implementation fits seeds 27101, 27102 and 27103, each with pooled training followed by AMI adaptation. It retains the original source weights, optimizer, learning rates, epoch budgets and early-stopping rules. Only the original fitting and calibration rows are read during training and selection. Averaging all three seed scores and applying the original calibration rule selects margin zero; the frozen full ensemble retains margin .01. Each seed also receives its own calibration-selected margin for descriptive reporting (Table 17). The complete calibration grids are included. Policies are recorded before evaluating the 4,096 sources; predictions are sealed before reference-text scoring.

Table 17: **Expert inputs under matched training and calibration.** Each entry is a word-error count. Seed rows use separately calibrated margins; the primary contrast uses the two three-seed ensembles. All seeds are shown, without evaluation-based selection.

	Full inputs		Masked inputs		
Seed	Margin	Errors	Margin	Errors	Difference
27101	0.02	8,839	0	9,167	+328
27102	0.005	8,858	0	9,131	+273
27103	0.01	8,847	0	9,106	+259
Ensemble	0.01	8,794	0	9,079	+285

The full ensemble has 285 fewer errors, or a 0.903-point WER reduction (95% group-bootstrap interval [0.263, 1.592]; exact two-sided group sign-flip $P = 0.0152$). Thirteen of 16 connected groups improve and three worsen (Table 18). The single contrast uses the original group definitions, 50,000 resamples and seed 260916, with an unadjusted P value. The full ensemble has 1,379 recoveries, 637 erasures and 1,204 insertions, versus 1,325, 960 and 1,112 after masking: $54 + 323 - 92 = 285$ fewer errors. This estimates the value of the explicit expert block under the shared training and calibration procedure, without distinguishing individual expert families.

Table 18: **Complete group counts for the expert-input contrast.** Positive differences favour full CARE-ASR. These paired counts reproduce the additional contrast's sign-flip test and bootstrap interval.

Source group	Words	Full CARE	Masked inputs	Difference
meeting:EN2001a	2,500	546	642	+96
meeting:EN2002a	1,722	695	712	+17
meeting:EN2004a	1,997	625	629	+4
meeting:EN2006a	1,422	592	610	+18
meeting:ES2003a	2,363	606	639	+33
meeting:ES2004a	2,348	607	605	-2
meeting:ES2008a	2,355	442	475	+33
meeting:ES2009a	2,113	551	563	+12
meeting:ES2010a	2,015	379	400	+21
meeting:ES2012a	2,554	662	627	-35
meeting:ES2016a	1,714	413	454	+41
meeting:IS1007a	1,784	510	515	+5
meeting:TS3005a	1,836	626	637	+11
meeting:TS3007a	1,575	543	529	-14
meeting:TS3011a	1,547	393	425	+32
meeting:TS3012a	1,720	604	617	+13
Total	31,565	8,794	9,079	+285

The supplement includes all six fitted checkpoints, epoch histories, normalization arrays, policy grids, per-source hypotheses, risks and alignments. Independent replay reproduces all six training stages exactly on the recorded setup, all 4,096 original CARE outputs and the complete additional analysis.

F.7 Linear combination of expert risks

Can a simple combination of the eight expert risks match joint candidate scoring? This additional comparison was specified after the earlier cohort and ablation results. It retains CARE, its candidates and all earlier tests, fits coefficients only on original fitting rows, and selects every policy using the 2,231 original AMI calibration rows.

Linear control. For expert-risk fields $v_{ij} \in \mathbb{R}^8$ (zero-based fields 16–23), set $d_{ij} = v_{ij} - v_{i0}$ and predict $\widetilde{r}_{ij} = d_{ij}^{\mathsf{T}}\beta$. Risks for non-native ROVER hypotheses retain the interpolation defined in Appendix C. There is no intercept or sign constraint. Fit scopes are pooled four-corpus fitting and AMI-only fitting. Each valid nonzero candidate receives weight u_i/k_i, where k_i counts such

candidates and $u_i = \mathrm{den}_i/(CD_{c(i)})$: $D_{c(i)}$ is its corpus's fitting reference-word total, C is four or one, and den_i is the stored relative-error denominator. Normalize these weights to sum to one and scale each field by its fitting-only weighted root-mean-square; values below 10^{-8} use scale one. Do not center or clip.

Minimize the weighted squared relative-error residual plus $\alpha\|\beta_{\mathrm{scaled}}\|_2^2$, for $\alpha \in \{0, 10^{-6}, 10^{-4}, .01, 1\}$. SVD uses relative cutoff 10^{-12} for the unregularized pseudoinverse. The two scopes and five strengths give ten coefficient sets. Each receives the original eight margin choices $0, .002, .005, .01, .02, .05, .1$, or no-change. Select among the 80 policies by calibration errors, switches, scope order, ridge order and margin order. The equal-average reference uses $\beta_e = 1/8$ on unscaled risks and the same calibration margin rule. Within this additional run, coefficients and policies are locked before opening cohort features, and predictions are sealed before reference scoring; the cohort's earlier results were already known.

Recognition results. Calibration selects the pooled linear fit with $\alpha = 1$ and margin .02; equal averaging uses margin .01. The selected raw-scale coefficients, in recorded expert order, are approximately $(.06295, .05664, .07157, .08396, .19163, .11831, .16461, .17861)$; full precision, all ten fits and all 80 calibration entries are supplied. Table 19 reports both policies. CARE has 252 fewer errors than the selected linear policy: 233 additional recoveries and 64 fewer insertions outweigh 45 additional erasures, giving $233 - 45 + 64 = 252$.

Table 19: **Expert-risk combination and candidate-bank capacity.** Counts use the same 4,096 recordings and 31,565 words. Recovery and erasure are relative to raw recognition. Only the learned linear comparison has an inferential test; equal averaging and the reference-assisted oracle are descriptive.

