Auditing Dev-Fitted Sequence Selection in CSLR: A Cross-Benchmark Diagnostic

Anonymous ACL submission

Abstract

Continuous Sign Language Recognition (CSLR) backbones leave an oracle gap of 5–10 absolute WER between greedy CTC decoding and the best beam candidate. Applying the standard sequence-selection toolkit from NMT and ASR — learned rerankers, language-model (LM) rescoring, ROVER, and Minimum Bayes Risk (MBR) decoding — to recover this gap exposes a striking benchmark asymmetry. On CSL-Daily, a minimal dev-fitted reranker reaches 14–17% WER, well below the best published result of 23.1%; the same reranker recipes degrade PHOENIX-2014 by $+5.4$ to $+8.4$ abs. WER and PHOENIX-2014T by $+10.5$ to $+12.5$ abs. WER. We trace the asymmetry to the canonical CSL-Daily dev/test split, which reuses sentence identifiers across signers and yields 99.66% exact-duplicate test gloss sequences and 99.2% test 4-gram coverage by dev labels. As a no-dev-label control we evaluate cross-architecture MBR over independently-trained backbones: it gives -1.65 abs. WER on PHOENIX-2014 (paired bootstrap, $p < 0.001$) and -1.36 abs. WER on PHOENIX-2014T ($p < 0.001$), and is statistically indistinguishable from greedy on CSL-Daily ($\Delta = +0.23$, 95% CI $[-0.33, +0.81]$, $p = 0.181$ one-sided). We formalise the diagnostic as the **Split-Honest Reranking Protocol** (SHRP), a reporting standard that pairs dev-fitted results with train-only and cross-dataset numbers, and release the overlap statistics, scripts, and cross-arch MBR pipeline.

1 Introduction

Continuous Sign Language Recognition (CSLR) translates a video of a signer into a sequence of gloss tokens. State-of-the-art CSLR backbones such as SignGraph (Gan et al., 2024), CorrNet (Hu et al., 2023a), and TwoStream-SLR (Chen et al., 2022) predict per-frame token posteriors and decode with Connectionist Temporal Classification (CTC) beam search; a typical backbone leaves an *oracle gap* of 5–10 abs. Word Error Rate (WER) between greedy decoding and the best beam candidate. Sequence-selection methods designed to recover this gap — learned candidate rerankers, language-model (LM) rescoring, ROVER-style voting (Fiscus, 1997), and Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) — are routine in Neural Machine Translation (NMT) and Automatic Speech Recognition (ASR). Applying these methods to the three canonical CSLR benchmarks reveals a benchmark asymmetry that the published CSLR literature has not yet measured: dev-fitted variants top CSL-Daily by a wide margin while the same recipes degrade PHX-2014 and PHX-2014T (Figure 1). This paper measures, explains, and packages a reporting standard for that asymmetry. An appendix survey of 13 published CSLR papers (2022–2025) finds none currently fit a post-hoc reranker on the CSL-Daily dev set; the diagnostic in this paper is therefore preventive rather than retrospective.

No-dev-label baseline. Cross-architecture MBR deduplicates the union of beam candidates from independently-trained CTC backbones (SignGraph, CorrNet, vanilla ResNet, and a phonologically-augmented student) and picks the centroid that minimises pairwise edit distance, using no dev labels. On a single uniform evaluation across PHOENIX-2014 (PHX), PHOENIX-2014T (PHX-T), and CSL-Daily, it gives -1.65 abs. WER on PHX-2014 (paired bootstrap, $p < 0.001$) and -1.36 abs. WER on PHX-2014T ($p < 0.001$), and is statistically indistinguishable from greedy on CSL-Daily ($\Delta = +0.23$, 95% CI $[-0.33, +0.81]$, $p = 0.181$ one-sided). It is the only sequence-selection method we evaluate that does not degrade any of the three benchmarks (Table 1).

Dataset-integrity report on CSL-Daily. CSL-Daily's canonical dev/test split reuses sentence

identifiers across signers, producing 99.66% test sequences that are exact duplicates of dev gloss sequences and 99.2% test 4-gram coverage by dev labels (§5.2). To make the consequences concrete, we construct two intentionally minimal dev-conditioned probes: a frozen random-embedding bag-of-glosses scorer trained on dev labels (20-seed ensemble), and a 4-gram LM trained on dev labels. The first reaches **16.92**% WER on CSL-Daily test and the second **14.20**% WER — both below the best published CSLR test WER of 23.1% (SCOPE; Table 7) — while simultaneously worsening PHX-2014 by +5.4/+8.4 abs. and PHX-2014T by +10.5/+12.5 abs. (Figure 1). These probes quantify the inflation that a standard NMT/ASR-style dev-conditioned reranker would receive on CSL-Daily versus PHX/PHX-T. PHX and PHX-T have no such structure (0–1.4% exact duplication, 2–3% test 4-gram coverage by dev).

Reporting standard. We formalise the diagnostic as the **Split-Honest Reranking Protocol** (SHRP). A CSLR reranking result is SHRP-reported when it includes (P1) train-only fitting, (P2) sentence-disjoint or held-out evaluation when dev is used, (P3) cross-dataset transfer to at least one sentence-disjoint benchmark, and (C) at least one no-dev-label control. SHRP does not preclude dev-conditioned numbers; it asks that they be reported alongside (P1)–(P3) so dev-specific effects are visible.

