

Whitening and Fine-Tuning Are Substitutes: One-Shot Indian Sign Language Recognition

Galiveeti Poornima

Dept. of Engineering and Technology
S-VYASA Deemed to be University
Bengaluru, India

varaprabha.lita@gmail.com

Sandeep Mathias

Dept. of Computer Science and Engg.
Presidency University
Bengaluru, India

sandeepalbert@

presidencyuniversity.in

Abstract

We describe our submission to the WSLP 2026 shared task on Isolated Sign Language Recognition. The task is one-shot retrieval over 4,361 Indian Sign Language word classes with roughly one training clip per class. Our final system placed 3rd on top-1 accuracy at 0.6825 and 2nd on top-10 at 0.7357. The system is deliberately simple. It combines MediaPipe key-points, a graph attention backbone pretrained on a different ISL corpus, PCA whitening, and nearest-exemplar cosine retrieval against an enriched gallery. Our main finding concerns how two very differently priced interventions relate. PCA whitening of the frozen cross-corpus encoder is worth +18.85 points on a leave-one-out gallery evaluation. Fine-tuning that encoder end-to-end is worth a comparable amount. The two do not compose. Once the encoder has been fine-tuned, whitening adds +0.38 points, an interaction of -18.5 . Whitening and fine-tuning act as substitutes because both correct the same embedding anisotropy, and a transform that is fitted in seconds captures most of what a day of GPU training buys. We also report fifteen interventions that produced no usable gain: twelve indistinguishable from zero and three actively harmful. These were evaluated with McNemar tests on a 2,030-query held-out split with a minimum detectable effect of 1.23 points. We argue that the 526-clip official test set is too small to support the comparisons participants routinely make with it.

1 Introduction

Indian Sign Language (ISL) serves a large signing population. It is supported by a small fraction of the annotated data available for the sign languages that drive most published progress. The WSLP 2026 Isolated Sign Language Recognition shared task confronts this shortage directly (WSLP 2026 Shared Task Organisers, 2026). Participants receive roughly 4,398 training clips spanning about 4,361

word classes. That is approximately one example per class.

This is not a low-data variant of a standard classification problem. At a mean of 1.01 clips per class, it is a one-shot retrieval problem. Any method whose gains depend on gradient signal per class is structurally disadvantaged.

Our approach was shaped by an early observation. We applied a graph attention backbone (Patra et al., 2024), pretrained on FDMSE-ISL, to WSLP data without adaptation. Under plain cosine retrieval it scored 0.2510 top-1. This falls far below what the encoder’s in-domain performance would predict. The natural diagnosis is that the features are out of domain and need fine-tuning. That diagnosis turned out to be substantially wrong. The features were largely intact. Their geometry was not. A few high-variance directions dominated every cosine comparison. A single closed-form whitening transform moved the system from 0.2510 to 0.6198 on the official test split without touching a model weight.

This paper reports that result and what followed from it. We make five contributions.

1. A shared task system reaching 3rd place at 0.6825 top-1 and 2nd on top-10, built from four components and no ensemble of convenience.
2. An ablation isolating the whitening gain to the variance term rather than to the removal of nuisance axes. Whitening strength moves accuracy by tens of points, while the number of discarded leading components is nearly irrelevant at the operating point.
3. A 2×2 analysis showing that whitening and fine-tuning act as substitutes. Whitening is worth +18.85 points on the pretrained encoder and +0.38 on the fine-tuned one, an interaction of -18.5 .

4. Evidence that the graph attention backbone depends architecturally on arm geometry. Masking the input to hands alone collapses its accuracy, where an equivalently trained temporal convolutional network is essentially unaffected.
5. Fifteen negative results at stated statistical power, together with an argument that the official test set is too small to adjudicate them.

2 Task and Data

2.1 Task definition

Given a query video of a single isolated sign, a system predicts the word class. Submissions provide a ranked list of ten candidates per query. Scoring is by top-1 accuracy, with top-5 and top-10 also reported.

2.2 Splits

The organisers provide approximately 4,398 training clips over 4,361 classes, 109 validation clips, and 526 test clips. The validation split is far too small to gate decisions. A single clip is worth 0.92 points. We did not use it for model selection.

2.3 Gallery enrichment

The task is retrieval rather than classification. Additional labelled clips can therefore be added to the gallery at inference time without retraining. With organiser sanction we added clips from CISLR (Joshi et al., 2022). These contributed 3,758 additional prototypes, taking the gallery from 4,398 to 8,156 clips. We verified that no `uid` overlapped between the added clips and either the validation or the test split. Even after enrichment, 53.4% of classes remain singletons.