Method	Errors	WER (%)	Recovery	Erasure	Insertions
CARE-ASR	8,794	27.860	1,379	637	1,204
Linear expert combination	9,046	28.658	1,146	592	1,268
Equal expert average	9,044	28.652	1,218	627	1,303
Candidate oracle	6,657	21.090	2,414	225	514

The single declared additional contrast uses the original 16 connected groups: the 95% interval from 50,000 paired-group bootstrap draws is [0.464, 1.141] WER points; the exact two-sided sign-flip P is 0.000427. These use seed 260916 and a separate single-contrast analysis, leaving earlier multiplicity corrections unchanged. Thirteen groups improve and three worsen (Table 20). This comparison does not separate richer inputs from nonlinear scoring; Appendix F.10 subsequently tests the input difference with matched nonlinear training.

Slots and available transcripts. Across the cohort, all 32,768 expert slots reproduce a native scoring-text sequence; their later duplicates are invalidated. Thus the 16-slot interface supplies at most eight distinct hypotheses, not sixteen distinct recognition outputs. The numbers of recordings with one through eight unique valid token sequences are respectively 212, 482, 626, 686, 698, 514, 479 and 399. The source-wise oracle $\sum_i \min_{j \in V_i} ED(y_i, h_{ij})$ gives 6,657 errors (21.090% WER). It uses reference labels and is a hindsight lower bound, not a deployable selector. The package provides the per-source candidate counts, oracle choices, both learned-policy outputs and fixed analysis code.

F.8 Candidate-pipeline sensitivity to ZipEnhancer exclusion

The saved ZipEnhancer outputs have the implementation issue documented in Appendix C.2. We test dependence on that path by removing it from every corpus-specific bank and retaining all other paths, including BSRNN-Flow where originally present. This sensitivity analysis follows the original AMI cohort results. It retains all 9,121 original fitting and 3,087 calibration sources and all 4,096 evaluation sources in their existing 16 groups, with 31,565 reference words; no source is excluded based on its outcome.

Matched rebuilding and refitting. Both ROVER orders and all bank-dependent transcript, expert and edit fields are recomputed. Unchanged acoustic caches are reused only for matching waveform,

Table 20: **Source-group counts for expert-risk combination.** Positive differences favour CARE-ASR. All groups enter the paired analysis; no model or margin is selected using these counts.

Source group	Words	CARE-ASR	Linear	Equal average	Linear minus CARE
meeting:EN2001a	2,500	546	594	593	+48
meeting:EN2002a	1,722	695	719	724	+24
meeting:EN2004a	1,997	625	648	648	+23
meeting:EN2006a	1,422	592	596	600	+4
meeting:ES2003a	2,363	606	621	621	+15
meeting:ES2004a	2,348	607	604	612	-3
meeting:ES2008a	2,355	442	455	453	+13
meeting:ES2009a	2,113	551	546	544	-5
meeting:ES2010a	2,015	379	408	403	+29
meeting:ES2012a	2,554	662	661	664	-1
meeting:ES2016a	1,714	413	430	433	+17
meeting:IS1007a	1,784	510	531	525	+21
meeting:TS3005a	1,836	626	637	630	+11
meeting:TS3007a	1,575	543	567	563	+24
meeting:TS3011a	1,547	393	416	420	+23
meeting:TS3012a	1,720	604	613	611	+9
Total	31,565	8,794	9,046	9,044	+252

text, condition and recognizer settings. The fixed 16-slot interface retains the two ROVER outputs, five direct native paths and eight refitted expert choices; former ZipEnhancer slot 7 is invalid, zero-featured and excluded from aggregation and deduplication. The risk and target anchor remains raw-first ROVER in slot 0; the fixed reverse-ROVER comparator is slot 1. The reduced AMI bank has at most seven distinct scoring sequences.

We refit 24 upstream expert models (four feature families, two objectives and three seeds) and perform 12 downstream fitting stages: pooled fitting and AMI adaptation for both CARE and matched MSE at each of the three original seeds. Both objectives retain the original 110-field scorer, 21,601-parameter capacity, optimization schedule and calibration rule; no additional default or residual-policy search is introduced. The linear control follows Appendix F.7: ten coefficient fits and 80 calibration choices from the eight refitted risk inputs. All choices use original fitting/calibration data, with downstream and linear margins selected on the 2,231 original AMI calibration sources. Evaluation predictions are sealed before opening their references.

The selected ensemble margins are .02 for CARE and .01 for MSE; the linear policy selects pooled coefficients with $\alpha = .01$ and margin .002. Table 24 gives their calibration counts and the complete group-level evaluation data. Full-precision coefficients, selected rows and source hashes are indexed in `generated/zip_sensitivity_details.json`. The original records are `zip_exclusion_sensitivity/predict/DOWNSTREAM_POLICY_LOCK.json` and `zip_exclusion_sensitivity/fit/POLICY_LOCK.json`.

Fixed comparisons and interpretation. The primary contrast is refitted CARE versus recomputed reverse ROVER; the CARE–MSE and CARE–linear contrasts form a separate two-test Holm family. Each uses all 65,536 two-sided group sign assignments and 50,000 paired-group bootstrap resamples with seed 260916. Intervals are unadjusted 95% percentiles. CARE has 712 fewer errors than reverse ROVER and 271 fewer than the linear control; 15 and 16 groups improve, respectively. The MSE difference is 60 errors, with interval [-0.016, 0.399] WER points and Holm $P = 0.106201$ (Table 22). The reduced pipeline preserves the advantage over voting and linear combination, while matched MSE remains close. Because exclusion changes the bank and its fitted models together, this analysis measures pipeline sensitivity, not the tail defect's isolated causal effect or the performance of corrected ZipEnhancer outputs.

Table 21: **Recognition after excluding ZipEnhancer and refitting.** All rows use the same 4,096 AMI recordings. Recovery and erasure are relative to unchanged raw recognition. The candidate oracle is reference-assisted and not deployable.