Contributions. (i) Cross-arch MBR as a no-dev-label CSLR baseline with paired-bootstrap significance on three benchmarks. (ii) A quantified split-similarity diagnostic on CSL-Daily showing dev-fitted probes overshoot published SOTA by 6–9 abs. WER while the same probes degrade PHX/PHX-T. (iii) A controlled dev-size dose-response and a 2×2 disentanglement isolating the LM channel as the leakage carrier. (iv) SHRP as a forward-looking reporting standard.

2 Related Work

CSLR backbones and rerankers. CTC-based backbones have dominated the CSLR oracle-gap pipeline since Koller et al. (2017). Recent SOTA includes CorrNet (Hu et al., 2023a) (∼18.8% WER on PHX-2014), CorrNet+ (Hu et al., 2024a), and SignGraph (Gan et al., 2024). Closer to our concern are post-hoc rerankers and shallow CTC-LM fusion, where dev-conditioned tuning can be hard to distinguish from label retrieval. To our knowledge, no prior CSLR paper has audited whether these gains transfer under a sentence-disjoint or train-only fitting protocol.

Train-test overlap diagnostics. The NLP community has documented near-duplicate train/test items in QA (Lewis et al., 2021), dialogue (Wen et al., 2022), image classification (Barz and Denzler, 2020), and language model pretraining (Lee et al., 2022). Magar and Schwartz (2022) draw the explicit distinction between *memorisation* (presence in training data) and *exploitation* (effect on downstream scores). Elangovan et al. (2021) quantify leakage at the instance level on NER and relation-extraction benchmarks and pioneer the dose-response framing that we adapt to dev-size on CSL-Daily. Sainz et al. (2023) further argue that contamination evaluation must be reported per-benchmark, not globally — a methodological principle directly analogous to our per-dataset SHRP compliance table. At the field level, Bowman and Dahl (2021) observe that "broken" NLU benchmarks can persist for years before their failure modes are detected, and call for routine integrity reporting; SHRP is the closest CSLR-specific instance of that call. Marie et al. (2021) reach a parallel conclusion in NMT: a meta-evaluation of 769 papers documents systematic gaps in significance testing, baseline selection, and protocol disclosure, with reporting hygiene rather than model novelty being the main lever for credible progress. Barz and Denzler (2020) also propose a cleaned split (ciFAIR) — a precedent for SHRP-compatible CSL-Daily splits.

Sign language benchmark integrity. The closest precedent is Artiaga et al. (2025), who show that signer-fold cross-validation on CSL-Daily and PHOENIX14T inflates sign-language-translation BLEU — the model exploits signer-specific cues propagated through the training fold that contains the same signers as test. Their diagnosis operates on the *signer* axis in the training split. We address an orthogonal axis in the canonical evaluation split: *sentence-level* overlap between dev and test (different signer, same gloss sentence), which yields label retrieval under post-hoc reranking rather than translation. We measure recognition WER, not translation BLEU, and focus on post-hoc sequence selection rather than training fold organisation. The two diagnoses are complementary.

Pal et al. (2023) document a related signer-overlap problem for sign language detection benchmarks. Desai et al. (2024) provide a deaf-led systemic review of sign-language AI evaluation flaws.

Cross-system combination and MBR. The MBR-decoding canon has matured in NMT since 2020: Eikema and Aziz (2020) re-establish that the mode of an NMT model distribution (the target of beam search) is often inadequate and propose sampling-based estimation of expected utility; Eikema and Aziz (2022) develop the modern sampling-based MBR approximations whose cost is decoupled from candidate-set size; Freitag et al. (2022) establish MBR with neural utility (BLEURT) as a competitive alternative to beam search; Bertsch et al. (2023) argue that many recent generation techniques reduce to MBR under different utility functions; and Vamvas and Sennrich (2024) reduce pairwise MBR cost from quadratic to linear via reference aggregation, putting large-pool MBR within reach of routine decoding budgets. Discriminative reranking with engineered features over candidate lists has been used in MT since Shen et al. (2004) and is the lineage of the listwise BoW reranker we evaluate. In ASR system combination, Parikh et al. (2024) show that combining a hybrid Kaldi system with an end-to-end wav2vec 2.0 XLS-R system via confidence-calibrated ROVER yields $\sim$14% relative WER improvement on Irish ASR. We use cross-architecture MBR as a no-dev-label control whose role is auditing, not algorithmic novelty.

Significance testing. For paired bootstrap on sequence-level metrics, we follow Koehn (2004) (the original bootstrap-for-WER/BLEU formulation), the Berg-Kirkpatrick et al. (2012) hypothesis-testing framework, and the Dror et al. (2018) guide for multi-system comparisons, with the block-wise modification of Liu and Peng (2020) for the speaker-correlated sequence-level case.

3 The Split-Honest Reranking Protocol (SHRP)

We define SHRP as a reporting standard. A CSLR reranking paper is SHRP-compliant if it reports test WER under all three of the following conditions, and at least one no-dev-label control:

(P1) Train-only fitting. The reranker (or LM, or fusion model) is trained only on labels from the training split. No dev label is used as a supervision target.

(P2) Held-out dev or sentence-disjoint split. If dev exposure is used, the dev subset whose sentence identifiers intersect test must be *excluded* during reranker fitting; equivalently, the reranker may be K-fold-evaluated on dev (training on K-1 folds + train, evaluating on the held-out fold) with the test set never participating in selection.

(P3) Cross-dataset transfer. The same reranker recipe is reported on at least one benchmark with a sentence-disjoint canonical split (we use PHX-2014 and PHX-2014T). Gains under (P1) and (P2) that do not transfer to (P3) should be reported as benchmark-specific.