3 Related Work

Indian Sign Language has seen steady growth in corpora but remains thin relative to ASL or DGS. INCLUDE (Sridhar et al., 2020) covers a small vocabulary at high depth. CISLR (Joshi et al., 2022) covers roughly 4,700 words at approximately 1.5 videos per word, which makes it explicitly a one-shot resource. FDMSE-ISL (Patra et al., 2024) inverts the trade-off with 2,002 words recorded by 20 signers. iSign (Joshi et al., 2024) assembles a broader benchmark suite. Closest to our setting, Vandendriessche et al. (2025) treat one-shot ISLR as a retrieval problem and argue that sign-level

representations are what make such systems functional, which is the framing we adopt here. The WSLP 2026 task sits at the CISLR end of this spectrum. The methods that succeed there differ accordingly from those developed on depth-rich corpora such as WLASL (Li et al., 2020) or PHOENIX14T (Camgoz et al., 2018).

Skeleton-based recognition avoids the appearance overfitting of RGB pipelines and is the dominant approach for low-resource sign languages. HWGAT (Patra et al., 2024) applies windowed graph attention over an upper-body skeleton graph and reports gains over prior keypoint models across several corpora. Temporal convolutional networks (Bai et al., 2018) remain a strong and considerably simpler baseline. Our fusion results reinforce that point.

The degradation of cosine similarity under uneven embedding variance is well established outside sign language work. PCA whitening is standard in image retrieval (Jégou and Chum, 2012). Removing dominant directions improves word vectors (Mu and Viswanath, 2018). Contextual encoders are known to occupy narrow cones (Ethayarajh, 2019; Su et al., 2021). Our contribution is not the technique. It is the magnitude of the effect in the cross-corpus one-shot setting, and the finding that whitening substitutes for fine-tuning rather than compounding with it.

4 System

4.1 Keypoint extraction

We extract pose, hand and face landmarks with MediaPipe Holistic (Lugaresi et al., 2019), using `model_complexity=2` and detection and tracking thresholds of 0.1. We retain the 29 keypoints used by the backbone’s transform chain and resample every clip to 192 frames.

Two practical details cost us time. Detection thresholds of 0.0 produce a degenerate matrix and crash the legacy `mp.solutions` API, so 0.1 is the safe floor. We also pin MediaPipe to version 0.10.14, because landmark conventions are not stable across releases.

The cached tensor layout is $(N, 192, 64, 2)$. This is four windows of 16 keypoints with duplication, not 29 independent keypoints. The four windows are head with right arm and right hand, head with left arm and left hand, head with right arm and left hand, and head with left arm and right hand. Horizontal flip augmentation therefore requires swap-

ping the first two windows, swapping the second two, and swapping the eyes. Negating the x coordinate alone yields anatomically invalid poses. We flag this because the failure is silent. Training proceeds and loss decreases. We verified our permutation empirically, as described in Section 10.

4.2 Backbone and fine-tuning

We use HWGAT (Patra et al., 2024), a hierarchical window-based graph attention network pretrained on FDMSE-ISL. We fine-tune end-to-end with an ArcFace objective (Deng et al., 2019) for 12 epochs at batch size 8 with gradient accumulation 8 and horizontal flip probability 0.5. Training ran on a single RTX 5060, with peak memory of 7.67 GB out of 8.5 GB and 5 to 8 minutes per epoch.

One methodological point deserves emphasis. ArcFace loss fell steeply during fine-tuning, from 6.87 to 2.14, while held-out retrieval accuracy moved by fractions of a point. At approximately 1.87 clips per class after enrichment, the margin objective is easy to satisfy in ways that do not improve retrieval. Loss is not a usable proxy for retrieval quality at this data scale. Every promotion decision in this work was gated on held-out top-1 alone.

4.3 PCA whitening

Let $X \in \mathbb{R}^{N \times D}$ be the gallery embeddings. We compute the mean μ and the principal directions V of the centred matrix by SVD. We drop the leading d components, retain the next k , and rescale each retained axis by $\lambda_i^{-\alpha/2}$, where λ_i is the variance of that axis and α is the whitening strength. The transform $z = (x - \mu)W$, followed by L2 normalisation, is applied to both gallery and query embeddings. The construction follows standard practice in image retrieval (Jégou and Chum, 2012) and sentence representation learning (Su et al., 2021).

Two properties matter for the analysis that follows. First, the fit is computed on the gallery side only and then frozen. We never refit on test queries. Second, α is a continuous knob. At $\alpha = 0$ the transform is a pure rotation and leaves cosine similarity unchanged. At $\alpha = 1$ every retained axis is mapped to unit variance. This lets us attribute the gain in Section 7.