Method	Errors	WER (%)	Recovery	Erasure	Insertions
Raw	10,048	31.833	0	0	1,716
Raw-first ROVER	9,750	30.889	578	750	1,246
Reverse ROVER	9,614	30.458	1,023	1,003	1,302
Refitted CARE-ASR	8,902	28.202	1,450	776	1,244
Matched MSE	8,962	28.392	1,452	728	1,354
Linear expert combination	9,173	29.061	1,290	829	1,302
Candidate oracle	6,934	21.967	2,179	231	550

Table 22: **Additional pipeline-sensitivity contrasts.** Positive differences are comparator minus CARE in WER points. W/T/L gives improving, tied and worsening groups. The primary voting contrast is unadjusted; Holm correction applies only to the two secondary contrasts.

Comparator	ΔWER (pp)	95% interval	Exact P	Holm P	W/T/L
Reverse ROVER	+2.256	[1.746, 2.754]	0.000061	–	15/0/1
Matched MSE	+0.190	[-0.016, 0.399]	0.106201	0.106201	11/2/3
Linear experts	+0.859	[0.582, 1.166]	0.000031	0.000061	16/0/0

Table 23: **All fixed downstream seeds after ZipEnhancer exclusion.** Each seed uses its configuration's locked ensemble margin. These descriptive results neither select a seed nor introduce further policy searches. Recovery and erasure remain relative to raw recognition.

Method	Errors	WER (%)	Recovery	Erasure	Insertions
CARE-ASR, seed 27101	8,968	28.411	1,483	942	1,177
CARE-ASR, seed 27102	8,901	28.199	1,406	740	1,235
CARE-ASR, seed 27103	8,930	28.291	1,388	760	1,226
Matched MSE, seed 27101	8,921	28.262	1,423	693	1,319
Matched MSE, seed 27102	8,968	28.411	1,440	733	1,343
Matched MSE, seed 27103	8,947	28.345	1,506	758	1,363

Table 24: **Group-level errors and calibration settings for the reduced pipeline.** The upper panel contains all 16 connected meeting–speaker groups, each with 256 evaluation recordings. Identifiers name connected components, not single sessions. Words are reference-word counts; the remaining columns are total word errors. These rows reproduce the three group-level comparisons in Table 22. The lower panel gives settings selected using the original 2,231 AMI calibration recordings and 16,571 reference words. Switches count departures from raw-first ROVER; α is the linear fit's regularization parameter.

Source group	Words	Reverse ROVER	CARE-ASR	MSE	Linear
meeting:EN2001a	2,500	649	579	583	589
meeting:EN2002a	1,722	748	689	694	716
meeting:EN2004a	1,997	677	632	632	634
meeting:EN2006a	1,422	630	594	580	606
meeting:ES2003a	2,363	671	602	610	612
meeting:ES2004a	2,348	648	604	611	611
meeting:ES2008a	2,355	493	459	455	462
meeting:ES2009a	2,113	599	545	548	560
meeting:ES2010a	2,015	450	380	380	407
meeting:ES2012a	2,554	657	636	643	659
meeting:ES2016a	1,714	458	426	433	462
meeting:IS1007a	1,784	550	522	541	542
meeting:TS3005a	1,836	718	642	647	675
meeting:TS3007a	1,575	590	563	582	586
meeting:TS3011a	1,547	441	392	395	411
meeting:TS3012a	1,720	635	637	628	641
Total	31,565	9,614	8,902	8,962	9,173

Selector	Margin	Cal. errors	Switches	Linear scope	α
CARE-ASR	0.02	4,344	934	–	–
Matched MSE	0.01	4,383	1,123	–	–
Linear experts	0.002	4,461	1,098	pooled	0.01

F.9 EXPERT INFORMATION ACROSS SCORING OBJECTIVES

The previous input ablation retains the experts' chosen transcripts while removing explicit information about their judgements. We extend this control to MSE training, compare a linear selector with access to all 110 fields, and examine the expert chosen by original calibration. The protocol was specified after the cohort and earlier comparison results were known. New fitting and selection use the original roles; neither the original CARE policy nor earlier analyses change.

Matched MSE input control. The input intervention is identical to Appendix F.6: zero-based fields 16–47 and 79 are set to zero before fitting-only standardization. Candidate texts, valid slots, ordering and all other fields remain unchanged. The matched MSE implementation retains the same 21,601-parameter model, source weights, relative-error target, optimizer, early stopping and pooled-then-AMI schedules as the full-input MSE control. Seeds 27101, 27102 and 27103 each undergo both fitting stages. Original AMI calibration selects margin .005 for their mean risks and zero for each seed's descriptive policy. No seed is selected using evaluation errors.

The masked ensemble produces 9,164 errors, versus 8,855 with full MSE inputs (Table 25). The full-input reduction of 309 errors comprises 12 additional recoveries, 385 fewer erasures and 88 additional insertions. The three masked seed policies give 9,195, 9,174 and 9,136 errors, respectively. All six checkpoints, histories, calibration grids and seed predictions are supplied.

Full-representation linear control. The linear model follows the source weighting, anchored prediction, fitting-only RMS scaling, two fitting scopes, five ridge strengths and eight-margin selection procedure of Appendix F.7. Its inputs are now all 110 candidate-minus-anchor fields, rather than the eight risk differences. It has no intercept, clipping or additional nonlinear transform. All ten fits and 80 calibration settings are retained. Calibration selects AMI-only fitting, $\alpha = 1$ and margin .002, giving 8,935 evaluation errors. CARE's 141-error advantage consists of 90 extra recoveries and 74 fewer insertions, offset by 23 extra erasures. The comparison gives both models access to the same engineered fields; their losses, regularization and fitting procedures differ.

Table 25: **Expert information under two learning objectives and full-input linear selection.** All rows use 31,565 words. Recovery and erasure are relative to raw recognition. Masking removes explicit expert fields, not the chosen candidate texts. The single expert is chosen on original calibration data.