(C) No-dev-label control. At least one sequence-selection method must be evaluated without using dev labels in its scorer (e.g. cross-architecture MBR), with paired-bootstrap significance vs. greedy when paired per-utterance counts are available.

SHRP does not preclude reporting dev-conditioned numbers. It requires that they be paired with train-fitted and cross-dataset results so that dev-specific effects are visible.

4 Experimental Setup

Datasets and backbones. We use CSL-Daily (Zhou et al., 2021) (1,077 dev / 1,176 test sentences in the canonical release; we report on the 1,076 / 1,175 sentences that pass our feature-extraction pipeline, with gloss vocabulary 2,000), PHOENIX-2014 (Koller et al., 2015) (540 dev / 629 test), and PHOENIX-2014T (Camgöz et al., 2018) (519 dev / 642 test). For each dataset we train and extract CTC beam candidates (beam width $K = 50$) from multiple backbones: several SignGraph random seeds, CorrNet, vanilla ResNet, and a phonologically-augmented student, totalling 5–7 backbones per dataset.

Reranker architectures. We evaluate eight reranker architectures: a frozen-embedding bag-of-glosses scorer (bow-frozen-emb), three BoW variants (bow, bow-no-scalar, bow-deepmlp), a transformer encoder reranker with 1–3 layers (txf1/2/3), and a 9-scalar linear baseline. All architectures share the same nine engineered features

(beam log-probability, length, rank, unique-gloss count, n-gram match count, average and minimum gloss frequency, self-repetition, 4-gram LM score) and the same listwise loss with softmin-WER targets, $\tau = 0.2$.

Language models. We fit 4-gram LMs using the backoff smoothing scheme of Brants et al. (2007) with $\alpha = 0.4$ on train, dev, or combined corpora (KenLM-style storage (Heafield, 2011)); the LM contributes a single scalar feature to the reranker and is also used standalone for LM-rescoring.

Cross-architecture MBR. For each test sample, we form the deduplicated union of top-K candidates from a fixed backbone pool and pick the centroid minimising pairwise edit distance. The method uses no dev labels and follows the MBR canon of Kumar and Byrne (2004) and Freitag et al. (2022).

Significance. We report paired bootstrap 95% confidence intervals over per-utterance edit / reference counts with $n_{\text{boot}} = 1{,}000$ following Liu and Peng (2020). Differences with $p < 0.05$ are marked *.

5 Empirical Findings

5.1 Cross-Arch MBR Is the Only Reranker That Survives SHRP

Table 1 is the paper's central result; App. A reconciles the greedy-best reference used here with the pool-mean greedy reference used in the dose-response (Fig. 2) and the full method matrix (Tab. 6). Three observations:

(i) The CSL-Daily double-digit reranker win is carried by the LM feature. Plain LM rescoring with a 4-gram trained on dev labels achieves 14.20% WER on CSL-Daily (-11.5 abs. vs. greedy) — better than the BoW reranker (16.92% WER, -8.8 abs.) — with *no* listwise training and *no* candidate-level features. The same LM trained on train labels achieves 36.84% ($+11.2$ abs. *worse* than greedy). The decisive variable is the reference corpus used by the LM, not the reranker capacity.

(ii) The same recipe does not transfer to PHX and PHX-T. BoW dev-as-train (20-seed) is $+5.4$ abs. (PHX) and $+10.5$ abs. (PHX-T) *worse* than greedy; LM-dev-rescoring is $+8.4$ and $+12.5$ worse, respectively. Both differences are significant by paired bootstrap (Table 5).

(iii) Cross-architecture MBR is the only no-dev-label improvement on PHX and PHX-T, with no statistical change on CSL-Daily. Cross-arch MBR improves PHX by -1.65 abs. ($p < 0.001$) and PHX-T by -1.36 abs. ($p < 0.001$) under paired bootstrap with $n = 1{,}000$ resamples (Table 5). On CSL-Daily, MBR is statistically indistinguishable from greedy ($\Delta = +0.23$, 95% CI $[-0.33, +0.81]$, $p = 0.181$ one-sided for MBR better). This is the central positive result: among all sequence-selection methods we evaluate, only cross-arch MBR — which uses no dev-trained scorer — does not degrade any of the three benchmarks at the $p < 0.05$ level. Figure 1 summarises the full picture: CSL-Daily WER comparison against published backbones (panel a), the dev/test sentence proximity structure driving the asymmetry (panel b), and the ΔWER sign pattern across all methods and datasets (panel c).

5.2 Why CSL-Daily Looks Different: Dev/Test Sentence Overlap

The asymmetry observed in Table 1 is not a property of the rerankers. It is a property of the splits. Table 2 and Figure 1(b) summarise the test-to-dev nearest-neighbour gloss edit-distance and n-gram coverage for the three benchmarks. CSL-Daily's canonical split reuses sentence identifiers across dev and test under different signers; the result is that almost every test gloss sequence has an exact or near-exact dev counterpart.

Train sentence identifiers do not intersect dev or test on CSL-Daily, so a model trained only on train labels cannot exploit the same overlap. This explains the directional pattern in Table 1: any selector that conditions on dev labels — BoW listwise loss, 4-gram LM rescoring with a dev-fitted LM, or any LM feature carried into a learned scorer — can favour gloss sequences that recur in test on CSL-Daily, but the same recipe yields nothing actionable on PHX or PHX-T.