4.4 Retrieval

We score a query against every gallery clip by cosine similarity. For each word class we then take the maximum similarity over that class’s prototypes, rather than the similarity to their centroid.

This is nearest-exemplar matching rather than prototype matching (Snell et al., 2017). Class-max outperformed centroid averaging consistently.

The interpretation is straightforward. Prototypes for a class come from different signers with different execution styles. Averaging heterogeneous exemplars produces a centroid that resembles none of them. Under one-shot conditions with multi-signer galleries, the nearest exemplar is the right query.

4.5 Score fusion

Our final submission fuses the fine-tuned HWGAT with a temporal convolutional network (Bai et al., 2018) trained from scratch on the same key-points. Fusion is by z-normalised score averaging at weight 0.6. Section 9 explains why this addition moved the score when six others did not, and what that movement is worth.

5 Evaluation Protocol

The official test set contains 526 clips. Throughout, we write accuracies as percentages and differences between them as points. One clip is worth 0.19 points. The binomial standard error at $p = 0.68$ is approximately 2.03 points. Paired comparisons are more sensitive than that marginal figure suggests, but even under McNemar a difference below roughly 2.4 points is not detectable at $n = 526$ for systems as similar as ours. Many of the interventions we wanted to test have true effects well inside that band. Gating on the leaderboard would have meant promoting noise.

We therefore constructed a held-out split from the enriched gallery. Every class with at least two gallery clips donates one clip to a held-out query set. This yields 2,030 queries against a remaining 6,126-clip gallery. Whitening is fitted on the gallery side only, mirroring the submission pipeline exactly.

We compare systems with McNemar’s test (McNemar, 1947) on paired per-query outcomes. This is the appropriate test for two classifiers evaluated on the same items (Dietterich, 1998), and it is far more sensitive than comparing marginal accuracies. We state the sensitivity explicitly, since it is quoted throughout. Writing d for the discordant-pair rate between two systems, the difference detectable at $\alpha = 0.05$ and 80% power is $2.80\sqrt{d/n}$. The rate varies with how similar the two systems are. Our seed-varied models are the most similar pair we

Instrument	n	Primary use
Official test	526	Table 2, fusion
Held-out split	2,030	ablations, McNemar tests
Leave-one-out	8,156	Table 3, Fig. 1

Table 1: The three evaluation instruments used in this paper. Accuracies from different instruments are not comparable. Leave-one-out figures are the lowest of the three because the full gallery is dense with singleton classes, which is why whitening is worth 36.9 points on the official test but 18.9 points under leave-one-out.

System	top-1	top-5	top-10
Frozen HWGAT, raw cosine	0.2510	0.3441	0.3821
+ PCA whitening	0.6198	0.6958	0.7262
+ ArcFace fine-tune	0.6711	0.7205	0.7357
+ TCN score fusion	0.6825	0.7243	0.7357

Table 2: Official test leaderboard progression. Our final system ranked 3rd on the official leaderboard by top-1 accuracy, and 2nd by top-10 accuracy. The system ranked 2nd by top-1 reached 0.7319 on top-10.

compare, agreeing on 96.1% of held-out queries, which gives $d \approx 0.039$ and a minimum detectable effect of 1.23 points at $n = 2,030$. Comparisons between genuinely different systems have higher discordance. Across the pairs in this paper d reaches roughly 0.12, which raises the threshold to 2.15 points. We quote 1.23 as the best case available to us rather than a figure that applies uniformly, and the corresponding floors are 2.41 points on the 526-clip test split and 0.61 points under leave-one-out.

The instrument tracked the leaderboard closely. It gave 62.81% against an actual 61.98%, and 66.70% against an actual 67.11%. This gives us reasonable confidence that a null on the held-out split is also a null on the test set, rather than an artefact of a different distribution.

Table 1 names the three instruments and states which results depend on each. We flag one contamination case for honesty. A model trained on all 8,156 clips has seen every held-out query. Held-out numbers for that model are inflated and are not comparable to the rest of the table. We judge that model on the leaderboard alone.

6 Results

Table 2 gives the progression on the official test split. Two interventions produced gains the official test split can resolve, and a third, score fusion, was adopted for our submission but falls inside that split’s noise band, as Section 9 sets out. Fifteen

others produced no usable gain. Twelve were indistinguishable from zero and three were harmful, as Section 10 sets out.

One feature of our final standing is worth drawing out. Our top-1 of 0.6825 placed 3rd, but our top-10 of 0.7357 placed 2nd, ahead of the system that beat us on top-1 by 1.7 points. The spread between our top-1 and top-10 is 5.3 points. For the winning system it is 2.1. Our retrieval surfaces the correct class within ten candidates about as reliably as the best submission, and loses ground specifically in ordering the top of that list. This is what one would expect if the residual limitation is fine-grained ranking among near neighbours rather than gross representational failure. It is consistent with the whitening analysis in Section 7. Correcting global variance restores coarse retrieval, but the one-shot prototypes remain noisy enough that the top of the ranking stays unstable.