Selector	Errors	WER (%)	Recovery	Erasure	Insertions
CARE-ASR / full pairwise	8,794	27.860	1,379	637	1,204
Pairwise, expert inputs masked	9,079	28.763	1,325	960	1,112
Full MSE	8,855	28.053	1,332	576	1,279
MSE, expert inputs masked	9,164	29.032	1,320	961	1,191
Full-input linear	8,935	28.307	1,289	614	1,278
Calibration-selected expert	9,040	28.639	1,166	550	1,324

Joint analysis. Three new two-sided comparisons form a single Holm family: full versus masked MSE, CARE versus full-input linear, and the interaction

$$(L_{\text{masked pairwise}} - L_{\text{CARE}}) - (L_{\text{masked MSE}} - L_{\text{full MSE}}). \tag{9}$$

The interaction is computed within each of the original 16 connected groups. Each test enumerates all 2^{16} sign assignments; intervals are unadjusted percentiles from 50,000 paired-group bootstrap draws with seed 260916. Both direct comparisons pass Holm correction (Table 26). The interaction is $285 - 309 = -24$ errors: -0.076 WER points $[-0.491, 0.348]$, adjusted $P = 0.743805$. No interaction is detected, which does not establish equivalence between objectives. Earlier test families are unchanged.

Table 26: **Three additional expert-information contrasts.** Positive direct effects favour full MSE or CARE. A positive interaction denotes a larger expert-input benefit under pairwise training. All intervals are unadjusted 95% group-bootstrap intervals; Holm correction covers these three tests only.

Contrast	Δ WER	95% interval	Holm P	Groups $+/0/-$
Full versus masked MSE	0.979	[0.519, 1.435]	0.004211	13/0/3
CARE versus full-input linear	0.447	[0.160, 0.767]	0.017822	12/1/3
Objective–input interaction	-0.076	[-0.491, 0.348]	0.743805	8/0/8

Calibration-selected standalone expert. Among the eight archived expert policies, original AMI calibration selects the fewest word errors, then the fewest changes from raw, then fixed expert order. The selected decision-global expert has 4,378 errors over 16,571 calibration words and 1,154 changes from raw. Its original policy returns the valid non-raw view with minimum predicted risk when that risk is below $-.002$, and raw otherwise; ties follow view order. No raw-relative length guard is active for this expert. The supplementary policy records and archived-risk replay specify this standalone rule separately from the final default-policy stage. Its unchanged predictions give 9,040 assessment errors. CARE removes 246 errors (0.779 WER points; descriptive 95% group-bootstrap interval [0.449, 1.135]), with no additional inferential test.

The outputs differ on 831 of the 4,096 recordings. Among these 831 recordings, CARE has fewer word errors on 290, more on 213, and the same number on 328. The remaining 3,265 recordings have identical outputs. CARE adds 213 recoveries and 87 erasures while removing 120 insertions, giving $213 - 87 + 120 = 246$ fewer errors.

The common-bank controls support using expert information beyond the chosen transcripts. The MSE result shows this benefit with ordinary risk regression, while the full-input linear result goes beyond the earlier eight-risk comparison. Table 27 reports every group used in the additional analysis.

Table 27: **Complete group counts for the expert-information analyses.** Every group contains 256 recordings. P-mask and M-mask are pairwise and MSE selectors with the 33 expert fields masked. Linear uses all 110 fields; Expert is the original-calibration-selected standalone policy. All entries other than Words are word errors.

Group	Words	CARE	P-mask	MSE	M-mask	Linear	Expert
meeting:EN2001a	2,500	546	642	578	646	590	590
meeting:EN2002a	1,722	695	712	700	730	720	738
meeting:EN2004a	1,997	625	629	632	639	633	626
meeting:EN2006a	1,422	592	610	591	607	600	601
meeting:ES2003a	2,363	606	639	614	640	618	618
meeting:ES2004a	2,348	607	605	606	635	615	614
meeting:ES2008a	2,355	442	475	445	465	447	463
meeting:ES2009a	2,113	551	563	547	576	547	546
meeting:ES2010a	2,015	379	400	379	410	386	389
meeting:ES2012a	2,554	662	627	656	648	662	678
meeting:ES2016a	1,714	413	454	422	457	423	438
meeting:IS1007a	1,784	510	515	526	510	525	533
meeting:TS3005a	1,836	626	637	617	640	625	643
meeting:TS3007a	1,575	543	529	543	547	544	564
meeting:TS3011a	1,547	393	425	402	418	409	400
meeting:TS3012a	1,720	604	617	597	596	591	599
Total	31,565	8,794	9,079	8,855	9,164	8,935	9,040

Independent checks reproduce the calibration choices, predicted actions, source-level alignments, counts and three joint tests. The supplementary package supplies the fitted models, complete calibration grids, sealed predictions and replay tools.

F.10 NONLINEAR SELECTION FROM EXPERT RISKS ALONE

The input controls in Appendix F.9 show that explicit expert fields help, but leave open whether a nonlinear combination of the risks alone can recover the full model's performance. We therefore retain the eight candidate-specific expert risks (zero-based fields 16–23) and set the other 102 fields to zero before fitting the scaler and before inference. The candidate bank, slot order, validity masks and supervision are unchanged. Risk predictions themselves contain learned upstream information; this is a restriction on downstream inputs, not an end-to-end system without acoustic information.

Matched fitting and calibration. The control uses the same 21,601-parameter graph, independent candidate encoding and pairwise objective as the matched full-input Edit refit. Zeroing inputs preserves the graph, but reduces its varying input directions. We reuse the original fitting function, corpus and source weights, scaler rule and optimizer settings: pooled learning rate .001, at most 40 epochs and patience 7, followed by AMI adaptation at .00001, at most 24 epochs and patience 5. Batch size is 256, weight decay .0001 and gradient clipping 5. All three seeds (27101–27103) undergo both stages, with no seed selection. All 9,121 fitting and 3,087 calibration sources retain their original roles; AMI contributes 7,095 fitting and 2,231 calibration sources.