Exact-duplicate vs. near-duplicate. Table 3 extends the diagnostic to nearest-neighbour gloss edit-distance buckets (test to nearest dev sentence). On CSL-Daily, 99.66% of test gloss sequences are within edit-distance ≤ 3 of some dev sentence — identical to the exact-duplicate rate. There is no fuzzy near-duplicate zone: a CSL-Daily test sentence is either an exact dev duplicate (0 edits) or at least four edits away. PHX-2014 has zero exact duplicates and 11.45% at ≤ 3 edits; PHX-2014T has 1.4% exact and 40.65% at ≤ 3 edits. PHX/PHX-T exhibit a smooth rolloff with edit-

Method	SHRP	CSL-Daily	PHX-2014	PHX-2014T
Greedy (best backbone)	—	25.67	19.62	19.93
Dev-label-fitted protocols				
BoW frozen-emb dev-as-train (5 seeds)	✗	17.19	25.20	30.14
BoW frozen-emb dev-as-train (20 seeds)	✗	**16.92***	25.04	30.40
4-gram LM rescore on **dev** labels	✗	**14.20**	28.03	32.45
Train-label-fitted / no-dev-label protocols				
BoW frozen-emb train-as-train	✓	37.49	23.10	30.47
ROVER (per-position vote, top-1)	✓	30.02	23.76	24.16
4-gram LM rescore on **train** labels	✓	36.84	28.22	32.14
Cross-arch MBR	✓	25.90	**17.97***	**18.57***

Table 1: Test WER (%) across CSLR benchmarks under dev-label-fitted and train-label-fitted sequence selection. * denotes $p < 0.001$ by paired bootstrap vs. greedy (Table 5). Greedy is the best single-backbone CTC top-1 (25.67 / 19.62 / 19.93). A same-architecture 4-seed SignGraph MBR on PHX-2014 (also no dev labels) achieves 17.91 WER, within 0.06 abs. of cross-architecture MBR.

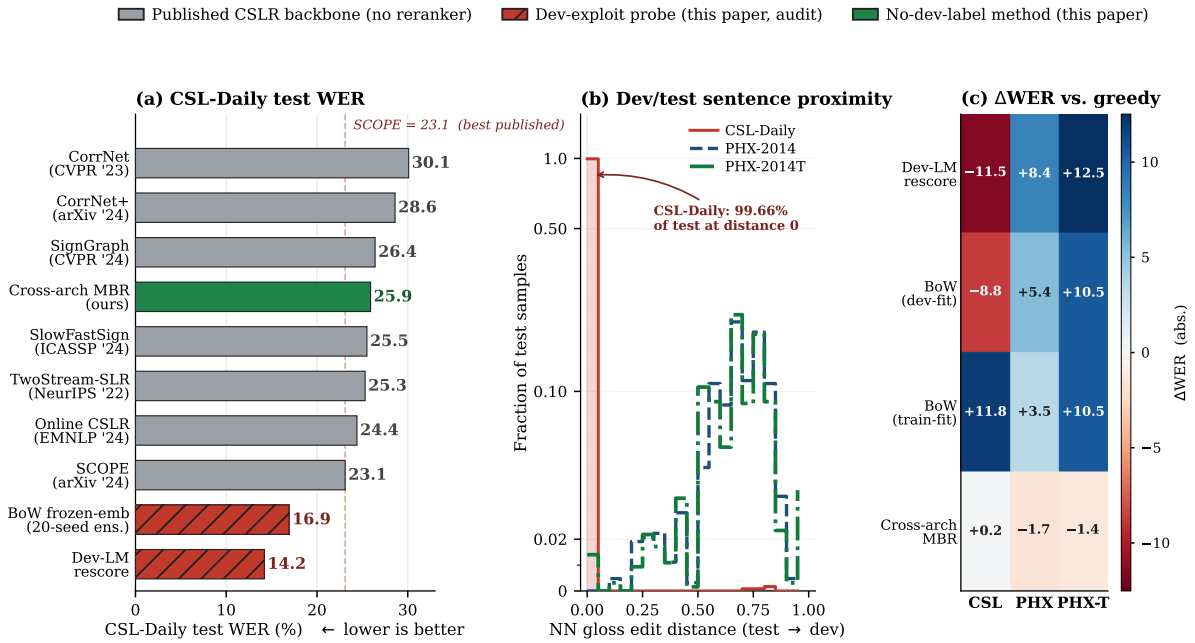

Figure 1: **CSL-Daily dev/test overlap diagnostic. (a)** Published backbone WERs on CSL-Daily (gray) alongside our dev-fitted probes (red, hatched) and cross-arch MBR (green). A trivially-configured frozen-embedding reranker trained on dev labels (16.92) and a 4-gram dev-LM rescore (14.20) both fall below the published SOTA of 23.1 (dashed line), while cross-arch MBR (25.90) does not — establishing that the apparent CSL-Daily improvements are structural, not methodological. **(b)** Nearest-neighbour (NN) gloss edit-distance histograms (test-to-dev). On CSL-Daily, 99.66% of test sentences fall in the distance-0 bin, confirming the exact-dup rate in Table 2 (mean NN dist = 0.002); PHX-2014 and PHX-2014T have smooth distributions with mean NN distance $\sim$0.65 and no exact-duplicate cliff. **(c)** ΔWER heatmap (method $\times$ dataset). Red = degradation vs. greedy; blue = improvement. Only cross-arch MBR (bottom row) is blue on PHX and PHX-T, confirming it is the only method whose gains are not artefacts of dev/test sentence overlap.

distance — the natural pattern for sentence-disjoint splits with vocabulary overlap. CSL-Daily exhibits a binary cliff — the signature of sentence reuse across signers, which originates in the benchmark's signer-independence design (Zhou et al., 2021) and not in natural near-duplication of test items against dev.