The three interventions we adopted share a property. Each changed the representation. Whitening changed its geometry, fine-tuning changed the encoder, and the TCN introduced a second architecture with different failure modes. Every inference-side method that reused an existing representation returned zero, regardless of sophistication. That set includes CSLS, query expansion, reranking, test-time augmentation, and alternative inference rules. We return to this pattern in Section 11.

7 Variance Equalisation

Whitening the frozen cross-corpus encoder moved test top-1 from 0.2510 to 0.6198. This section asks what the transform is actually correcting, and how its contribution relates to the contribution of fine-tuning.

7.1 Rescaling versus axis removal

The obvious explanation is that whitening removes a handful of nuisance directions. Signer identity, framing and lighting all plausibly dominate the leading principal components of a keypoint encoder. That explanation is wrong here, and the whitening grid says so (Figure 1). Sweeping the whitening strength α at fixed d moves leave-one-out top-1 by 29.0 points. Sweeping d at the operating point of $\alpha = 1.0$ moves it by 0.4 points, with $d = 2$, $d = 4$ and $d = 8$ falling within 0.08 points of each other.

The one regime where discarding components does matter is $\alpha = 0$, a pure rotation with no rescal-

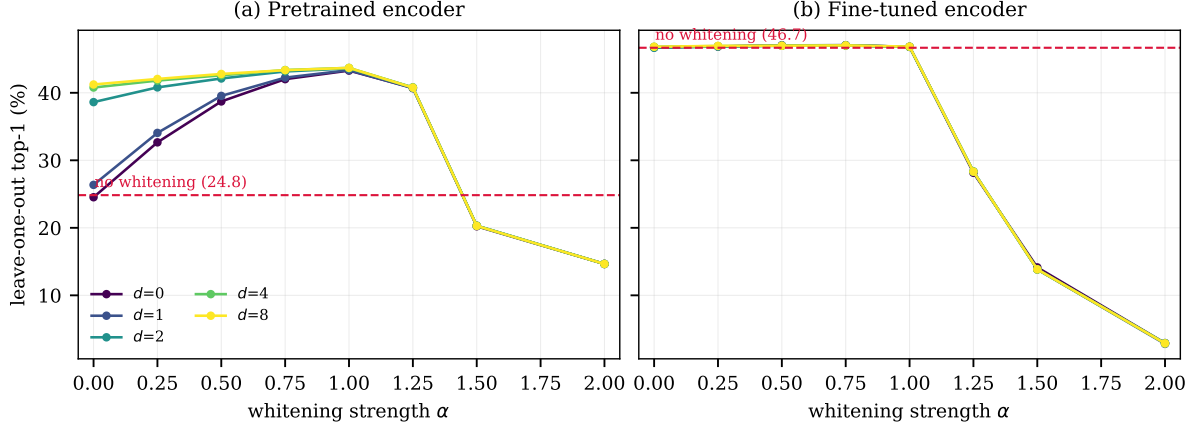


Figure 1: Leave-one-out top-1 against whitening strength α , with one line per number of discarded leading components d . Dashed line is the unwhitened baseline. On the pretrained encoder (a) the curves rise steeply with α and converge, so rescaling rather than deletion carries the gain, and the exception at $\alpha = 0$ is where deletion is the only correction available. On the fine-tuned encoder (b) whitening is flat until it becomes harmful, which is the substitution effect of Section 7.2 seen directly. Both encoders collapse under over-whitening.

ing, where moving from $d = 0$ to $d = 8$ is worth 16.7 points. This is the exception that confirms the rule. When no rescaling is applied, deleting the dominant axes outright is the only available way to stop them from controlling the inner product. Once α is large enough to rescale them, deleting them adds almost nothing.

The pathology is therefore global scale imbalance across the whole space, rather than a few contaminated axes. This mirrors the anisotropy reported for contextual word embeddings (Ethayarajh, 2019), where a small number of dominant directions likewise distort cosine geometry and can be corrected by removing or rescaling them (Mu and Viswanath, 2018). Cosine similarity implicitly assumes comparable variance across dimensions. When a cross-corpus encoder has a steep variance spectrum, a small number of directions dominate every inner product. The remaining dimensions carry the lexical distinctions, and they are effectively ignored. Equalising variance restores their contribution.