Original AMI calibration selects the margin from $\{0, .002, .005, .01, .02, .05, .1, \text{no change}\}$ by word errors, switches from slot zero, then grid order. The risk-only ensemble selects .02, with 4,416 calibration errors over 16,571 words and 822 switches. Full-input calibration replays the archived .01 choice. The matched full-input refit has different weights from the original CARE checkpoints, but reproduces every original CARE transcript on all 4,096 evaluation recordings; it is not counted as a second replication. The primary comparison uses that unchanged CARE output.

Table 28: **Full representation versus eight expert risks under matched nonlinear training.** Both use the same candidate bank and pairwise fitting procedure. The ensemble is the primary comparison; individual seeds are descriptive, not independent source replications. Recovery and erasure are relative to raw recognition.

Selector	Errors	WER (%)	Recovery	Erasure	Insertions
Full CARE / ensemble	8,794	27.860	1,379	637	1,204
Eight-risk nonlinear / ensemble	9,004	28.525	1,214	591	1,295
Separately calibrated seeds (descriptive)					
Full-input seed 27101	8,839	28.003	1,361	693	1,175
Risk-only seed 27101	9,016	28.563	1,185	567	1,302
Full-input seed 27102	8,858	28.063	1,368	690	1,204
Risk-only seed 27102	9,026	28.595	1,197	598	1,293
Full-input seed 27103	8,847	28.028	1,366	720	1,161
Risk-only seed 27103	8,983	28.459	1,274	621	1,304

Outcomes and analysis. The risk-only ensemble makes 9,004 errors (28.525% WER), compared with CARE's 8,794 (27.860%). CARE's 210-error reduction comprises 165 additional recoveries and 91 fewer insertions, offset by 46 additional erasures. All three paired-seed differences favour full inputs (Table 28). Of 901 recordings with different outputs, CARE has fewer errors on 319 and more on 232; the other 350 have equal error counts.

The protocol specifies one comparison on the already-observed cohort, with fitting and policy selection confined to the original roles. For all 16 meeting–speaker groups, we enumerate the 2^{16} sign assignments and use 50,000 paired-group bootstrap draws with seed 260916. The WER reduction is 0.665 points, with an unadjusted 95% interval of [0.339, 1.013] and two-sided $P = 0.001221$; 13 groups improve and three worsen. These are follow-up results, not a new prospective confirmation. The single-contrast test is separate from the earlier families and does not correct across the research history.

The contrast establishes the value of the remaining input fields jointly within this matched procedure. It does not identify an individual field block as the source of improvement. Together with the 33-field expert-input ablations, it supports using expert judgements in combination with candidate evidence, rather than relying on either chosen texts or eight risks alone.

Table 29: **All source groups for the nonlinear risk-only comparison.** Each group contains 256 recordings. Positive differences are risk-only minus CARE and favour the complete representation; the reference-word denominators reproduce the pooled and bootstrap results.

Group	Words	Full CARE	Eight-risk	Error difference	Δ WER (pp)
meeting:EN2001a	2,500	546	597	+51	+2.040
meeting:EN2002a	1,722	695	719	+24	+1.394
meeting:EN2004a	1,997	625	645	+20	+1.002
meeting:EN2006a	1,422	592	595	+3	+0.211
meeting:ES2003a	2,363	606	621	+15	+0.635
meeting:ES2004a	2,348	607	599	-8	-0.341
meeting:ES2008a	2,355	442	456	+14	+0.594
meeting:ES2009a	2,113	551	547	-4	-0.189
meeting:ES2010a	2,015	379	398	+19	+0.943
meeting:ES2012a	2,554	662	659	-3	-0.117
meeting:ES2016a	1,714	413	429	+16	+0.933
meeting:IS1007a	1,784	510	527	+17	+0.953
meeting:TS3005a	1,836	626	634	+8	+0.436
meeting:TS3007a	1,575	543	559	+16	+1.016
meeting:TS3011a	1,547	393	412	+19	+1.228
meeting:TS3012a	1,720	604	607	+3	+0.174
Total	31,565	8,794	9,004	+210	+0.665

The release provides all six fitted checkpoints, calibration grids, sealed predictions and original execution records. The independent verifier replays masked-input inference and calibration, recomputes word alignments and group statistics, and checks file hashes without refitting. Cached inputs shared with the other controls are referenced by content hash rather than duplicated.

G EXPERT CROSS-FITTING AND SHARED-RISK FUSION

These two follow-up studies address the construction of expert training features and the use of shared quality scores for transcript fusion. They retain the existing 4,096-recording AMI cohort, its 16 meeting–speaker connected groups and 31,565 reference words. Their protocols were fixed after earlier cohort results were available. All fitting, early stopping and policy selection use the original fitting or calibration roles; all assessment sources are retained. Each study declares two comparisons and applies Holm correction within that pair. Neither changes the original three-test prospective analysis.

G.1 AMI FITTING-SOURCE CROSS-FITTED EXPERT INPUTS

Source assignment. The 7,095 AMI fitting sources are assigned to three folds by their source-ID meeting keys. Fold 0 contains IB4010, IB4001 and IS1008c (2,479 sources); fold 1 contains IB4011, IB4003, IS1008b and IS1008a (2,394); fold 2 contains IB4004, IB4002 and IS1008d (2,222). All speaker intervals within a key stay together. The folds share speakers and are not meeting–speaker-component-disjoint. They separate each source's own labels from the expert that generates its training features.

For each fold, upstream fitting and feature normalization exclude its complete AMI source allocation. The remaining AMI fitting sources and all original non-AMI fitting sources are retained. The 3,087 original calibration sources determine early stopping and expert action thresholds. The eight expert families and three seeds in Appendix C give 72 fits across the three folds, with the original architecture, losses, source weighting, optimizer and training budget. TF32 is disabled. Each AMI fitting source receives exactly one three-seed out-of-fold prediction per expert family. Held-fold labels do not select the generating expert, scaler, checkpoint or threshold.