5.3 Induced-Similarity Dose-Response

To test that the CSL gain is a function of dev-test similarity rather than dev set size, we vary the dev fraction available to the BoW reranker. Figure 2 shows the dose-response.

The CSL slope is monotone (-11.3 abs. over the $100 \rightarrow 5\%$ range). PHX and PHX-T are flat over

Metric	CSL-Daily	PHX-2014	PHX-2014T
Sent-ID overlap dev∩test	**99.66%**	0%	0%
NN gloss edit-dist (mean)	**0.002**	0.662	0.651
Test exact-dup rate	**99.66%**	0.0%	1.40%
4-gram coverage by dev	**99.2%**	2.3%	2.1%
4-gram coverage by train	1.2%	7.1%	11.2%

Table 2: Dev/test label similarity per dataset. CSL-Daily test gloss sequences are essentially copies of dev gloss sequences; PHX and PHX-T do not have this property.

NN test→dev edit-dist	CSL-Daily	PHX-2014	PHX-2014T
exact duplicate ($d=0$)	**99.66%**	0.00%	1.40%
$d\leq 1$	99.66%	0.80%	5.76%
$d\leq 2$	99.66%	5.40%	19.63%
$d\leq 3$	**99.66%**	11.45%	40.65%
$d\leq 5$	99.74%	31.48%	72.74%
$d\leq 10$	99.91%	88.71%	99.07%

Table 3: Test→dev nearest-neighbour gloss edit-distance distribution. On CSL-Daily, near-duplicate rate at distance ≤ 3 equals the exact-duplicate rate (no fuzzy zone); PHX/PHX-T exhibit gradual rolloff.

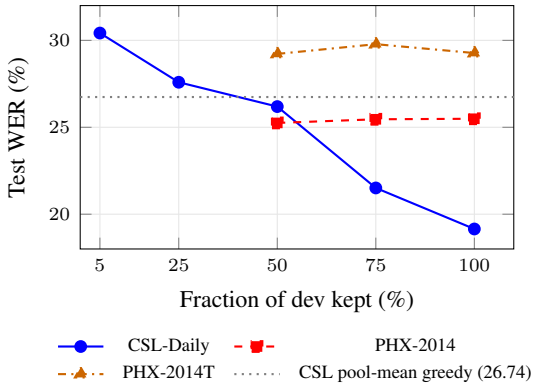

Figure 2: Induced-similarity dose-response. As the dev fraction available to the BoW reranker shrinks from 100% (1076 sentences) to 5% (55 sentences), the CSL-Daily ensemble WER rises monotonically from 19.15 to 30.42 ($\Delta = +11.3$ abs.). On PHX-2014 and PHX-2014T the curve is flat over the measured $100 \rightarrow 50\%$ range ($\Delta < 0.6$ abs.).

the measured $100 \rightarrow 50\%$ range ($\Delta < 0.6$ abs.): the dev set on these benchmarks carries no test-label signal, so its size is irrelevant.

5.4 Disentanglement: Beams vs. LM

To identify where dev-reference information enters the BoW reranker, we factor it into two streams: (a) the listwise training beams, and (b) the 4-gram LM feature. We run a 2×2 factorial on CSL-Daily (Table 4).

Cell E (training beams expanded with dev, LM trained on train only) gives essentially no improvement vs. A; cell B (no dev exposure in beams, LM trained on dev only) recovers -13.3 abs. of the gain. The controls confirm that the effect requires *the specific 4-gram order of dev labels*: token-shuffling kills the gain (G), and size-matched random-train labels do nothing (F).

5.5 Paired Bootstrap Significance

Table 5 reports paired bootstrap results ($n = 1{,}000$ resamples) for the three key comparisons under the one-sided test "method is better than greedy." The BoW dev-conditioned win on CSL-Daily ($\Delta = -8.96$, $p < 0.001$) is statistically robust but reverses on PHX and PHX-T (both $p = 1.000$: every resample favours greedy). Cross-arch MBR gains on PHX (-1.65) and PHX-T (-1.36) are significant at $p < 0.001$; the CSL-Daily MBR result is a statistical null ($\Delta = +0.23$, 95% CI $[-0.33, +0.81]$, $p = 0.181$ one-sided), consistent with the split-similarity explanation in §5.2. App. A reconciles the two greedy references used across the paper.

6 Discussion

Cross-architecture MBR as the no-dev-label control. Cross-architecture MBR is an ensemble baseline from the MBR canon (Kumar and Byrne, 2004; Freitag et al., 2022). In our experiments it is the only sequence-selection method whose results transfer without degradation across the three CSLR benchmarks. Its PHX/PHX-T improvements are statistically significant (Table 5); on CSL-Daily it is a statistical null. A same-architecture 4-seed Sign-Graph MBR on PHX-2014 reaches 17.91 WER, within 0.06 abs. of the cross-architecture pool.