Over-whitening is harmful, and sharply so. Accuracy peaks at $\alpha = 1.0$ on the pretrained encoder and then falls away, reaching 20.3 points at $\alpha = 1.5$ and 14.7 at $\alpha = 2.0$, well below the unwhitened baseline of 24.8. Pushing genuinely low-variance noise axes to unit scale amplifies them. The same collapse appears on the fine-tuned encoder, and at a lower threshold. Reducing dimensionality to 256 or 128 also costs 2 to 3 points, which confirms that the discriminative signal is distributed rather than

Encoder	no whitening	whitening	gain
Pretrained	24.84	43.69	+18.85
Fine-tuned	46.67	47.05	+0.38

Table 3: Leave-one-out top-1 (%) over the 8,156-clip gallery. Whitening is worth +18.85 points on the pre-trained encoder and +0.38 on the fine-tuned one, an interaction of -18.5 . The fine-tuned row is inflated by training contamination and should not be read as a clean estimate of the effect of fine-tuning.

concentrated.

7.2 Substitution with fine-tuning

If whitening and fine-tuning correct the same anisotropy, they should not compose. We tested this with a 2×2 over pretrained and fine-tuned encoders crossed with whitening, evaluated throughout by leave-one-out retrieval over the 8,156-clip gallery so that every cell comes from one instrument (Table 3).

Whitening is worth +18.85 points on the pre-trained encoder and +0.38 points on the fine-tuned one. The interaction is -18.5 . Practically the whole value of the transform disappears once the encoder has been fine-tuned. Panel (b) of Figure 1 shows the same result as a nearly flat line. Across every setting of d and every α up to 1.0, whitening moves the fine-tuned encoder by well under a point, so the shallow optimum reported below sits inside a band that is immaterial in practice.

We are deliberately careful about what this table does and does not license. The two within-row

comparisons are clean, because each contrasts an encoder with itself. The across-row comparison is not, because the fine-tuned model was trained on these gallery clips and has seen every query. Its row is inflated by an unknown amount. We therefore make no claim that either intervention has the larger main effect. Our claim is about their relationship, which the within-row contrasts establish on their own: the marginal value of whitening collapses by 98% once fine-tuning has been applied.

We should be explicit about the alternative reading this leaves open. A model that has memorised the gallery depends less on cosine geometry for retrieval, and would therefore show a flat whitening response whether or not its embedding space is still anisotropic. The flat fine-tuned row is thus consistent with substitution and with saturation by memorisation, and this table cannot separate the two. The decisive experiment is a fine-tuned encoder evaluated on queries it has never seen, on the official test split or on a held-out split withheld from training. We regard that as the first thing to run in any extension of this work.

The point can be put more strongly. On the instrument that produced it, the +0.38 residual sits below the 0.61-point detection floor derived in Section 5. Whitening the fine-tuned encoder does not merely help less. It produces no difference this evaluation can resolve. The contrast with the pre-trained row is stark, since +18.85 clears the same floor by a factor of thirty.

The direction of the asymmetry is what makes this practically useful. Both interventions correct the same pathology, but they cost very different amounts. The whitening transform is a single SVD fitted in seconds on cached embeddings. The fine-tune is roughly an hour of GPU time in our setup and considerably more on larger backbones. When two interventions substitute for each other, the rational order is to try the cheap one first and measure before paying for the expensive one.

A dose-response pattern in the optimal whitening strength corroborates the substitution account, though for reasons given below it is not independent evidence for it. The pretrained encoder prefers aggressive correction at $d = 4$ and $\alpha = 1.0$. The fine-tuned encoder peaks milder, at $d = 0$ and $\alpha = 0.75$, and tolerates over-whitening far worse. A model trained on all 8,156 clips prefers milder correction again. The optimum moves toward weaker correction as the encoder acquires in-domain data, which is what a better-conditioned

embedding space would produce. We note that the middle point of this gradient is not resolvable on this instrument, since the fine-tuned encoder’s curve is shallow enough over $\alpha \leq 1.0$ that the location of its peak sits inside the same band as the residual itself. The gradient therefore rests on its endpoints.

We state the limit of this argument rather than leaving it to be found. All three optima are measured by leave-one-out over the gallery, and across these three encoders the amount of in-domain training data and the degree of contamination with respect to that evaluation increase together. An encoder that has memorised more of the gallery depends less on cosine geometry and would also prefer a milder correction, so the gradient on this instrument is consistent with the substitution account and with contamination alike. Separating them would require sweeping α on the official test split, whose queries no encoder trained on. We did not run that sweep. We therefore present the gradient as consistent with the substitution account rather than as independent support for it.

8 Input Masking and Arm Geometry

Sign phonology decomposes signs into handshape, location, movement and orientation. We asked which of these the backbone actually relies on. To do so we restricted the input to hand keypoints alone and fine-tuned under that mask.