Rebuilding the downstream bank. All 80 transcript and expert fields are recomputed from the fold-generated risks and expert actions. The 30 acoustic and edit fields are reused only for the same source, exact candidate text and voting anchor. Expert outputs select existing native hypotheses, so the unique candidate text sets remain unchanged. Validity, valid-slot targets and errors are checked exactly. Non-AMI fitting rows and every calibration row retain their original values; calibration and assessment use the original full-fit expert features. This is an AMI-specific sensitivity analysis, not cross-fitting of every training corpus or of the recognizer.

Three downstream versions are fitted from the same random initializations: the original in-sample feature bank, the cross-fitted AMI bank, and the cross-fitted bank with fields 16–47 and 79 masked. The mask removes explicit expert inputs, not their candidate texts. All use three seeds, the original 21,601-parameter independent-context model, pairwise loss, pooled fitting and AMI adaptation schedules. This gives 18 downstream training stages. Scaling is recomputed from each version's fitting candidates after masking. The original AMI calibration pool independently selects each ensemble margin, without assessment-based variant or seed selection.

Results. The original-bank refit reproduces 8,794 errors. Cross-fitted expert inputs give 8,793, while masking their explicit fields gives 9,033 (Table 30). The information contrast contains 76 additional recoveries, 154 fewer erasures and 10 fewer insertions, summing to 240 fewer errors. All three paired seed comparisons favour full inputs; Table 32 reports every seed at its version's selected ensemble margin.

Both planned contrasts support a positive difference: 689 fewer errors than reverse ROVER and 240 fewer than the matched masked version (Table 31). Relative to the original-bank refit, the one-error difference is 0.003 WER points with a descriptive 95% interval of $[-0.347, 0.304]$. No equivalence margin was specified, and this interval is not an equivalence test. Similar total WER also does not imply identical error composition: the cross-fitted model has 107 more erasures and 101 fewer insertions than the original-bank refit, along with seven more recoveries.

Table 30: **Recognition after cross-fitting AMI expert inputs.** Word errors, WER and error transitions over the same 31,565 words. In-sample refit and cross-fitted models share training budgets. Recovery and erasure use raw recognition as the reference prediction.

System	Errors	WER (%)	Recovery	Erasure	Insertions
Original CARE	8,794	27.860	1,379	637	1,204
Reverse ROVER	9,482	30.040	1,011	876	1,285
In-sample refit	8,794	27.860	1,379	637	1,204
OOF full	8,793	27.857	1,386	744	1,103
OOF expert-masked	9,033	28.617	1,310	898	1,113

Table 31: **Two planned cross-fitting contrasts.** Positive WER differences favour the cross-fitted full model. Intervals use 50,000 paired-group bootstrap draws; exact two-sided tests enumerate 2^{16} group sign assignments. Holm correction applies to these two tests.

Comparison	Δ (pp)	95% interval	Exact p	Holm p	W/T/L
OOF full vs. Reverse ROVER	+2.183	[+1.557, +2.760]	0.000061	0.000122	15/0/1
OOF full vs. OOF expert-masked	+0.760	[+0.228, +1.282]	0.016235	0.016235	12/1/3

Positive Δ: lower WER for the first system; W/T/L: connected-group wins/ties/losses.

Table 32: **Seed-level cross-fitting results.** Assessment WER uses 31,565 words. Each version averages all three seeds for its primary policy; individual rows use that same ensemble-selected margin. Margins are chosen on the original 2,231 calibration sources with 16,571 words.

System	Seed	Ensemble τ	Errors	WER (%)
In-sample refit	27101	0.01	8,804	27.892
In-sample refit	27102	0.01	8,861	28.072
In-sample refit	27103	0.01	8,847	28.028
OOF full	27101	0.002	8,861	28.072
OOF full	27102	0.002	8,864	28.082
OOF full	27103	0.002	8,863	28.079
OOF expert-masked	27101	0	9,145	28.972
OOF expert-masked	27102	0	9,119	28.890
OOF expert-masked	27103	0	9,217	29.200

Each seed uses its variant's ensemble-selected calibration margin; seeds are descriptive.

Table 33: **Source groups for the cross-fitting analysis.** Each group contains 256 recordings. All entries except Words are word errors. Refitted denotes the original-bank execution control; OOF denotes fold-generated AMI fitting features.

Connected group	Words	Refit	OOF full	OOF masked	Reverse ROVER
meeting:EN2001a	2,500	546	585	649	677
meeting:EN2002a	1,722	695	682	707	724
meeting:EN2004a	1,997	625	624	628	674
meeting:EN2006a	1,422	592	593	615	632
meeting:ES2003a	2,363	606	598	626	673
meeting:ES2004a	2,348	607	627	608	646
meeting:ES2008a	2,355	442	434	473	501
meeting:ES2009a	2,113	551	547	562	579
meeting:ES2010a	2,015	379	381	401	449
meeting:ES2012a	2,554	662	652	631	647
meeting:ES2016a	1,714	413	422	450	466
meeting:IS1007a	1,784	510	505	509	532
meeting:TS3005a	1,836	626	608	632	662
meeting:TS3007a	1,575	543	535	520	590
meeting:TS3011a	1,547	393	397	419	420
meeting:TS3012a	1,720	604	603	603	610
Total	31,565	8,794	8,793	9,033	9,482

System columns report error counts, not percentages.

G.2 QUALITY-RANKED ROVER WITH SHARED CANDIDATE RISKS

Quality estimates can determine the order of ROVER inputs (Jalalvand et al., 2015). Here we share the original CARE ensemble risks with fusion, isolating the output rule without training a different quality estimator. This is a quality-ranked ROVER adaptation, not a reproduction of the estimator used by Jalalvand et al. (2015).