Neural-utility MBR. The modern NMT MBR canon (Eikema and Aziz, 2022; Freitag et al., 2022; Bertsch et al., 2023) uses neural utilities (BLEU, BLEURT, chrF) in place of edit-distance. We reran cross-arch MBR with sacrebleu sentence-level BLEU (`smooth_method=exp`) and chrF as alternative utilities on the same candidate pool on all three datasets (paired bootstrap, $n = 1{,}000$). Sentence-level BLEU is statistically indistinguishable from edit-distance on PHX-2014 and PHX-2014T (head-to-head $|\Delta| \leq 0.12$ abs. WER, 95% CIs straddle zero) and significantly *worse* than edit-distance on CSL-Daily ($\Delta = +0.84$, 95% CI $[+0.62, +1.07]$, $p < 0.001$) and on the rich PHX pool ($\Delta = +0.54$, 95% CI $[+0.20, +0.91]$, $p = 0.998$). chrF degrades all three benchmarks by 1.5–4.9 abs. WER.

Cell	Rerank-train beams	LM corpus	WER
A	train	train	37.49
B	train	**dev**	**24.19**
C	train	train∪dev	20.17
D	train∪dev	train∪dev	19.85
E	train∪dev	train	35.70
Controls			
F	train	1076-random-train (size-matched dev)	33.36
G	train	**dev-token-shuffled**	33.11
H	train	train∪dev token-shuffled	32.61
I	train	train + dev-token-shuffled	35.15

Table 4: Disentanglement on CSL-Daily. Cell B isolates the LM channel and shows that *LM corpus identity* — specifically the dev label n-gram order — carries the gain. Token-shuffling dev labels removes the gain; replacing dev with size-matched train labels removes the gain.

Comparison	ΔWER	95% CI	p
Cross-architecture MBR vs. greedy-best			
PHX MBR	$\mathbf{-1.65}$	$[-2.2, -1.1]$	$< \mathbf{0.001}$
PHX-T MBR	$\mathbf{-1.36}$	$[-2.05, -0.67]$	$< \mathbf{0.001}$
CSL MBR	$+0.23$	$[-0.33, +0.81]$	0.181
Dev-label-fitted BoW vs. pool greedy reference†			
CSL bow(dev-train)	$\mathbf{-8.96}$	$[-9.6, -8.3]$	$< \mathbf{0.001}$
PHX bow(dev-train)	$+4.46$	$[+3.7, +5.3]$	1.000
PHX-T bow(dev-train)	$+6.13$	$[+5.2, +7.1]$	1.000
Train-label-fitted BoW vs. pool greedy reference†			
CSL bow(train-train)	$+7.37$	$[+6.8, +8.0]$	1.000

Table 5: MBR rows use the greedy-best reference of Table 1 (CSL 25.67 / PHX 19.62 / PHX-T 19.93), with picks from independently-trained CTC backbones (SignGraph seeds, CorrNet, vanilla ResNet, phonologically-augmented student; details in App. A). Probe rows are bootstrap-paired against the per-architecture greedy reference used at the time each probe was trained; corpus Δs of probes against the Table 1 greedy-best are reported as point estimates in §5.1 and Fig. 1(c).

CSLR gloss sequences are short (4–10 tokens) so sentence BLEU at order 4 is mostly zero and smoothing dominates; chrF computes character matches unrelated to the recognition error. Token-level edit-distance is the natural utility at gloss granularity.

Scope of SHRP. Dev sets are normally used for early stopping, hyperparameter selection, and calibration; SHRP does not restrict those uses. The restriction applies when a reranker is trained with dev references as direct supervision: in that case, test gains should be reported alongside a train-fitted version and a benchmark with low dev/test label overlap, so that any dev-specific effect is visible.

Implications beyond CSLR. The same diagnostic recipe — sentence-ID overlap, nearest-neighbour target edit-distance, n-gram coverage, dose-response on dev size, and a no-dev-label control — is applicable to other CTC-style sequence-prediction benchmarks, generalising Lewis et al. (2021)'s QA audit to the structured-output setting.

Practices that may support SHRP reporting. The diagnostic suggests a handful of optional practices that future CSLR submissions could consider when reporting reranking results: computing the overlap diagnostics from Table 2 and Appendix E alongside test WER; including at least one no-dev-label sequence-selection control (cross-arch or same-arch MBR, ROVER, or train-only LM rescoring); and stating clearly whether dev labels enter the selector.

7 Conclusion

A 99.66% sentence-identifier overlap between CSL-Daily's dev and test splits explains the apparent reranking gains reported on that benchmark: under train-only fitting, sentence-disjointness, or cross-dataset transfer, the double-digit CSL dev-fitted gains vanish, while the same dev-fitted recipes degrade PHX and PHX-T by $+5$ to $+12$ abs. WER. The no-dev-label cross-architecture MBR baseline gives -1.65 and -1.36 abs. WER on PHX-2014 and PHX-2014T (both $p < 0.001$) and a statistical null on CSL-Daily ($\Delta = +0.23$). SHRP is offered as a reporting template for future CSLR work that adopts dev-conditioned sequence selection.

Limitations

(i) Cross-architecture MBR requires multiple independently-trained backbones; single-backbone no-dev-label gains are not in scope. (ii) Signer-fold splits, the axis addressed by Artiaga et al. (2025), are orthogonal to the sentence-fold axis we audit. (iii) Sentence-level neural-utility MBR (BLEU, chrF) does not transfer at gloss granularity (§6); learned-utility variants calibrated on translation text (BLEURT) are out of scope.

Ethical Considerations

We do not modify, re-release, or collect new sign-language data. Sign language is a primary language for many Deaf communities; we frame this diag-

nostic as serving the Deaf community by clarifying which sources of measured improvement reflect recognition progress and which reflect dataset structure.