HWGAT collapsed, losing approximately 20 points as recorded in Table 5. A TCN trained from scratch on identical masked inputs was essentially unaffected.

One confound must be stated before this comparison can be read. HWGAT was pretrained on full-body skeletons and then fine-tuned under the mask, whereas the TCN was trained from scratch under it. A model trained from scratch shapes its filters around whatever inputs exist, while a pretrained model arrives with an attention structure organised around groupings it no longer receives. The contrast may therefore be measuring pretrained-then-deprived against trained-under-deprivation rather than architectural dependence as such. Separating the two would require training HWGAT from scratch under the mask, which we did not have the compute budget to add.

Subject to that caveat, we read the result as a statement about the architecture rather than about ISL. The windowed graph attention structure is

built around groupings of head, arm and hand. The arm keypoints are not merely informative features. They are the scaffolding that the attention mechanism is organised around. Remove them and the inductive bias is not degraded but invalidated. The TCN treats keypoints as undifferentiated channels and simply routes around the missing input.

This bears on transfer to other sign languages. An architecture whose performance is contingent on a specific articulator grouping will transfer unpredictably to corpora with different framing conventions, different use of signing space, or seated rather than standing capture. We encourage reporting masked-input results as a routine robustness check in sign language recognition papers. The check is cheap and it is diagnostic.

9 Score Fusion

We attempted seven fusions and exactly one improved the score (Table 4). Whether that improvement is real is the subject of the closing paragraph of this section. The predictor of success was not architecture family, not individual member accuracy, and not fusion weight. It was pairwise top-1 agreement between members.

Members agreeing on 88 to 92% of queries contributed nothing or hurt. They fail on the same clips, so averaging their scores mostly averages their noise. The TCN agreed on 60 to 64% of queries and was worth +1.14 points.

Diversity alone is not sufficient either. Transformers agreeing on only 48 to 52% added nothing. Those members were both diverse and weaker. Score fusion cannot extract signal from a member that is wrong in novel ways without also being right in novel ways.

The useful summary is that ensemble headroom saturated after one architecturally distinct member. Fusion weight was insensitive across the range 0.6 to 0.9, which is itself evidence that the fused signal was not delicately balanced.

One consequence of Section 5 should be stated plainly rather than left for a reader to derive. The +1.14 point fusion gain was measured on the official test split, whose detection floor we put at 2.41 points. The fusion we adopted for our final submission therefore sits inside the noise band of the instrument that selected it. We report it because it is what we submitted, not because we can demonstrate that it helped. This is an uncomfortable illustration of the argument we make about the

526-clip split, and it applies to our own decisions as readily as to anyone else’s. The 1.7-point gap between our top-1 and the second-placed system falls inside the same band.

10 Negative Results

Table 5 lists them. Five deserve comment.

10.1 The largest observed differences

An ArcFace head trained on frozen features gained +0.69 points and proxy classification lost 0.54 points. These are the largest differences we observed among the null interventions, and both are smaller than the 1.23-point minimum detectable effect derived in Section 5, so neither is distinguishable from zero on this instrument. We report them separately for transparency rather than because they were detected. We note that the two quantities are not interchangeable. A minimum detectable effect states what a design would catch four times in five at a given power. Whether a particular observed difference reaches significance depends on the discordant-pair counts, and under McNemar a difference of this size can fall either side of the threshold. We therefore report both differences without asserting either outcome, and release the paired per-query outcomes so that the test can be recomputed. Neither changed a decision we made and we shipped neither.

10.2 Post-hoc similarity methods

CSLS (Conneau et al., 2018), alpha query expansion (Radenović et al., 2019) and rank fusion (Cormack et al., 2009) all reweight or reproject the same similarity structure. Once whitening had corrected the geometry, there was nothing left for them to correct. Their usual gains in the retrieval literature may be partly gains against uncorrected anisotropy. If so, they are redundant with whitening rather than additive to it.

10.3 Flip test-time augmentation

Mirrored queries scored 67.09% against 67.04% for originals, a difference of 0.05 points. The encoder is essentially handedness-invariant, so no residual variance remains for test-time augmentation to average away. This follows directly from training with flip probability 0.5.

The result also verifies the window-aware permutation described earlier. An incorrect permutation would produce anatomically invalid poses and a collapsed score, not parity.

Fusion	agreement	Instrument	Δ top-1
HWGAT seed variants (3)	88–92%	test	−0.19
HWGAT + full-data HWGAT	not measured	test	−0.19
Pretrained + fine-tuned	not measured	held-out	0.00
+ from-scratch TCN	60–64%	test	+1.14
+ 3 TCNs, 2 transformers	48–52%	test	0.00
+ restricted-view spec.	not measured	test	0.00

Table 4: Fusion outcomes, ordered by when they were attempted. Only the first architecturally distinct member improved the score, and that improvement sits below the test split’s detection floor (Section 9).