Common inputs and fusion. The native views are raw, FlowSE, MetricGAN+, PASE, RE-USE and ZipEnhancer. Each takes the risk of its first valid identical scoring-text slot in the CARE bank. Sort by increasing risk, breaking ties by original native-view order. Identical native views retain their original votes; duplicated expert-output slots add no voters. The primary fusion uses all six ranked views and the original NIST ROVER binary with equal-frequency voting, synthetic CTM intervals and the original boundary anchors. Its arguments are `-m avgconf -a 1.0 -c 0.0 -f 0 -s`, without time-distance alignment. Before the comparison, both original fixed orders were replayed exactly on all 2,231 calibration and 4,096 assessment sources.

A second rule calibrates the number of retained views. It compares $K = 0$ (raw-first ROVER fallback), $K = 1$ (the lowest-risk native transcript) and $K = 2, \ldots, 6$ (fusion of the first K ranked views). Selection minimizes original calibration word errors, breaking ties by smaller K. There is no additional rejection margin or learned confidence weight. The calibration grid selects $K = 1$ (Table 36); this policy performs selection rather than multi-view fusion.

Table 34: **Selection and fusion with shared quality scores.** Exact assessment counts over 31,565 words. Calibrated $K = 1$ returns one native transcript. Original CARE retains its original weights and margin.

System	Errors	WER (%)	Recovery	Erasure	Insertions
Original CARE	8,794	27.860	1,379	637	1,204
Quality-ranked fusion (all six)	9,343	29.599	1,014	744	1,281
Native selection ($K = 1$)	8,817	27.933	1,430	684	1,231

Table 35: **Shared-risk fusion comparisons.** Positive differences favour original CARE. Both contrasts use the same 16-group sign-flip and 50,000-draw bootstrap procedure, with a separate two-test Holm correction.

Comparison	Δ (pp)	95% interval	Exact p	Holm p	W/T/L
CARE vs. six-view fusion	+1.739	[+1.030, +2.476]	0.000336	0.000671	14/0/2
CARE vs. native selection	+0.073	[-0.057, +0.221]	0.356079	0.356079	9/2/5

Positive Δ: lower WER for the first system; W/T/L: connected-group wins/ties/losses.

Results. Six-view ranked fusion produces 9,343 errors, compared with CARE's 8,794. CARE gains 365 recoveries, removes 107 erasures and removes 77 insertions: $365 + 107 + 77 = 549$. Fourteen source groups improve. Against calibrated native selection, CARE has 51 fewer recoveries, 47 fewer erasures and 27 fewer insertions, for $-51 + 47 + 27 = 23$ fewer errors. Its interval includes zero and the adjusted test is not significant (Table 35). The comparisons therefore distinguish the gain over six-view ranked fusion from the much smaller difference against calibrated single-view selection.

Verification and release scope. The complete cross-fitting archive retains the 72 upstream fits, fold assignments, rebuilt training bank, 18 downstream stages, calibration grids, source predictions and hash manifests. Independent checks verify checkpoint hashes and initialization chains, replay the 72 upstream and nine adapted downstream models, recompute fold exclusion and normalization, reconstruct the feature bank, and reproduce source counts and group statistics. The compact submission modules provide downstream assessment replay and statistical verification; large upstream fitting arrays and weights remain in the separately preserved full archive. The fusion module supplies native hypotheses, shared risks, calibration choices, original ROVER code and recorded outputs. Its numerical verification also reconstructs source-level word alignments. No new waveform enhancement or recognition is needed to replay these saved-candidate comparisons.

Table 36: **Complete fusion calibration grid.** Only original AMI calibration data determine K. The two assessment comparisons are both reported, independent of their outcomes.

K	Errors	WER (%)	S	D	I	Changes	Policy
0	4,748	28.652	1,852	2,121	775	0	Raw-first fallback
1	4,339	26.184	1,732	1,872	735	1,080	Selected
2	4,422	26.685	1,950	1,480	992	1,032	–
3	4,424	26.697	1,832	1,769	823	835	–
4	4,502	27.168	1,841	1,773	888	725	–
5	4,557	27.500	1,791	1,967	799	610	–
6	4,624	27.904	1,809	2,022	793	479	Fixed primary

Selection: minimum calibration errors, then minimum K; $K = 6$ is the fixed primary.
Calibration: 2,231 sources, 16,571 words; changes are relative to raw-first ROVER.

Table 37: **Source groups for shared-risk fusion.** Each group contains 256 recordings; errors and reference-word counts reproduce the aggregate results and uncertainty estimates.

Connected group	Words	CARE	All-six fusion	Selected $K = 1$
meeting:EN2001a	2,500	546	667	540
meeting:EN2002a	1,722	695	720	709
meeting:EN2004a	1,997	625	655	628
meeting:EN2006a	1,422	592	627	592
meeting:ES2003a	2,363	606	654	607
meeting:ES2004a	2,348	607	631	605
meeting:ES2008a	2,355	442	493	444
meeting:ES2009a	2,113	551	576	545
meeting:ES2010a	2,015	379	432	380
meeting:ES2012a	2,554	662	640	669
meeting:ES2016a	1,714	413	467	411
meeting:IS1007a	1,784	510	524	521
meeting:TS3005a	1,836	626	675	620
meeting:TS3007a	1,575	543	562	545
meeting:TS3011a	1,547	393	421	397
meeting:TS3012a	1,720	604	599	604
Total	31,565	8,794	9,343	8,817

System columns report error counts, not percentages.

H SUPPLEMENTARY CLASSIFICATION RESULTS

The earlier classification CARE model selects raw, enhanced or probability-mean predictions using protection and rescue scores. Here, a receiver is the frozen classifier applied to the raw or enhanced input. Keyword spotting uses separate random forests, and vision uses a spatial event network. These experiments share the selective-intervention perspective but use different fitted models, action spaces and metrics from CARE-ASR.