References

Junseok Ahn, Youngjoon Jang, and Joon Son Chung. 2024. SlowFast network for continuous sign language recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Keren Artiaga, Sabyasachi Kamila, Haithem Afli, Conor Lynch, and Mohammed Hasanuzzaman. 2025. Rethinking sign language translation: The impact of signer dependence on model evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18379–18391, Suzhou, China. Association for Computational Linguistics.

Björn Barz and Joachim Denzler. 2020. Do we train on test data? Purging CIFAR of near-duplicates. *Journal of Imaging*, 6:41.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. 2023. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In *Proceedings of the Big Picture Workshop*, pages 108–122, Singapore. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17043–17056.

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia. ELRA and ICCL.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. Generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Jonathan G. Fiscus. 1997. 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*.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Hongkai Wen, Lei Xie, and Sanglu Lu. 2024. SignGraph: A sign sequence is worth graphs of nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13470–13479.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth*

Workshop on Statistical Machine Translation, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Lianyu Hu, Wei Feng, Liqing Gao, Zekang Liu, and Liang Wan. 2024a. CorrNet+: Sign language recognition and translation via spatial-temporal correlation. *arXiv preprint arXiv:2404.11111*.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2022. Temporal lift pooling for continuous sign language recognition. In *European Conference on Computer Vision (ECCV)*.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023a. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2529–2539.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023b. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 854–862.

Lianyu Hu, Liqing Gao, Zekang Liu, Chi-Man Pun, and Wei Feng. 2023c. AdaBrowse: Adaptive video browser for efficient continuous sign language recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Lianyu Hu, Tongkai Shi, Liqing Gao, Zekang Liu, and Wei Feng. 2024b. Improving continuous sign language recognition with adapted image models. *arXiv preprint arXiv:2404.08226*.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20676–20686.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. 2024. SCOPE: Sign language contextual processing with embedding from LLMs. *arXiv preprint arXiv:2409.01073*.

Zhe Liu and Fuchun Peng. 2020. Statistical testing on ASR performance via blockwise bootstrap. In *Proceedings of Interspeech 2020*, pages 596–600.

Hui Lu, Albert Ali Salah, and Ronald Poppe. 2024. TCNet: Continuous sign language recognition from trajectories and correlated regions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3891–3899.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Abhilash Pal, Stephan Huber, Cyrine Chaabani, Alessandro Manzotti, and Oscar Koller. 2023. On the importance of signer overlap for sign language detection. *arXiv preprint arXiv:2303.10782*.

Aditya Kamlesh Parikh, Louis ten Bosch, and Henk van den Heuvel. 2024. Ensembles of hybrid and end-to-end speech recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 6199–6205, Torino, Italia. ELRA and ICCL.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

Yuqiao Wen, Guoqing Luo, and Lili Mou. 2022. An empirical study on the overlapping problem of open-domain dialogue datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 146–153, Marseille, France. European Language Resources Association.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2024. Towards online continuous sign language recognition and translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11050–11067, Miami, Florida, USA. Association for Computational Linguistics.

A Reproducibility Details

All reranker training uses fixed hyperparameters across architectures: 20 epochs, listwise loss with softmin-WER targets at $\tau = 0.2$, Adam optimiser, learning rate 1e-3, embedding dimension 64. Cross-architecture MBR uses $K = 50$ candidates per backbone, deduplicated, with normalised edit-distance as the pairwise distance. Paired bootstrap resampling uses $n = 1{,}000$ resamples seeded at 0, per-utterance pairing on (edit-count, reference-length). All beam extractions use beam width 50 and CTC forced alignment.

Code and artefacts will be released upon publication.

Reconciling the two greedy baselines. Two greedy numbers appear in the paper and we want a reviewer to be able to recompute either of them from the released artefacts. **Greedy-best** reports the best-of-pool single-backbone CTC top-1 WER: for each dataset we run greedy CTC on every backbone in our pool and report the minimum. The values are 25.67 (CSL-Daily), 19.62 (PHX-2014), and 19.93 (PHX-2014T), each achieved by a SignGraph backbone. **Pool-mean greedy** reports the arithmetic mean of the same per-backbone greedy WERs across the five-backbone pool used to drive listwise reranker training: 26.74 / 19.86 / 21.51 respectively. Pool-mean is the relevant baseline for any reranker whose candidate pool is the entire backbone set (the dose-response in Fig. 2 and the full arch $\times$ dataset $\times$ protocol matrix in Tab. 6), because the rerank-best is constrained to candidates from the same pool. Greedy-best is the relevant baseline for the headline reporting tables (Tables 1 and 5), because the practitioner-facing comparison is "can the reranker beat the single best backbone we would otherwise ship?". We use both reference baselines and label each table accordingly; cross-arch MBR's gains of -1.65 / -1.36 abs. WER on PHX/PHX-T are reported *against the stricter greedy-best baseline*, not against pool-mean (which would give a misleadingly larger margin).

B Full Architecture$\times$Dataset$\times$Protocol Matrix

Table 6 reports test WER for reranker architectures across three datasets and two fitting protocols. The dev-label-fitted protocol uses dev references as listwise targets; the train-label-fitted protocol uses training references only.