Intervention	Instrument	Δ top-1
CSLS hubness correction ($k=5,10,20$)	LOO	+0.02
Alpha query expansion ($k=2,3,5$)	LOO	$\leq +0.02$
Whitening re-sweep, fine-tuned feats	LOO	0.00
Extended fine-tuning (12→15+ ep)	held-out	+0.15
Two-encoder fusion	held-out	0.00
Flip TTA	held-out	+0.05
Retrain on all 8,156 clips	test	≈ 0.00
Seed-varied HWGAT ensemble	test	−0.19
Nine-model ensemble vs. four	test	0.00
Fusion weight 0.6 → 0.9	test	0.00
<i>Largest observed differences, still below threshold</i>		
ArcFace head on frozen features	held-out	+0.69
Proxy classification vs. retrieval	held-out	−0.54
<i>Harmful</i>		
Hand-only HWGAT fine-tune	held-out	≈ -20
Dimensionality reduction (256, 128)	LOO	−2 to −3
Over-whitening (α 1.25–2.0)	LOO	−2 to −3

Table 5: Interventions producing no usable gain, with the instrument each was measured on. Every difference in the first block is smaller in magnitude than the minimum detectable effect on the instrument that measured it, so none is distinguishable from zero. Those effects are 1.23 points on the held-out split, roughly 2.4 points on the official test split, and 0.61 points under leave-one-out over the 8,156-clip gallery. The two entries in the second block are the largest differences observed and remain below threshold. Failing to detect an effect is not evidence that no effect exists.

10.4 Train and gallery allocation

Training on all 8,156 clips rather than the 6,126-clip split produced no leaderboard change. The 2,030 held-out clips contributed about as much sitting in the gallery as they would have as training examples. For one-shot retrieval at this scale, the allocation of a clip between training and gallery is close to immaterial.

10.5 Coverage

Leave-one-out over the enriched gallery, bucketed by class depth, gave 33.2% top-1 for classes with one remaining example and 38.8% for classes with five or more. The ceiling at high depth is low and the relation is non-monotonic, with classes holding two examples scoring below classes holding one. That inversion is counterintuitive, since more examples per class should help, and we attribute it to the confound set out next rather than to any

property of the retrieval rule.

We read this as consistent with the representation rather than data quantity being the binding constraint, and we stop short of asserting it. Class depth is not randomly assigned. How many clips a word appears in reflects how common, how short and how canonical its sign is, so difficulty is confounded with depth and a shallow gradient cannot isolate representation quality on its own. The non-monotonic dip is most likely a symptom of the same confound, since the classes that happen to hold exactly two clips are not a random sample of the vocabulary. Establishing the causal claim would require holding class difficulty fixed while varying depth, which the available annotations do not support.

11 Discussion

A pattern runs through the null results. Every intervention that reused an existing representation at inference time returned nothing, and every intervention that changed the representation returned something. The largest difference among the representation-reusing interventions was +0.69 points from training a new head on frozen features, which is below the detection threshold on the instrument that measured it. This suggests a triage rule for one-shot retrieval under compute constraints. Before spending time on inference-side sophistication, establish whether the representation has slack in it. Whitening is a cheap probe for exactly that.

The 526-clip test set cannot distinguish systems separated by less than roughly 2.4 points, a floor that encompasses most of the gaps on the final leaderboard and our own fusion gain. Shared tasks in low-resource sign language processing face a real tension here, because test data is expensive to collect and annotate. We nonetheless encourage organisers to report a noise floor alongside the leaderboard, and participants to build a held-out instrument rather than iterating against a split that cannot support it. We suspect that a meaningful fraction of reported improvements in this space fall within noise, a concern raised more generally for NLP evaluation by [Card et al. \(2020\)](#).

Our findings concern one backbone, one corpus and one sign language, and we have not tested the whitening result beyond that setting. The mechanism it rests on, anisotropy in a cross-corpus embedding space, is not ISL-specific and is well documented elsewhere, so we expect the effect to appear wherever a pretrained encoder is applied across corpora. We state that as an expectation rather than a demonstrated result. The magnitude will depend on how anisotropic the source encoder happens to be. The masking result is explicitly architecture-specific, and we make no claim beyond HWGAT.