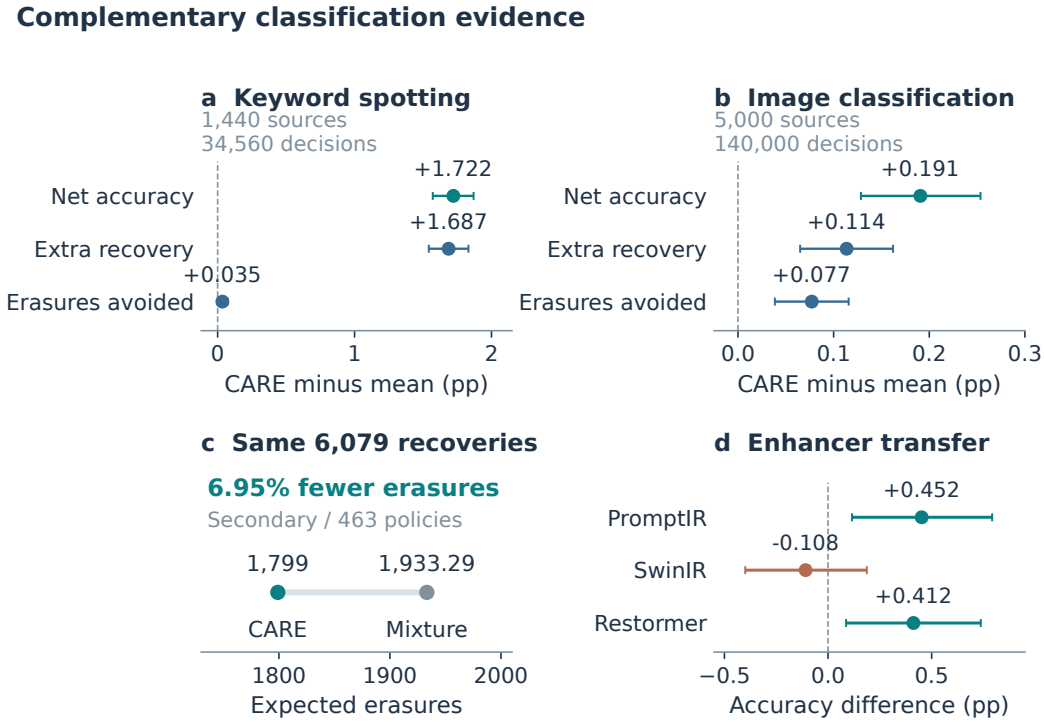


Figure 4: **Selective intervention in classification.** a,b, Accuracy, recovery and erasure differences relative to probability averaging, with the recorded marginal 95% paired-source bootstrap intervals. c, Post-evaluation exact-recovery matching on ImageNetV2 using a mixture among 463 frozen comparator points. d, Enhancer substitution with the previously fitted controller held fixed; SwinIR was evaluated before Restormer.

Keyword spotting. The reserved evaluation contains 1,440 Speech Commands recordings (Warden, 2018), four frozen receivers and six environments, yielding 34,560 decisions. Compared with probability averaging, classification CARE adds 595 correct decisions: 583 extra recoveries and 12 fewer erasures. The accuracy gain is 1.72164 percentage points, with a source-cluster 95% interval of [1.57118, 1.86921]; the smaller erasure difference has an interval including zero. Relative to output-only selection, the gain is 140 decisions. Against shuffle-refit, CARE has three additional correct decisions, 22 fewer erasures and 19 fewer recoveries. Using the entire enhanced view gives 82.211% accuracy and 383 erasures; CARE gives 79.381% and 132. Accuracy, recovery and preservation thus favour different policies; these comparisons do not hold recovery equal.

Image classification. The reserved visual evaluation uses 5,000 ImageNetV2 images (Recht et al., 2019), four frozen receivers and seven raw–PromptIR conditions (Potlapalli et al., 2023), yielding 140,000 decisions. CARE improves probability averaging from 63.31357% to 63.50429% accuracy: 267 extra correct predictions, including 159 extra recoveries and 108 fewer erasures. The accuracy difference is 0.19071 points [0.12857, 0.25357]. Extra recovery is 0.11357 points [0.06500, 0.16214], and erasure reduction is 0.07714 points [0.03857, 0.11571]. Relative to shuffle-refit,

CARE has 264 fewer erasures and 376 fewer recoveries, or 112 fewer correct predictions. Better Aggregation–partial, adapted from Shanmugam et al. (2021), gives 63.62143% accuracy and 2,826 erasures. The shuffle comparison concerns preservation within the prespecified tolerance of 0.15 percentage points in absolute accuracy difference, not equal recovery.

Recovery matching and enhancer substitution. Post-evaluation matching selects a two-policy mixture on the lower recovery–erasure hull of 463 frozen comparator operating points. Linear interpolation to CARE's 6,079 recoveries assigns weights $470/1661$ and $1191/1661$ to Better Aggregation–partial and output-only fallback, whose recovery counts are 7,270 and 5,609. This mixture minimizes expected erasure at CARE's expected recovery count, giving 1,933.29 erasures versus CARE's 1,799. The difference is 0.09592 points, with a fixed-mixture interval of [0.05893, 0.13359]. Reselecting the mixture in 2,000 source resamples gives [0.04165, 0.14116]. Selection uses the same evaluation outcomes. The earlier tolerance-selected comparison had a 13-erasure difference with an interval crossing zero; the separate Flowers102 comparison did not establish recovery equivalence or erasure superiority.

The previously fitted visual controller, normalization and policy mixture were held fixed while the enhancer changed on 1,000 previously used images, giving 8,000 decisions. The accuracy difference against output-only selection is -0.10828 points [-0.39952, 0.18708] for SwinIR (Liang et al., 2021) and $+0.41212$ points [0.08722, 0.73706] for the subsequent Restormer evaluation (Zamir et al., 2022). Restormer's two-target 97.5% interval is [0.03962, 0.78278]. The supplementary files retain the original estimands, complete comparator records and source-level intervals for these classification analyses.