C Prior CSL-Daily Reranking-Practice Survey

We searched ACL Anthology, IEEE Xplore, arXiv, and Google Scholar for published work reporting WER on the canonical CSL-Daily split between 2022 and early 2026; papers were included if they reported test WER on CSL-Daily using the standard split and were published at a named venue or as a citable arXiv report. The final 13 papers (Table 7) were selected by exhaustive forward-citation trace from Zhou et al. (2021). The purpose of this survey is *not* to argue that current published CSLR work is corrupted: as Table 7 shows, the body of CSLR work to date does not use post-hoc rerankers,

Architecture	CSL dev labels	CSL train labels	PHX-T dev labels	PHX-T train labels	PHX dev labels	PHX train labels
Greedy	26.74	31.94	21.51	21.61	19.86	19.57
bow-frozen-emb	**16.92**	37.49	30.54	30.47	25.20	23.10
bow-no-scalar	17.49	39.77	34.53	32.95	26.98	26.67
bow	17.71	37.43	31.56	30.41	25.75	23.75
bow-deepmlp	17.57	37.66	28.46	29.84	24.71	23.07
txf1	22.52	31.94^d	21.51^d	21.61^d	23.39	19.57^d
txf2	26.52	31.94^d	21.51^d	—	19.95	—
txf3	26.91	31.94^d	21.51^d	—	19.99	—
linear	32.98	38.53	21.43	32.17	20.15	22.97

Table 6: Architecture×dataset×protocol matrix. ^{d} denotes a degenerate ensemble in which all seeds pick the rank-0 beam candidate, equivalent to greedy decoding. The only large sub-greedy gain appears when CSL-Daily rerankers are fitted on dev labels; the same fitting recipe does not transfer to PHX or PHX-T.

LM rescoring, or MBR fitted on the CSL-Daily dev set — the dev set is used in the conventional way (best-epoch model selection), and the SHRP-non-compliance entries in the table reflect the absence of a reporting standard rather than any wrongdoing. We include the survey as a baseline measurement for monitoring: as CSLR follows NMT and ASR in adopting the standard reranking toolkit (e.g. shallow CTC-LM fusion, MWER training as already used by SCOPE, and listwise candidate selection), Table 7 is the row against which future SHRP compliance can be tracked. The diagnostic experiments in §5.1 show that absent SHRP, a single dev-conditioned reranker can move the apparent CSL-Daily WER record from 23.1 to ∼14–17 without solving any recognition problem.

D Variance Across 20 Seeds

We re-ran bow-frozen-emb on each dataset under the dev-as-train protocol with seeds 0 – 19 in two independent 10-seed blocks. The block-to-block ensemble standard deviation on PHX-2014T was ± 0.13 abs. (blocks: 30.14 and 30.40); on CSL-Daily ± 0.14 (17.19 and 16.92); on PHX-2014 ± 0.10 (25.20 and 25.04). The headline numbers in Table 1 are therefore robust to seed variance well below the magnitudes of the dataset-level differences.

E N-gram Coverage Detail

Table 8 reports test n-gram coverage by dev and by train across $n \in \{1, \ldots, 4\}$ for all three datasets. CSL-Daily's 4-gram coverage by dev is 99.2% while train covers only 1.2%; on PHX and PHX-T this asymmetry is inverted (train covers more than dev). This single contrast underlies all dev-conditioned reranker gains documented in the main text.

n-gram order	CSL-Daily	PHX-2014	PHX-2014T
Test n-gram coverage by dev (%)			
1-gram	99.8	69.2	70.3
2-gram	99.4	32.6	31.0
3-gram	99.3	8.1	6.8
4-gram	**99.2**	2.3	2.1
Test n-gram coverage by train (%)			
1-gram	100.0	91.7	95.4
2-gram	41.2	66.2	74.8
3-gram	7.2	24.4	33.8
4-gram	**1.2**	7.1	11.2

Table 8: Test set n-gram coverage by dev and by train, for $n \in \{1, 2, 3, 4\}$. On CSL-Daily, dev covers 99.2% of test 4-grams while train covers only 1.2%. On PHX and PHX-T this asymmetry is inverted (train covers more than dev).

Paper	Venue	Year	WER	Dev-fit	99.66%?	PHX/PHX-T?
TwoStream-SLR (Chen et al., 2022)	NeurIPS	2022	25.3	No	No	Yes
TLP (Hu et al., 2022)	ECCV	2022	32.5	No	No	Yes
CorrNet (Hu et al., 2023a)	CVPR	2023	30.1	No	No	Yes
SEN (Hu et al., 2023b)	AAAI	2023	30.7	No	No	Yes
AdaptSign (Hu et al., 2024b)	arXiv	2024	29.6	No	No	Yes
AdaBrowse(+) (Hu et al., 2023c)	ACM MM	2023	30.7	No	No	Yes
CoSign-2s (Jiao et al., 2023)	ICCV	2023	27.2	No	No	PHX only
CorrNet+ (Hu et al., 2024a)	arXiv	2024	28.6	No	No	Yes
TCNet (Lu et al., 2024)	AAAI	2024	29.3	No	No	Yes
SlowFastSign (Ahn et al., 2024)	ICASSP	2024	25.5	No	No	Yes
SignGraph (Gan et al., 2024)	CVPR	2024	26.4	No	No	Yes
Online CSLR (Zuo et al., 2024)	EMNLP	2024	24.4	No	No	Yes
SCOPE (Liu et al., 2024)	arXiv	2024	23.1	MWER only	No	No

Table 7: Prior-work survey: 13 published CSL-Daily CSLR papers (2022–2025). No paper to date fits a post-hoc reranker on the CSL-Daily dev set. SCOPE's MWER is a training-time objective, not a post-hoc reranker. The published best test WER stands at 23.1; the dev-conditioned probes in Table 1 and Figure 1(a) reach 14–17 on CSL-Daily, showing the WER level a single dev-conditioned reranker could appear to deliver without SHRP reporting.