12 Conclusion

We reached 3rd place on the WSLP 2026 ISLR shared task with a system whose largest single component is a closed-form linear transform. Whitening a frozen cross-corpus encoder and fine-tuning it end-to-end turn out to be substitutes rather than complements. Each corrects the same embedding anisotropy, and applying either one collapses the marginal value of the other, with an interaction of −18.5 points. Fifteen further interventions, several

of them standard practice in retrieval, produced no usable gain. Twelve were indistinguishable from zero at 1.23-point detection sensitivity and three were harmful. For researchers working on under-represented sign languages with limited compute, the practical takeaway is that two interventions with very different price tags can buy the same thing, so the cheap one is worth measuring before the expensive one is paid for.

Limitations

Our results come from a single sign language, a single pretrained backbone and a single shared task. We cannot separate properties of HWGAT from properties of ISL keypoint data without replication on other encoders and corpora.

The held-out split is drawn from the same gallery distribution as the training data and shares signers with it. It may therefore understate the difficulty of genuinely held-out signers. The close agreement with leaderboard scores is reassuring but not decisive.

Our 2×2 is evaluated by leave-one-out retrieval over the gallery, and the fine-tuned model was trained on those clips. That row is inflated by contamination, so the table supports our claim about the relationship between whitening and fine-tuning, which rests on within-row contrasts, but not any claim about which intervention has the larger main effect. Separating the two main effects cleanly would require training a fine-tuned encoder that has never seen the evaluation clips at all, which we did not have the compute budget to add. We also note that leave-one-out accuracy over the full gallery is systematically lower than official test accuracy, because the gallery contains many singleton classes, so numbers from the two instruments should not be compared directly.

Our negative results are powered to detect effects above 1.23 points on the held-out split and 0.61 points under leave-one-out. We cannot rule out smaller true effects, several of which might in principle compose.

Gallery enrichment used a second corpus whose recording conditions differ from the task data. We did not isolate how much of the enrichment gain comes from coverage and how much from domain diversity.

Finally, we report single-run results for the fine-tuning experiments. We did not have the compute budget for multi-seed confidence intervals on every

configuration. The seed-variance experiments we did run were aimed at ensembling rather than at characterising run-to-run spread.

Artifact and Code Availability

All corpora used are publicly released research resources, obtained from their original distributors and used within their stated terms. The backbone architecture and its pretrained weights are publicly available from the authors of the original work. Our contribution consists of keypoint extraction and caching scripts, the whitening fit and sweep, the retrieval and fusion code, the held-out evaluation harness including the McNemar procedure, and the configuration files needed to reproduce every number reported here. These will be released under an open licence on acceptance, and are available from the authors on request in the meantime.

Ethics Statement

Sign language recognition systems built without the involvement of Deaf communities risk encoding hearing assumptions about how signing works. They also risk producing tools that serve researchers more than signers. Our work is a technical contribution to a shared task and makes no claim to deployment readiness. Top-1 accuracy near 0.68 on isolated signs is far from adequate for any communicative application, and presenting it as such would be harmful.

References

- Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. [An empirical evaluation of generic convolutional and recurrent networks for sequence modeling](#). *arXiv preprint arXiv:1803.01271*.
- Necati Cihan Camgoz, 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)*, pages 7784–7793.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With little power comes great responsibility](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274. Association for Computational Linguistics.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699.
- Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 55–65. Association for Computational Linguistics.
- Hervé Jégou and Ondřej Chum. 2012. Negative evidences and co-occurrences in image retrieval: The benefit of PCA and whitening. In *European Conference on Computer Vision (ECCV)*, pages 774–787.
- Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. [CISLR: Corpus for Indian Sign Language recognition](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10357–10366, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. [iSign: A benchmark for Indian Sign Language processing](#). In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.
- Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1469.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A framework for building perception pipelines](#). *arXiv preprint arXiv:1906.08172*.
- Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

- Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*.
- Suvajit Patra, Arkadip Maitra, Megha Tiwari, K. Kumaran, Swathy Prabhu, Swami Punyeshwarananda, and Soumitra Samanta. 2024. [Hierarchical windowed graph attention network and a large scale dataset for isolated Indian sign language recognition](#). *arXiv preprint arXiv:2407.14224*.
- Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2019. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. INCLUDE: A large scale dataset for Indian sign language recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1366–1375.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. [Whitening sentence representations for better semantics and faster retrieval](#). *arXiv preprint arXiv:2103.15316*.
- Tom Vandendriessche, Mathieu De Coster, Annelies Lejon, and Joni Dambre. 2025. [Representing signs as signs: One-shot ISLR to facilitate functional sign language technologies](#). *arXiv preprint arXiv:2502.20171*.
- WSLP 2026 Shared Task Organisers. 2026. Findings of the WSLP 2026 shared task on isolated sign language recognition. In *Proceedings of the 2nd Workshop on Sign Language Processing (WSLP 2026)*. To appear.